Understanding the function of the transcriptional activator *ZFY* in relation to its structure and evolution.

A thesis submitted to the University of Kent for the degree of
Doctor of Philosophy in Cell Biology

July 2024

Isabella Garcia

Division of Natural Sciences

I Declaration

No part of this thesis has been submitted in support of an application for any degree or qualification at the University of Kent or any other University or institute of learning.

Isabella Garcia
July 2024

II Acknowledgements

First of all, I would like to acknowledge and thank all three of my supervisors: Dr Peter Ellis (Primary supervisor), Dr Lee Larcombe (CASE supervisor) and Dr Tim Fenton (Co-supervisor) all of whom have been vital to this journey. Thank you, Peter, for your constant patience, guidance and incredible knowledge. Thank you, Lee, for going above and beyond and providing me with so much support throughout my PhD but especially towards the end. I cannot explain how grateful I am for all the opportunities you have given me. Finally, thank you, Tim for your constant support and guidance.

Thank you to my crazy lab group (Marie Claire, Liv, Sally, Sarah, Nella, Nicole, Carla, Richard, Corey, Frances and Claudia) who have supported me throughout my four years and have definitely made it an interesting but amazing experience. I could not have done it without all of you, although I am grateful that I am no longer social sec.

Thank you to my SoCoBio buddies (Laura, Paige, Vicky, Chloe and Kseniia) for all the amazing experiences we have had together throughout our PhDs. Paige thank you for being my travel buddy and exploring Thailand with me, but also for being one of my biggest hype women. Laura, we have definitely seen each other at our worsts, but even with all the tears we have somehow done it. Thank you, Vicky, for putting up with me and living with me multiple times. You always knew how to comfort me. Thank you to all my other friends who have seen me through my PhD at all the good and bad points.

A massive thank you to the Apexomics team who have seen me through the last 6 months of my PhD and the not so fun thesis writing. Thank you for all the coffee breaks and lunch walks.

I also want to thank my family for always being supportive and helping me through this challenging yet rewarding journey.

I cry a lot, but I am so productive, it's an art. Taylor Swift

III Contents

I Declaration 2
II Acknowledgements 3
III Contents 4
IV Abbreviations 8
V Abstract 10
Chapter 1: Introduction 11
Overview Thesis Goals 11
1.1 Overview of the Sex Chromosomes 11
1.1.1 Sex Chromosome Evolution and the Specialised Gene Content of the Y 12
1.1.2 Sex Chromosome Dosage Compensation 14
1.1.3 Consequences of Sex Chromosome Aneuploidy 15
1.1.4 Structure and Gene Content of the Human Y Chromosome 16
1.2 Spermatogenesis 19
1.2.1 Overview of Premeiotic Spermatogenesis 19
1.2.2 Overview of Meiosis 20
1.2.3 Overview of Spermiogenesis 22
1.2.4 Meiotic Sex Chromosome Inactivation During Meiosis 23
1.2.5 Evolution of Spermatogenic Function 24
1.3 Cancer-Testis Antigens 26
1.4 Differential Prevalence of Cancers between Males and Females 28
1.5 Preliminary Observation: *ZFY* is potentially Associated with Cancer 29
1.6 Zinc Finger Proteins 29
1.6.1 The Structure of Zinc Finger Proteins 30
1.6.2 The Biological Function of Zinc Finger 32
1.7 Zinc Finger Y-Chromosomal Protein 33
1.7.1 The Structure of *ZFY* 35
1.7.2 Known Biological Functions if *ZFY* 38
1.7.2.1 Meiotic Functions of *ZFY* 38
1.7.2.2 Post-Meiotic Functions of *ZFY* 40
1.7.3 Alternative Splicing Concerning *ZFY* 41
1.7.3.1 *RBMY*, an Overview 41
1.7.3.2 Why *RBMY*? 43
1.7.4 *ZFX*, the X-Chromosome Homologue of *ZFY* 45
1.7.5 *ZFY*'s Role in Y Chromosome Evolution 46
1.7.6 *ZFY* is a Possible Proto-Oncogene 46
1.7.6.1 Head and Neck Cancer 48
1.7.6.2 HPV-negative vs HPV-positive HNSCC 49
1.7.6.3 HNSCC Male Prevalence 50
1.8 Project Outline & Aims 50
Chapter 2: Tracing the Evolution of *ZFY* from an Autosomal Gene to a Y Chromosome Gene 53
2.1 Introduction 53
2.1.1 Phylogenetics – a Powerful Scientific Field 53
2.1.2 Detecting Selection Via Analysis of Mutation Rates 55
2.1.3 Understanding *ZFY* Evolution: Consequences of Y Linkage 56
2.1.4 Understanding *ZFY* Evolution: Structural Considerations 57
2.2 Materials and Methods 59
2.2.1 Nucleotide and Protein Sequence Collection 59
2.2.2 Sequence Alignment 59
2.2.3 Phylogenetic Tree Construction 61
2.2.4 Conserved Domain Analysis 62
2.2.5 Geneconv Gene Conversion Tool 64
2.2.6 Ancestral Sequence Reconstruction 65
2.3 Results 66

2.3.1 *ZFY* is Evolving More Rapidly than *ZFX*, Particularly in Rodents 66
2.3.2 Possible Gene Conversions Identified by Nucleotide Phylogeny 69
2.3.3 Defining the Functional Domains of *ZFY* 71
2.3.4. Specific Domain Selection Analysis 73
2.3.4.1 The *ZFY* Coding Domain Sequence Remains Under Negative Selection 73
2.3.4.2 The *ZFY* and *ZFX* Coding Domain Sequences Show Similar Selection Pressures 75
2.3.4.3 The Analysis of the *ZFY* Acidic Activating Domain and DNA Binding Domains Evolution 76
2.3.5 Exon 7 is Subject to Gene Conversion and Rapid Evolution 78
2.3.6 Genetic Exchange Between *ZFY* and *ZFX* Could Have Led to Their High Homology, but they Continue to Persist for Male and Female Function Respectively 82
2.3.6.1 Global Fragment Analysis Reveals Possible Gene Conversions Across a Range of *ZFY*/*ZFX* Species 82
2.3.6.2 Calibrating the Phylogeny Using Known Species Divergence Times 85
2.3.7 Ancestral Reconstruction of *ZFY*/*ZFX* Ancestors 88
2.3.7.1 Tracing Back to the Last Common Ancestor of *ZFY*/*ZFX* Following the Marsupial/Eutherian Divergence 88
2.4 Discussion 99
Chapter 3: Unravelling *ZFY*'s Splicing Mechanism – Exploring the Role of *RBMY* 102
3.1 Introduction 102
3.1.1 Alternative Splicing Increases Genome Complexity 102
3.1.2 Alternative Splicing of *ZFY* Forms a Testis-Specific Short Variant 104
3.2 Materials and Methods 106
3.2.1 RNA-Seq Data Analysis Looking into Splicing Variations 106
3.2.2 Cancer Cell Line Maintenance 108
3.2.3 Reverse Transcription Polymerase Chain Reaction 108
3.2.4 GFP-Splicing Reporter Preparation 111
3.2.4.1 Lysogeny Broth Media and Agar 111
3.2.4.2 DNA Transformation 111
3.2.4.3 Plasmid DNA Miniprep & Reporter Cloning 111
3.2.5 GFP-Splicing Reporter Mammalian Cell Transfection 112
3.2.5.1 HEK293 Cell Maintenance 113
3.2.5.2 Lipofectamine 3000 Transfection 113
3.2.5.3 Cell Fixing & Harvesting 114
3.2.5.4 Slide Preparation & Fluorescence Microscopy 114
3.2.6 Flow Cytometry 114
3.2.7 GFP-Splicing Reporter Polymerase Chain Reaction 115
3.3 Results 117
3.3.1 Human and Mouse Splicing Event is not Directly Detected in Short Read RNA-Seq 117
3.3.2 The Eutherian Testis-Specific Isoform *ZFYS* is Conserved in Opossum 121
3.3.3 A Potential Alternative Short ZF* Splice Form is Observed in Chicken Testis 123
3.3.4 *ZFYS* and *RBMY* are Expressed in a Head and Neck Squamous Cell Carcinoma 125
3.3.5 Direct Testing of *RBMY* Regulation of *ZFY* Splicing in a Model System 128
3.4 Discussion 134
Chapter 4: Overexpressing *ZFY* in HEK293 Cells to Conduct Transcriptomic Analysis 137
4.1 Introduction 137
4.1.1 The Growing Transcriptomic Field 137

4.1.2 Gene Overexpression System 138
4.1.3 The Current *ZFY* Functional Knowledge 139
4.2 Materials and Methods 142
4.2.1 DNA Constructs and Transformations 142
4.2.2 Plasmid DNA Miniprep 143
4.2.3 Mammalian Cell Line 143
4.2.3.1 Lipofectamine 3000 Transfection 143
4.2.3.2 Cell Microscopy 144
4.2.3.3 Flow Cytometry 144
4.2.4 Western Blotting 145
4.2.5 RNA Extraction 147
4.2.6 RNA-Seq Data Collection and Preparation 147
4.2.7 DESeq2 149
4.2.8 *ZFX* Differential Expression Data 150
4.2.9 Gene Ontology 151
4.2.10 Primer Design and Selection 151
4.2.11 Primer Checking and cDNA Synthesis 153
4.2.12 pGEM-T Easy Vector Ligation and Transformation 154
4.2.13 Quantitative Reverse-Transcription Polymerase Chain Reaction 154
4.2.14 Cancer Dataset and Cancer Correlation Analysis 155
4.2.15 Leukaemia Cell Lines 157
4.2.16 RNA Extraction Part 2 157
4.2.17 Cancer Cell Line Primers 158
4.2.18 Primer Optimisation and RT-qPCR 158
4.3 Results 160
4.3.1 Nuclear Localisation of GFP-Tagged Constructs 160
4.3.2 High Transfection Efficiency Achieved 161
4.3.3 Western Blot Confirmation of Successful Transformation 163
4.3.4 Good QC Analysis and Alignment Rates 165
4.3.5 *ZFY* Overexpression Does not Silence the X Chromosome 166
4.3.6 DESeq2 Differential Gene Expression Analysis 167
4.3.6.1 Dataset Dispersion 167
4.3.6.2 Sample Clustering 168
4.3.6.3 Differential Gene Expression Contrasts 175
4.3.6.4 Differential Gene Lists 177
4.6.7 Bioinformatics Validation 179
4.3.8 Gene Ontology and Enrichment Analysis 181
4.3.9 Cancer Correlation Analysis 184
4.4 Discussion 202
4.4.1 A Potential Extracellular Matrix Function 203
4.4.2 *ZFYL* has a Potential Neuronal-Like Function in Sperm 205
4.4.3 Gene Ontology Potentially Confirms *ZFY* as a Potential Cancer-Testis Gene 205
4.4.4 The ZF Family Appears to Target the *WNT*-Signalling Pathway 206
4.4.5 *ZFY* has a Weak Cancer Correlation 210
4.4.6 A Potential Feedback Mechanism with *RBMY* 211
Chapter 5: Utilising Proteomics to Understand the Role of *ZFY* Through the Identification of its Interacting Partners 212
5.1 Introduction 212
5.1.1 The lac Promoter as a Useful Target 212
5.1.2 *ZFYS*' Structure in Relation to its Role as a Transcription Factor 214
5.2 Materials and Methods 216
5.2.1 DNA Construct Design 216
5.2.2 Restriction Digest 217
5.2.3 *E. coli* Competent Cells 217
5.2.4 Transformation of Plasmids into Competent Cells 217

5.2.5 Plating Assay in Bl21(DE3) Competent Cells 218
5.2.6 Protein Growth and Induction 218
5.2.7 Cell Lysis 219
5.2.8 Freezing-Thawing Protocol 219
5.2.9 Nickel Column Chromatography 220
5.2.10 Dialysis 220
5.2.11 Anion Exchange Chromatography 220
5.2.12 Sodium-Sodecyl-Sulfate-Polyacrylamide Gel Electrophoresis 221
5.2.13 Western Blotting 222
5.2.14 Mass Spectrometry 223
5.2.14.1 Protein Identification 223
5.2.14.2 Intact Mass Spectrometry 224
5.2.14.3 Top-Down Sequencing 224
5.2.15 GFP Pull-Down 224
5.2.15.1 Mammalian Cell Line 225
5.2.15.2 Lipofectamine 3000 Transfection 225
5.2.15.3 Cell Harvesting and Lysate Preparation 225
5.2.15.4 GFP-Trap Agarose Protocol 226
5.2.15.5 Proteomics via Mass Spectrometry 227
5.3 Results 228
5.3.1 BL21 Expression System 0.4mM IPTG 3-hour Induction 228
5.3.2 BL21 Expression System 0.8mM IPTG Overnight Induction 232
5.3.3 Mass Spectrometry Confirmation 236
5.3.4 Rosetta Cell Expression System 242
5.3.5 Anion Exchange Chromatography 244
5.3.6 Further Buffer Optimisation 247
5.3.7 Codon Optimisation 251
5.3.8 Anti-*ZFY* Antibody Testing 255
5.3.9 BL21 PlysS Freeze-Thaw Lysis Method 256
5.3.10 GFP-Pull Down 259
3.3.11 Proteomics 260
5.4 Discussion 269
Chapter 6: Overall Discussion & Future Work 277
6.1 Phylogenetic Analysis of *ZFY* Reveals Strong Negative Selection, specific Conserved Motifs in the Acidic Domain, and accelerated Evolution in Rodents .. 277
6.2 The Testis-Specific Splicing *ZFY* is Conserved in Non-Eutherian Species and is likely to be Regulated by *RBMY* 279
6.3 *ZFYS* is a “Weaker Version” of *ZFYL*, but both have Important Spermatogenic Roles 281
6.3.1 The Extracellular Matrix if Crucial during Spermatogenesis 282
6.3.1.1 Fertilisation and the Egg Extracellular Matrix 285
6.3.1.2 The Tumour Microenvironment Shaping and Cancer Progression 286
6.3.2 *WNT* Signalling: a Key Cascade Regulating Development 287
6.3.2.1 *WNT* Signalling and Spermatogenesis 288
6.4 *ZFYL* Shows Enrichment of Presynaptic Function Pathways 289
6.5 *ZFYS* Activates a Key Cancer Pathway Driver 290
6.6 Proteomics Analysis Shows links to DNA and RNA Metabolism 291
6.7 Suggestions for Future Works 292
6.8 Final Conclusions 295
Chapter 7: References 297
Chapter 8: Supplementary Data 331

IV Abbreviations

9aaTAD	Nine Amino Acid Transactivation Domain
AAD	Acidic Activating Domain
AZF	Azoospermia factor
BCA	Bicinchoninic Acid
BC-KA	Bonferroni-corrected Karlin-Altschul P-values
BIC	Bayesian Information Criterion
BLAST	Basic Local Alignment Search Tool
Bp	Base Pairs
BSA	Bovine Serum Albumin
CCLE	Cancer Cell Line Encyclopaedia
cDNA	Complementary Deoxyribonucleic acid
CDS	Coding Domain Sequence
ChIP(-Seq)	Chromatin immunoprecipitation (-sequencing)
CT	Cancer-Testis
DAPI	4′,6-diamidino-2-phenylindole
DBD	DNA-Binding Domain
ddH2O	Deionised Water
DKO	Double Knock-Out
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
DTT	Dithiothreitol
E. Coli	*Escherichia Coli*
EDTA	Ethylenediaminetetraacetic acid
FBS	Fetal Bovine Serum
FDR	False Discovery Rate
FITC	Fluorescein Isothiocyanate
FSC	Forward Scatter Area
GFP	Green Fluorescent Protein
HEK293	Human Embryonic Kidney 293 Cells
HNC	Head and Neck Cancer
hnRNP	Heterogenous Nuclear Ribonucleoproteins
HNSCCs	Head and Neck Squamous Cell Carcinomas
HPV	Human Papillomavirus
IPTG	Isopropylthio-β-galactoside
Kb	Kilobase
KRAB	Krüppel-Associated Box
L2FC	Log2 Fold Change
LacI	Lac Repressor Protein
LB	Lysogeny broth
MEGA	Molecular evolutionary genetic analysis
$MgCl_2$	Magnesium Chloride
mRNA	Messenger Ribonucleic acid
MSCI	Meiotic Sex Chromosome Inactivation
MUSCLE	Multiple Sequence Comparison by Log-Expectation

NaCl	Sodium Chloride
NCBI	National Centre for Biotechnology Information
ncRNA	Non-Coding Ribonucleic Acid
NGS	Next Generation Sequencing
OPSCC	Oropharyngeal squamous cell carcinoma
PAH	Polycyclic Aromatic Hydrocarbons
PAR	Pseudoautosomal Region
PBS	Phosphate Buffered Saline
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PFA	Paraformaldehyde
PMSF	Phenylmethylsulfonyl Fluoride
PVDF	Polyvinylidene Fluoride
QC	Quality Control
qPCR	Quantitative Polymerase Chain Reaction
RBMY	RNA-binding motif Y chromosome
RIPA	Radioimmunoprecipitation assay
RNA	Ribonucleic acid
RNA-Seq	RNA Sequencing
RRM	RNA Recognition Motif
rRNA	Ribosomal Ribonucleic Acid
RT	Room Temperature
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SCC-H	Side Scatter Height
SDS	Sodium Dodecyl Sulfate
SDS-Page	Sodium Dodecyl Sulfate–Polyacrylamide Gel Electrophoresis
snRNP	Small Nuclear Ribonucleic Proteins
SRY	Sex-Determining Region Y
SSC-A	Side Scatter Area
TAF	TBP-Associated Factor
TBP	TATA-Binding Protein
TBS	Tris-Buffered Saline
TBST	Tris-Buffered Saline with Tween
TF	Transcription Factor
TFIID	Transcription Factor II D
TPM	Transcript Per Million
tRNA	Transfer Ribonucleic acid
ZF	Zinc Finger
ZFY	Zinc Finger Y-Chromosomal Protein

V Abstract

The Y chromosome is a small, gene-poor chromosome that is enriched with repetitive sequences. This is due to its inability to undergo recombination, which leads to genetic degeneration over evolutionary time. Consequently, it is sometimes regarded as a “functional wasteland”. The genes that persist on the Y chromosome are essential for sex determination and male germ cell development. Among these, the transcription factor *ZFY* is one of very few genes consistently present on the Y chromosome in almost all eutherian species, indicating that it must have essential functions in men. A further distinctive feature of *ZFY* is the presence of two distinct developmentally regulated splice variants; a ubiquitous full-length major variant and a testis-specific minor short variant. This thesis seeks to understand the evolution, structure and function of *ZFY*, building on recent theoretical advances in understanding the mechanisms of transcription factor activity, and on up-to-date transcriptomic and proteomic experimental techniques.

This thesis comprises four results chapters that collectively probe different aspects of *ZFY* structure and function. First, a phylogenetic analysis defines conserved versus rapidly evolving regions of *ZFY* and relates this to newly predicted functional motifs within the acidic domain. Secondly, cross-species examination of *ZFY* splicing data reveals that the testis-specific splicing pattern predates its recruitment to the Y chromosome, and potentially implicates *RBMY* as a potential splicing factor involved. Thirdly, RNA-Seq analysis highlights the core downstream pathways regulated by each splice isoform of *ZFY*, and finally pull-down proteomics identifies a range of potential interacting partners.

Overall, the results identify a potential novel feedback loop regulating *ZFY* splicing during testis development and suggest several key pathways that *ZFY* may regulate including *WNT* signalling, ErbB signalling and extracellular matrix remodelling. Whilst an earlier observation that the short *ZFY* form is mis-expressed in some cancers was replicated, a wider role for *ZFY* as a cancer-testis gene was generally not supported.

1. Chapter 1: Introduction

Overall Thesis Goals

This thesis aims to investigate the structure and function of *ZFY*, a zinc finger transcription factor present on the Y chromosome believed to play an important role in regulating apoptosis and sperm development in the testis. Its possible role(s) in other tissues are hitherto largely uncharacterised, and analyses of its evolutionary history have not been grounded in an up-to-date understanding of its domain structure and function.

In this introduction, the broad evolutionary history of the sex chromosomes will be discussed and related to the potential roles of Y-linked genes in spermatogenesis and cancer. Then, the biology of transcriptional regulation by zinc finger proteins, and what is known to date about *ZFY* function in light of this will be addressed.

1.1 Overview of Sex Chromosomes

Sex chromosomes are defined as chromosomes whose complement (copy number) varies between the sexes in a dioecious species (Abbott *et al*., 2017);(Palmer *et al*., 2019). Within each species, sex determination is controlled by genes borne on the sex chromosomes that initiate a cascade of sex-specific gene expression, thus directing embryonic development towards either a male or female phenotype. These key sex-determining genes are located in a major sex-determining region that may also harbour sexually antagonistic genes: that is, genes that are advantageous for one sex but detrimental to the opposite sex (Abbott *et al*., 2017).

The first observation of a sex chromosome was made in 1891 by Hermann von Henking (Carey *et al*., 2022). Lacking the ability to detect the cytologically minute Y chromosome, he noted the presence of an isolated chromosome that was not consistently passed in the gametes and named it “X” to represent the unknown (Carey *et al*., 2022). Subsequently, the first discovery of paired sex chromosomes that specifically correlated with organism sex was made by Nettie Stevens in 1905, observing that male mealworm cells carried a single smaller chromosome when compared to the female cells that carried equal size chromosomes (Stevens, 1905);(Furman *et al*., 2020). Following the discovery of the sex chromosomes, it has been made clear that they exhibit major interspecific and intraspecific diversity and that a wide variety of sex chromosome systems have evolved independently in different taxa (Bachtrog *et al*., 2014);(Furman *et al*., 2020).

Multiple different types of sex chromosome systems are found throughout the tree of life (Hake & O’Connor, 2008);(Ezaz *et al*., 2006). Amongst animals, the most common is male heterogamety, as seen in marsupials and eutherian mammals. In this system,

males have an XY sex chromosome complement (copy number) while females are XX. In this type of system, the Y chromosome is passed clonally from father to son, is never present in a female body, and does not undergo recombination over much of its length. This lack of recombination has profound consequences for its structure and gene content, leading to progressive degeneration of its functional content over evolutionary time.

1.1.1 Sex Chromosome Evolution and the Specialised Gene Content of the Y

Sex chromosomes can be either homomorphic or heteromorphic (Furman & Evans, 2018);(Furman *et al*., 2020). Whilst homomorphic sex chromosomes have very little divergence between the pairs, heteromorphic chromosomes show a degree of genetic divergence. These can include SNPs, inversions, and/or deletions distinguishing the sex chromosomes (Furman *et al*., 2020). This latter model is followed by many sex chromosome systems and is clearly evident by the shared gene content observed in XY (marsupial and eutherian), and ZW (avian) UV (algae and bryophytes) systems. However, these systems differences have resulted in broad evolutionary and genomic implications (Bachtrog *et al*., 2011). Recently there has been increasing evidence that some sex chromosome systems have arisen independently, and these species do not share a common ancestor with X or Z (Bachtrog *et al*., 2011);(Furman *et al*., 2020).

The prevailing theoretical model is that approximately 165 million years ago, the therian X and Y chromosomes originated from an ancestral autosomal pair through the emergence of SRY gene as a master sex-determining locus. This divergence of SRY from its X homologue SOX3 occurred subsequent to the split of monotremes from the eutherian and marsupial lineages (Holmlund et al., 2023). Subsequently the mammalian sex chromosomes evolved via a series of inversions on the Y chromosome, each event further suppressing X-Y recombination around the sex-determining locus, expanding the non-recombining region and allowing further differentiation to proceed (B. Lahn & Page, 1999);(Vicoso, 2019) (**Figure 1.1**). Overall, there have been four major inversion events during primate evolution. The first major event occurred subsequent to the marsupial/eutherian split and involved translocation of autosomal material into the PAR as well as subsequent recruitment from the PAR into the non-recombing region. Thus, there are a number of genes (including ZFY) that are autosomal in marsupials but sex-linked in eutherians. Subsequent inversions primarily recruited material from the PAR into the non-recombining region of the Y, leading to successive "strata" of XY divergence, with the fourth event occurred during recent primate evolution. The four potential inversion

events have been proposed to drive the evolution of the Y chromosome and enable differentiation from the X chromosome. (B. Lahn & Page, 1999);(Vicoso, 2019). In addition to inversion events that drive X/Y divergence, a large block of autosomal material was recruited to the sex chromosomes subsequent to the marsupial/eutherian split, via translocation of autosomal material into the PAR followed by subsequent recruitment from the PAR into the non-recombing region. Thus, there are a number of genes (including ZFY) that are autosomal in marsupials but sex-linked in eutherians.

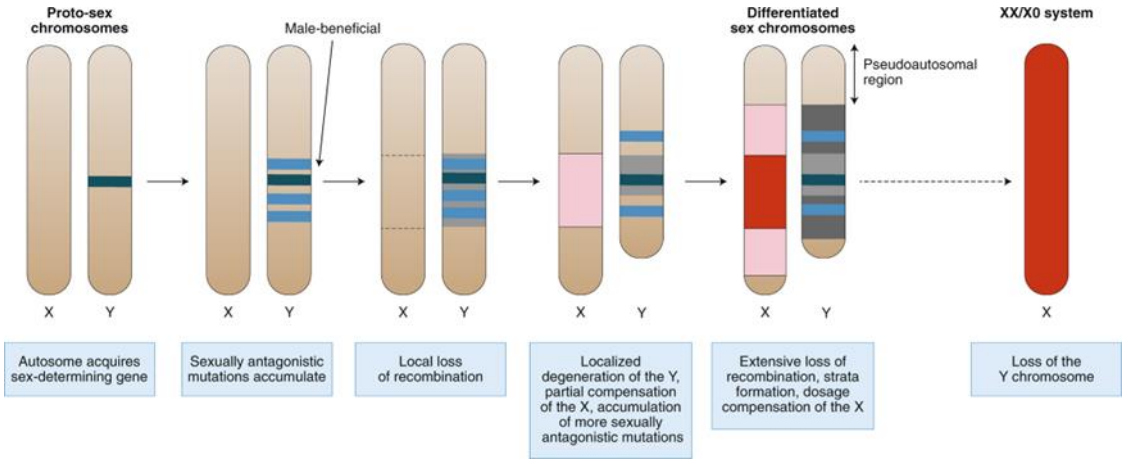


Figure 1.1: The evolution of heteromorphic sex chromosomes (Vicoso, 2019)**.** The male-determining gene (SRY), depicted as a dark blue line, is acquired first. Blue lines indicate the evolution of sexually antagonistic mutations advantageous to males. Pink regions are undergoing the process of acquiring dosage compensation, while red regions are fully dosage compensated. The order of events goes from left to right.

The Y chromosome harbours two categories of genes: (1) genes that directly benefit males, such as those involved in regulating spermatogenesis, and (2) dose-sensitive X-Y homologous genes, whose expression from both sex chromosomes is crucial in somatic tissues of both males and females(Ellis & Affara, 2009);(Colaco & Modi, 2018);(Subrini & Turner, 2021). These gene types become enriched as a result of the gradual degradation of Y genes over evolutionary time, with only those genes conferring sufficient selective advantage being retained. Consequently, the majority of the genes on the ancestral proto-sex chromosomes are lost while genes beneficial to spermatogenesis and male development accumulate. Moreover, any novel function benefiting males will undergo strong selection pressure due to its exclusive existence in men, leading to the acquisition of new male-specific functions on the Y chromosome as it diverges from the X (Ellis & Affara, 2009);(Colaco & Modi, 2018);(Subrini & Turner, 2021).

1.1.2 Sex Chromosome Dosage Compensation

While males inherit one X chromosome and one Y chromosome, females inherit two copies of the X chromosome leading to an imbalance in alleles between the sexes (Sidorenko *et al*., 2019). This has led to the evolution of dosage compensation mechanisms which randomly inactive one of the X chromosomes during early female embryonic development (Shvetsova *et al*., 2018);(Sidorenko *et al*., 2019). X chromosome inactivation serves to prevent the double expression of genes in females in comparison to males (Shvetsova *et al*., 2018). Both X chromosomes have an equal chance of being silenced, and once silenced, this state remains stable throughout all subsequent cell generations (Panning, 2008). This generates a mosaic of cells in females in which either the maternally inherited or paternally inherited X is silenced.

Although X chromosome inactivation is observed in various placental mammals, with studies primarily focusing on mice and humans, monotremes lack extensive X inactivation, and marsupials exhibit paternally imprinted X inactivation (Ercan, 2015). In placental mammals, the trigger for X chromosome inactivation is *Xist*, while this gene is not conserved in marsupials (Furlan & Galupa, 2022). *Xist* is the X-inactive-specific transcript and one of the first long noncoding RNAs identified in the early 1990s (Borsani *et al*., 1991);(Brockdorff *et al*., 1991);(Loda & Heard, 2019). For X chromosome inactivation to occur the *Xist* gene is transcribed into RNA and spreads coating the entire X chromosome (Panning, 2008). *Xist* RNA recruits a plethora of chromatin-modifying factors leading to the structural reorganisation of the X chromosome resulting in the silencing of its >1,000 genes (Loda & Heard, 2019). This spreading event leads to the transformation of the inactivated X chromosome into an organised heterochromatic structure known as the "Barr body". However, ~12%-20% of human genes do escape gene silencing and are therefore expressed from both the active and inactive X chromosome and are thought to play a key role in female development and disease susceptibility. For example, human XO female embryos survive to term but are affected by Turner's syndrome (Loda & Heard, 2019). This is due to the deficiency of those X chromosomal genes which escape inactivation and for which two copies are necessary for normal female development.

ZFY has been found to play a role in meiotic sex chromosome inactivation (MSCI) (see section 1.7.2.1), and as a result, abnormal expression of *ZFY* due to sex chromosomal abnormalities can lead to spermatogenesis failure.

1.1.3 Consequences of Sex Chromosome Aneuploidy

Individuals with an XO chromosome complement develop as females due to the lack of the testis-determining gene SRY. This condition is commonly known as Turner's syndrome (Wilson, 1906);(Furman *et al*., 2020). Turner's syndrome was first described in 1938 as a consequence of a lack of Barr body (Mueller & Young, 1995). The presentation of Turner's syndrome can begin during pregnancy or later during adulthood with the main medical problems being short stature and ovarian failure (Ranke & Saenger, 2001). Other issues include lymphatic and skeletal abnormalities, with >60% of patients presenting with lymphoedema of the heads, feet and neck regions and is now used as a key diagnostic indicator (Atton *et al*., 2015). It is now more common for Turner's syndrome to be detected during the second trimester of pregnancy by an ultrasound (Mueller & Young, 1995). In Turner syndrome patients, the absence of SRY results in a female phenotype. However, the lack of the second sex chromosome in addition leads to phenotypes such as neck webbing, lymphedema, and horseshoe kidney. This implies that the genes responsible for these phenotypes are X-Y homologous and are present on both the X and Y chromosomes, suggesting that the X copy evades X chromosome inactivation.

Other circumstances where there is a loss or gain of sex chromosomes have also been noted. Another common example is Klinefelter's syndrome which was first described in 1942 (Mueller & Young, 1995);(Lanfranco *et al*., 2004). Klinefelter males present with an additional X chromosome (XXY), this has been identified in 1 in 1000 males making it a relatively common condition. Common symptoms of Klinefelter's syndrome include mild learning difficulties, being taller than average, gynaecomastia and being infertile (Mueller & Young, 1995). Trisomy X is the presence of an additional X chromosome (XXX) that has been identified in 0.1% of all females (Tartaglia *et al*., 2010). Generally, these individuals lack physical abnormalities resulting in only approximately 10% of cases being diagnosed. Key although minor physical characteristics include epicanthal folds, hypertelorism, upslanting palpebral fissures, clinodactyly, overlapping digits, pes planus, and pectus excavatum (Tartaglia *et al*., 2010). However, due to the surplus of X chromosomes, individuals with an extra X chromosome are more susceptible to autoimmune diseases (Loda & Heard, 2019). The upregulation of *Xist* leads to the silencing of two X chromosomes, potentially resulting in cell death (Loda & Heard, 2019).

1.1.4 Structure and Gene Content of the Human Y Chromosome

The human X and Y chromosomes both contain regions known as the pseudoautosomal regions (PAR) **(Figure 1.2)** and these regions still recombine during male meiosis ensuring to keep X-Y nucleotide sequence identity (B. Lahn & Page, 1999). The two pseudoautosomal regions are known as PAR1 and PAR2 and behave as autosomes (Mangs & Morris, 2007). PAR1 is a much larger region spanning 2.6Mb, whilst PAR2 only spans 320kb. PAR2 has been found to be non-essential for fertility and exhibits a much lower pairing and recombination frequency. However, many genes found in PAR1 escape X inactivation (Mangs & Morris, 2007). There are other regions on the X and Y chromosomes where this recombination event is suppressed, and these regions have become highly differentiated during evolution (B. Lahn & Page, 1999). This has resulted in only a few similarities persisting in these chromosomes. Most of these X-Y gene pairs are located on the short arm of the X-chromosome and are concentrated towards the distal end. Whilst, on the Y chromosome, the genes act as singletons and are dispersed throughout the euchromatic portion of the Y chromosome (B. Lahn & Page, 1999).

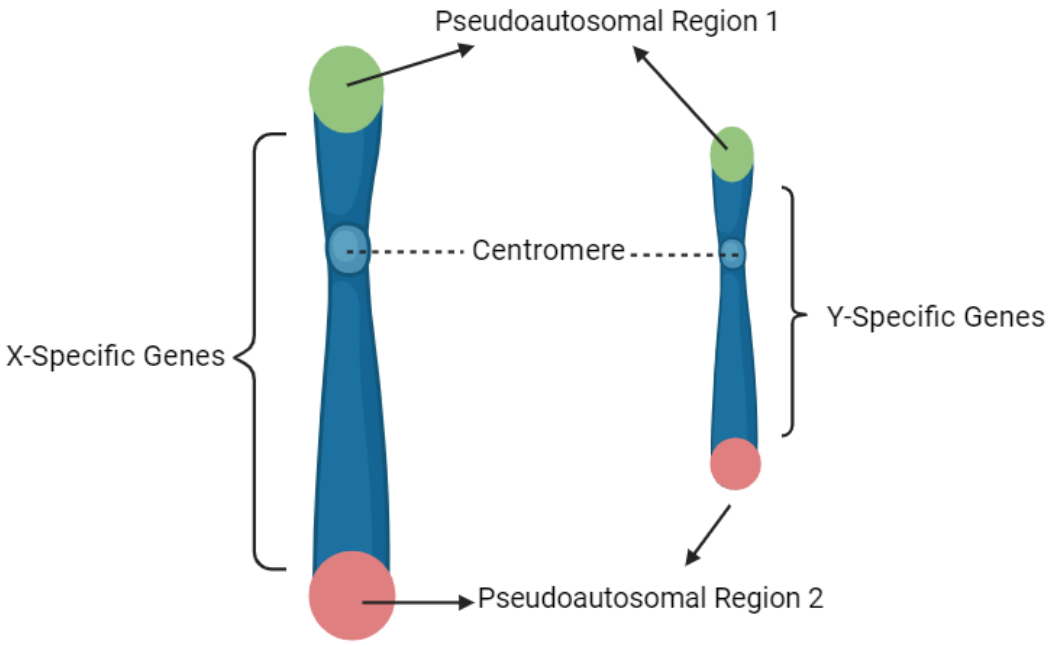


Figure 1.2: A schematic diagram demonstrating the location of the pseudoautosomal regions on both the X and Y chromosomes. These pseudoautosomal regions are homologous between the X and Y chromosome and can recombine acting in an autosomal fashion. Left chromosome: larger X chromosome, Right chromosome: smaller Y chromosome. Not to scale.

The Y chromosome has been conserved across nearly all eutherian and marsupials, with only a few mammals noted as Y-less (Holmlund *et al.*, 2023). In the human

genome, the Y chromosome is one of the smallest chromosomes with a size of ~60Mb (Quintana-Murci & Fellous, 2001). Previous papers have described the Y chromosome as wimpy and fragile due to it rapidly losing genes and shrinking PARs (Holmlund *et al*., 2023);(Ruiz-Herrera *et al*., 2022). These events are causing the Y functions to shift to autosomal chromosomes (Holmlund *et al*., 2023). At the current rate of decline, there is a possibility it may cease to exist within the next 10 million years (Holmlund *et al*., 2023). It was originally thought for many decades that the Y chromosome was a gene desert, with the only notable gene being SRY (Quintana-Murci & Fellous, 2001);(Holmlund *et al*., 2023). Nevertheless, advancements in genomics have revealed that the Y chromosome contains numerous genes essential for male reproduction. The functions of many of these Y chromosome genes remain undiscovered and require further exploration. The challenge in assigning spermatogenic functions to Y-specific genes arises from the limitations of current conventional methods for function mapping. Other problems are associated with the highly repetitive sequence structure of the Y chromosome (Holmlund *et al*., 2023).

The AZF regions on the Y chromosome are critical for male fertility, as they harbour genes necessary for spermatogenesis (Navarro-Costa, Plancha, *et al*., 2010). The first paper to name the AZF regions was published in 1996 where their investigations into Yq11 microdeletions associated with male sterility confirmed the presence of an AZF locus only in the distal Yq11 region (Vog *et al*., 1996). There are three AZF regions, AZFa, AZFb, and AZFc. **(Figure 1.3)**.

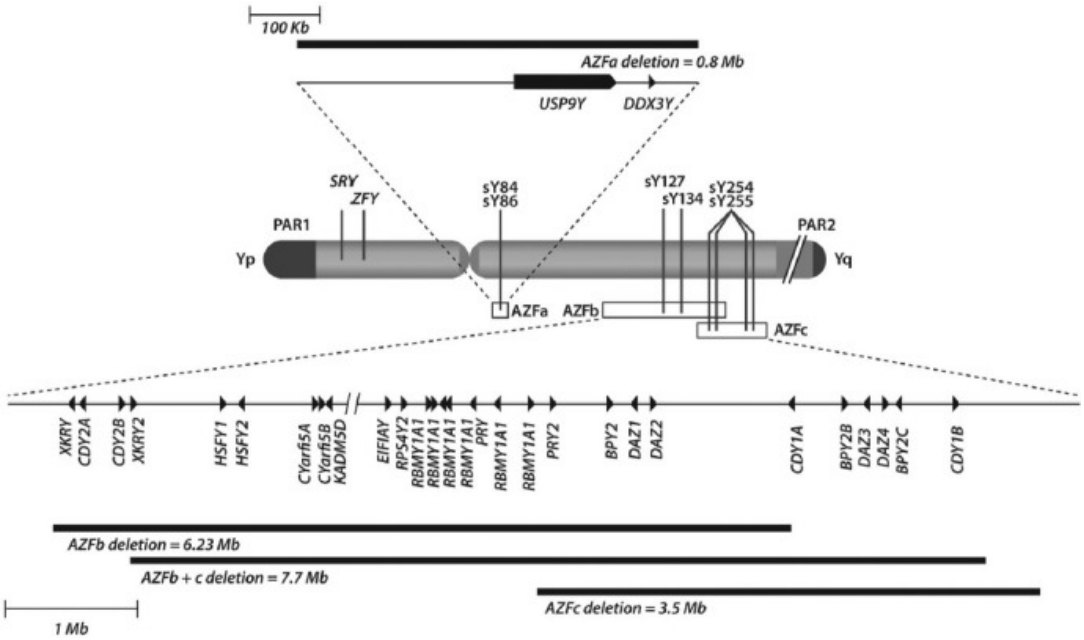


Figure 1.3: Schematic diagram of the Y chromosome AZF regions and gene content reproduced from (Esteves, 2015). This figure highlights the AZF loci on the

Y chromosome, consisting of the subregions AZFa, AZFb, and AZFc, and shows the genes located within each subregion.

Among these regions, complete deletions of the AZFa region have been demonstrated to have the most severe impact on spermatogenesis, leading to a Sertoli cell-only phenotype **(Table 1.1)** (Dicke *et al*., 2023). Sertoli cell-only syndrome is characterised by azoospermia in which the seminiferous tubules are lined only by Sertoli cells resulting in spermatogenesis failure (Gashti *et al*., 2021). AZFa contains single-copy genes;*USP9Y* and *DDX3Y*, with *DDX3Y* encoding a testis-specific RNA helicase. Functional data and studies suggest that *DDX3Y* may be the key spermatogenesis gene within AZFa (Dicke *et al*., 2023). Dicke and colleagues identified four potential pathogenic loss-of-function variants in *DDX3Y* through the sequencing of more than 1,600 infertile men, with one mutation identified as being *de novo.* Testicular biopsy for three of these cases was performed and identified that the patients suffered from Sertoli Cell-Only phenotype. This confirmed *DDX3Y* as the key spermatogenic factor in AZFa and should be used within a diagnostic workflow (Dicke et al., 2023).

AZFb spans a total of 6.23Mb and maps to ~18.1-24.7Mb of the Y (Navarro-Costa, Plancha, *et al*., 2010). The complete deletion of this extended genomic region on the Y chromosome leads to the loss of numerous Y genes including six protein-coding Y genes and is linked to meiotic arrest (Vogt *et al*., 2008). Deletions of the AZFb region account for 15% of the Y chromosome microdeletions, with AZFb microdeletions resulting in spermatogenesis arrest and azoospermia **(Table 1.1)** (Layman, 2012). The six protein-coding genes include *EIFA1Y, HSFY, PRY*, *RBMY1*, *RPS4Y* and *KDMD*(Vogt *et al*., 2021). In this thesis, *RBMY* is of interest because of its exclusive expression in male germ cells, particularly in premeiotic germ cells, with connections to functions in spermatogenesis and sperm motility (Vogt *et al*., 2021).

The AZFc region is located in the palindromes P1-P3 and consists of genes critical for sperm production (Rhie *et al*., 2023). This region is the cause of genomic variation across the male population, yet the consequences on spermatogenesis are unclear (Navarro-Costa, Gonçalves, *et al*., 2010). The AZFc and AZFb regions overlap at the proximal end of the AZFc region and the distal end of the AZFb region. Evidence indicates that AZFc deletions lead to significant spermatogenic defects, resulting in men with these deletions having a reduced sperm concentration of less than 1 million sperm/ml compared to the normal concentration of over 20 million sperm/ml **(Table 1.1)**. Deletions in the AZFc region make up approximately 60% of all documented AZF deletions, underscoring their importance (Navarro-Costa, Gonçalves, *et al*., 2010).

The emergence of AZFd remains a subject of controversy (Kent-First et al., 1999);(Yu *et al.*, 2015). While some studies suggest that the AZFd locus cannot be located between the AZFb and AZFc regions as currently understood, while other studies have reported deletions in AZFd in relation to male fertility (Yu *et al.*, 2015). Overall, when considering the structure of the Y chromosome, the AZFd region seems unlikely. Overall, the key spermatogenic genes are located in the AZF regions in the Yq11 of the Y chromosome, and deletions of these regions are detrimental to spermatogenesis and thus male fertility.

Table 1.1: AZF microdeletions reproduced from (Dobbs *et al.*, 2018). Microdeletion frequency in the different AZF regions and the prognosis of these deletions.

Mutation	Frequency	Genes affected	Prognosis
AZFa	Rare	DDX3Y and DBY or USP9Y	Sertoli only syndrome: no sperm
AZFb	Rare	RBMY1 PRY	No sperm
AZFc	10% of men with nonobstructive azoospermia, 1:4000 overall (Hotaling, 2014)	DAZ, gr/gr	Rarely sperm in ejaculate, 71.4% chance of finding sperm on mTESE. (Stahl et al., 2010)

1.2 Spermatogenesis

Given the enrichment of spermatogenesis genes on the Y chromosome, in understanding Y gene function it is important to consider the cellular and molecular events occurring during spermatogenesis.

1.2.1 Overview of Premeiotic Spermatogenesis

Spermatogenesis is defined as the process in which spermatogonia forms spermatozoa and begins during puberty in males **(Figure 1.4).** The process can be split into three distinct phases: (1) the proliferative phase (cells undergo rapid division), (2) the meiotic phase (genetic recombination and segregation), and finally (3) the differentiation of round spermatids into spermatozoa (Russell *et al.*, 1990). This entire process takes approximately 74 days in humans (Tenorio *et al.*, 2016). From the age of puberty, males produce millions of sperm per day and the process of spermatogenesis continues throughout adulthood (Russell *et al.*, 1990).

Spermatogonia are the starting point of spermatogenesis and are self-renewing mitotic stem cells located within the testis. From spermatogonia arises sperm necessary for male reproduction and fertility (Russell *et al*., 1990);(Wang *et al*., 2001). In males, there are three types of spermatogonia: stem cell spermatogonia, proliferative spermatogonia, and differentiating spermatogonia (Russell *et al*., 1990). Both stem cell and proliferative spermatogonia are also known as undifferentiated spermatogonia. Spermatogonia are located within the seminiferous tubules in contact with Sertoli cells. Once, undifferentiated spermatogonia differentiate these mature spermatogonia divide and enter the meiotic phase, forming the young primary spermatocytes, also known as preleptotene spermatocytes (Russell *et al*., 1990);(Cannarella *et al*., 2020).

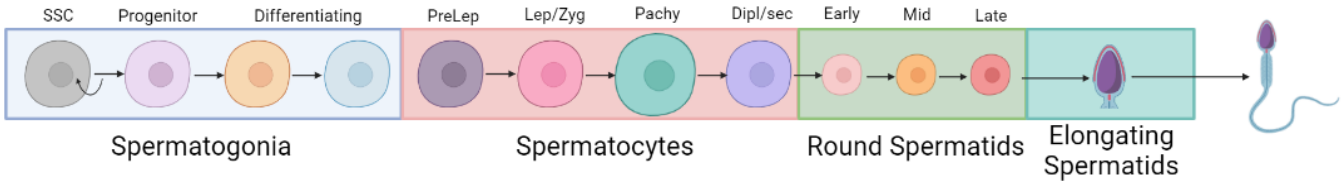


Figure 1.4: A schematic diagram showing the process of spermatogenesis from spermatogonia to sperm. Spermatogenesis is a long process in mammals and one cycle takes approximately 74 days in humans. Recreated from paper (Griswold, 2016).

1.2.2 Overview of Meiosis

The main stages of meiosis include genetic recombination, chromosome halving, and then increasing germ cell number resulting in haploid spermatids **(Figure 1.5)** (Russell *et al*., 1990).

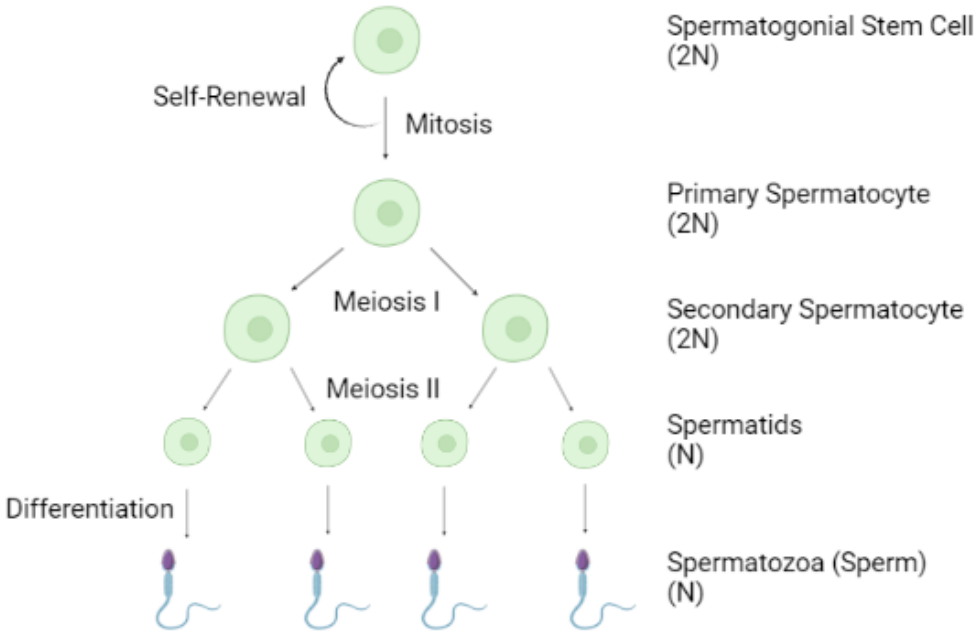


Figure 1.5: A schematic diagram showing the stages of spermatogenesis alongside the chromosome number present at each stage. Spermatogonial stem cells continuously undergo mitosis to ensure constant supply throughout adulthood, however, some of these stem cells undergo meiosis beginning the process of making functional spermatozoa with half the number of chromosomes compared to the stem cells following two stages of meiosis.

Prophase is the first meiotic division and is long-lasting, lasting roughly three weeks. Cell size and nuclei size increase during prophase. These changes in nuclei size and morphology are the basis of the stages of meiotic prophase. The prophase stage can be divided into five further subsections; preleptotene, leptotene, zygotene, pachytene, and finally diplotene **(Figure 1.4)**. Preleptotene spermatocytes persist for approximately one day, while the meiotic division itself takes less than an hour and these spermatocytes require close contact with Sertoli cells and the epithelium (Cannarella *et al*., 2020). The preleptotene to leptotene transition is defined by the cells moving away from the tubule and forming a rounded form along with the nuclei rounding. Homologous chromosomes during zygotene become paired via the synaptonemal complex, a tripartite structure. At the end of zygotene, the chromosomes become fully paired and this generally lasts around 1.5-2 weeks. Crossing over occurs which is the process of genetic recombination, resulting in the final spermatids containing unique genetic material (Zickler & Kleckner, 1998). During pachytene the nucleoli grow larger and sex vesicle forms. When the cells enter diplotene, the synaptonemal complex dissociates and the chromosome pairs separate except at the chiasmata region (Zickler & Kleckner, 1998). These diplotene

cells are also known as the largest primary spermatocytes. Following diplotene, the remainder of the first meiotic division occurs very quickly and passes through metaphase, anaphase, and telophase to complete the first meiotic division forming secondary spermatocytes (Gilbert, 2000) **(Figure 1.4/1.5)**. Secondary spermatocytes are very short-lived cells and rapidly enter the second meiotic division resulting in haploid spermatids (Gilbert, 2000).

1.2.3 Overview of Spermiogenesis

Once meiosis is complete, haploid spermatids enter the spermiogenic phase. It is during this phase that the flagellum develops, and this is composed of the midpiece, principal, and end pieces **(Figure 1.6)** (Russell *et al*., 1990). During flagellum development, mitochondria are recruited forming the middle piece of the flagellum and the outer dense fibres produce the midpiece as well as the principal piece. Another vital feature of spermatids is the acrosome, an essential system for egg penetration resulting in the restoration of the diploid condition (Berruti & Paiardi, 2011). This development is slow and is not fully complete until the very last stage of spermiogenesis. During these last stages of spermatid development, the spermatids undergo nuclear shaping, nuclear condensation, and cytoplasm elimination. This ultimately results in fully morphologically complete sperm that are shed into the lumen of the seminiferous tubules. From there they progress to the epididymis where they undergo final maturation in preparation for fertilisation (Russell *et al*., 1990).

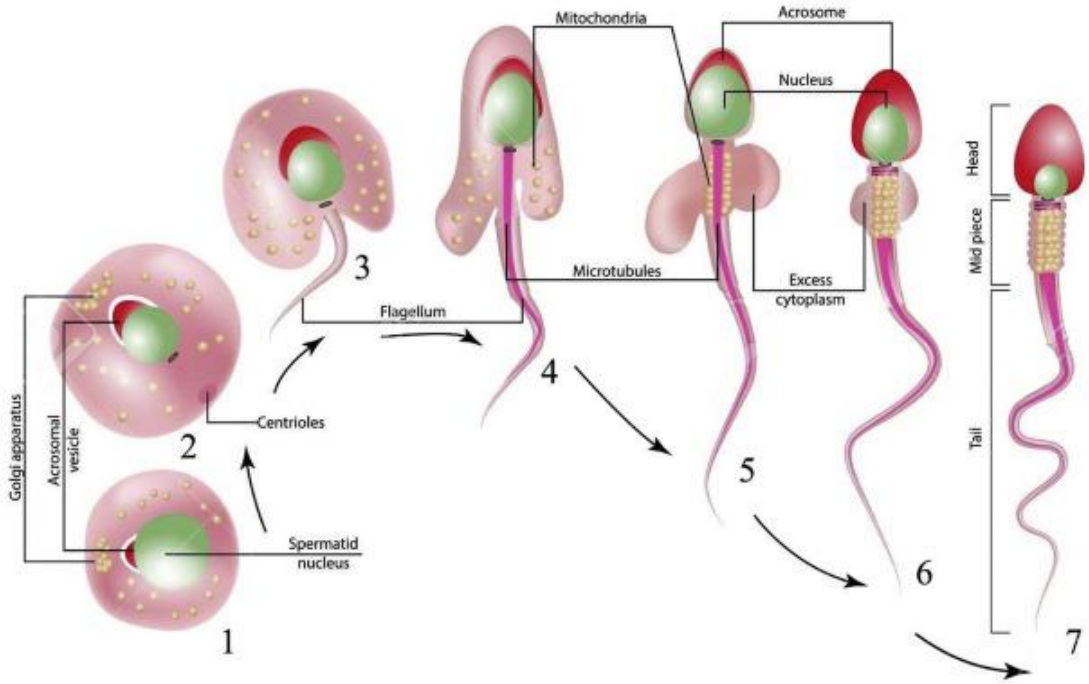


Figure 1.6: Morphologic changes at spermiogenesis (Oehninger & Kruger, 2021)**.** The cytoskeletal networks of Sertoli cells and germ cells are critical for the morphological changes that occur during spermiogenesis. (1) In the round spermatids

the acrosomal vesicle begins to form and this starts when the proacrosomal vesicles move from the trans-Golgi stacks to the nucleus. (2) In advanced round spermatids the acrosomal vesicle continues to develop and the Golgi apparatus begins to cluster initiating nucleus condensation. (3,4) Spermatid elongation starts with the formation of the head-to-tail coupling apparatus and the reshaping of the nucleus. Mitochondria are also organised in preparation for loading onto the outer dense fibres. (5,6) The elongated spermatids display a distinct acrosome, highly condensed nucleus, and excess cytoplasm (Oehninger & Kruger, 2021).

1.2.4 Meiotic Sex Chromosome Inactivation During Meiosis

Meiotic sex chromosome inactivation (MSCI) occurs during spermatogenesis and is a key contributor to a checkpoint process which eliminates germ cells with aberrant synapsis during mid-pachytene (Vernet, Mahadevaiah, de Rooij, *et al*., 2016). MSCI is a process of transcriptional silencing that results in the compartmentalisation of the X and Y chromosomes into the sex body (or XY-body) and the exclusion of RNA polymerase: a state that persists through the rest of pachytene and diplotene **(Figure 1.7)** (Turner, 2007). This compartmentalisation is mediated through chromatin condensation events (Cloutier & Turner, 2010). MSCI is necessary for the successful completion of spermatogenesis (Alavattam *et al*., 2022). Therefore, if germ cells during spermatogenesis fail to undergo MSCI they will be eliminated via apoptosis during that mid-pachytene stage of meiosis and is therefore a factor of meiotic sterility (Vernet, Mahadevaiah, de Rooij, *et al*., 2016). This process in some ways conceptually resembles that of X chromosome silencing in female somatic cells to equalise X chromosome dosage between males and females (Panning, 2008). However, it is mechanistically distinct in that the trigger for transcriptional silencing during pachytene is the lack of synaptic pairing between the axes of the diverged X and Y chromosomes (Royo *et al*., 2010). MSCI is thus seen specifically in males since they carry unpaired chromosomes, unlike females with paired X chromosomes. Where chromosomal rearrangements are present that impede synapsis, affected autosomal segments may be silenced in either male or female meiosis – in this case, the silencing is known as meiotic silencing of unpaired chromatin (MSUC) (Manterola *et al*., 2009);(Royo *et al*., 2010).

As mentioned above in 1.2.1, the earliest identifiable cell type in men is the spermatogonial stem cells, which undergo mitosis to ensure sufficient self-renewal. At this stage in spermatogenesis, genes present on both the X and Y chromosomes are both transcriptionally active (Turner, 2007). The germ cells enter meiosis and during this many events occur in which again both X and Y chromosomes are still transcriptionally active and continue to remain, so until after the zygotene-to-pachytene transition when meiotic synapsis is complete and both the X and Y

chromosome are silenced and compartmentalised into the sex bodies, peripheral nuclear subdomain **(Figure 1.7)**. Following the formation of sex bodies MSCI persists continuing through into pachytene and diplotene. After exiting meiosis, most X and Y genes remain repressed whilst some are reactivated during spermatid formation (Turner, 2007).

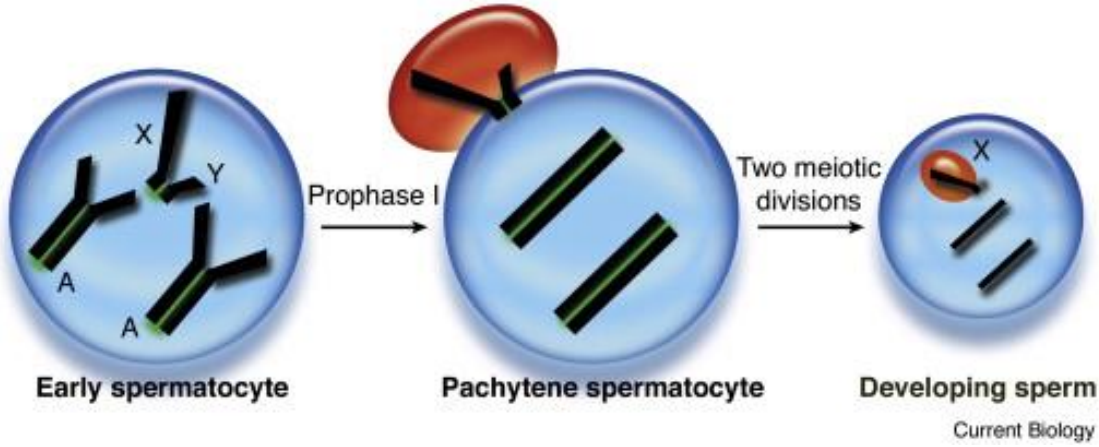


Figure 1.7: Meiotic sex chromosome inactivation schematic (Cloutier & Turner, 2010)**.** Early spermatocytes contain homologous chromosomes (black) which begin to synapse (shown in green). At this early stage, the autosomes (A) and the X and Y chromosomes are transcriptionally active. The autosomes are fully synapsed in pachytene spermatocytes and active, whilst the X and Y chromosomes only synapse at a small region of homology. This means that the exposed, unsynapsed regions are subject to MSCI resulting in the formation of the sex body (red). This means that the resulting developing sperm consists of a transcriptionally repressed X chromosome (Cloutier & Turner, 2010).

1.2.5 Evolution of Spermatogenic Function

There is rapid evolution of both the morphological and molecular aspects of spermatogenesis in mammals (Murat *et al*., 2023). These changes are probably due to evolutionary pressures on men to be reproductively successful. It has been found that the rapid evolution of the testes is accelerated by fixation rates of the following; gene expression changes, amino acid substitutions, and new genes in late spermatogenic stages (Ramm *et al*., 2014). It is thought that these changes could have been pushed by haploid selection, chromatin remodelling, and reduced pleiotropic constraints. Genes across species have been identified to show temporal expression changes whilst others showed conserved expression controlling ancestral spermatogenic processes. As a result of these changes, traits across mammals such as testis size, sperm production rates, sperm morphologies, and other cellular traits vary (Ramm *et al*., 2014). This is still true between closely related species. Gene expression comparison has shown this high evolutionary rate in testes is possibly due

to purifying selection. It was also suggested that testis-specific expression tends to be enriched with genes under positive selection, with new genes emerging during evolution predominately in the testis both likely contributing to the rapid phenotypic evolution (Murat *et al*., 2023).

Looking specifically at chromatin remodelling during spermatogenesis has shown that this process leads to leaky transcription within the genome, which could be a causing factor of transcription resulting in the frequent emergence of new testis-expressed genes and alternative exon splicing events during evolution, such as those of *ZFY* (Soumillon *et al*., 2013);(Murat *et al*., 2023). Sex chromosomes were the result of the differentiation of ancestral autosomes, and this differentiation resulted in the emergence of MSCI in both eutherians and marsupials (Turner, 2015). This emergence increased the gene copy number to substitute for parental genes located on the X chromosome during meiosis under X chromosome inactivation to compensate for dosage. Even with this X chromosome dosage compensation evolution has led to many testis-expressed genes located on the X chromosome (Murat *et al*., 2023).

Not only is there an evolutionary change in the genome but there are also physical changes associated with the sperm themselves. But why is this since all sperm function to fertilise an egg yet there is such a great sperm morphology difference across mammals (Ramm *et al*., 2014)? These changes have been suggested to be a result of sperm competition, which occurs when two or more males compete to fertilise a female's eggs (Ramm *et al*., 2014). One observed morphological difference in sperm across species is sperm length. The current explanation of this evolutionary difference is the environment where fertilisation occurs (Kahrl *et al*., 2021). For instance, species with longer sperm typically deposit sperm directly into the female, while species with shorter sperm often inhabit aquatic environments where external fertilisation occurs (Kahrl *et al*., 2021).

Spermatogenesis is a complex and dynamic cellular process, as evidenced by a study examining the single-cell transcriptome throughout mammalian spermatogenesis (Hermann *et al*., 2018). This analysis profiled gene expression from spermatogonial stem cells through spermatids. This type of information-rich, single-cell analysis of spermatogenesis provides valuable resources for investigating male meiosis, testicular cancer, male infertility, and contraceptive development.

ZFY has been shown to be vital for spermatogenesis in mice, but a former master's student in the Fenton-Ellis lab identified the possible expression of the short-testis-specific *ZFY* splice-variant in a head and neck cancer cell line, leading to the

hypothesis that *ZFY* may also have a cancer-specific role outside of the testis, which was investigated in this thesis.

1.3 Cancer-Testis Antigens

A cancer-testis antigen refers to a distinct category of tumour-associated proteins, typically found in male germ cells but not in somatic cells yet observed with irregular expression across various cancer types (Salmaninejad *et al*., 2016). Many of the cancer-testis antigens are encoded on the X chromosome and have historically been termed "X-CT genes" to differentiate from those encoded on autosomes (non-X-CT genes). (Feichtinger *et al*., 2012);(Nin & Deng, 2023). Due to their naturally reduced expression profile, they are being investigated as possible cancer biomarkers and targets for immunotherapy strategies (Salmaninejad *et al*., 2016).

The first human cancer-testis antigen was discovered in 1991 by Thierry Boon and colleagues using cDNA expression (van der Bruggen *et al*., 1991);(Ward *et al*., 2016). Initially categorised as a melanoma antigen, the gene was later named *MAGE1*, followed by the discovery of two additional family members, *MAGE2* and *MAGE3* (Nin & Deng, 2023). Since then, significant progress in high-throughput PCR and sequencing methodologies has unveiled more than 200 cancer-testis antigens. However, the mechanisms underlying the activation of these genes during tumorigenesis remain poorly understood (Nin & Deng, 2023).

Common patterns are observed in both tumour progression and germ cell growth and development, including invasion, migration, apoptosis resistance, immune subversion and angiogenesis (Salmaninejad *et al*., 2016);(X.-F. Li *et al*., 2020). Cancer-testis antigens are crucial for fertility in the male phenotype, with knockout experiments showing impaired fertility. Links have been made to functions including sperm metabolism, sperm RNA regulation, sperm movement and meiosis in sperm cells, however, they also are key players in tumour invasion and metastasis (Salmaninejad *et al*., 2016). This might also be attributed to the immune privilege of the testicular region generated by the blood-testis barrier, leading to the immune system failing to recognise cancer-testis antigens as self-proteins (Sammut *et al*., 2014);(Jay *et al*., 2021). When these antigens are expressed outside the testes, they trigger an immune response, fostering a cancer-specific reaction. Consequently, this supports the justification for employing them as targets for immunotherapy (Jay *et al*., 2021).

Multiple studies have shown that cancer-testis antigens drive multiple cellular pathways leading to cancer phenotypes in human cells suggesting an involvement in initiating or reactivating hallmarks of cancers **(Figure 1.8)** (Nin & Deng, 2023). Take, for instance, numerous cancer-testis antigens like *SPO11*, *TEX15*, and *SYCP1/3*,

which play roles in meiosis, a predictable association considering that germ cells are the sole site of meiotic cell division. These genes have distinct functions within meiosis, including involvement in DNA damage and repair response pathways during meiotic division, meaning that aberrant expression can potentially lead to abnormal chromosome segregation, aneuploidy, genomic instability, and mutations, characteristic features of cancer. Consequently, their expression in somatic or cancer cells has been linked to tumorigenesis, cancer advancement, or resistance to therapy. The MAGE family has demonstrated involvement in apoptosis regulation by impeding p53 promoter binding (Mei *et al.*, 2020);(Nin & Deng, 2023). In multiple myeloma, *MAGE3* was discovered to hinder apoptosis by activating p53-dependent BAX. This further underscores the link between cancer-testis expression and a pivotal cancer hallmark: resistance to apoptosis (Nin & Deng, 2023). This further implies that genes involved in controlling meiotic chromosome behaviour, regulating germ cells, and directing gametogenic development could potentially exhibit significant cancer-causing effects if they are expressed abnormally in non-reproductive somatic cells (Sammut *et al.*, 2014).

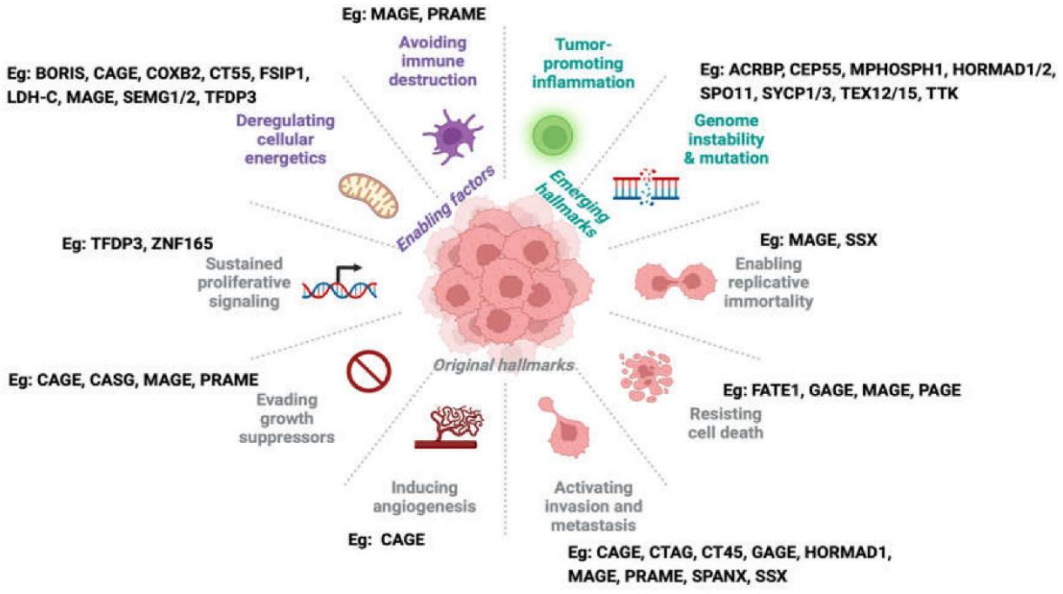


Figure 1.8: The Hallmarks of Cancer (Nin & Deng, 2023)**.** Examples of cancer-testis antigens with roles associated with each cancer hallmark.

Consequently, due to the close connection between the roles of cancer-testis genes and the hallmarks of cancer ectopic activation of a germline programme may drive cells towards tumorigenesis. This could potentially also explain the differential prevalence of certain cancers in men and women due to the presence of cancer-testis antigens on the different sex chromosomes. This highlights the need for further investigations into this class of tumour-associated proteins to develop further cancer-

testis antigen anticancer strategies. Especially given the limited understanding of the temporal relationship between meiotic gene expression and cancer development, a critical question arises: are meiotic genes activated late in the development simply due to the accumulation of cellular disruption and/or DNA damage, or do these genes actively drive cancer advancement?

1.4 Differential prevalence of cancers between males and females

Cancer susceptibility and cancer survival have been linked to sex differences, which have been reported to affect mutational burden, DNA repair, epigenetics, metabolism, tumour suppressor activity, cell cycle regulation and immunity (Rubin, 2022). These sex differences are different to sexual dimorphisms which refer to features such as ovaries and testes and differences between sexes such as height (Cook *et al*., 2012). Therefore, genetic, epigenetic and gonadal hormone actions are sexual differentiations which can lead to differences across disease prognosis, diagnosis and pathogenesis (Cook *et al*., 2012). However, what these differences mean for cancer risk, treatment response and prognosis are currently unknown and require further investigations (Rubin, 2022).

While behaviours (smoking & alcohol), anthropometrics (body mass index), lifestyle (physical activity & diet), and demographic factors between males and females have been investigated and could explain the male predominance in at least 21 cancer sites, these cannot be the sole factors at play (Jackson *et al*., 2022). Three potential explanations to explain this predominance are currently available.

One possible explanation is that there may be X-linked tumour suppressor genes that escape X inactivation (Rubin, 2022). Multiple essential tumour suppressors are located within the non-pseudo-autosomal region of the X chromosome, resulting in a lack of genetic and function equilibration between males and females. Examples include *KDM6A, ATRX, DDX3X* and *TLR6/7*, with a higher dose in women. These genes have important epigenetic and tumour suppressor functions including immune surveillance and cancer cell elimination, indicating that women have greater protection compared to men (Rubin, 2022). The second explanation is Y-linked proto-oncogenes, which are inherently male-specific and can only be activated to trigger tumorigenesis in males such as *TSPY* and *RBMY* which have both been found to be ectopically expressed in diseased somatic cells (Kido & Lau, 2015). Since these genes are Y chromosome-specific they could potentially influence the development, progression and outcomes of male-predominated cancers (Kido & Lau, 2015). Finally, as mentioned in section 1.3 male germline genes have been shown to be wrongfully activated in cancers encoding cancer-testis antigens thus driving tumorigenesis

(McFarlane *et al*., 2014). The interest surrounding this category of cancer antigen has increased since their cancer-restricted profile makes them potentially a beneficial cancer biomarker for diagnosis, prognosis and subsequent immunotherapy (McFarlane *et al*., 2014).

1.5 Preliminary Observation: *ZFY* is Potentially Associated with Cancer

Given the above links between cancer and spermatogenesis, in searching for genetic factors that could explain the excess male prevalence of certain cancers, it makes sense to focus on sex linked genes, testis-expressed genes, and in particular genes that are known to have a role in apoptosis and/or chromatin remodelling. This thesis focuses on a zinc finger protein, *ZFY*, that falls within all three of those categories. *ZFY* is a Y-linked transcription factor with links to germ cell development, meiosis and apoptosis (see below), but its exact function remains poorly understood. Previous work in the Ellis-Fenton lab at the University of Kent showed altered splicing of *ZFY* in Human Papillomavirus (HPV) negative oropharyngeal squamous cell carcinoma (OPSCC) cell lines (Trujillo, 2019). Given the connection to apoptosis, this could potentially explain the male prevalence of head and neck squamous cell carcinomas (HNSCCs).

1.6 Zinc Finger Proteins

The Zinc finger protein family has been identified to play a significant role in spermatogenesis (Vickram *et al*., 2021). This group of transcription factors have been shown to have a critical role in proliferation and differentiation during spermatogenesis.

In the late 1980s the first Zinc finger, Transcription Factor IIIa (TFIIIa) was discovered in the *Xenopus laevis* (Cassandri *et al*., 2017). This resulted in the downstream discovery of a new group of transcriptional activator proteins containing 30 amino acid repeat regions. Currently, 30 types of zinc-finger proteins have been designated based on their zinc-finger domain structure (Cassandri *et al*., 2017).

The classical C_2H_2 (Krüppel-type) zinc finger proteins form the largest mammalian regulatory protein family making up 1% of the total mammalian proteins (Iuchi, 2001) and almost half of the annotated transcription factors in the human genome (Emerson & Thomas, 2009). Their vast presence is demonstrated by 133 different C_2H_2 cDNA identified in the brain alone (Iuchi, 2001). Zinc finger proteins are involved in a range of cellular activities, these include development, differentiation, and tumour suppression. Zinc fingers seem to have a vital role which explains why they are also found in lower eukaryotes and prokaryotes (Iuchi, 2001).

Great variation is seen across the zinc-finger family, with differences in structure identified. Differences in cysteine/histidine combinations lead to the identification of non-classical types of zinc finger proteins **(Figure 1.9)** (Cassandri *et al*., 2017).

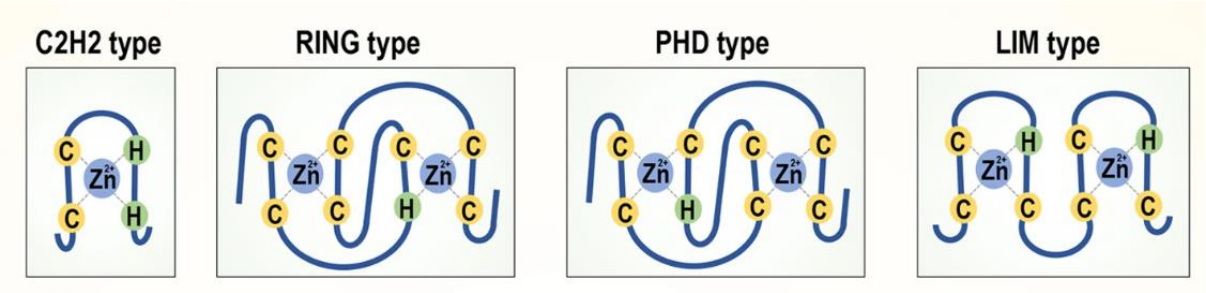


Figure 1.9: A schematic representation of the zinc finger domain subtype structures (Cassandri *et al*., 2017)**.** RING-type: really interesting new gene, PHD-type: plant homeodomain and LIM-type: Lin-ll, Isl-1, and Mec-3.

1.6.1 The Structure of Zinc Finger Proteins

The majority of zinc finger proteins consist of an N-terminal protein interacting domain, and a C-terminal DNA binding domain (DBD) which is generally where the zinc finger motifs are located (Emerson & Thomas, 2009). The N-terminal interacts with other proteins in order to regulate transcription, whilst the C-terminal region binds to DNA. Within the N-terminal domain around 40% of the human ZF members also contain an N-terminal Krüppel-associated box (KRAB) domain. This KRAB domain functions to repress transcription via the recruitment of KAP-1, which results in chromatin modification and gene silencing (Emerson & Thomas, 2009). Like other proteins, the binding potential depends on the amino acid sequence of the finger domains in the C-terminal and also the linkers between the fingers (Iuchi, 2001). The number of zinc fingers in tandem can vary from one to more than thirty (Iuchi, 2001). In contrast to the KRAB-ZF family of transcriptional repressors, other ZF proteins contain acidic activating domains (AADs) and promote rather than repress transcription (Emerson & Thomas, 2009). AADs have been identified as disordered regions of transcription factors that bind the coactivators needed for transcription, for example, the Med15 subunit of Mediator (Sanborn *et al*., 2021);(Staller *et al*., 2021). In general, there seems to be a lack of sequence conservation within AADs, yet they still manage to make specific protein-protein interactions with the necessary transcriptional machinery (Melcher, 2000). However, there is conservation in the proposed targets of AADs. Research has demonstrated a two-step mechanism whereby activator regions first stimulate the assembly of the transcriptional machinery on DNA, followed by overexpression of the activator regions themselves, which inhibits their initial stimulatory effect through a negative feedback mechanism.

Therefore, the AAD region of transcription factors is vital for the recruitment of the necessary transcriptional machinery (Melcher, 2000).

It has been found that many mammalian transcription factors interact specifically with transcription factor II D (*TFIID*) composed of the TATA-binding protein (*TBP*) and the TBP-associated factors (TAFs) (S. Piskacek *et al*., 2007). Within the AAD regions, nine amino acid transactivation domain (9aaTAD) motifs can be identified, and these are thought to be the specific binding regions within the domain (M. Piskacek *et al*., 2016). The 9aaTAD domains are recognised by transcriptional machinery from yeast to man, and there are many prediction algorithms available to locate these motifs within a sequence (M. Piskacek *et al*., 2016). 9aaTADs have been found to be required for the function of the AADs for many eukaryote transcription factors and are vital for the transactivation function of many transcription factors (S. Piskacek *et al*., 2007). The vast majority of DNA-binding motifs in eukaryotes consist of zinc finger domains (Isernia *et al*., 2020). A zinc finger domain consists of a zinc ion, coordinated by cysteines and histidines in varying combinations **(Figure 1.10)** (Cassandri *et al*., 2017). This forms a complex and compact ββ α-structure consisting of two β-sheets and one α-helix which was identified by crystallographic studies (Isernia *et al*., 2020).

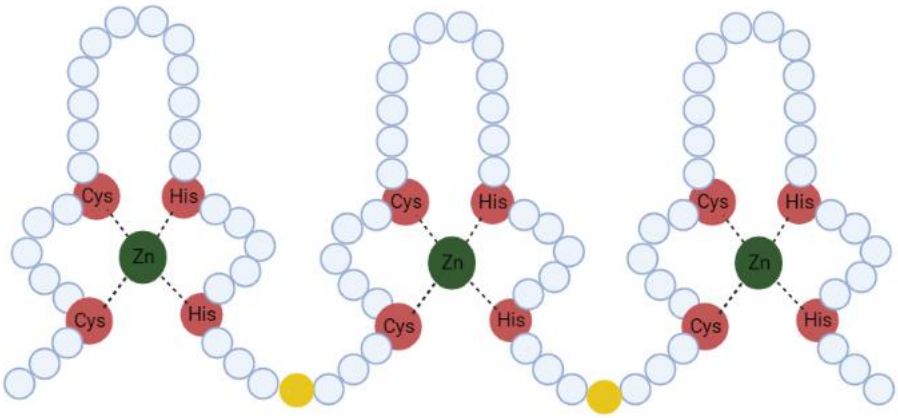


Figure 1.10: A schematic diagram of C_2H_2 zinc finger domains in tandem. Zinc fingers consist of the zinc ion core surrounded by two cysteine and two histidine residues.

Given the focus on *ZFY*, this project focuses on the C_2H_2 group of zinc finger proteins. The consensus sequence of C_2H_2 fingers is $CX_2CX_3FX_5LX_2HX_3H$, consisting of hydrophobic residues contained within the core except for the two cysteines and two histidine residues (X. Li *et al*., 2022). It is believed that the C_2H_2 domain motif targets a three-base pair sequence (X. Li *et al*., 2022). Three main C_2H_2 subgroups have been identified via their number and pattern, these three groups are; triple-C_2H_2, multiple-adjacent-C_2H_2, and separated-paired C_2H_2 finger proteins (Iuchi, 2001). The

various C_2H_2 zinc finger subgroups have different binding capabilities. The triple-C_2H_2 and multiple-adjacent-C_2H_2 subgroups can bind multiple different ligands, whereas the separated-paired fingers subgroup binds its target using a single finger pair. The amino acid residues in the alpha helices of the zinc fingers allow high affinity binding to their target DNA segments, which is their primary role. The zinc fingers can then control the transcription of target genes along with other participating factors (Iuchi, 2001).

It has been suggested that the greater the number of zinc fingers, the more specific the affinity is for ligands (Iuchi, 2001). It is also possible that proteins containing a large number of tandem zinc fingers bind only one specific target rather than multiple targets using different subsets of the available zinc fingers (Emerson & Thomas, 2009). The size and diversity of the zinc finger protein family is surprising large even with the ancestral size being small (Emerson & Thomas, 2009). The KRAB-specific domain was identified to have first arisen in the tetrapod vertebrates (Emerson & Thomas, 2009);(Bellefroid *et al*., 1991);(Birtle *et al*., 2006) with this event being recognised as a key subject of the vertebrate lineage-specific expansion (Emerson & Thomas, 2009).

The zinc finger structure provides the framework for versatile gene targeting allowing the zinc finger family group to have a diverse array of functions.

1.6.2 The Biological Functions of Zinc Finger Proteins

The functions of zinc finger proteins are reliant on the presence of zinc finger domains as they bind to target promoter regions, enabling the subsequent activation or inhibition of the target (X. Li *et al*., 2022). Zinc finger proteins have a wide array of roles within the human body, including involvement in processes like development, differentiation, metabolism, transcriptional and post-transcriptional regulation, activation, protein degradation, and signal transduction. The varied combinations and functions exhibited by these zinc finger proteins underscore their versatility within the biological system (S. Liu *et al*., 2022);(X. Li *et al*., 2022);(Iuchi, 2001). Many noted zinc finger protein functions relate to cellular biological processes with links to cancer progression, tumour invasion, and metastasis (S. Liu *et al*., 2022).

Some Zinc finger proteins have been identified to have physiological roles in the skin such as *KLF4*, a C_2H_2-type transcription factor (Cassandri *et al*., 2017). In a mouse study knocking down KLF4, a crucial role in keratinocyte differentiation was identified as the absence of the protein resulted in altered skin barrier formation (Segre *et al*., 1999). *SLUG*, another C_2H_2-type transcription factor has also been shown to be involved in adipocyte differentiation, with other zinc finger transcription factors noted

to have important roles in the intestine, muscles, and cellular stemness regulation (Cassandri *et al*., 2017);(Antonio Pé rez-Mancera *et al*., 2007).

Although zinc finger proteins are vital for health, they have also been shown to have an important role in diseases, such as cancer onset and progression (Cassandri *et al*., 2017). This is because zinc finger proteins are involved in key cancer progression pathways ranging from carcinogenesis to metastasis formation. Evidence suggests that zinc finger proteins can recruit chromatin modifiers and function as structural proteins regulating cancer cell migration and invasion. For example, *ZNF281* has been identified as a major player in tumorigenesis and tumour invasion due to its role in the DNA damage response process and the epithelial-mesenchymal transition (Cassandri *et al*., 2017);(Pieraccioli *et al*., 2016). Many other zinc finger proteins have been identified as cancer-driving.

1.7 Zinc Finger Y-Chromosomal Protein

Zinc Finger Y-Chromosomal protein (*ZFY*) is encoded by the *ZFY* gene located on the Y chromosome (Yp11.2) and *ZFY* protein belongs to the Krüppel-type family of C_2H_2-type zinc finger proteins (Koopman *et al*., 1991). An X chromosome homologue of the gene exists in females and is known as *ZFX*. *ZFX* and *ZFY* appear to have diverged from a common ancestral gene prior to the diversification of placental mammals, as evidenced by their sequence homology. However, *ZFX* and *ZFY* have evolved distinct functional roles based on their locations on the X and Y chromosomes, respectively (Schneider-Gadicke *et al*., 1989). *ZFY* is a highly conserved transcription factor expressed in many tissues in placental mammals and has been suggested to have a role in spermatogenesis regulation, but these have not yet been fully established and understood (Romano *et al*., 2017);(Holmlund *et al*., 2023). *ZFY* was previously thought and suggested to be the sex-determining gene, however, this was later disproved and SRY was identified as the sex-determining gene. Due to the lack of sex determination function, the *ZFY* gene has very much disappeared from public interest with research taking a back seat for almost two decades. Although it is not the sex-determining gene, *ZFY* has been recognised as pivotal in safeguarding the Y chromosome. *ZFY* acts as the overseer of meiotic surveillance, thus contributing to the preservation of the Y chromosome and averting its potential extinction (Waters & Ruiz-Herrera, 2020);(Holmlund *et al*., 2023).

In humans, two splice variants have been identified; the full-length version (*ZFYL*) containing 8 exons (7 coding exons, **Figure 1.11**) expressed ubiquitously and a testis-specific short-spliced version (*ZFYS*) lacking the third exon (second coding exon –

573bp) (Decarpentrie *et al*., 2012). This short transcript was identified via reverse transcription polymerase chain reaction (RT-PCR) (Decarpentrie *et al*., 2012).

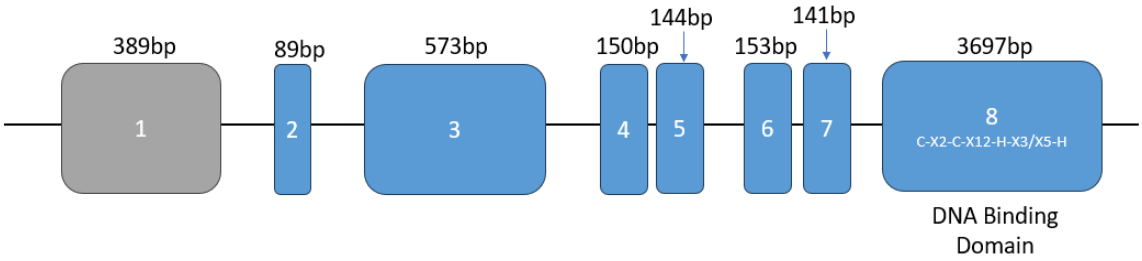


Figure 1.11: A schematic diagram of the Human *ZFY* gene transcript. *ZFY* consists of one non-coding exon shown in grey and 7 coding exons shown in blue. *ZFY* also consists of a large N-terminal acid activating domain spanning exons 2 to 6 and a C-terminal DNA binding domain located in exon 8. The introns are highlighted as a black line. The total gene including introns is 47.13kb but the exon length is only 5.336kb.

Roles of both variants have been speculated but no studies have confirmed their separate functions. Prediction methodology has suggested that *ZFY* is located intracellularly within the nucleoli and nucleoplasm (*ZFY*: The Human Protein Atlas, 2024). The exact function of these different variants is generally unknown, and both variants seem to have different expression patterns within males.

ZFY has been most closely investigated in mouse models where two paralogous Y-linked copies have been identified; *Zfy1* and *Zfy2* (Holmlund *et al*., 2023). *Zfy1* is similar to human *ZFY* and produces both long and short isoforms, whilst *Zfy2* expresses almost exclusively the long isoform. In mice *Zfy1* and *Zfy2* are expressed in spermatocytes during meiotic prophase I leptotene and zygotene stages (Vernet, Szot, *et al*., 2014). Subsequently, both genes are then silenced at the onset of MSCI, this silencing is essential for pachytene progression (Royo *et al*., 2010). A secondary surge in expression post-meiosis was identified. Following meiosis, *Zfy2* exhibits greater expression in spermatids compared to *Zfy1* (Vernet, Szot, *et al*., 2014);(Decarpentrie *et al*., 2012);(Holmlund *et al*., 2023). Studies in mice have demonstrated a shift in promoter activation following meiosis. Prior to MSCI, the expression of *Zfy1* and *Zfy2* is regulated by their homologous *Zfy* promoter. However, after meiosis, this regulation switches, and *Zfy2* expression is driven by a potent, spermatid-specific promoter possibly explaining the shift in *Zfy2* expression post-meiosis. A splicing pattern similar to what has been demonstrated in humans is also seen in mice. Two *Zfy1* transcripts have been identified; a short transcript missing exon 4 and a long transcript retaining exon 4. Exon 4 is homologous to exon 3 in human *ZFY* which is alternatively spliced out to form the short isoform and is therefore

equivalent to that of the mouse short *Zfy* transcript, with both short transcripts also being confirmed as testis-specific (Decarpentrie *et al*., 2012). The short *Zfy1* transcript was most prominent in spermatocytes, whilst the long was prominent in spermatids. A short *Zfy2* transcript was also identified but the expression of this transcript is at very low levels compared to the long *Zfy2* form (Holmlund *et al*., 2023). In humans, northern blots and RT-PCR have confirmed that the *ZFYS* form is expressed pre-meiotically, and its expression is testis-specific, whilst the *ZFYL* form has been identified to be expressed post-meiotically, but its expression is ubiquitous (Decarpentrie *et al*., 2012). This splicing pattern has been identified in other animals, such as sheep suggesting that *ZFY* splicing is a highly conserved event (Holmlund *et al*., 2023).

The full-length mouse isoform has been shown to possess transactivation ability within a yeast reporter system, whilst the short-spliced version has no such detectable ability shown in **Figure 1.13** (Vernet, Mahadevaiah, de Rooij, *et al*., 2016). It is suspected that the majority of *Zfy* transactivation in mice comes from the *Zfy2* isoform.

Further work looking into mouse *Zfy* has also shown that *Zfy1* and *Zfy2* both trigger germ cell apoptosis and that subsequent silencing of both the *Zfy1* and *Zfy2* are necessary for pachytene progression (Vernet, Szot, *et al*., 2014);(Vernet, Mahadevaiah, de Rooij, *et al*., 2016). The differences between mouse and human *ZFY,* such as the existence of two paralogs and their restricted testis expression, causes some downstream problems when looking into their function and clinical relevance therefore, throughout this work, the focus has not been on the mice *Zfy* forms as identifying possible clinical relevance in humans was a key aim.

ZFY structure follows the classical Zinc finger transcription factor structure, and it is presumed to have major importance in the roles it performs.

1.7.1 The Structure of *ZFY*

The overall structure of *ZFY* is one of a transcription factor as it contains a large N-terminal AAD and a C-terminal DBD. Within the DBD, *ZFY* has thirteen zinc fingers, specifically C_2H_2-type zinc fingers. Because of these identified structures, it is believed that *ZFY* is a possible eukaryotic transcription factor.

The *ZFY* zinc fingers follow the C-X2-C-X12-H-X3/X4-H pattern, therefore slightly differing from other C-X2-C-X12-H-X3-H zinc fingers. In this sequence structure; X is an amino acid, and the number indicates the number of residues, with C indicating a cysteine residue and H indicating a histidine residue. The number of amino acids between the histidines in *ZFY* zinc fingers alternate between three and four amino

acids which is expected to be the result of a duplication event. The C-X2-C-X12-H-X3-H zinc finger pattern is commonly referred to as the poly-F as there are normally at least 4 zinc finger repeats in tandem however, *ZFY* contains 13 zinc fingers in tandem (Emerson & Thomas, 2009).

ZFY as mentioned consists of eight exons, of which seven are coding exons. The large AAD is located within exons two to six, whilst the DBD is encoded by a single exon, exon eight. *ZFY* in total is 801 amino acids long (90.5KDa) with a very negative predicted charge of -16 and an isoelectric point between 5.65-5.99. However, the short-spliced variant which was identified via reverse transcription polymerase chain reaction (RT-PCR), lacks the second coding exon which encodes half the acidic domain and hence why it is referred to as *ZFYS* form **(Figure 1.12)**. *ZFYS* is missing the 573bp second coding exon, which makes the short form a total of 191 amino acids shorter than that of the *ZFYL* form **(Figure 1.12)**.

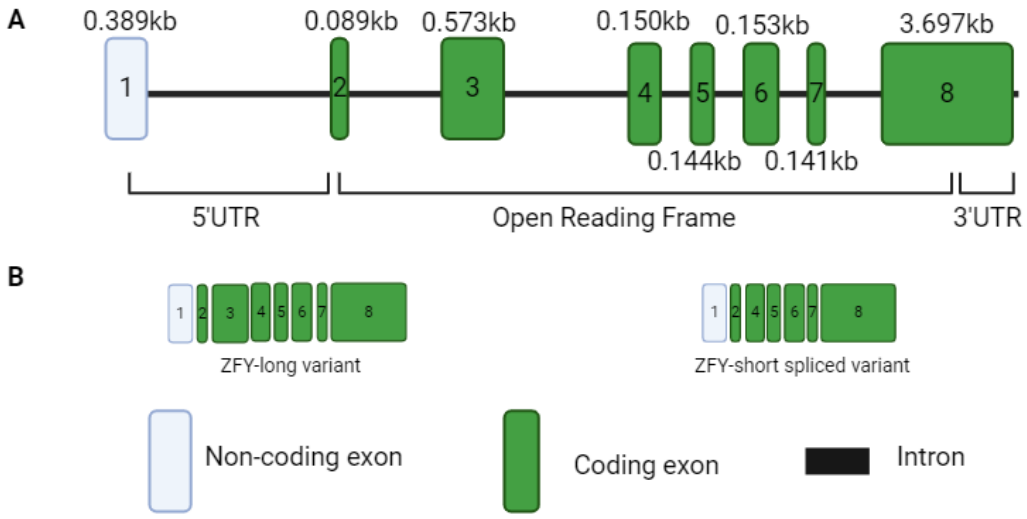


Figure 1.12: A schematic comparing the two isoforms of *ZFY*. A: Genome organisation of *ZFY*. **B:** The two ZFY transcripts following splicing in which *ZFY-short* lacks the third exon which is 0.573kb indicated by the cross, making the total exon length 4.753kb.

As previously mentioned in section 1.5.1, AAD and DBD in combination are hallmarks of eukaryotic transcription factors. The AAD and DBD of *ZFY* are separated by a short basic nuclear localisation signal in exon 7 implying that *ZFY* is a nuclear protein (Koopman *et al*., 1991). The AAD is thought to bind and recruit the necessary transcriptional machinery due to its highly negative charge. It is presumed due to these features in combination that *ZFY* acts like a conventional transcriptional factor and interacts with TFIID via TATA box binding protein (TBP)-associated factor (TAF9), a known transcriptional cofactor (S. Piskacek *et al*., 2007). Protein-protein interactions form between this cofactor and 9aaTADs, initiating and regulating

transcription machinery. *GAL4*, a transcription factor in yeast has been identified to contain these 9aaTADs and it has been demonstrated that these motifs are vital for *GAL4*'s transactivation activity. Experiments carried out with the *ZFYL* acidic domain lacking the DBD have shown transactivation activity when fused to the Gal4-DBD *S. cerevisiae* reporter system (S. Piskacek *et al*., 2007);(Decarpentrie *et al*., 2012). Whilst *ZFYS* showed no transactivation activity possibly as a result of lacking part of the acidic domain due to the splicing event, there could be other factors at play. It was therefore predicted, that *ZFYS* could have a direct or competitive repressor function that blocks *ZFYL* functions **(Figure 1.13)** (Mardon *et al*., 1990). As a result, *ZFYS* has been designated as the inactive isoform, while *ZFYL* is recognised as the active isoform.

However, *ZFY* transactivation has yet to be investigated within a mammalian expression model system, and it is possible that the events under these circumstances may differ from those seen in the yeast reporter model system.

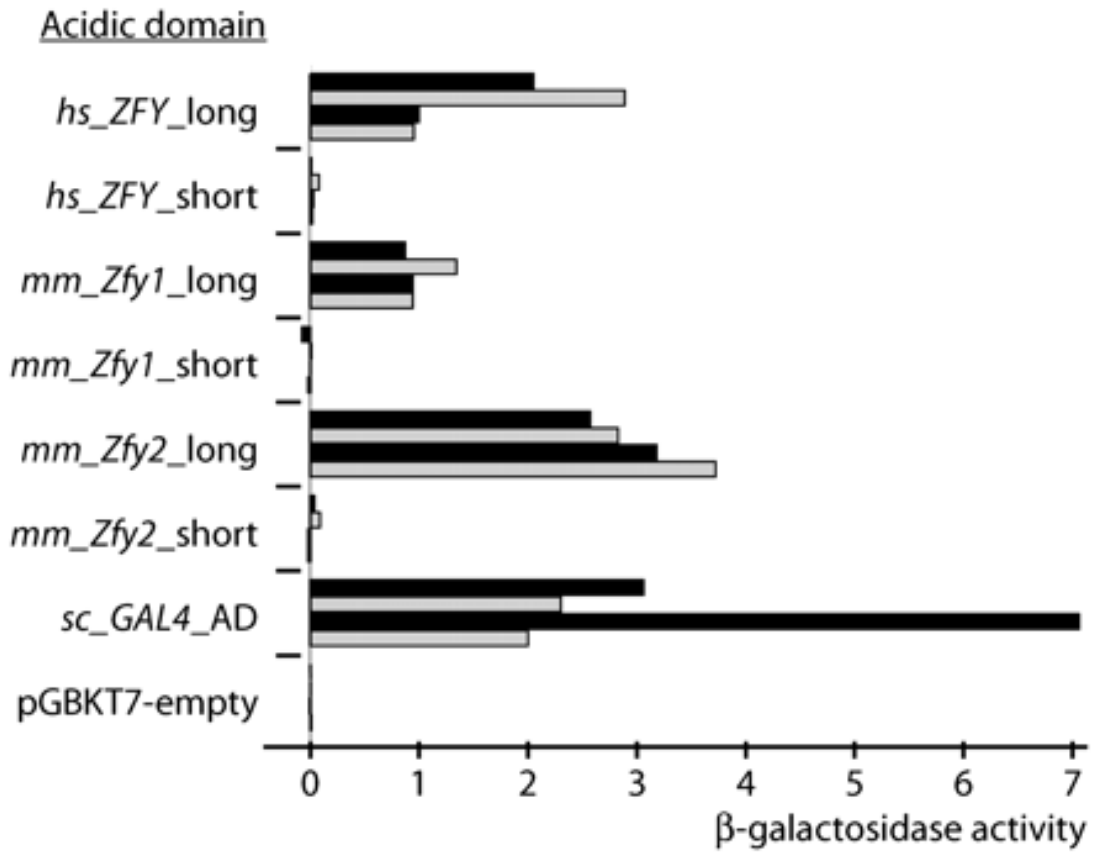


Figure 1.13: *ZFY* Transactivation activity (Decarpentrie *et al*., 2012)**.** β-galactosidase activity was measured and compared between the pGBKT7 negative control and the fusion acidic domain from *S. cerevisiae* (sc) GAL4-AD positive control. *ZFY* isoforms with long or short acidic domains from humans (hs) or mouse (mm) were tested in duplicate from two independent transformations (four bars). The short acidic domains consistently failed to transactivate in yeast cells.

Located at the C-terminal end of *ZFY* the DBD consists of thirteen tandem zinc fingers. This region has demonstrated significant conservation among all *ZFY* variants and their homologs. When examining the genetic similarity between human *ZFY* and mouse *ZFY*1 and *ZFY*2, there is a minimum of 70% sequence identity (Holmlund *et al*., 2023). This similarity also holds for human *ZFX*. It is suggested that this DBD is vital for *ZFY* function (Holmlund *et al*., 2023).

Deciphering the role of *ZFY* is complicated by the existence of both long and short isoforms generated by alternative splicing, a process that is still poorly understood. A component of this thesis will therefore examine the splicing event that produces these *ZFY* isoforms and potential gene/s that may regulate this process.

1.7.2 Known Biological Functions of *ZFY*

Although *ZFY* is expressed widely in many tissues in most mammalian species, its functions in non-testicular tissues remain completely unexplored. In 1990, researchers investigated the case of a human X,t(Y;22) female with a deletion of the *ZFY* gene, as it was thought that *ZFY* and *ZFX* might be associated with Turner syndrome (Page et al., 1990). Despite the absence of *ZFY*, the individual showed no somatic features of Turner syndrome, raising questions and creating uncertainty about the role of ZFY in somatic tissues (Page et al., 1990).

The studies that have been performed to date have focused on its multiple roles during spermatogenesis. This work was recently reviewed by Holmlund *et al* (Holmlund *et al*., 2023) and studies fall into two classes; meiotic and post-meiotic.

Firstly, work in the early 2000s focused on chromosomally variant mice lacking most or all of the Y chromosome, to which *Zfy* was then added back as a transgene to determine which function(s) were restored by the presence of *Zfy*. More recently, CRISPR technology has allowed the use of gene targeting to produce specific knockouts of *Zfy1*, *Zfy2* or both, and characterise the testicular phenotypes.

Collectively, these studies have indicated that mouse *Zfy* functions at several stages in spermatogenesis, with specific roles related to unpaired germ cell removal, control of MSCI during meiosis I, progression of meiosis II, and promoting spermiogenesis (Holmlund *et al*., 2023).

1.7.2.1 Meiotic Functions of *ZFY*

ZFY has been suggested to have apoptotic control at both the MSCI checkpoint as well as the late spindle assembly (metaphase of the first meiotic division) checkpoint (Vernet, Mahadevaiah, de Rooij, *et al*., 2016). This means that incorrect expression of *ZFY* can cause germ cell death. The genomic location of *ZFY* on the Y

chromosome implies a negative feedback loop, which has not yet been established as direct or indirect. Y chromosome location is a sensor for the failure or success of MSCI, with juvenile mice lacking *Zfy* showing delayed onset of MSCI in spermatocytes (Vernet, Mahadevaiah, de Rooij, *et al*., 2016). The *ZFY* loop results in *ZFY* genes repressing their expression during the transition into pachytene of meiosis and then reactivating themselves in spermatids. But like with much of *ZFY*, this mechanism is not understood. However, models suggest that continuous stimulation of *ZFY* expression in lagging cells could stimulate the completion of MSCI and then the ceasing of *ZFY* transcription would allow for prophase to proceed as normal and spermatogenesis would continue (Vernet, Mahadevaiah, de Rooij, *et al*., 2016). However, if *ZFY* expression continues for a prolonged period, MSCI would fail, and cells would undergo apoptosis (Vernet, Mahadevaiah, de Rooij, *et al*., 2016). An example of MSCI leakage is observed in XYY males, where impaired Y chromosome silencing leads to arrested mid-pachytene spermatocytes that ultimately undergo apoptosis (Royo *et al*., 2010);(Decarpentrie *et al*., 2012);(Vernet, Mahadevaiah, de Rooij, *et al*., 2016). Male mice carrying an autosomal *Zfy1* or *Zfy2* transgene were found to wrongly express *Zfy1/2* during pachytene, leading to extensive germ cell apoptosis. This affirmed that the pachytene expression of *Zfy1/2* genes induces a stage IV block in XY males (Royo *et al*., 2010). In summary, *ZFY* serves three key functions in MSCI: (1) it acts as an MSCI activator, (2) functions as a progress sensor due to its Y chromosome location, and (3) acts as an executor for mis-expressed pachytene stage cells in the event of MSCI failure (Holmlund *et al*., 2023). With other functions linked to quality control and meiosis progression.

During the first meiotic metaphase in males, there also needs to be efficient apoptotic elimination of univalent chromosomes (Vernet *et al*., 2011). Mouse studies demonstrated that XSxr^{b}O male mice with deleted Y short-arm genes experience spermatogonial arrest. However, reintroducing *Eif2s3y* overcomes this spermatogonial arrest but does not result in the anticipated elimination of first meiotic metaphase spermatocytes in response to the univalent. Subsequently, it was discovered that *Zfy2*, but not *Zfy1*, was capable of restoring the apoptotic response to X univalence. This suggests that *Zfy2* is necessary to trigger the response to univalent chromosomes at meiotic metaphase 1 (Vernet *et al*., 2011).

Further research has also revealed that *Zfy1* and *Zfy2* are involved in facilitating the successful transition into the second meiotic stage (Vernet, Mahadevaiah, *et al*., 2014). Both genes are expressed during the interphase stage, situated between the two meiotic divisions, a phase exclusive to male meiosis. This indicates that both *Zfy1*

and *Zfy2* help promote the second meiotic division (Vernet, Mahadevaiah, *et al*., 2014).

The role of *ZFY* splice variation in regulating these functions remains somewhat unclear, however, it has been shown that *ZFYS* is expressed predominantly pre-meiotically within the testis, whilst *ZFYL* is expressed predominantly post-meiotically suggesting that these two variants have different possible functions within spermatogenesis (Decarpentrie *et al*., 2012). If *ZFYS* is indeed a competitive inhibitor of *ZFYL* activity, one possibility is that the purpose of the short form is to block the activity of the long form until its expression is required later on in spermatogenesis. This is consistent with the generally greater effect of mouse *Zfy2* in promoting all the phenotypes outlined above, and in particular with the work by Nakasuji *et al* that specifically knocked out the long splice variant while leaving the short splice variant untargeted (Nakasuji *et al*., 2017).

To recap, mouse *Zfy1* and *Zfy2* are pivotal in enhancing meiotic quality control during pachytene, with *Zfy2* additionally influencing processes during the first meiotic metaphase. Moreover, they are involved in initiating the second meiotic stage, underscoring the significance of *ZFY* in various stages of spermatogenesis.

1.7.2.2 Post-Meiotic Functions of *ZFY*

As mentioned above, *ZFY* expression increases in germ cells as the cells begin to enter meiosis, but then is silenced during MSCI during pachytene, it has been found that in mice *Zfy2* is reactivated strongly in spermatids. Does this therefore suggest a spermatid function (Decarpentrie *et al*., 2012)?

Previous findings have suggested that mouse *Zfy2* plays a role in the formation of spermatozoa from haploid round spermatids (Yamauchi *et al*., 2015). This paper, therefore, identified a novel role of *Zfy2* in spermatogenesis and fertilisation (Yamauchi *et al*., 2015);(Vernet, Mahadevaiah, Decarpentrie, *et al*., 2016). Following the reactivation and continuation of *ZFY* in secondary spermatocytes, mouse round spermatids were identified to have *Zfy2* dominance, and *Zfy2* was identified as a key contributor to the transition of round spermatids to spermatids undergoing sperm morphogenesis **(Figure 1.15)** and function during assisted fertilisation (Yamauchi *et al*., 2015);(Vernet, Mahadevaiah, Decarpentrie, *et al*., 2016). This *Zfy2* dominance in spermatids could be a result of *Zfy2* containing a spermatid-specific promoter derived from an X-linked CYPT gene, a spermatid-specific gene family (Vernet *et al*., 2012);(Yamauchi *et al*., 2015);(Vernet, Mahadevaiah, Decarpentrie, *et al*., 2016). *Zfy1* lacks this upstream promoter and is expressed at lower levels in spermatids. This presence of a spermatid-specific promoter could therefore explain the elevated levels

of *Zfy2* transcripts in spermatids, explaining the secondary spike in *ZFY* expression post-meiotically in spermatogenesis (Decarpentrie *et al*., 2012).

Vernet *et al* further confirmed the role of *Zfy2* in sperm morphogenesis promotion after adding the *Zfy2* transgene to $X^{E}Y^{*X}$Sry males in which the only Y genes present are Sry and the X-located *Eif2s3y* (Vernet, Mahadevaiah, Decarpentrie, *et al*., 2016). The introduction of the *Zfy2* transgene advanced spermiogenic progression, whereas the addition of a *Zfy1* transgene had no discernible impact on spermiogenic progression, resulting in the continued failure of sperm elongation. This further suggests that *Zfy2* through its Cypt-derived promoter promotes spermatid elongation (Vernet *et al*., 2012);(Vernet, Mahadevaiah, Decarpentrie, *et al*., 2016).

Both *Zfy1* and *Zfy2* have been continuously proven to be vital to male fertility, with knockout experiments showing abnormalities in their sperm including defects in morphology, motility, capacitation, acrosome reaction, and oocyte activation, as well as chromosomal aberrations (Nakasuji *et al*., 2017);(Yamauchi *et al*., 2022).

1.7.3 Alternative Splicing Regulation in Relation to *ZFY*

Splicing is the removal of introns from pre-mRNA before the exons are joined for translation into a functional protein (Baralle & Baralle, 2005). This process occurs in the spliceosome and is regulated by a variety of RNA-RNA, RNA-protein, and protein-protein interactions to ensure that introns are removed, and exons are joined precisely in the correct order. However, on occasion errors occur and mutations can result in complete exon skipping, intron retention, or can introduce a new splice site (Baralle & Baralle, 2005).

ZFY undergoes splicing producing a testis-specific short isoform, missing an entire exon. The splicer of *ZFY* is currently unknown. In this thesis, we propose that *RBMY* acts to prevent the inclusion of the *ZFY* core acidic domain exon, resulting in the identified short isoform.

1.7.3.1 *RBMY*, an Overview

RNA-binding motif on the Y chromosome (*RBMY*) is located on the Y chromosome in the AZFb region on the Yq11 and encodes a germ-cell-specific protein with functions associated with spermatogenesis (Tsuei *et al*., 2004).

30 copies of *RBMY* genes and pseudogenes have been identified and possibly many are still unidentified, but only a few have been identified as functional (Tsuei *et al*., 2004). Due to the multicopy nature of *RBMY* defining the specific function in spermatogenesis has been difficult (Brriton-Jones & Haines, 2000). Immunohistochemistry has shown the *RBMY1A1* protein is localised within the

nucleus of human male germ cells indicating its testis-specific function (*RBMY1A1*, 2024). The expression of *RBMY1A1* in normal tissues is restricted to the testis, with the median expression noted as 11.47 TPM (GEPIA2, **Figure 1.14**). Whilst expression was not identified elsewhere in the body as shown in **figure 1.14** below. Alongside its specific expression in male germ cells, *RBMY1A1* is highly conserved on the Y chromosome throughout evolution suggesting it is of high importance in spermatogenesis (Brriton-Jones & Haines, 2000).

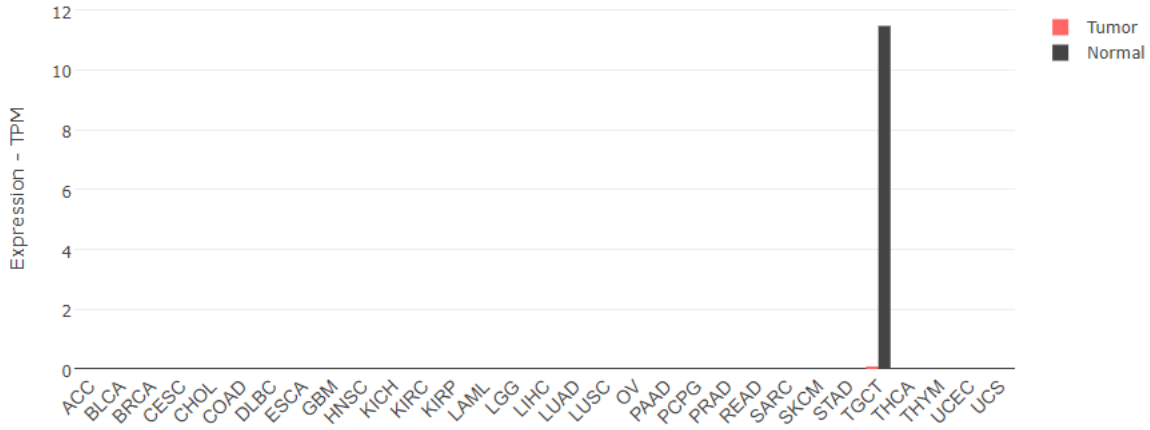


Figure 1.14: Gene expression profile of *RBMY* across all tumour samples and paired normal tissues (GEPIA2). The height of the bars represents the median transcript per million (TPM) expression.

RBMY1A1 structure consists of a C-terminal protein interaction repeat domain which is enriched in serine, arginine, glycine, and tyrosine (Navarro-Costa, Plancha, *et al*., 2010). This is probably the region that controls the regulatory function of *RBMY1A1*. *RBMY1A1* is localised to domains enriched in pre-mRNA splicing components, likely due to the RNA recognition motif (RRM). It has been suggested that *RBMY1A1* modulates pre-mRNA splicing regulators such as SR and SR-related proteins, which modulate splice site selection through their RRM domain. These findings have confirmed *RBMY1A1* function in splicing and mRNA metabolism but further studies looking to identify *RBMY1A1*-interacting proteins have shown potential alternative roles. *RBMY1A1* was found to interact with the steroidogenic acute regulatory (*STAR*) protein, a mitochondrial protein regulating steroid hormone biosynthesis, and the testis-signal transduction and activation of RNA (*T-STAR*) protein. These interactions suggest that *RBMY1A1* is also involved in aspects of meiotic and pre-meiotic regulation through the formation of protein complexes (Navarro-Costa, Plancha, *et al*., 2010);(Stocco, 2001). There has been partial success in identifying *RBMY1A1* RNA targets. Results indicate a complex system as the RRM domain seems to bind RNA with low and high affinity (Navarro-Costa, Plancha, *et al*., 2010). Alongside this,

RBMY1A1 also displays a unique complex RNA recognition mechanism consisting of two steps. Firstly, *RBMY1A1* interacts with a sequence-specific site, and this is then followed by a conformational modification. This suggests a high plasticity concerning RNA partners. Murine studies have indicated 12 potential mRNA targets for *RBMY1A1*, with many of them being expressed in the testis (Navarro-Costa, Plancha, *et al*., 2010).

Overall, *RBMY* is a key Y chromosome gene with functions relating to sperm development specifically in splicing regulation and gene expression during spermatogenesis. This makes *RBMY* a gene of interest for *ZFY* splicing and is why this thesis looks into this relationship.

1.7.3.2 Why *RBMY*?

In the paper that first discovered the *ZFY* alternative splicing, they looked at two human patients with early meiotic arrest (Decarpentrie *et al*., 2012). Patient APOP12 had no AZF deletion and is therefore expected to show normal *RBMY* expression (Decarpentrie *et al*., 2012). This patient's *ZFYS* expression was near normal levels **(Figure 1.15)** (Decarpentrie *et al*., 2012). Patient Ste363 had an AZFb deletion, which is known to eliminate *RBMY* expression, and was found to have much lower levels of *ZFYS* in the testis **(Figure 1.15)** (Decarpentrie *et al*., 2012). Thus, although the histological phenotype is somewhat similar in both patients, *ZFYS* expression is much lower in the patients lacking *RBMY*.

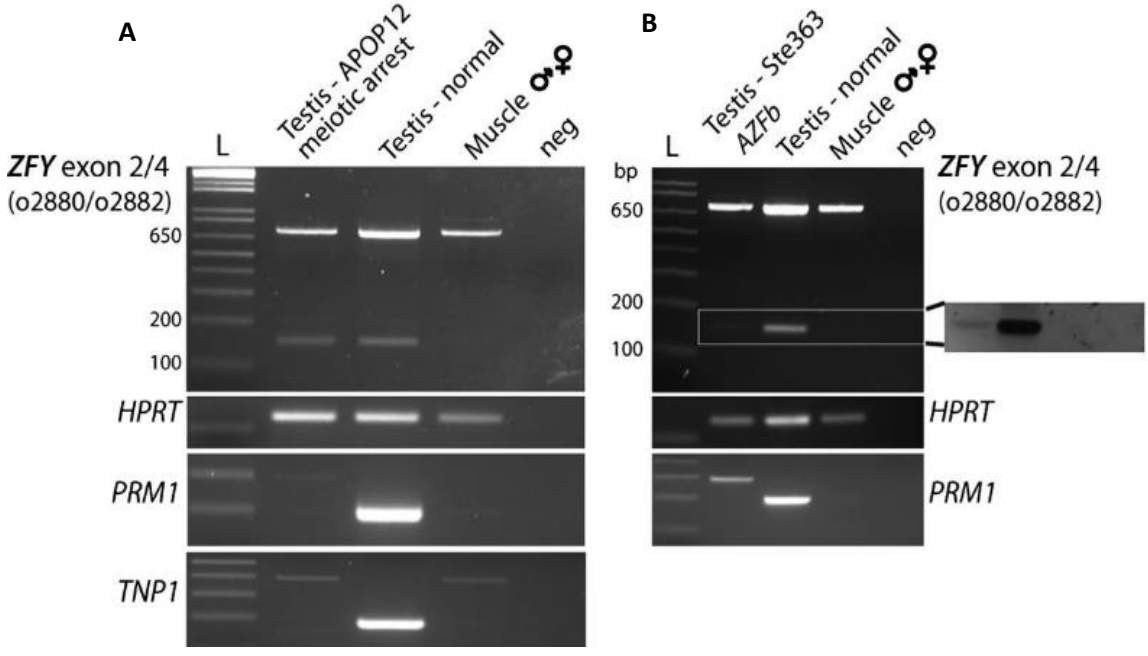


Figure 1.15: RT-PCR analysis of *ZFY* transcripts (Decarpentrie *et al*., 2012). RT-analysis of *ZFY* in human tissues of patient **(A)** APOP12 and **(B)** Ste363. The primers

were designed from exons 2 and 4, to include exon 3 (coding exon 2) which is spliced out forming the short *ZFY* isoform. PRM1 (Protamine 1), TNP1 (Transition protein 1) are spermatid specific and were used to confirm the absence of post-meiotic germ cells. HRPT = Housekeeping gene used as a positive control.

Further circumstantial evidence is provided by expression timing. *RBMY*, like *ZFYS*, is testis-specific and is located in the nucleus of spermatogonia and spermatocytes, but not spermatids. This matches the predominant expression of *ZFYS* before meiosis and *ZFYL* post-meiosis in spermatids. Moreover, the phenotype of men carrying an AZFb deletion on their Y chromosome (and therefore lacking *RBMY*) is typically a complete mid-meiotic arrest. This is unlike the phenotype seen in other Y chromosome deletions not involving AZFb and is similar to the phenotype observed in mice transgenically overexpressing *Zfy1* or *Zfy2* during pachytene. If *ZFYS* is indeed a lower-activity variant that acts as an antagonist to *ZFYL*, then selective deficiency of *ZFYS* (triggered by absence of *RBMY*) might result in the same phenotype as *ZFYL* overexpression.

Mechanistically, functional studies have investigated *RBMXL2*, a closely related gene to *RBMY*. The *RBMXL2* gene is located on chromosome 11, with the protein only being expressed in the testis (Ehrmann *et al*., 2019). Mouse studies knocking down *RBMXL2* showed a block in spermatogenesis, with adult mice only having very few spermatids and no elongated spermatids suggesting that *RBMXL2* loss prevents sperm production, and this is possibly due to a developmental block during meiosis. It was found that the mouse germ cells could develop as far as meiosis but then would undergo cell death via apoptosis during diplotene. With enrichment of *RBMXL2* binding in both alternative exon sequences and in regions near regulated splice sites, as opposed to non-regulated events, was also noted. The study that *RBMXL2* has a key role in meiotic transcriptome control, and thus, loss of *RBMXL2* can result in male infertility due to germ cell type-specific cryptic splicing. This was confirmed by gene ontology analysis showing that many genes controlled by *RBMXL2* splicing had functions in spermatogenesis, meiosis, and germ cell development. To further this 25/186 *RBMXL2* target genes were identified to result in infertility if the whole gene was deleted. Importantly for this hypothesis, it was shown that *RBMXL2* acts by suppressing the use of a specific subset of acceptor splice sites, triggering the skipping of specific exons in the testis. This suggests that the closely related *RBMY* (the N-terminal RRM is 77.2% similar to *RBMXL2* (Ehrmann *et al*., 2019)) may also act to block specific acceptor splice sites, i.e. that specific exons will be skipped in premeiotic male germ cells expressing *RBMY*. In the context of the hypothesis, we proposed that premeiotic *RBMY* expression could trigger skipping of the exon 3 of *ZFY* in spermatogonia.

1.7.4 *ZFX*, the X Chromosome Homologue of *ZFY*

ZFX is the X-linked homologue of *ZFY* and has a similar structure. *ZFX* is located on the short arm of the X chromosome in the region of Xp21.3 and p22.1 (Schneider-Gadicke *et al*., 1989). Evidence has suggested these two genes have diverged from a common ancestor due to their similarities in exon sequences and organisation. Within the C-terminal DBD, both *ZFY* and *ZFX* encode a total of 13 zinc fingers and 99% of their zinc finger amino acids are identical suggesting functional similarities **(Figure 1.16)**. Like *ZFY*, *ZFX* zinc fingers follow the same general zinc finger structure of C_2H_2. Due to their high similarity, *ZFY* and *ZFX* likely bind to the same DNA sequences and targets. Northern blot experiments within human-rodent hybrids suggest that *ZFX* escapes X-inactivation, as a higher level of *ZFX* transcripts was identified with an increasing number of X chromosomes (Schneider-Gadicke *et al*., 1989).

ZFX has been studied highly in relation to multiple different human cancers, with implications in the initiation or progression of human cancers including; breast, colorectal, glioma, renal carcinoma, prostate, and many more (Ni *et al*., 2020). Specifically, high expression of *ZFX* has been correlated with poor survival rates in cancer patients. This discovery leads to the thinking that *ZFX* does not seem to be a tumour-type-specific oncogene but rather it contributes to metaplastic transformation caused by tumour-promoting changes in the transcriptome. Again, like *ZFY* the mechanism by which *ZFX* influences these changes has not been determined (Ni et al., 2020).

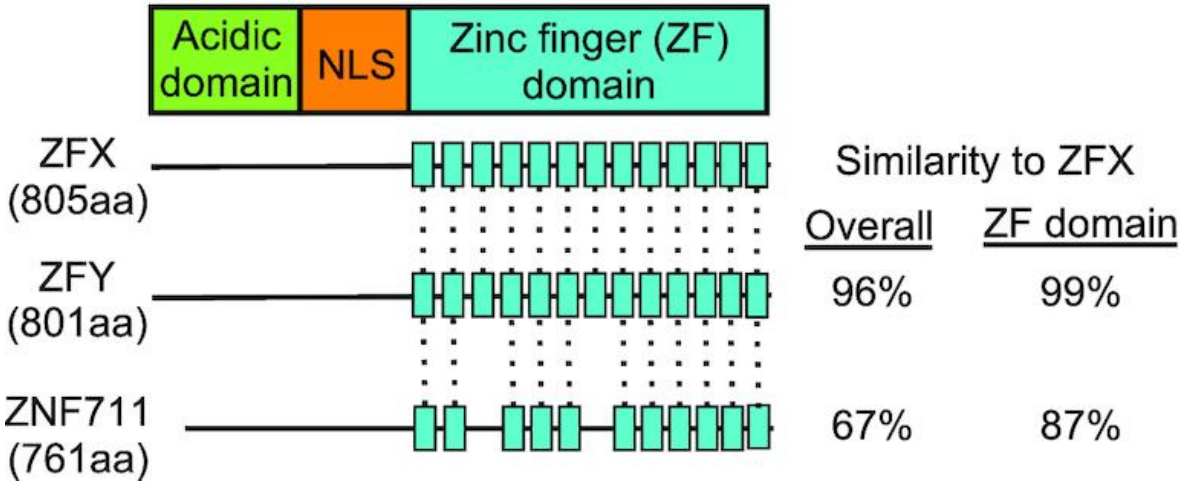


Figure 1.16: Schematic diagram showing the conservation between *ZFY*, *ZFX*, and ZNF711 (Ni *et al*., 2020). This family of zinc finger proteins is highly conserved with 96% overall sequence identity between *ZFX* and *ZFY*. The sequence conservation is even higher in the zinc finger (ZF) domain critical for function, approaching 100%. *ZNF711* is a more distant zinc finger protein family member with

lower but still high similarity to both *ZFX* and *ZFY*. However, *ZNF711* lacks two of the zinc fingers present in the ZF domain of *ZFX*/*ZFY*.

Ni *et al* characterised the transcription factors that bind downstream of the *ZFX* CpG island promoter start site, providing both RNA-sequencing and ChIP-seq data (Ni *et al*., 2020). These datasets are of interest given the structural similarities between *ZFY* and *ZFX* despite their divergent functions. Analysing these *ZFX*-associated data may provide insights into *ZFY* transcriptional regulation.

1.7.5 *ZFY*'s Role in Y Chromosome Evolution

As mentioned earlier, the Y chromosome has earned the reputation of being fragile and vulnerable due to the loss of genes and PAR regions (Holmlund *et al*., 2023). It is widely believed that meiosis plays a significant role in the evolutionary dynamics of the Y chromosome (Holmlund *et al*., 2023).

The continuous evolutionary pressure for the persistence of the Y chromosome is deemed to be down to the existence of a meiotic executioner (Waters & Ruiz-Herrera, 2020);(Holmlund *et al*., 2023). The necessary criteria for this meiotic executioner are: (1) it must be encoded on the Y chromosome, (2) essential for fertility or survival, and (3) causes lethality at the pachytene stage during MSCI (Holmlund *et al*., 2023). *ZFY* is therefore a top contender for this meiotic executioner in eutherians. Evidence from mice (Vernet, Mahadevaiah, de Rooij, *et al*., 2016), horses (Ruiz *et al*., 2019), pigs (Barasc *et al*., 2012), and rodents where the Y chromosome had been lost through Y transposition to the X chromosome, *ZFY* had been discovered to be situated on the X chromosome (Holmlund *et al*., 2023). Additionally, after sequencing the Y chromosome of the giant panda, researchers identified an independent gene conversion occurring on exon 7 (Holmlund *et al*., 2023). This gene conversion was further identified in various other mammalian lineages. This implies that such a mechanism might serve to safeguard the functionality of *ZFY* on the Y chromosome in mammals (Holmlund *et al*., 2023). In particular, gene conversion affecting the DNA binding domain would serve to homogenise this region between X and Y homologues, ensuring that both continue to target the similar sets of downstream genes even as the acidic domain (controlling the strength and direction of transcriptional regulation) diverges between X and Y gene copies.

1.7.6 *ZFY* is a Possible Proto-Oncogene

While extensive research has examined *ZFY*'s roles in spermatogenesis and other male-specific functions, studies have also emerged investigating *ZFY* as a candidate

cancer-testis gene. This is based on findings that *ZFY* is expressed in certain cancers yet displays more highly in the testis than in normal tissues.

An oncogene results in the uncontrollable division of cells, potentially resulting in cancer (*Oncogene*, 2024). However, before the transition into an oncogene through mutation, the gene is known as a proto-oncogene and functions in regulating normal cell division.

Experiments looking at *ZFY* expression, have shown unexplained expression of *ZFY* in cancer cells. It is thought that *ZFY* could possess indirect oncogenic activity (Tricoli & Bruce Bracken, 1993). The zinc finger motifs have shown the potential to regulate the expression of other target genes which suggests a potential role in malignancy development. Through RT-PCR and northern blots *ZFY* for example was identified in 20 of 31 prostate adenocarcinomas, and then further southern blot analysis showed that the Y chromosome segment containing *ZFY* was not lost from a majority of the tumour cells. Unlike in benign hyperplastic tissue where *ZFY* was not identified to be expressed. This suggested that *ZFY* can become transcriptionally active in human tumours resulting in the dysregulation of other vital growth control genes contributing to malignancy formation (Tricoli & Bruce Bracken, 1993). At the time this data was published, the *ZFYS* splice variant had not been identified.

According to the human protein atlas, *ZFY* has low cancer specificity but has been detected in many cancer types with suggestions to being a prognostic marker in head and neck cancers (*ZFY*: The Human Protein Atlas, 2024) **(Figure 1.17)**. Overall, there is an ongoing need to identify new cancer biomarkers and potential therapeutic targets. *ZFY* represents a prospective cancer-testis gene based on its restricted expression profile and evidence of expression in certain tumour types. Further research is required to evaluate *ZFY* as a potential diagnostic marker or treatment target in specific cancers.

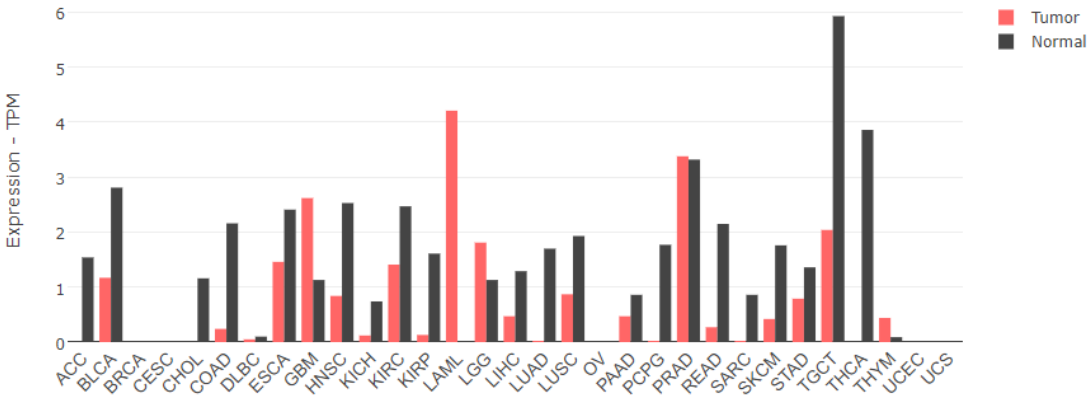


Figure 1.17: Gene expression profile of *ZFY* across all tumour samples and paired normal tissues (GEPIA2). The height of the bars represents the median transcript per million (TPM) expression.

1.7.6.1 Head and Neck Cancer

Previous work by a student in the Ellis-Fenton lab showed expression of the short-spliced variant in HPV-negative oropharyngeal cancer squamous cell carcinoma (OPSCC) cell lines, which is intriguing given the excess male prevalence of head & neck squamous cell carcinomas (HNSCCs) (Trujillo, 2019).

Head and Neck cancer (HNC) as of 2022 is the seventh most common cancer globally (Gormley *et al*., 2022). The incidence of HNCs has risen steadily, with a 36.5% increase observed over the past decade **(Figure 1.18)**. Current projections indicate this rising incidence will continue (Sabatini & Chiocca, 2020). However, the mortality rate is also increasing whilst survival rates remain static. More specifically, approximately 90% of NHC are squamous cell carcinomas, which are when cancer arises from the epithelial lining of the oral cavity, pharynx, and larynx (Gormley *et al*., 2022). Two very distinct oncogenic pathways have been shown to drive HNSCC carcinogenesis; either chemical carcinogens or HPV infection (Powell *et al*., 2021).

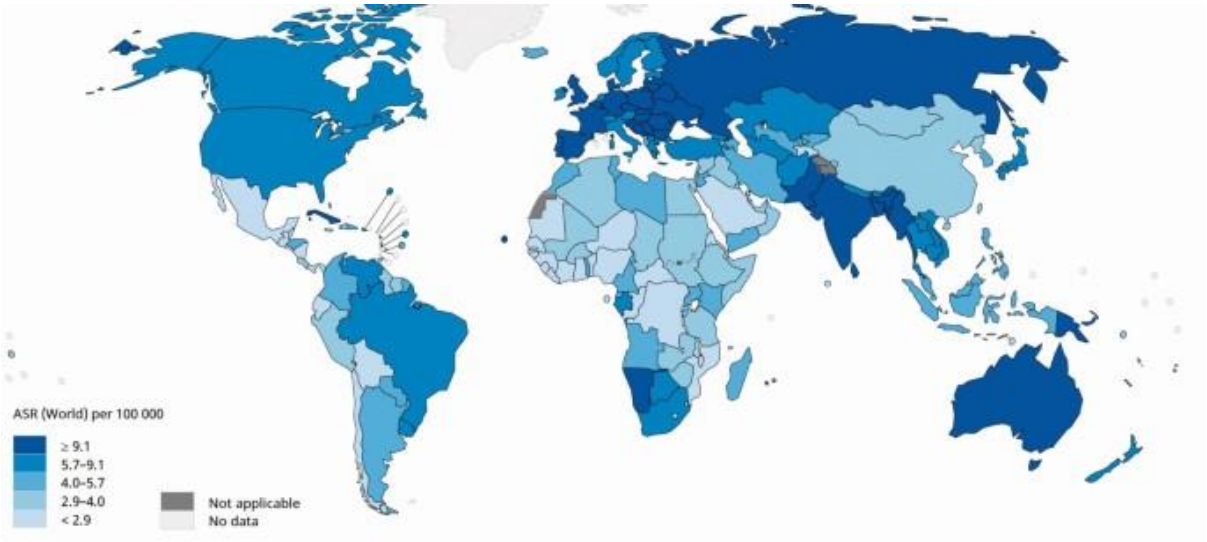


Figure 1.18: The global age-standardized incidence rate of head and neck cancer (Gormley *et al*., 2022)**.** This global incidence includes the lip, oral cavity, oropharynx, hypopharynx and larynx cancer sites.

The increasing burden of HNSCCs has been correlated to tobacco-derived carcinogens, excessive alcohol consumption, or both (Johnson *et al*., 2020). Genetic risk factors have also been shown to contribute to HNSCC including Fanconi anemia, a rare disease associated with impaired DNA repair. But tumours arising from the oropharynx epithelial lining have been linked to infection with oncogenic strains of

HPV, normally HPV-16 but also possibly HPV-18. Comparisons between HPV-positive and HPV-negative HNSCCC showed different gene expressions and mutational and immune profiles (Johnson *et al*., 2020).

1.7.6.2 HPV-Negative vs HPV-Positive HNSCC

HPV-negative HNSCC is largely caused as a consequence of tobacco which consists of more than 5,000 different chemicals. Of those 5,000 chemicals, the cancer-causing ones have been identified as polycyclic aromatic hydrocarbons (PAHs), including benzo(a)pyrene and nitrosamines (Johnson *et al*., 2020). These carcinogens undergo metabolic activation which can result in DNA mutations and other genetic abnormalities. An emerging tobacco-induced carcinogenesis mechanism points towards the alteration of microRNA (miRNA) expression (Powell *et al*., 2021). Tobacco could be altering miRNA regulation resulting in major signalling changes and metabolic processes. Alcohol is another key risk factor and is thought to enhance the exposure of epithelial cells to carcinogens. The oral microbiome could also be negatively impacted by chronic alcohol consumption, with links to decreasing the *Lactobacillus* abundance. This potentially leads to the growth of other oral microbiome pathogens which could promote carcinogenic effects (Powell *et al*., 2021).

HPV-negative HNSCC has been associated with poor oral hygiene and lower socioeconomic status (Sabatini & Chiocca, 2020). These tumours exhibit significant genomic complexity, frequently harbouring mutations in the TP53 tumour suppressor and cell cycle regulators. HPV-negative status correlates with poorer prognosis, and standard treatment involves cisplatin and radiation therapy. HPV-positive HNSCC has a better prognosis and is more susceptible to treatment by radiation and anticancer drugs (Sabatini & Chiocca, 2020).

HPV-positive HNSCC is mostly caused by HPV-16, with HPV-18, HPV-31, HPV-33, and HPV-52 being detected in only a small group of patients. The first discovery of HPVs oncogenic role was over 40 years ago by Zur Hausen who identified a possible link between HPV and cervical cancer onset (Sabatini & Chiocca, 2020). HPV-16 is a small, double-stranded, circular DNA virus that can integrate its viral genome into the human genome. The key players involved are the seven early genes (E1-E7) and the late genes (L1 and L2). The late genes are responsible for encoding the capsid proteins, whilst the early genes are responsible for replication and transcription of the genome. Specifically, E6 and E7 have been identified as essential for the oncogenic transformation in the host cells. E6 results in the oncogenic transformation by interacting with p53 to form a complex promoting the ubiquitylation and proteasomal degradation of p53 (C. Zhou & Parsons, 2020). This results in the loss of cell cycle

regulation at the G1/S and G2/M checkpoints leading to genomic instability, accumulation of chromosomal aberrations, unchecked tumour cell proliferation, and eventual tumour formation (C. Zhou & Parsons, 2020). Whilst E7 works by strongly binding to RB1, a cell cycle regulator retinoblastoma-associated protein. This leads to the proteasomal destruction of RB1 resulting in further feedback upregulation of a commonly identified gene in oropharyngeal tumours known as p16INK4A (Johnson *et al*., 2020).

A higher male-to-female ratio of HNSCC prevalence has been reflected, however, the exact mechanisms and causes remain generally unknown.

1.7.6.3 HNSCC Male Prevalence

A study between 1995 and 2012 comparing OPSCC showed a 2-fold prevalence in males (D'Souza *et al*., 2017);(Sabatini & Chiocca, 2020). However, HPV was still identified to be a major driver of OPSCC in both sexes, with 62% of males and 56% of females testing HPV-positive (Sabatini & Chiocca, 2020). Many potential suggestions have been made for gender differences seen in HPV-positive tumours. These include differences in sexual behaviour and lifestyle differences between the genders, such as smoking and alcohol consumption. Other potential factors include hormonal factors such as oestrogen-related and progesterone-related factors which could play a role in cancer protection specifically in females. However, it is likely that intrinsic biochemical and molecular differences between males and females, independent of sex hormone influences, may impact tumorigenesis in distinct ways and will therefore be affected by viruses in different ways (Sabatini & Chiocca, 2020). The Human Protein Atlas identifies *ZFY* as a favourable prognostic marker in head and neck squamous cell carcinomas. However, this data should be interpreted with caution, as the analysis does not adequately distinguish between *ZFY* and its X chromosome homolog *ZFX*. Investigating *ZFY* as a potential cancer-testis antigen is a focus of this thesis project.

1.8 Project Outline & Aims

The current understanding of human *ZFY* variants is limited, leading to several unanswered questions which form the basis of this research where we set out to better understand the structure/function relationships that underpin *ZFY* gene function and have sculpted its evolution. Using modern 'omics-scale analyses at the genomic, transcriptional, transcriptomic and proteomic level we aimed to produce a cohesive set of interlinked experiments that address multiple facets of the structure/function relationships of this long-neglected gene and uncover novel

insights into its evolutionary history and mechanism of action in the testis and beyond. The progress of *ZFY* research over the past ~50 years (summarised in **Figure 1.19**) has left numerous unanswered questions, which have largely shaped the research aim of this thesis.

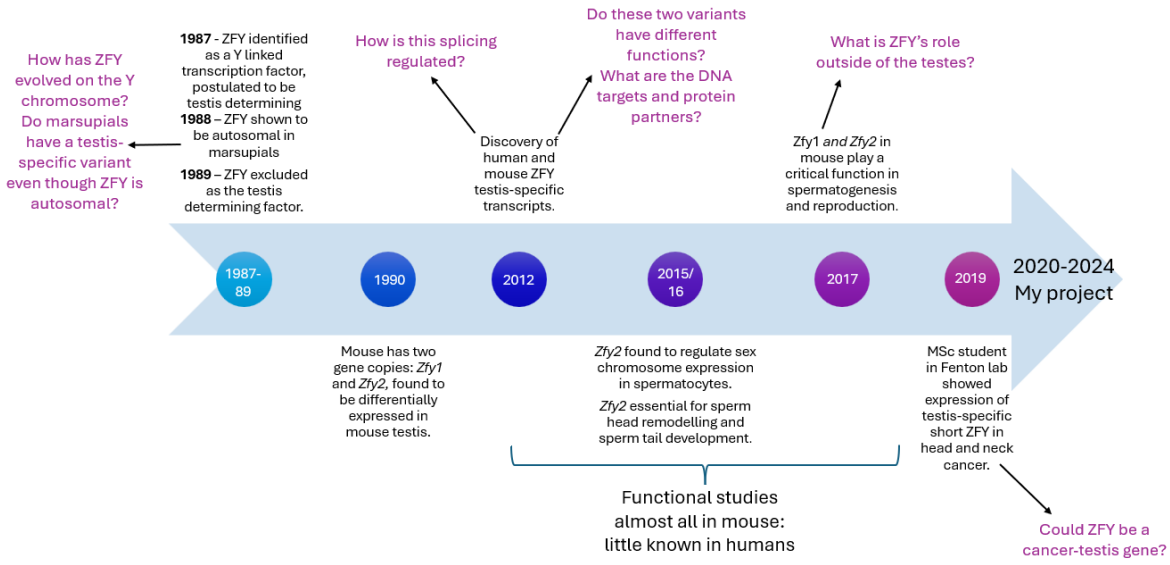


Figure 1.19: A timeline showing the scientific discoveries of *ZFY* over the last ~50 years summarised in this thesis introduction. The questions that have remained unanswered are highlighted in purple and make up this research project.

To address this aim, the research has been broken down into specific objectives to:

- Explore the evolutionary history of *ZFY* and how this relates to its predicted functional domain organisation.
- Investigate how *ZFY* splicing is regulated to produce the short and long structural isoforms observed in testes, and whether this regulation is evolutionarily conserved.
- Determine the downstream transcriptional targets of the short and long isoforms of *ZFY*, and how disruption of this translational programme may relate to its putative role as an oncogene.
- Establish whether *ZFYS* and *ZFYL* have distinct interacting partners that co-regulate distinct targets, or do they share a common interactome.

These four objectives are explored across 4 results chapters in this thesis as described below.

Using evolutionary analysis **Chapter 2** investigates how *ZFY* has evolved from an autosomal gene to a sex chromosome gene at the point of the placental mammal divergence. Using a wide range of animal species DNA and protein sequence conservation was analysed to define portions of the *ZFY* open reading frame that are undergoing positive versus negative selection and interpret this in relation to protein structural features and to specific evolutionary transitions.

Chapters 3, 4 and 5 utilise commercially synthesised *ZFY* constructs to investigate the following:

Chapter 3 used RNA-Seq data to investigate whether splice variation in *ZFY* is conserved in species with autosomal Zf*, and how regulation of splicing variation in *ZFY* results in the second shorter variant. Using a modified GFP reporter system we aimed to test *in vitro* whether *RBMY* expression causes exon skipping of the third exon producing the short isoform

Chapter 4 aimed to identify transcriptomic changes caused by the overexpression of *ZFYL* and *ZFYS* in a mammalian system using genome-wide RNA-Seq. A cell culture model was used to carry out global transcription profiling of the consequences of *ZFYS* and *ZFYL* overexpression to identify target genes and pathways, especially, any cancer-targeting pathways or oncogenes.

Chapter 5 aimed to use protein purification and/or immunoprecipitation to express and purify the human *ZFYL* and *ZFYS* acidic domains (lacking DNA binding domains) in bacterial culture and perform preliminary structural characterisation to assess protein folding. Then using the purified protein, perform pull-down experiments to identify potential binding partners of both variants. However, following months of work and unsuccessful attempts a mammalian pull-down system was utilised to identify direct protein targets of both *ZFY* variants.

2. Chapter 2: Tracing the Evolution of *ZFY* from an Autosomal Gene to a Y Chromosome Gene

2.1 Introduction

2.1.1 Phylogenetics – A Powerful Scientific Field

Identifying and understanding the evolution of genes is vital to scientific discovery and insight into a gene's function (Munjal *et al*., 2018). Phylogenetics is a powerful scientific field used to uncover the evolutionary relationships between - and histories of - species or specific genetic sequences. Due to the constant development of new sequencing technologies and analytical methods, the evolutionary history of a gene can be reconstructed from comparisons of orthologues across many species. Phylogenetic sequence analysis relies on the extraction of DNA, RNA or protein sequences to build a phylogeny of sequences based on the similarities and differences between species or within a species (Munjal *et al*., 2018). The basis of an evolutionary model assumes a tree-like structure known as a "phylogenetic tree", a model widely used to explore hypotheses (Huson & Bryant, 2006). A phylogenetic tree presents in a graphical form the history of a gene or taxon **(Figure 2.1)** (Scott & Baum, 2016). The tips of the tree represent the species, populations, individuals or genes under investigation. The tips are linked by branches which indicate the amount of genetic change that has occurred, with longer branches representing a greater amount of genetic change. A node is made when genetic isolation has resulted in a lineage split, giving rise to two or more new lineages. The branches then all converge on one point, known as the root of the tree which represents the last common ancestor (Scott & Baum, 2016). Accurate rooting of tree models is vital for the accurate interpretation of ancestral changes across sequences, but unrooted trees can still be very useful for showing the distance between sequences (Kinene *et al*., 2016).

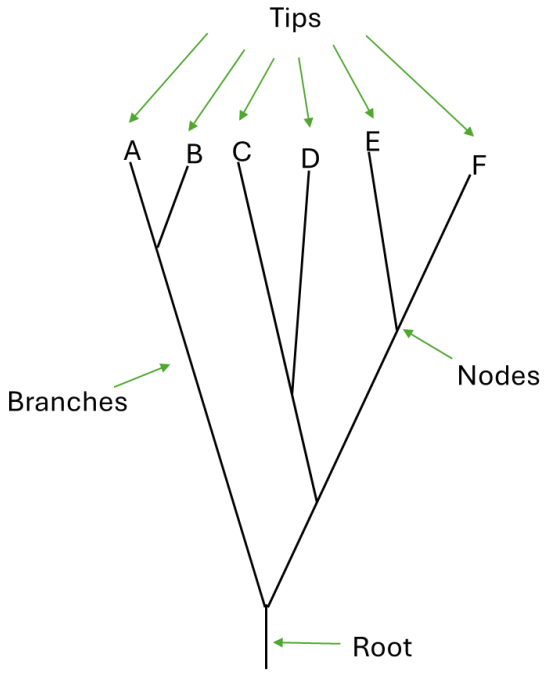


Figure 2.1: The characteristics of a phylogenetic tree. Edited from (Scott & Baum, 2016).

A phylogenetic tree is a great way to start an evolutionary investigation but then understanding the selection pressures on the gene sequences is vital. Mutations in the gene sequences can range from single nucleotide changes to insertions, deletions, duplications or conversions (Anisimova & Liberles, 2012). These changes can be either beneficial or catastrophic to survival. DNA mutation rates vary within the genome and among species, for example, small-bodied mammal species have been shown to have higher rates of molecular evolution compared to their larger relatives (Kumar & Subramanian, 2002);(Bromham, 2009). This is potentially linked to smaller mammals having more generations per unit of time (Bromham, 2009). Within a species genome, there are regions of mutational hot and cold spots, with genes in mutational hot spots benefitting from higher mutation rates allowing for more flexible responses to the changing environment (Chuang & Li, 2004). While genes within mutational cold spots require protection from deleterious mutations (Chuang & Li, 2004). These mutational variations within the genome have to be considered when looking at ancestry hence the development of several statistical best-fit models from Jukes-Cantor to Tamura-Nei and beyond when making phylogenetic trees (Posada & Crandall, 2001). These allow for the selection of the best model of nucleotide substitution and should be routine in phylogenetic analysis (Posada & Crandall, 2001).

2.1.2 Detecting Selection Via Analysis of Mutation Rates

Positive selection occurs when mutations are advantageous and are positively associated with fitness and survival (Agrawal *et al*., 2010);(Anisimova & Liberles, 2012). This promotes the emergence of a new beneficial phenotype. A beneficial phenotype within a population will eventually become fixed as individuals carrying the mutation will give rise to a larger number of offspring and the allele will quickly begin to spread through the population with each new generation (Desai & Fisher, 2007). In contrast, negative or purifying selection removes deleterious variants, and this selection type is essential for retaining genetic regions vital for proper function and survival (Anisimova & Liberles, 2012);(Pouyet *et al*., 2018). Any deleterious allele is selectively purged from the population and is not passed on to offspring as the carriers have fewer offspring than noncarriers (Rapaport *et al*., 2021). By calculating the mutation rates between coding DNA sequences, the method of selection can be identified (H.-S. Kim & Takenaka, 2000).

Mutations can be subdivided into synonymous or non-synonymous substitutions. Synonymous substitutions result in a change in the DNA sequence but do not change the encoded amino acid, whilst a non-synonymous substitution is a change in the DNA sequence that does change the encoded amino acid (A. L. Hughes *et al*., 2008). A higher ratio of non-synonymous substitutions to synonymous substitutions is indicative of positive selection and infers that the amino acid change is beneficial. Over time this change becomes fixed and essential for its continuation (A. L. Hughes *et al*., 2008). Evidence from yeast and bacteria work have shown growing evidence that synonymous mutations are less selectively neutral concerning fitness despite their lack of effect on the encoded amino acids (Lebeuf-Taylor *et al*., 2019);(Bailey *et al*., 2021). It has now been suggested that synonymous mutations play a larger role in adaptation than originally thought (Bailey *et al*., 2021). Synonymous mutations are proposed to modify mRNA structures, thereby influencing translation initiation, mRNA stability, or even protein folding thus affecting protein function as a result (Kristofich *et al*., 2018);(Lebeuf-Taylor *et al*., 2019). While most synonymous alterations are deemed neutral or only mildly harmful, their impacts may be amplified under intense selection pressure. This was shown by identified point mutations in an enzyme required for arginine and proline biosynthesis, synonymous mutations were shown to affect the growth both positively and negatively. These mutations were proposed to either disrupt the stability of a stem-loop structure at the start codon or impede start codon accessibility. Both scenarios would consequently impact translational efficiency, leading to reduced mRNA stability since ribosomes would be unable to shield mRNA from degradation (Kristofich *et al*., 2018). This has shown the potential

importance of synonymous mutations in evolution and how thus far they have been under-appreciated. Nevertheless, since the selective pressures acting on non-synonymous mutations are generally much stronger than those acting on synonymous mutations, the ratio of non-synonymous to synonymous changes (known as dN/dS comparison) can be used to infer the presence of positive or negative selection within any given genetic lineage (Kryazhimskiy & Plotkin, 2008).

2.1.3 Understanding *ZFY* Evolution: Consequences of Y Linkage

Y chromosome evolution is special with positive and negative selection acting differently on this unique chromosome. The Y chromosome undergoes erosion which has been linked to the reduced recombination between the X and Y chromosomes along the majority of their length (Engelstädter, 2008). This leads to the "hitchhiking effect" of deleterious mutations as the lack of recombination allows for the fixation of beneficial mutations on the Y chromosome, but they also bring with them deleterious mutations at other loci on the Y chromosome (Engelstädter, 2008). Another mechanism in which deleterious mutations accumulate on the Y chromosome is termed "Muller's ratchet". This process results in the irreversible fixation of deleterious mutations due to the absence of recombination explaining the rapid loss of the Y chromosome genes (Engelstädter, 2008);(Sakamoto & Innan, 2022). Finally, the accumulation of deleterious mutations on the Y chromosome can also be influenced by a reduction in population size because of "background selection" (Engelstädter, 2008). Background selection is more potent on the Y chromosome as there is reduced recombination, leading to the reduction in the genetic diversity across the chromosome and the accumulation of deleterious mutations instead (Wilson Sayres *et al*., 2014). This is further enhanced due to the low frequency of the Y chromosome in the population, as it is only found in males and thus further reduces the diversity (Wilson Sayres *et al*., 2014). All of these affect the Y gene evolution and explain the extremely low diversity of the entire human Y chromosome. Negative/purifying selection acts on the Y chromosome preserving both the number and the type of function-coding genes vital for male development (Engelstädter, 2008). Although it was originally thought that the X and Y chromosomes were homologous, HJ Muller inferred that the Y chromosome's permanent heterozygosity was because of the lack of recombination exchange between the X and Y chromosomes and its transmission solely through males (Muller, 1918);(Charlesworth, 2003). This reflects the uniqueness of the Y chromosome and its evolution.

ZFY is a highly conserved gene located on the Y chromosome in all eutherian mammals. However, in marsupial mammals, *ZFY* is located on an autosome rather

than the sex chromosomes (Sinclair, 1988);(Tucker *et al*., 2003). This indicates that after the evolutionary divergence of marsupials and eutherian mammals, *ZFY* became a sex-linked gene in eutherian mammals (Sinclair, 1988);(Tucker *et al*., 2003). There is substantial sequence homology between the X and Y chromosomes, suggesting a history of either genetic recombination or gene conversion between them (Pamilo & Bianchi, 1993). However, genetic recombination between sex chromosomes can only occur in pseudoautosomal regions where the sequences are homologous. In 1989, Schneider-Gädicke proposed that gene conversion between human *ZFX* and *ZFY* explains their high sequence homology (Schneider-Gadicke *et al*., 1989);(Pamilo & Bianchi, 1993).

The evolution of the Y chromosome has been said to be protected by *ZFY* due to its role in meiosis surveillance (Holmlund *et al*., 2023). Although *ZFY* functions are generally unknown, *ZFY* has continued to persist on the Y chromosome even after being noted as a “fragile” chromosome due to its rapid gene loss (Blackmon & Demuth, 2015);(Ruiz-Herrera *et al*., 2022). The loss of genes has been explained by its shrinking PARs, the accumulation of deleterious mutations and the reduced recombination. These are further exaggerated by the reduced number of Y chromosomes relative to autosomes in the population and therefore there is reduced diversity. This conservation of *ZFY* supports the notion that *ZFY* is essential for survival for reproduction as it continues to evolve and persist on the Y chromosome removing any detrimental mutations and maintaining its functions (Holmlund *et al*., 2023).

2.1.4 Understanding *ZFY* Evolution: Structural Considerations

A transcription factor requires specific domain regions to be able to bind and activate its targets, these domains include an AAD and a DBD **(Figure 2.2)** (Boija *et al*., 2018). The DBD is vital for binding specific targets and the three main classes in mammals are zinc finger domains, homeodomains and helix-loop-helix domains (Frietze & Farnham, 2011). The structure a DBD adopts determines the interactions of these domains with target DNA sequences (D. H. Gonzalez, 2016). Within a subclass of DBD similarities between transcription factors occur due to similar DNA-binding specificities even if they have differing downstream functions (D. H. Gonzalez, 2016). *ZFY*'s DBD consists of thirteen zinc fingers which are essential for recognising specific DNA sequences as the amino acids in each finger define its DNA recognition specificity (Cassandri *et al*., 2017). Changes in these amino acids would alter the DNA sequence recognition of the transcription factor (Cassandri *et al*., 2017). This is

possibly why during evolution in general these DNA binding regions are acted on by negative selection to ensure the gene's function is retained.

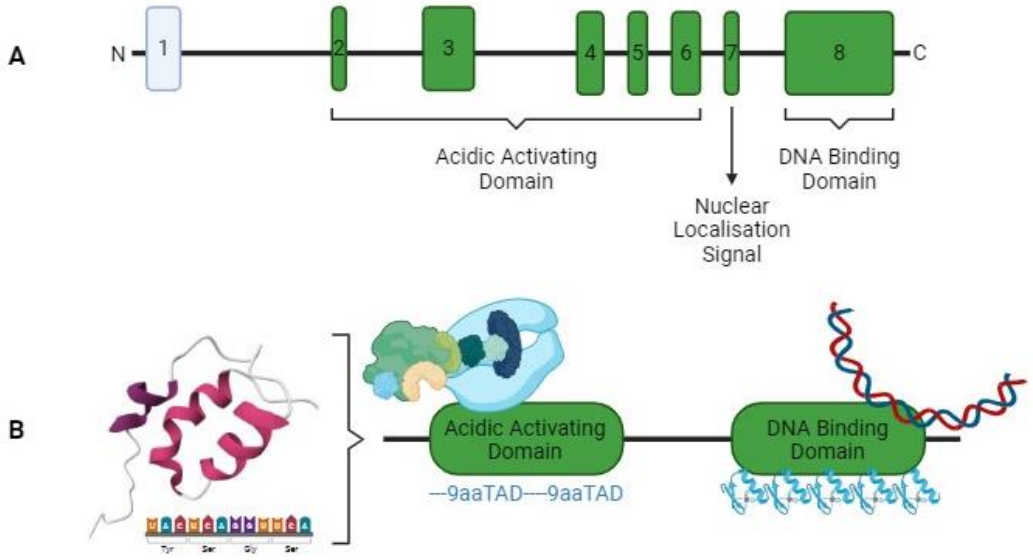


Figure 2.2: *ZFY* transcription factor domains. A: *ZFY* gene exon breakdown. The green boxes highlight the coding exons, and the light blue box highlights the non-coding exon. The acidic activating domain is located across coding exons 2-6 and exon 7 (coding exons 1-6) and the DNA binding domain is located in exon 8 (coding exon 7). These two domains are also linked by exon 7 (coding exon 6) containing the nuclear localisation signal. **B**: *ZFY* protein function. The acidic activating domain binds transcription machinery and consists of conserved 9aaTAD regions while the DNA binding domain uses thirteen zinc fingers to bind target DNA sequences. Note: The protein structure of *ZFY* is currently unknown.

The AAD binds coactivators required for transcription and is classified based on its amino acid composition (Triezenberg, 1995);(Staller *et al*., 2021). In general, these regions have been classified as having rich areas of glutamine, proline or serine and threonine (Triezenberg, 1995). Unlike the DBD, the AAD are more disordered and lack conservation (Melcher, 2000);(Kotha & Staller, 2023), however, the 9aaTAD regions of the AAD which have been identified as vital for transcription factor function seem to represent the more conserved regions within the AAD. It is thought that these 9aaTAD regions mediate conserved interactions with transcription cofactors (S. Piskacek *et al*., 2007), emphasising the need for the 9aaTADs to be under negative selection. Further analysis into these transcription factor-specific regions of *ZFY* is necessary to understand its conservation and explain its evolution.

This chapter aims to understand the conservation of *ZFY* as a Y-linked gene following the eutherian/marsupial divergence. By assessing 18 eutherian land mammals the aim is to explain the process of change from an autosomal gene to a sex-chromosome gene and identify any major changes in transcription-factor-specific regions that have resulted in *ZFY*'s current functions.

2.2 Materials and Methods

2.2.1 Nucleotide and Protein Sequence Collection

The National Centre for Biotechnology Information (NCBI) database was the primary source for collecting *ZFY* coding domain sequences (CDS). The NCBI search tool was used with the keywords "*ZFY*" and "zinc finger Y-chromosomal protein" to identify available sequences from placental mammals. Many NCBI entries were partial sequences, as Y chromosome sequencing lags behind that of the X chromosome. The lack of complete *ZFY* sequences for many species and the availability of only partial sequences significantly reduced the sample size, as incomplete sequences could not be included.

For species with complete *ZFY* sequences, the corresponding *ZFX* coding sequences were retrieved from the NCBI gene database using the search terms "*ZFX*" and "zinc finger X-chromosomal protein". If the *ZFX* sequence was not available, the species was removed. From the NCBI protein database, the corresponding protein sequences were then collected. The search found 18 placental land mammals with full *ZFY* and *ZFX* annotation (*supplementary* **Table 1** & **Table 2**).

Equus caballus (horse) was a mammal of interest, however complete *ZFY* nucleotide and protein sequences were not available on NCBI at the time. GTF files for the horse Y chromosome were obtained from Janečka and colleagues' paper (Janečka *et al*., 2018). The *ZFY* coordinates were extracted from the GTF file and used to extract the *ZFY* nucleotide sequence from the whole horse Y chromosome sequence in NCBI. The EMBOSS toolkit was then utilised to translate the nucleotide sequence into the corresponding *ZFY* protein sequence.

An additional 4 species where the nearest *ZFX/Y* homolog is autosomal were also identified and, sequences from these were used as outgroups for the analysis (*Supplementary* **Table 3**). Throughout the remainder of this chapter, these autosomal homologs are referred to as “Zf*” to indicate that they are not X- or Y-linked. Zf* sequences from fish species including *Danio rerio* (zebrafish), *Carcharodon carcharias* (Great White Shark), Taeniopygia guttata (Zebra finch) and *Takifugu rubripes* (Japanese puffer) were also originally collected but these were later excluded due to difficulties with sequence alignments.

2.2.2 Sequence Alignment

The collected sequences were aligned using the program “Molecular Evolutionary Genetic Analysis” (MEGA, V10.2.6, (Tamura *et al*., 2021)). MEGA was used for phylogenetic analysis, as it is an integrated tool for conducting both automatic and

manual sequence alignment, inferring phylogenetic trees, estimating rates of molecular evolution, and has many other features for evolutionary hypotheses.

As aligning placental mammals with outgroup species is a complex task, initially the placental mammals were aligned. Initial protein sequence alignment was performed using ClustalW (Thompson *et al*., 1994) in MEGA (V10.2.6, (Tamura *et al*., 2021)) with the default parameters shown in **Table 2.1**. The resulting alignments were then manually curated to improve accuracy. Separate protein sequence alignments were generated for *ZFY* sequences only, *ZFX* sequences only, and combined *ZFY*/*ZFX* sequences. Subsequently, separate alignments for autosomal mammals were created. The alignments were then combined using Clustal omega (V1.2.2, (Sievers *et al*., 2011)) using the default commands and the autosomal species from *supplementary* **Table 3** were seamlessly incorporated into the alignment without disturbing the pre-existing alignment.

Table 2.1: Default ClustalW Protein alignment paraments on MEGA. These parameters are considered the default for the protein alignment algorithm.

Pairwise Alignment	
Gap Opening Penalty	10.00
Gap Extension Penalty	0.10
Multiple Alignment	
Gap Opening Penalty	10.00
Gap Extension Penalty	0.20
Weight	
Protein Weight Matrix	Gonnet
Residue specific penalties	ON
Hydrophilic penalties	ON
Gap Separation Matrix	4
End Gap separation	OFF

Nucleotide alignment poses increased complexity, as MEGA does not make use of the underlying triplet codon structure of the genome when aligning protein sequences, leading to sequence misalignment. To address this issue, PAL2NAL-EMBL (v14, (Suyama *et al*., 2006)) was employed to convert the protein multiple sequence alignments into nucleotide sequence alignments. This tool considers and assigns the appropriate codon sequences to the DNA sequence. To preserve the open reading frames nucleotides are mostly moved in groups of three while managing frameshifts and indels. By inputting the multiple protein sequence alignment alongside the raw nucleotide sequences collected from NCBI, a more dependable and true-to-life nucleotide alignment was achieved.

Following this, both aligned protein and nucleotide sequences were constructed and were ready for subsequent analysis.

2.2.3 Phylogenetic Tree Construction

Maximum-likelihood tree construction followed sequence alignment to determine the evolutionary history of the gene. Initially, MEGA was employed for phylogenetic tree construction, due to its diverse array of phylogenetic analysis tools, including best-fit, pairwise distance, and mean substitution. The best-fit method incorporates a bootstrapping technique with 1000 iterations, this was selected to ensure the robustness of the phylogenetic analysis. However, this approach proved to be time-consuming and the IQ-TREE Web server (W-IQ, (Trifinopoulos *et al*., 2016)) was chosen as a more suitable tool for the analysis requirements.

IQ-TREE employs an ultrafast bootstrapping method, enabling swift and efficient tree construction (UFBoot, (Hoang *et al*., 2018)). The aligned FASTA files were uploaded, and a comprehensive report was generated featuring the best-fit model identified through Bayesian Information Criterion (BIC), an unrooted maximum likelihood tree, and a consensus tree. The default IQ-TREE parameters used are stated in **Table 2.2**. The BIC-best-fit model maximum likelihood tree was exported in Newick format and underwent editing on FigTree (v1.4.4, http://tree.bio.ed.ac.uk/software/figtree/). Basic edits in FigTree involved rooting the tree at the desired species, incorporating bootstrapped values from IQ-TREE, and adding a scale bar. Following these basic adjustments, the tree was exported as an SVG file for further refinement in Inkscape (v1.3.2, https://inkscape.org/release/inkscape-1.3.2/) a graphical editor facilitating image rendering. In Inkscape, additional graphical enhancements such as colouration, size adjustments, and species grouping were added.

Table 2.2: IQTREE default parameters used for both protein and nucleotide analysis. Ultrafast bootstrapping techniques were utilised, with a choice of 1000 iterations made to enhance reliability. The substitution model was configured to identify and apply the best-fit model. The algorithm autonomously identifies the optimal model for the provided sequences, facilitating the construction of an unrooted tree based on the identified model. This is an interactive web server.

IQ-Tree Options - Protein		
Substitution Model	Sequence type/Model	Amino acid
		Find best and apply
	Rate Heterogeneity	Create site rates file
	State Frequency	Estimated by Maximum likelihood
Brach Support	Bootstrap	Ultrafast 1000 replicates
Tree search	Perturbation strength	0.5
	# of unsuccessful iterations to stop	100
Reconstruct ancestral seqs	Root tree	None
	Tree Type	Unrooted
IQ-Tree Options - Nucleotide		
Substitution Model	Sequence type/Model	Nucleotide

		Find best and apply
Brach Support	Bootstrap	Ultrafast 1000 replicates
Tree search	Perturbation strength	0.5
	# of unsuccessful iterations to stop	100
Reconstruct ancestral seqs	Root tree	None
	Tree Type	Unrooted

2.2.4 Conserved Domains Analysis

After generating the sequence alignments, conserved domains within the *ZFY* coding region were analysed, specifically the AAD at the N-terminus and the DBD at the C-terminus alongside the full CDS. These sequences were analysed using Fast, Unconstrained Bayesian AppRoximation (FUBAR, (Murrell *et al*., 2013), https://www.datamonkey.org/fubar) to test for selection. FUBAR prevents the often-normal forcing of sites to belong to one class of unrealistic distribution constraints which often skews inference due to model misspecification by using an approximate hierarchical Bayesian method (Murrell *et al*., 2013). This allows for the identification of sites that are experiencing positive and negative/purifying selection in a much faster and more reliable way. The speed advantage of FUBAR functions well for larger datasets. FUBAR analysis is performed on nucleotide sequence and cannot be used for protein sequence analysis (Murrell *et al*., 2013).

FUBAR analysis was performed on the aligned CDS, but further domain analysis was performed for a more comprehensive analysis. The NCBI Conserved Domain Search tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) was utilised to identify the AAD coordinates in each species' *ZFY* sequence. The NCBI default options were utilised and are shown in **Figure 2.3**. The nucleotide FASTA files were submitted to the tool and the reported coordinates were used to extract the corresponding AAD sequences in the aligned files. These alignments were then submitted to the FUBAR web server. Furthermore, within the AAD, nine amino acid transactivation domains (9aaTAD) are vital to their function. To identify these regions within the sequences an online prediction tool (https://www.med.muni.cz/9aaTAD/) was used. A moderately stringent pattern search was performed as recommended for mammalian transcription factors. Based on the refined criteria, predicted 9aaTADs are identified with either 100% confidence (perfect match) or weaker confidence which are not 100% supported.

The NCBI conserved domain database could not reliably identify the full DBD coordinates. Therefore, to analyse the complete domain, the coordinates were selected starting after the end of the previously identified AAD region to the end of the coding sequence

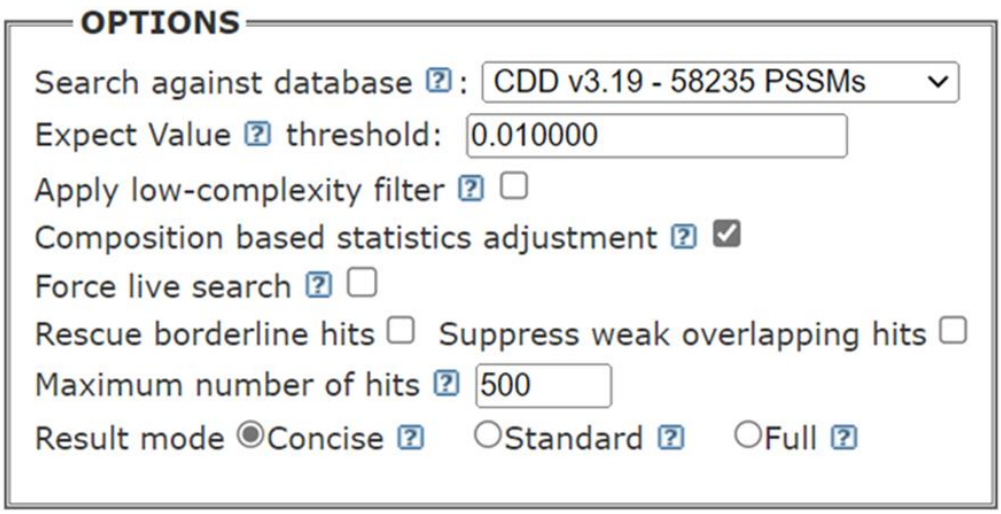


Figure 2.3: NCBI conserved domain search tool default options. These options allow for the collection of acidic transactivation domain coordinates for each species to allow further in-depth analysis.

Within the DBD, an in-depth analysis of the 13 zinc finger motifs was carried out. The zinc finger sequences were identified using the http://zf.princeton.edu tool (Persikov *et al*., 2009);(Persikov & Singh, 2014), which returns the top 13 hits for putative zinc finger motifs from a protein sequence. The zinc finger prediction tool requires protein sequences as the input and does not function with nucleotide sequences, precluding downstream FUBAR analysis. Therefore, after inputting the protein sequences and obtaining the predicted zinc finger motifs, these protein sequences were converted to coding nucleotide sequences using TBLASTN. This allowed for selection analysis on the nucleotide sequences encoding the zinc finger domains for each species.

FUBAR analysis was performed on the full CDS, AAD, DBD, and individual zinc fingers to assess selection. FUBAR evaluates selection on a per-site basis, identifying evidence for either pervasive diversifying or purifying selection (Murrell *et al*., 2013). This online tool utilises a Bayesian approach to analyse substitution rates across sites in a coding sequence alignment. FUBAR provides a more sophisticated selective pressure analysis than overall dN/dS ratios. FUBAR was also utilised due to its ability to analyse large alignments extremely fast over other programmes such as FEL (Murrell *et al*., 2013). By default, a posterior probability of 0.9 was set as >0.9 is strongly suggestive of positive selection.

The nucleotide FASTA files were uploaded for evolutionary analysis employing this Bayesian approach. The resulting data encompassed details related to selection pressure, graphs illustrating posterior rate distribution, and site-specific information.

2.2.5 Geneconv Gene Conversion Tool

Geneconv (v.1.81a) is a software that identifies the most likely possibility for aligned gene conversion events occurring between aligned sequences, as well as possible gene conversions outside of the provided alignment (Sawyer, 1999). These events are ranked by multiple-comparison corrected p-values and are listed in the data output. Predicted recombination between two aligned sequences is assessed based on BLAST-like statistics. This means that if two sequences are significantly similar, they are considered as undergoing possible gene conversion (Sawyer, 1999);(Jaya *et al*., 2023).

Geneconv identifies regions exhibiting large identical DNA stretches as potential conversions (Casola *et al*., 2012). These sites are identified by the comparison of the identical fragment lengths against the premutation of the observed alignment. Geneconv utilises two p-value methods to support the identification of a gene conversion; (1) a permutation method and (2) the Karlin & Altschul (Karlin & Altschul, 1990) method based on BLAST tools for DNA sequencing matching. The permutation method has been said to be more accurate whilst Karlin-Altschul is computationally much faster. The permutation procedure first identifies the highest-scoring fragments within the alignment both globally and pairwise (Sawyer, 1989). 10,000 permutations are the standard for Geneconv, meaning that the columns of the alignment are randomly permuted 10,000 times. A maximum fragment length score is calculated from this permuted array globally and pairwise again. A global permutation p-value is based on the number of permuted alignments that have a higher score than the original observed fragment, these p-values are globalised by multiple comparison corrections for all the sequence pairs within the alignment. Pairwise permutation p-values are not corrected for multiple comparisons across the sequence pairs. Using uncorrected p-values can introduce bias as it means gene conversions between more distantly related sequences with a higher density of different sites will not be significant in length against a background of more closely related pairs. This means there are normally fewer significant global fragments identified due to the more conservative methodology (Sawyer, 1989).

After excluding autosomal Zf* sequences from the pre-aligned nucleotide FASTA files, the remaining sequences of the placental mammal species were organised with alternating *ZFY* and *ZFX* sequences. This arrangement facilitated the grouping of data by Geneconv. Subsequently, the organised FASTA file was fed into the Geneconv program (v1.81a) via the command-line interface. By default, Geneconv uses N=10,000 permutations in the permutation procedure. The following options were set: /w123 initialised GENECONV's internal random-number generator to

guarantee that the program output will always be the same, on different computers. The option /lp tells GENECONV to produce lists of pairwise significant fragments and global lists. /b2 was used to make GENECONV consider consecutive pairs in this case the *ZFY* and *ZFX* pairing system.

Executing this command triggers Geneconv to generate an output comprising a FRAG file and a SUM file. The SUM file encompasses log run details and information about the inputted data, while the FRAG file contains information about potential gene conversions. This information includes species data and the corresponding coordinates where gene conversions have been detected. These nucleotide coordinates were utilised to pinpoint the relevant sequence, and through BLASTX analysis, these nucleotide sequences could be compared with the accession numbers of associated species to identify the protein sequence. This identification process allows the determination of gene conversion positions, specifically whether they occur within the DBD region.

2.2.6 Ancestral Sequence Reconstruction

The web server GRASP ((http://grasp.scmb.uq.edu.au/),(Foley *et al*., 2022)) was used to infer potential ancestral sequences for both *ZFY* and *ZFX*. GRASP can determine ancestral character states as well as identify the most strongly supported potential insertions and deletions. The aligned nucleotide sequences previously generated and Newick-formatted maximum-likelihood trees for *ZFY* only, *ZFX* only, and both *ZFY* and *ZFX*, were incorporated into GRASP for ancestral protein analysis. GRASP then generated the most probable ancestral sequence for each node of the tree exported for closer analysis.

2.3 Results

2.3.1 *ZFY* Is Evolving More Rapidly Than *ZFX*, Particular in Rodents

Phylogenetic tree construction was performed to determine the rates of genetic change in *ZFY* across placental mammals. Following *ZFY* and *ZFX* sequence alignment, phylogenetic trees were constructed to examine how *ZFY* is evolving over time for both protein sequences and nucleotide sequences (protein alignment can be found in *Supplementary* **Figure 1**). The aim was to examine the relationships between *ZFY* and *ZFX* sequences to identify genetic changes that have accumulated in *ZFY* over evolutionary timescales which have led to its distinction from *ZFX*.

The protein alignment was generated first due to protein sequences being easier to work with. The complete alignment consisted of 839 sites in total, with 390 conserved sites and 432 variable sites across all species, ranging from placental mammals to fish. There were 37 gaps in the alignment, which can be caused by indels that have occurred during evolutionary time within different species. (*Supplementary* **Figure 1)**.

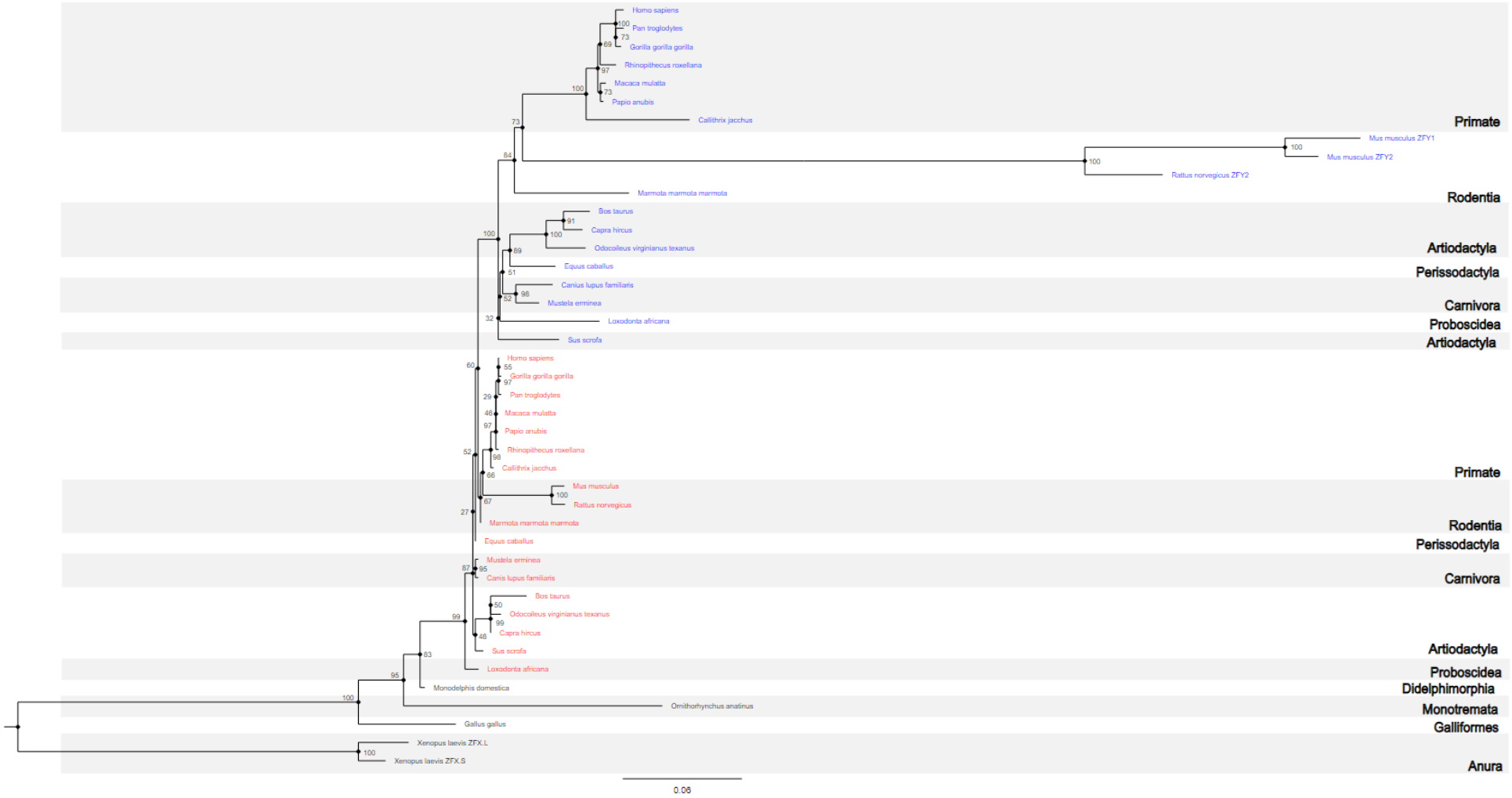


Figure 2.4: Phylogenetic Tree Construction of the ZFY and ZFX protein sequences. All placental mammals and marsupials are included with *Xenopus laevis* used to outgroup. ZFX is annotated as red and ZFY is annotated as blue. Numbers at the node represent the bootstrapping reliability. The scale bar below the tree shows the branch length, measured as substitutions per site. Bayesian information criterion (BIC) score: 16983.7526. Best-fit model according to BIC: HIVb+F+G4.

Figure 2.4 shows the protein phylogeny of the *ZFY* and *ZFX* sequences with *Xenopus laevis* (African clawed frog), *Gallus gallus* (chicken), *Ornithorhynchus anatinus* (platypus) and *Monodelphis domestica* (opossum) with a Zf* sequence as outgroups. The tree is rooted at the outgroup African clawed frog with the first nodes showing the divergence of birds, monotremes and marsupials, in all of which Zf* remains autosomal. Within eutherians, *ZFX* from *Loxodonta africana* (elephant) and then Artiodactyl and Carnivora *ZFX* appear to diverge immediately following the eutherian radiation, prior to the *ZFX*/*ZFY* split. However, these early nodes are not well supported by bootstrapping and are likely to be artefactual.

All eutherian *ZFY* sequences included trace back to one ancestral node supported by a strong bootstrap score of 100%. *ZFY* grouping follows the taxonomic expectations, with most orders of species grouping together with the exception of Artiodactyls. Sus scrofa (pig) seems to diverge away from the other Artiodactyls, however, this lineage break is not strongly supported, with the node only having a confidence value of 32%. This separation is not seen for the *ZFX* sequences. For the remaining *ZFY* sequences, the orders; Primates, Rodentia, and Carnivora cluster as expected with Perissodactyla and Proboscidea only having one species included in this analysis.

Comparing the *ZFX* and *ZFY* regions of the tree, the branch lengths are much longer for *ZFY* than *ZFX*, indicating accelerated evolution of the Y paralog compared to the X homolog. This has been previously observed (Tucker *et al*., 2003) and may be due to the reduced effective population size of the Y compared to the X chromosome. This observation is particularly pronounced for *ZFY* in Rodentia. Rodent *ZFY* has a much longer branch length indicating a large number of genetic changes over time making rodent *ZFY* very distinct both from other (non-rodent) *ZFY* and from *ZFX*. Furthermore, *Marmota marmota marmota* (Marmot) is seen to diverge away earlier than *Mus musculus* (mouse) and *Rattus norvegicus* (rat) with 84% branch support (UFboot). This could suggest that rat and mouse have a more recent common ancestor resulting in a separate rat/mouse rodent lineage before the divergence of the primates, however, with only 84% UFboot support it is not very likely as confident nodes are >95% for this metric. Overall, the rodent *ZFY* sequences appear to have undergone accelerated evolutionary rates compared to other eutherian groups like Primates. A similar acceleration is seen in *ZFX* evolution, with the *ZFX* branch for Rodentia being longer than the branches for other orders suggesting rapid evolution in the rodents. However, this is not as pronounced for *ZFX* as it is for *ZFY* and is also not seen in marmot.

At this level of comparison – i.e. aligning protein sequences - it is evident that the individual species' *ZFY* and *ZFX* sequences are not clustered together or intermingled

as would be expected from recurrent gene conversions. This implies that if there are gene conversion events happening between the two sequences, these are sufficiently small scale not to perturb the overall structure of the phylogenetic tree (see following section). Instead, *ZFY* and *ZFX* appear to be entirely distinct from each other across the whole protein level, diverging very early on, concurrently with the broader differentiation of X and Y nonrecombining regions. However, owing to the strong constraints placed on protein sequences by negative selection, more subtle evolutionary signatures may be obscured in this analysis, and to resolve these it is necessary to align the nucleotide sequences directly.

2.3.2 Possible Gene Conversions Identified by Nucleotide Phylogeny

Following the protein alignment described above, nucleotide alignments were generated specifically for the mammalian sequences. For this alignment, the nonmammalian relatives (frog and chicken) were omitted, and platypus was used as the outgroup due to the difficulty in generating nucleotide alignments over wider phylogenetic distances. Aligning the nucleotide sequences allows the comparisons to make use of information from synonymous changes that distinguish nucleotide sequences without altering the protein-coding sequence. The nucleotide alignment comprised of 2,475 sites, including 1,203 conserved sites and 1,218 variable sites, with 72 identified gaps.

This part of the thesis aimed to identify gene conversions which could have occurred resulting in the high homology of *ZFY* and *ZFX*. By looking at the nucleotide level the hope is to see sequences of high homology clustering together and thus potentially identify species where *ZFY* and *ZFX* have undergone gene conversions.

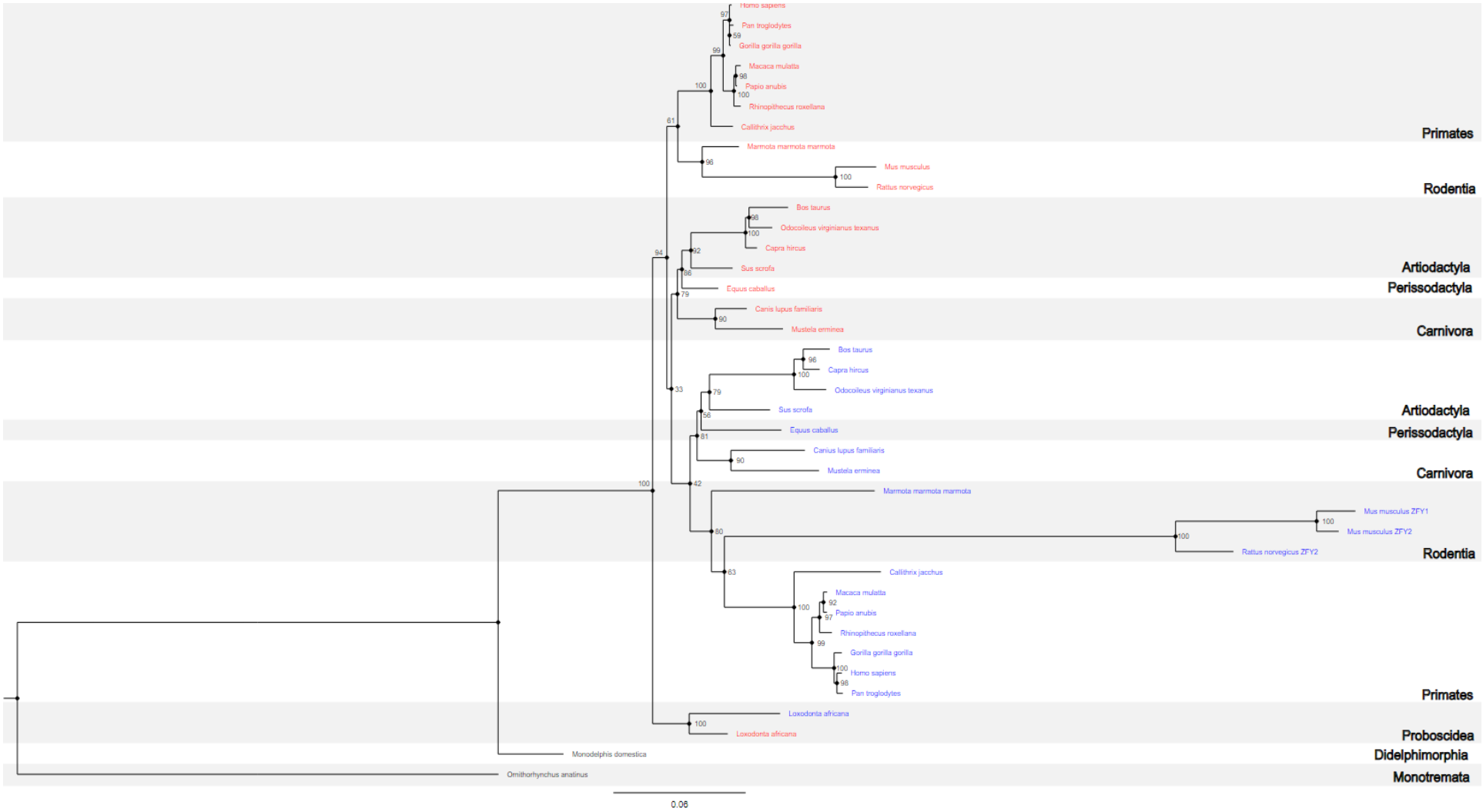


Figure 2.5: Phylogenetic Tree Construction of the ZFY and ZFX nucleotide sequences. All placental mammals and marsupials are included with *Ornithorhynchus anatinus* used to outgroup. ZFX is annotated as red and ZFY is annotated as blue. Numbers at the node represent the bootstrapping reliability. 0.06 shows the length of branch representing 0.06 substitution per site. Bayesian information criterion (BIC) score: 36504.4017. Best-fit model according to BIC: K2P+I+G4.

Figure 2.5 shows the nucleotide phylogeny of the *ZFY* and *ZFX* sequences with platypus as an outgroup. The main difference noted between the protein and nucleotide phylogenetic trees is the clustering of elephants. When looking at the nucleotide level, both elephant *ZFY* and *ZFX* are clustered together, suggesting there may have been an early gene conversion in this species. This branch is highly supported, with a branch support value of 100%. Following the early divergence of the elephants, the tree then breaks to separate *ZFX* in primates and rodents from the remaining *ZFX* and *ZFY* sequences (94%), which is again different to the protein clustering and discordant with the known phylogeny of mammals. These phylogenetic discrepancies are indicative of further potential lateral genetic transfer events (i.e. gene conversions). These are explored further in Section 2.3.5.

2.3.3 Defining the Functional Domains of *ZFY*

Investigations into individual structural and regulatory regions of *ZFY* and *ZFX* across eutherian mammals compared to outgroup species were performed. Key elements under investigation were the 9aaTAD motifs within the AAD, involved in the transcriptional activation of target genes (M. Piskacek, 2009) and the zinc finger domains within the DBD which bind the target DNA (Grover *et al*., 2010).

The general understanding of this class of transcription factors is that DBDs are generally more strongly conserved than AADs within mammals. This has led to a poorer understanding of AAD structure/function relationships because of its poor conservation and being more intrinsically disordered (Udupa *et al*., 2024). However, specific interaction motifs within the AAD, known as 9aaTADs have been identified as being more conserved across species suggesting they possess a high level of importance in the functioning of transcription factors (S. Piskacek *et al*., 2007);(M. Piskacek *et al*., 2016). To demonstrate the high conservation of both 9aaTAD motifs and zinc fingers across *ZFY*/*ZFX* across a plethora of species the coordinates of these regions were mapped across the exon shown in **Figure 2.6**.

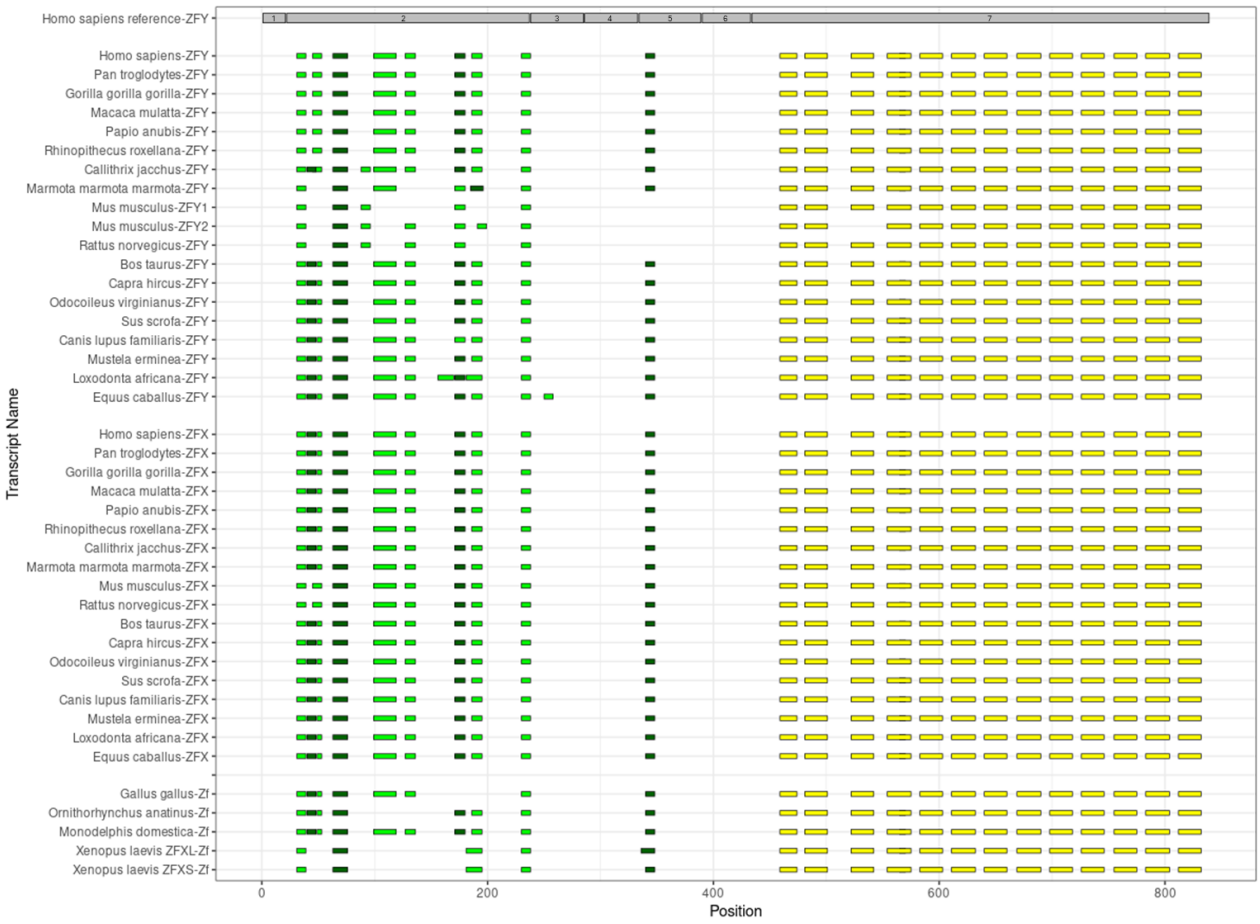


Figure 2.6: A plot showing the reference Homo sapiens *ZFY* transcript (coding exons 1-7) against the 9aaTAD motifs and zinc fingers across all the species included in the analysis. Green: 9aaTAD motifs that were identified but are not 100% confident, Dark Green: 9aaTAD motifs that were identified with 100% confidence & Yellow: zinc finger domains. Exon 2 is the exon that is alternatively spliced in testis.

As seen in **Figure 2.6** there is a great conservation of both 9aaTAD motifs and zinc finger domains which corresponds to their vital importance in both *ZFY* and *ZFX* activity. The majority of the 9aaTAD motifs are located in exon 2, which could explain the difference in transcriptional factor activity between *ZFYS* and *ZFYL* since 9aaTAD motifs are vital for binding the transcriptional machinery. In exon 5 there is a single highly conserved 9aaTAD motif which is constitutively present in all species except mouse and rat. This could be related to the somatic functions of *ZFY* as mouse and rat *Zfy* are testis specific. Furthermore, mouse *Zfy2* has a very high transactivation ability despite the loss of some of the lower confidence 9aaTAD motifs. This could suggest that these motifs are not key to *ZFYS*' transactivation and may be false positive predictions. It was also noted that these motifs were not present in *Xenopus* and are thus not highly conserved in general. When analysing the 9aaTAD motifs between ZFX and ZFY, only 13.5% of the species exhibit identical 9aaTAD mapping,

as shown in **Figure 2.6**. Whilst there is a greater variation across the 9aaTAD motifs, which is expected due to the less organised conservation of the AAD, the zinc fingers are constant throughout the placental and autosomal Zf* (98% of the species have 13 zinc finger domains) with the exception of the mouse *Zfy2* paralogue which is missing zinc finger 3. These zinc finger domains are vital to target binding and suggest that they are under pressure to remain fixed to continue to function as necessary.

2.3.4 Specific Domain Selection Analysis

Further to phylogenetic tree construction, an in-depth analysis of the underlying selection pressure on each region of *ZFY* was investigated. Selection pressure analysis was performed using FUBAR, which investigates the non-synonymous vs synonymous changes at the nucleotide level. This kind of probabilistic analysis is useful for identifying sites that are evolving under selection in protein-coding genes and is much more advantageous to other existing software which are simply too slow due to the size of the dataset (Murrell *et al*., 2013).

2.3.4.1 The *ZFY* Coding Domain Sequence Remains Under Negative Selection

The aligned nucleotide sequences of the coding domain of *ZFY* were inputted into the FUBAR software to obtain selection pressure information across the entire protein-coding gene. An analysis comparing *ZFY* and *ZFX* DNA alignments was also performed.

A

FUBAR found evidence of	
pervasive positive/diversifying selection at:	1 site
pervasive negative/purifying selection at:	533 sites
Total sites	823
Total sequences	21

*with a posterior probability of 0.9

B

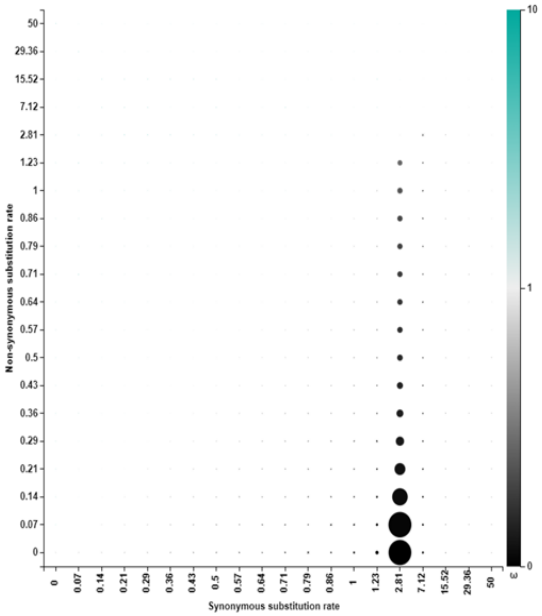


Figure 2.7: Alignment-wide *ZFY* sequence selection pressures. A: A table presenting the selection pressures identified by FUBAR at sites across the coding

domain. **B:** A graph showing the posterior distribution looking at the synonymous vs non-synonymous substitution rate across the coding domain**.** The dot's size correlates with the posterior weight assigned to the grid point, while its colour reflects the intensity of selection (ratio dN/dS). A posterior probability of 0.9 was selected to ensure reliability. A nucleotide alignment consisting of 21 sequences was inputted into the software. The outrgoup species included were platypus and opossum, with eutherians expanding out to primates.

Presented in **Figure 2.7** is the analysis of 823 protein sites performed by FUBAR. FUBAR only found evidence of 1 site under positive selection, with 533 sites identified as being under negative selection. With the posterior probability threshold set at 0.9, the evolutionary pressures on the remaining 289 *ZFY* sites could not be conclusively classified as either positive or negative selection or the sites are neutral. The 1 site identified to be under positive selection is strongly upheld to a posterior probability of 0.98. Looking at the posterior distribution alignment wide in **Figure 2.7B** it is clear that the synonymous substitution rate exceeds the non-synonymous substitution rate in *ZFY*, with DNA level changes not affecting *ZFY*'s protein sequence.

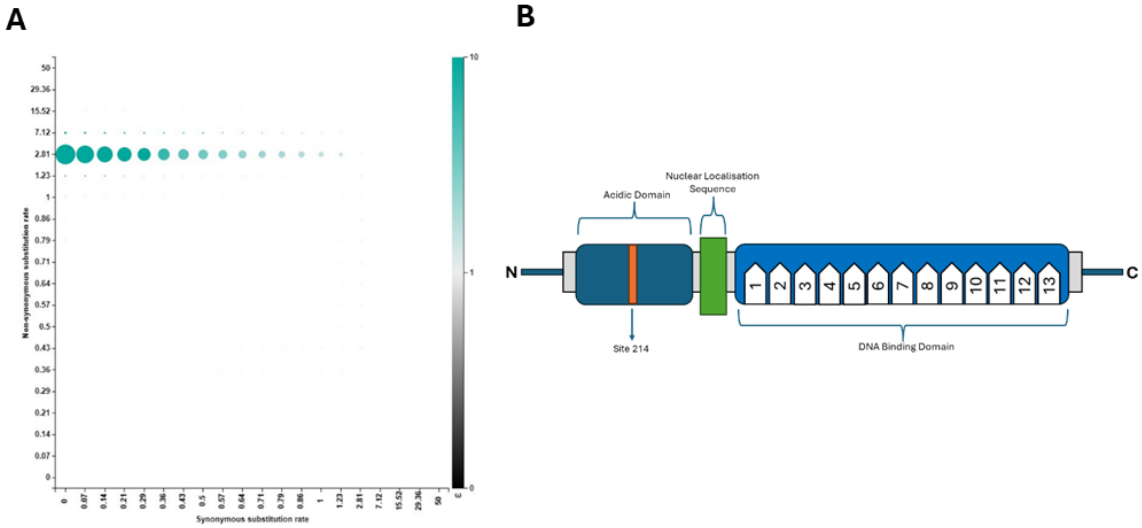


Figure 2.8: Alignment-wide *ZFY* sequence selection pressure at site 214 .A: Site 214 posterior distribution indicative of positive selection at work. **B:** Schematic diagram of site 214's location in *ZFY*'s AAD.

After the identification of 1 site under positive selection, **Figure 2.8A** shows the posterior distribution across the site. It is clear in comparison to the alignment-wide distribution plot, that this site is undergoing positive selection. It is evident that this site's non-synonymous substitution rate exceeds its synonymous substitution rate, confirming the presence of positive selection at this site. Site 214 is an amino acid located in the AAD portion of *ZFY* and is not within a predicted 9aaTAD region which are known regions of greater conservation due to their functional importance in the AAD **(Figure 2.8B)**. Site 214 encodes the amino acid Alanine (A) in the primates

except marmot where A is replaced with Glycine (G) with this amino acid also being present in other species such as the artiodactyls. Across the rodents, this site encodes different amino acids with the marmoset encoding Serine (S), mouse encoding Isoleucine (I) and rat encoding valine (V), this could indicate that this site is under differing pressures in these species that is beneficial to the rodent family. Other species with an S at site 214 include pig, Mustela, horse and opossum. Dog is the only species where site 214 encodes Asparagine (N). Though the rodents show changes in the encoded amino acid they are not the only species undergoing changes at this site.

2.3.4.2 The *ZFY* and *ZFX* Coding Domain Sequences Show Similar Selection Pressures

Following an initial selection pressure analysis on *ZFY* sequences in isolation, the *ZFY* and *ZFX* combined nucleotide alignment was inputted into the FUBAR software to see if any significant distinctions between them are identifiable that are leading to their continuation on the Y and X chromosomes respectively.

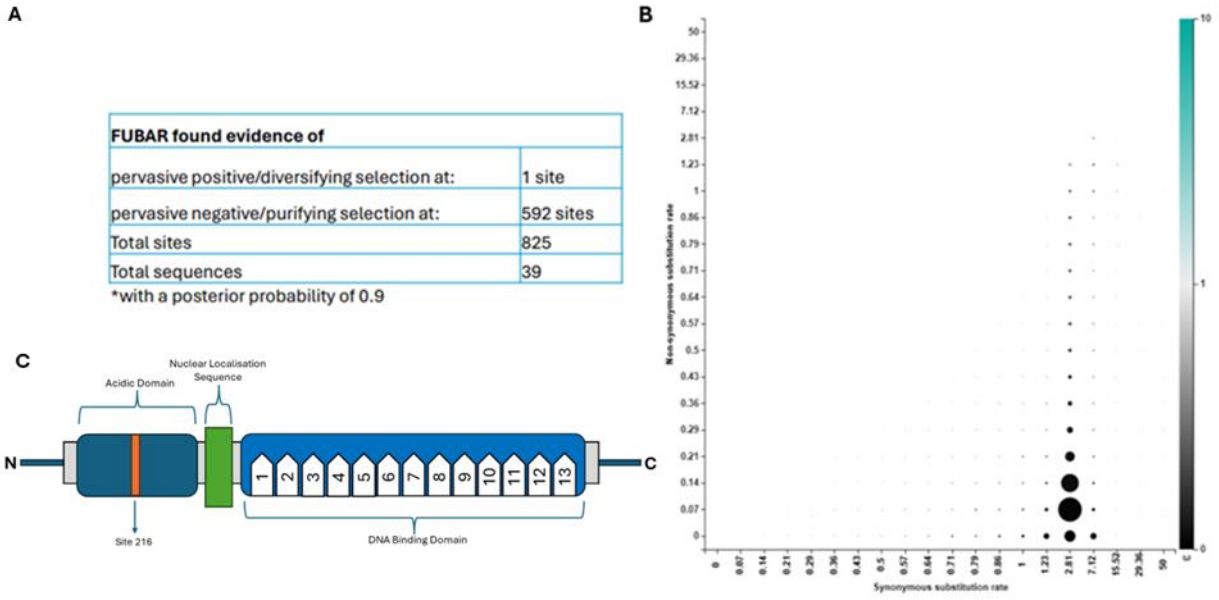


Figure 2.9: Alignment-wide *ZFY* and *ZFX* sequence selection pressures. A: A table presenting the selection pressures identified by FUBAR at sites across the coding domain. **B:** A graph showing the posterior distribution looking at the synonymous vs non-synonymous substitution rate across the coding domain. **C:** Schematic diagram of site 216's location in *ZFYS* AAD. A posterior probability of 0.9 was selected to ensure reliability. A nucleotide alignment consisting of 39 sequences was inputted into the software. Species used in this analysis included platypus as the monotreme specie, opossum as the marsupial specie with eutherians expanding out to primates.

The alignment wide posterior distribution of *ZFY* and *ZFX* combined in **Figure 2.9** highly resembles **Figure 2.8** which looked at *ZFY* in isolation. Again, one sole site was found to be under positive selection in both *ZFY* and *ZFX*, and this is site 216, an amino acid located in the acidic activating domain and not within a known 9aaTAD. Further inspection showed that this site is the same site identified in **Figure 2.8** from the *ZFY*-specific analysis and is spliced out in *ZFYS*. This change in site number is due to the addition of more sequences and has shifted the alignment slightly. This site in all the primate *ZFX* sequences encodes G, alongside the artiodactyls. In the *ZFX* rodents, the marmot encodes S, the mouse encodes N, and the rat encodes A, all distinct from the *ZFY* rodents. However, the rat *Zfx* site corresponds to the primate *ZFY* sequences. Similar to marmot, pig, dog, mustela, elephant and horse *ZFX* sequences encode S which is concordant with mouse evolving faster than rat and marmot.

Both *ZFY* and *ZFX* appear to undergo evolutionary changes influenced by negative selection, as indicated in **Figure 2.9** where the count of synonymous substitutions surpasses non-synonymous substitutions, signifying their crucial biological role. The shared identification of a specific site in both analyses reveals that this site is subject to positive selection in both *ZFY* and *ZFX*. Given its location within the acidic domain region, this suggests a potential functional significance in relation to *ZFY*/*ZFX* and their evolution across various species.

2.3.4.3 The Analysis of the *ZFY* Acidic Activating Domain and DNA Binding Domains Evolution

Following coding domain-wide analysis of the selection pressures acting on *ZFY* and *ZFX*, a look into the AAD and DBD of *ZFY* was carried out.

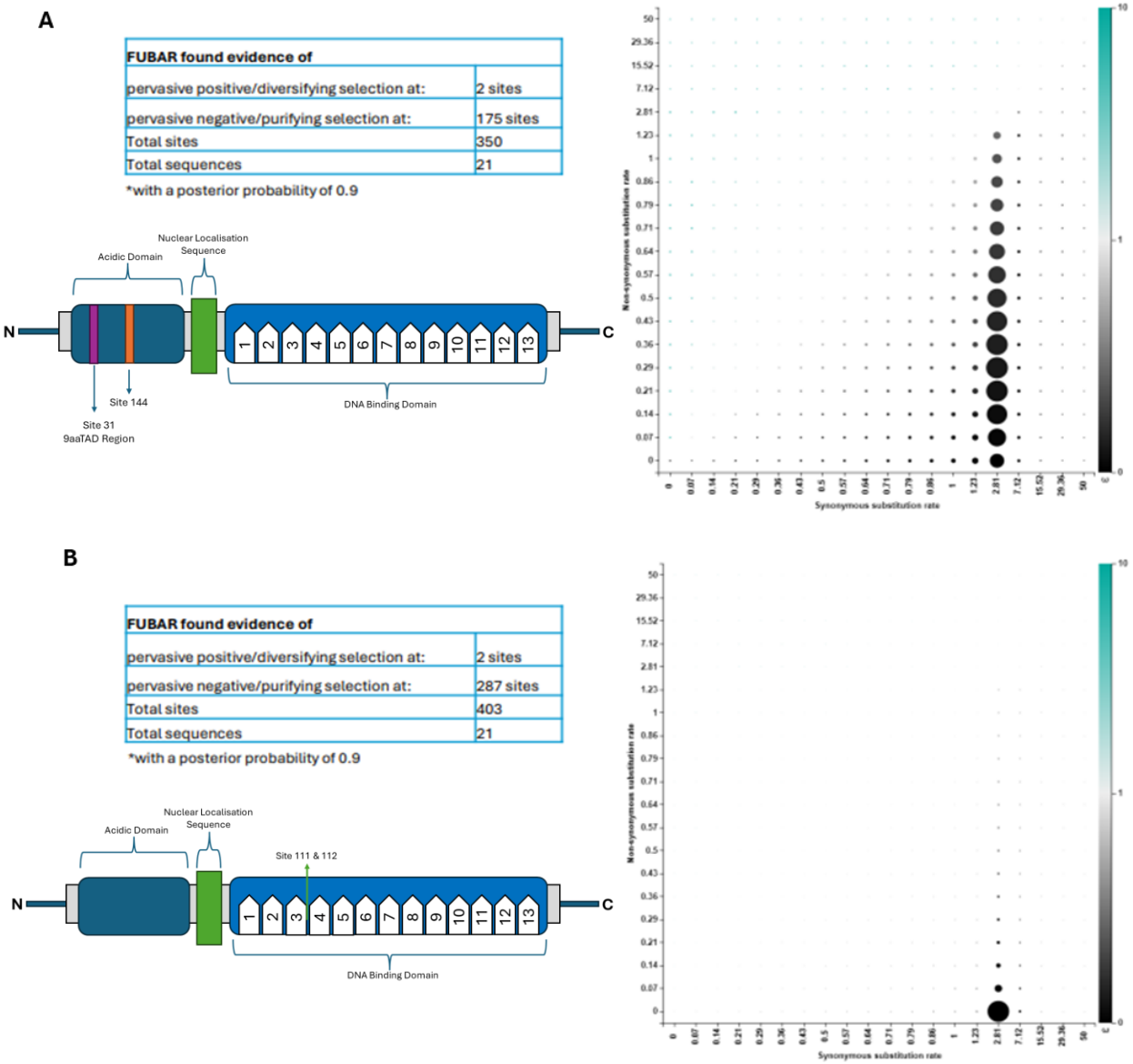


Figure 2.10: *ZFY* specific posterior distribution of (A): the acidic activating domain and **(B):** the DNA binding domain of *ZFY* indicating either positive or negative selection. A posterior distribution of 0.9 was set.

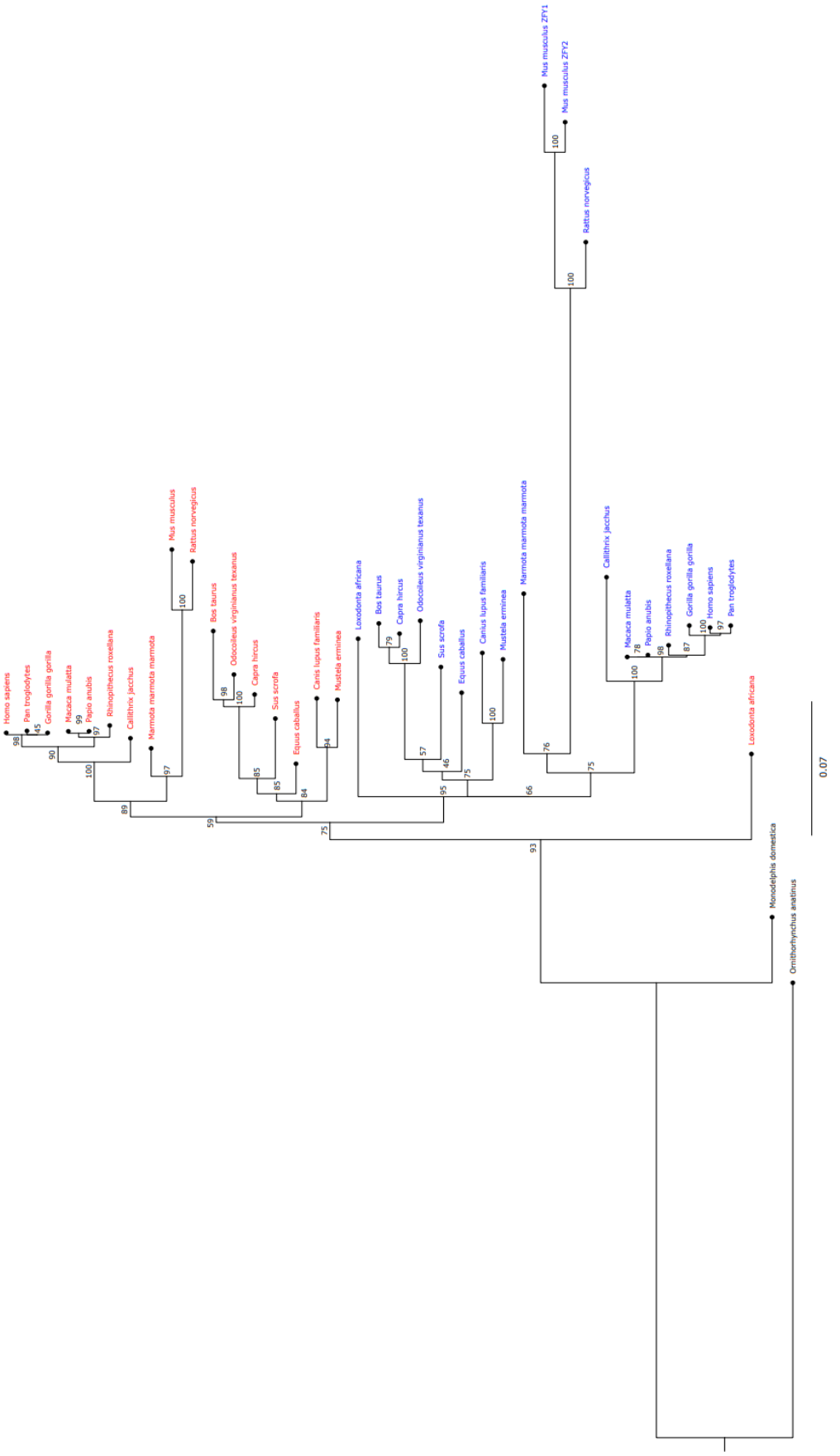
Fubar analysis in **Figure 2.10** shows that both the *ZFY* DBD and AAD have 2 sites identified as being under positive selection based on these sites having a posterior probability greater than the limit of 0.9. Many of the sites in the AAD and DBD were under negative selection, ensuring the conservation of the protein structure and function in these transcription factor-specific regions. However, sites 31 and 144 in the AAD were identified as being under positive selection. The remaining sites cannot be defined as under positive or negative selection or are neutral. Site 31, identified as undergoing positive selection, resides within a predicted 9aaTAD region pinpointed by a tool referenced in section 2.2.1. However, it is worth noting that this particular 9aaTAD region doesn't perfectly match the tool's prediction, thus hindering absolute confirmation. Site 31 interestingly is mostly conserved in the primates and rodents

(mouse and rat) with the site encoding Aspartate (D) while the majority of changes are seen in the artiodactyl (S), carnivore species (G, D) and marmot (N). Conversely, site 144 doesn't fall within a 9aaTAD but aligns with the previously identified site that underwent positive selection in the whole sequence alignment analysis. The two sites (111 & 112) identified as under positive selection in the zinc finger DBD are located in the region between zinc finger 3 and 4. All *ZFY* primates encode AN at these sites, and changes are seen in the rodents (marmot = NN, mouse and rat = VN). Many of the other sequences encode variations of AN, with some encoding A and a different second protein, whilst others encode an alternative protein and then N. Elephant isn't included in this rule and encodes SS at these identified sites. Though these sites are undergoing positive selection, the selection of each species seems to differ except for the primates. Zinc fingers and 9aaTADs are thought to be the more conserved regions of the DBD and AAD respectively, which would explain why these sites largely remain under negative selection to ensure the function of the protein remains intact.

2.3.5 Exon 7 is Subject to Gene Conversion and Rapid Evolution

Given the evidence for gene conversion seen in **Figure 2.5** and since gene conversions typically affect short regions of DNA and thus likely only convert part of the *ZFX*/*ZFY* sequence in any given species, this chapter sought to explore this further by looking separately at the phylogenetic history of different domains of the gene. Following on from the whole-gene analysis presented above, separate phylogenetic trees were constructed looking either at the transactivation domain (coding exons 1-6) or DNA binding domain (coding exon 7) of the gene.

A


Mus musculus ZFY1
Mus musculus ZFY2
Rattus norvegicus
Mus musculus
Rattus norvegicus
Bos taurus
Odocoileus virginianus texanus
Capra hircus
Sus scrofa
Equus caballus
Canis lupus familiaris
Mustela erminea
Loxodonta africana
Bos taurus
Capra hircus
Odocoileus virginianus texanus
Sus scrofa
Equus caballus
Canis lupus familiaris
Mustela erminea
Marmota marmota marmota
Callithrix jacchus
Macaca mulatta
Papio anubis
Rhinopithecus roxellana
Gorilla gorilla gorilla
Homo sapiens
Pan troglodytes
Homo sapiens
Pan troglodytes
Gorilla gorilla gorilla
Macaca mulatta
Papio anubis
Rhinopithecus roxellana
Callithrix jacchus
Marmota marmota marmota
Loxodonta africana
Monodelphis domestica
Ornithorhynchus anatinus
0.07

B

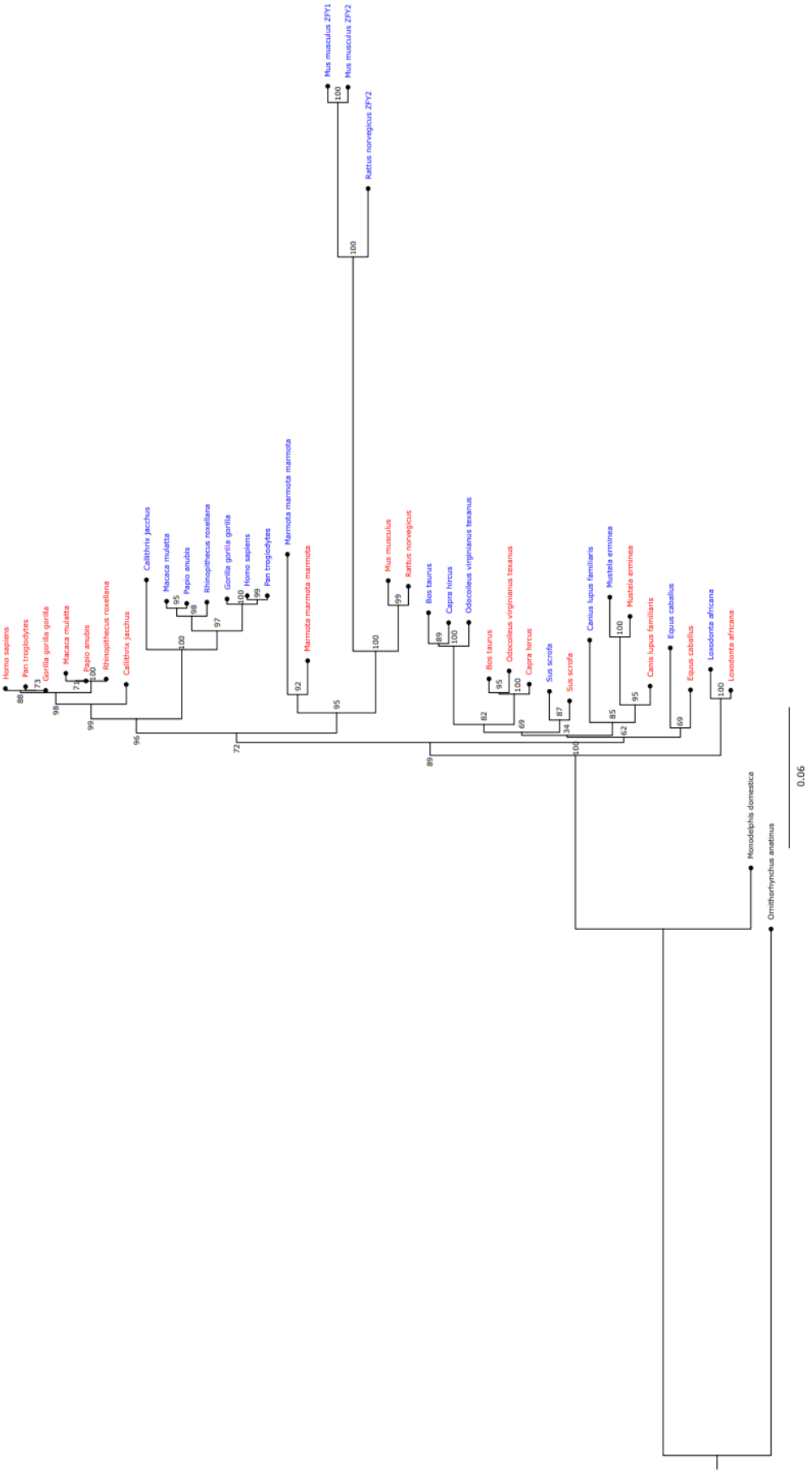

Mus musculus ZFY1
Mus musculus ZFY2
Rattus norvegicus ZFY2
Homo sapiens
Pan troglodytes
Gorilla gorilla gorilla
Macaca mulatta
Papio anubis
Rhinopithecus roxellana
Callithrix jacchus
Marmota marmota marmota
Mus musculus
Rattus norvegicus
Bos taurus
Capra hircus
Odocoileus virginianus texanus
Sus scrofa
Canis lupus familiaris
Mustela erminea
Equus caballus
Loxodonta africana
Monodelphis domestica
Ornithorhynchus anatinus
0.06

Figure 2.11: A: Nucleotide phylogenetic tree construction of *ZFY* coding exons 1-6, best-fit model according to BIC = K2P+G4 (BIC score: 19168.9192)**. B: Phylogenetic tree construction of *ZFY* coding exon 7**, best-fit model according to BIC = TIM3e+I+G4 (BIC score: 17083.7858)**.** Both trees are rooted at platypus for standardisation and the amount of genetic change for the branch length is noted for each tree.

The accumulation of DNA changes across coding exons 1-6 of *ZFY* during its transition from an autosomal gene to a sex-linked gene at the marsupial/eutherian divergence is shown in **Figure 2.11A**. This phylogenetic tree emphasises the first six coding exons, which encode the N-terminus AAD and mirrors the divergence pattern observed in the nucleotide phylogenetic tree of the entire sequence depicted in **Figure 2.5**. Figure **2.11B** illustrates the DNA changes in coding exon 7 only which encodes the DBD.

Separating the two functional domains of the gene in this way shows that in **Figure 2.11A**, there is a clear separation between *ZFY* and *ZFX* sequences and that within the *ZFX* and *ZFY* sub-trees, there are no discordances with the expected phylogeny of mammalian orders – thus we can conclude that there are few if any gene conversion events within this region of the gene. In contrast, when attention shifts to the nucleotide alignment of coding exon 7 in **Figure 2.11B**, the narrative takes a different turn with a series of potential gene conversions in elephant, horse, pig, marmot and stoat. These are identified by the clustering of these species' *ZFY* and *ZFX* sequences together, indicating they are highly similar in this exon. Overall, *ZFY* and *ZFX* sequences are no longer grouping separately but are intermingled, with many of the individual species or taxa clustering their *ZFY* and *ZFX* sequences together.

The overall picture is therefore one of recurrent gene conversion occurring specifically within exon 7, leading to the differences in phylogenetic association between, **Figures 2.11A** and **2.11B**, with **Figure 2.5** representing a confounded amalgamation of the signal across different regions of the gene. This restriction of conversion events to exon 7 aligns with the imperative to preserve the functionality of the DBD, and furthermore suggests there may have been evolutionary pressure to ensure the X and Y paralogs maintain identical (or closely similar) DNA binding domains within each species. Thus, *ZFX* and *ZFY* likely share a common set of downstream targets, but their effects on those targets may differ due to the accumulated changes in the activation domain (exons 1-6) that have not been homogenised between X and Y copies. Interestingly, however, the acceleration of *ZFY* evolution seen in **Figure 2.4** and **Figure 2.5** is seen in both the activation domain and DBD. In particular, in exon

7, in each paired set of *ZFY*/*ZFX* genes (partially or wholly homogenised at the root of the pair by gene conversion), the following Y branch is longer than the X branch.

2.3.6 Genetic Exchange Between *ZFY* and *ZFX* Could Have Led to Their High Homology, but They Continue to Persist for Male and Female Function Respectively

It is thought that the divergence pattern of mammalian *ZFY* and *ZFX* must be down to X and Y chromosome conversions (or recombinations) due to their not completely independent separation from each other (Pamilo & Bianchi, 1993). Following on from the identification of recurrent gene conversion in exon 7, Geneconv was utilised to identify the potential breakpoints for the gene conversions between aligned DNA sequences (Sawyer, 1989). Geneconv is able to identify possible gene conversions within and outside the alignment. Inner fragment conversion is when a possible gene conversion between ancestors of two given sequences in the aligned file is identified, while an outer sequence fragment conversion is when there is evidence of a possible gene conversion outside of the alignment or was within the alignment but then was later destroyed by continuous genetic mutations or gene conversion (Sawyer, 1989).

2.3.6.1 Global Fragment Analysis Reveals Possible Gene Conversions Across a Range of *ZFY*/*ZFX* Species

Global fragment analysis reveals gene conversions with p-values corrected for both the sequence length and sequence number. Because of the more stringent correction for global analysis compared to pairwise analysis, fewer conversions are normally detected. Geneconv only outputs suspected gene conversions if they pass at least one of the p-values for significance.

Table 2.3: Global inner fragment analysis. The table displays potential conversions identified using Geneconv. Permutation (sim) p-value and Bonferroni-corrected (BC) KA p-values are listed for each identified conversion. Num. Poly = Number of Polymorphisms, Num. Diffs= Number of Differences, Total Diffs = Total number of differences. * Indicates significant p-values <0.05 highlighting the cell background.

Sequence Names	Sim Pvalue	BC KA Pvalue	Aligned Offsets			Num. Poly	Num. Diffs	Total Diffs	Mismatch Penalty
			Begin	End	Length				
Rattus norvegicus *ZFY*2; Rattus norvegicus *ZFX*	0.0311*	0.05350	2353	2426	74	21	0	437	None

Loxodonta Africana *ZFY*; Loxodonta africana *ZFX*	0.0316*	0.05354	1931	2189	259	82	0	123	None
Mustela erminea *ZFY*; Mustela erminea *ZFX*	0.0430*	0.07910	2032	2279	248	79	0	123	None

Table 2.3 shows the global inner fragments that have possibly undergone gene conversion making their DNA homology significantly higher than expected. Geneconv identified three potential gene conversions between *ZFY* and *ZFX* in rat, elephants and *Mustela erminea* (stoat). A significant permutation p-value (0.0311) was calculated for a global fragment between rat *Zfy* and *Zfx* at the offsets 2352-2426 in the alignments which corresponds to a region within zinc fingers 12 and 13 of the DBD **(Table 2.3)**. This means 311 out of the 10,000 permutations across the global alignment had fragments as long or longer than this sequence pair. This tract had a similarity of 74 nucleotides with 21 polymorphic sites identified between the two rat sequences. Overall, 437 sites were identified to be different between the two aligned sequences. No internal mismatches were identified in this tract and there was no applied mismatch penalty. This is indicative of a potential gene conversion; however, the evidence of this gene conversion is not significant for the BC KA p-value but indicates something of interest. Whilst BC-KA p-values are less accurate, and this tract is suggestive of a gene conversion additional evidence would be required for confirmatory purposes. However, when looking at the less stringent pairwise analysis, this fragment has a permutation score of p=0.0037 (37 permutations out of 10,000 have a longer fragment) (*Supplementary* **Table 4**) and passes KA p-value testing, but as mentioned previously this kind of testing is not corrected for multiple-comparison and only looks at the specific sequence pair and not all the potential pairs leading to bias.

Another significant global permutation p-value (0.0316) was calculated for a 259-nucleotide global fragment between elephant *ZFY* and *ZFX* at the offsets 1931-2189 in the alignment **(Table 2.3)**. This region encompasses zinc fingers 7 through 10 in the DBD of *ZFY*. Of the 259 nucleotides, 82 polymorphic sites were identified between the two elephant sequences. No internal mismatches were identified in this tract and no applied mismatch penalty. This is indicative of a potential gene conversion but like the rat global fragment, the conversion is not significant for the BC-KA p-value. A significant pairwise permutation p-value (0.0016) and a significant KA p-value (0.00282) were identified for this fragment alongside two other fragments identified as significant by pairwise analysis (*Supplementary* **Table 4**). A 207-nucleotide

fragment spanning zinc fingers 10 through 13 and a 165-nucleotide fragment spanning zinc fingers 4 and 5 were also identified as possible regions of gene conversions through pairwise analysis, these need to be treated with caution as they are not corrected based on the global alignment (*Supplementary* **Table 4**). These finding of potential gene conversion in elephants supports the branching of the phylogenetic tree in **Figures 2.5** and **12.1B**.

Finally, a global inner fragment at the offsets 2032-2279 in the stoat alignment has a significant similarity p-value (0.0430) identifying it as another possible region suspected of a gene conversion. This identified region encompasses zinc fingers 9 through 11, with 79 sites identified to be polymorphic. Overall, 123 sites were identified as being different, but no internal mismatches were identified. No mismatch penalty was present. Like the elephant, another inner fragment was identified by pairwise analysis spanning zinc fingers 6 through 8, but global correction removed this identified fragment (*Supplementary* **Table 4**).

Overall, all three global inner fragments are indicative of gene conversion, but due to failing BC-KA p-value significance but passing the more conservative permutation testing, whilst suggestive they cannot be fully trusted and therefore cannot be confirmed as gene conversions. However, they do further emphasise that the DBD is undergoing gene conversions as seen in **Figure 2.11B**.

Global outer-segment fragment analysis was also performed by Geneconv and aimed to identify possible gene conversions that have occurred outside of the alignment or are hidden by a vast number of mutations or conversions. An outer fragment is defined by Geneconv as a maximal DNA length where a single sequence contains all unique nucleotides at each polymorphic site in the alignment. This fragment is unique in the alignment and is not shared with any other sequence. This could indicate a gene conversion from outside of the chromosome.

Table 2.4: Global outer-segment fragment analysis. The table displays potential conversions identified using Geneconv. Permutation (sim) p-value and Bonferroni-corrected (BC) KA p-values are listed for each identified conversion. Num. Poly = Number of Polymorphisms, Num. Mats= Number of nonunique sites within the fragment, Total Mats = Total number of nonunique sites for that sequence. * Indicates significant p-values <0.05.

Sequence Names	Sim Pvalue	BC KA Pvalue	Aligned Offsets			Num. Poly	Num. Mat	Total Mats	Mismatch Penalty
			Begin	End	Length				
Bos taurus *ZFY*	0.0000*	0.00000*	1050	1067	18	14	0	842	None
Bos taurus *ZFY*	0.0045*	0.00831*	1035	1048	14	8	0	842	None

Two potential global outer-segment fragments were pinpointed in the AAD of cow *ZFY* as possible regions of gene conversions and are highlighted in **Table 2.4**. These two short fragments are located in the AAD of the cow and have been classed as unique to the cow *ZFY* sequence, as these specific sequences are not detectable elsewhere in the global alignment. Both identified fragments exhibit significant p-values under permutation and BC KA metrics indicating strong evidence of a conversion within cow *ZFY* but potentially from outside of the Y chromosome. This could indicate that cow *ZFY* is slowly developing novel functions through alterations in the AAD responsible for binding transcriptional machinery.

Pairwise outer-sequence fragments were also identified in the acidic domain of marmot and mouse *ZFY*, which corroborates with their rapid evolutionary divergence (*Supplementary* **Table 5**). Whilst, both p-value metrics provide significant support for these unique fragments between sequence pairs, they are not unique across the global alignment and therefore are not recognised by the global comparison. This indicates that whilst marmot *ZFY* and mouse *Zfy2* have short unique fragments with their *ZFX* homologue, these fragments are not unique to the *ZFX*/*ZFY* family. So, while marmot and mouse *ZFY* display species-specific accelerated evolution against their *ZFX* ortholog, they do not possess substantial activating domain uniqueness globally.

2.3.6.2 Calibrating the Phylogeny Using Known Species Divergence Times

The next analysis delved further into the acceleration of *ZFY* evolution in rodents by computing the substitution rate per million years for every ancestral branch. By generating a time tree ((https://timetree.org/), (Kumar *et al*., 2022)) the number of years between each node in millions of years was gathered. The Newick trees contained the branch length in substitutions per site between nodes. By using these, the number of substitutions per site per millions of years was calculated for each branch. This would allow confirmation of which nodes are undergoing greater genetic changes such as the rodents.

Table 2.5: Timetree of *ZFY* nucleotide coding domain divergence. Through Timetree.org, the age of each node in million years (MYA) was identified. Substitutions per site for each branch were then determined using IQtree. By combining these data points, the substitutions per site per million years were calculated for each branch. Substitution columns are coloured by value highlighting low (light-coloured) and high (dark-coloured) substitution rates. Note: *Equus caballus* is excluded as the *ZFY* sequence does not group as expected making node calculations difficult.

Starting Node	End Node	Time of Starting Node (MYA)	End of Starting Node (MYA)	Time Covered by Branch (MYA)	Substitutions per Site	Substitutions per site per Myr
Theria	*Monodelphis domestica*	160	0	160	0.0271	0.000169
Theria	Eutheria	160	99.2	60.8	0.0663	0.00109
Eutheria	*Loxodonta africana*	99.2	0	99.2	0.0498	0.000502
Eutheria	Boreoeutheria	99.2	94	5.2	0.0081	0.001558
Boreoeutheria	Euarchontoglires	94	87.2	6.8	0.0089	0.001309
Boreoeutheria	Artiodactyl - Caniformia	94	76	18	0.0038	0.000211
Euarchontoglires	Rodentia	87.2	68.8	18.4	0.0064	0.000348
Euarchontoglires	Simiiformes	87.2	42.9	44.3	0.0348	0.000786
Rodentia	*Marmota marmota marmota*	68.8	0	68.8	0.0645	0.000938
Rodentia	*Rattus norvegicus – Mus musculus*	68.8	13.1	55.7	0.1859	0.003338
Rattus norvegicus – Mus musculus	*Rattus norvegicus*	13.1	0	13.1	0.0255	0.001947
Rattus Norvegicus – Mus musculus	*Mus musculus*	13.1	0	13.1	0.0607	0.004634
Simiiformes	*Callithrix jacchus*	42.9	0	42.9	0.0381	0.000888
Simiiformes	Catarrhini	42.9	28.82	14.08	0.0082	0.000582
Catarrhini	Cercopithecidae	28.82	17.75	11.07	0.0035	0.000316

Cercopithecidae	*Rhinopithecus roxellana*	17.75	0	17.75	0.0056	0.000315
Cercopithecidae	*Papio anubis – Macaca mulatta*	17.75	10.45	7.3	0.0018	0.000247
Papio anubis – Macaca mulatta	*Papio anubis*	10.45	0	10.45	0.0014	0.000134
Papio anubis – Macaca mulatta	*Macaca mulatta*	10.45	0	10.45	0.0015	0.000144
Catarrhini	Hominidae	28.82	8.6	20.22	0.0102	0.000504
Hominidae	*Gorilla gorilla gorilla*	8.6	0	8.6	0.0034	0.000395
Hominidae	*Homo sapiens – Pan troglodytes*	8.6	6.4	2.2	0.0013	0.000591
Homo sapiens – Pan troglodytes	*Homo sapiens*	6.4	0	6.4	0.0021	0.000328
Homo sapiens – Pan troglodytes	*Pan troglodytes*	6.4	0	6.4	0.0026	0.000406
Artiodactyl - Caniformia	Caniformia	76	45.1	31	0.015	0.000484
Caniformia	*Canis lupus familiaris*	45.1	0	45.1	0.0313	0.000694
Caniformia	*Mustela erminea*	45.1	0	45.1	0.0398	0.000882
Artiodactyl - Caniformia	Artiodactyl	76	61.8	14.2	0.0066	0.000465
Artiodactyl	*Sus scrofa*	61.8	0	61.8	0.0268	0.000434
Artiodactyl	*Odocoileus virginianus texanus – Bos taurus – Capra hircus*	61.8	22.79	39.01	0.0371	0.000951

Odocoileus virginianus texanus – Bos taurus – Capra hircus	*Odocoileus virginianus texanus*	22.79	0	22.79	0.0147	0.000645
Odocoileus virginianus texanus – Bos taurus – Capra hircus	*Bos taurus – Capra hircus*	22.79	22	0.79	0.004	0.005063
Bos taurus – Capra hircus	*Bos taurus*	22	0	22	0.012	0.000545
Bos taurus – Capra hircus	*Capra hircus*	22	0	22	0.0074	0.000336

Table 2.5 shows the accumulation of substitutions across the branches seen in **Figure 2.5,** which highlights the rapid Rodentia evolution. Within the Rodentia clade, substantial changes in substitution rates occur at each diverging node, indicating that they are accumulating a higher number of substitution changes over time, driving their rapid evolution away from other species. This corresponds with the extensive branch lengths seen for this group at both the protein **(Figure 2.4)** and nucleotide **(Figure 2.5)** levels.

2.3.7 Ancestral Reconstruction of *ZFY*/*ZFX* Ancestors

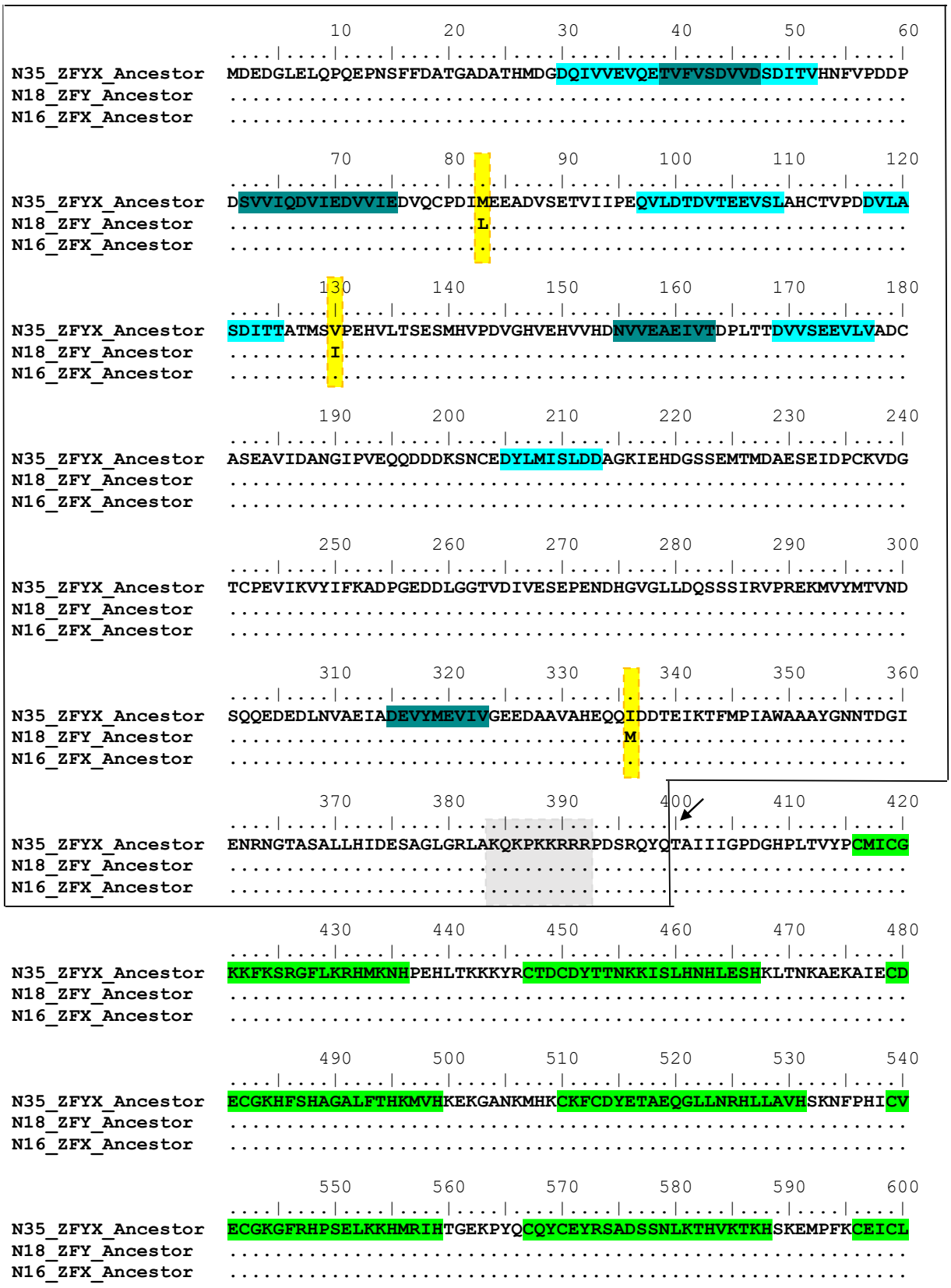
In an attempt to determine the ancestral sequences of significant nodes in the *ZFY*/*ZFX* evolution, ancestral reconstruction was performed using the GRASP tool. The key nodes of interest for sequence reconstruction included: the last common *ZFY* ancestor before divergence, the final ancestral *ZFX*, the last shared rodent *ZFY* progenitor, and the earliest primordial *ZFY*/*ZFX* predecessor preceding separation.

GRASP performs the reconstruction analysis by constructing phylogenetic trees with labelled nodes corresponding to an ancestral sequence it predicted. The nodes and the predictive sequences are labelled with the % confidence determined by GRASP.

2.3.7.1 Tracing Back to the Last Common Ancestor of *ZFY*/*ZFX* following the Marsupial/Eutherian Divergence

Ancestral reconstructed sequences at phylogenetically relevant nodes were selected from the GRASP output. For this construction analysis, the last *ZFY*/*ZFX* ancestor was selected and compared against the predicted final ancestral *ZFX* preceding

sexual differentiation, as well as the first *ZFY* sequence following divergence onto the Y chromosome. This comparison would hopefully provide some insight into the genetic changes occurring following the divergence of *ZFY* and *ZFX* on the Y- and X-chromosome respectively.


10 20 30 40 50 60
N35_ZFYX_Ancestor MDEDGLELQPQEPNSFFDATGADATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDDP
N18_ZFY_Ancestor
N16_ZFX_Ancestor
70 80 90 100 110 120
N35_ZFYX_Ancestor DSVVIQDVIEDVVIEDVQCPDIMEEADVSETVIIPEQVLDTDVTEEVSLAHCTVPDDVLA
N18_ZFY_Ancestor L
N16_ZFX_Ancestor
130 140 150 160 170 180
N35_ZFYX_Ancestor SDITTATMSVPEHVLTSESMHVPDVGHVEHVVHDNVVEAEIVTDPLTTDVVSEEVLVADC
N18_ZFY_Ancestor I
N16_ZFX_Ancestor
190 200 210 220 230 240
N35_ZFYX_Ancestor ASEAVIDANGIPVEQQDDDKSNCEDYLMISLDDAGKIEHDGSSEMTMDAESEIDPCKVDG
N18_ZFY_Ancestor
N16_ZFX_Ancestor
250 260 270 280 290 300
N35_ZFYX_Ancestor TCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDHGVGLLDQSSSIRVPREKMVYMTVND
N18_ZFY_Ancestor
N16_ZFX_Ancestor
310 320 330 340 350 360
N35_ZFYX_Ancestor SQQEDEDLNVAEIADEVYMEVIVGEEDAAVAHEQQIDDTEIKTFMPIAWAAAYGNNTDGI
N18_ZFY_Ancestor M
N16_ZFX_Ancestor
370 380 390 400 410 420
N35_ZFYX_Ancestor ENRNGTASALLHIDESAGLGRLAKQKPKKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICG
N18_ZFY_Ancestor
N16_ZFX_Ancestor
430 440 450 460 470 480
N35_ZFYX_Ancestor KKFKSRGFLKRHMKNHPEHLTKKKYRCTDCDYTTNKKISLHNHLESHKLTNKAEKAIECD
N18_ZFY_Ancestor
N16_ZFX_Ancestor
490 500 510 520 530 540
N35_ZFYX_Ancestor ECGKHFSHAGALFTHKMVHKEKGANKMHKCKFCDYETAEQGLLNRHLLAVHSKNFPHICV
N18_ZFY_Ancestor
N16_ZFX_Ancestor
550 560 570 580 590 600
N35_ZFYX_Ancestor ECGKGFRHPSELKKHMRIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCEICL
N18_ZFY_Ancestor
N16_ZFX_Ancestor

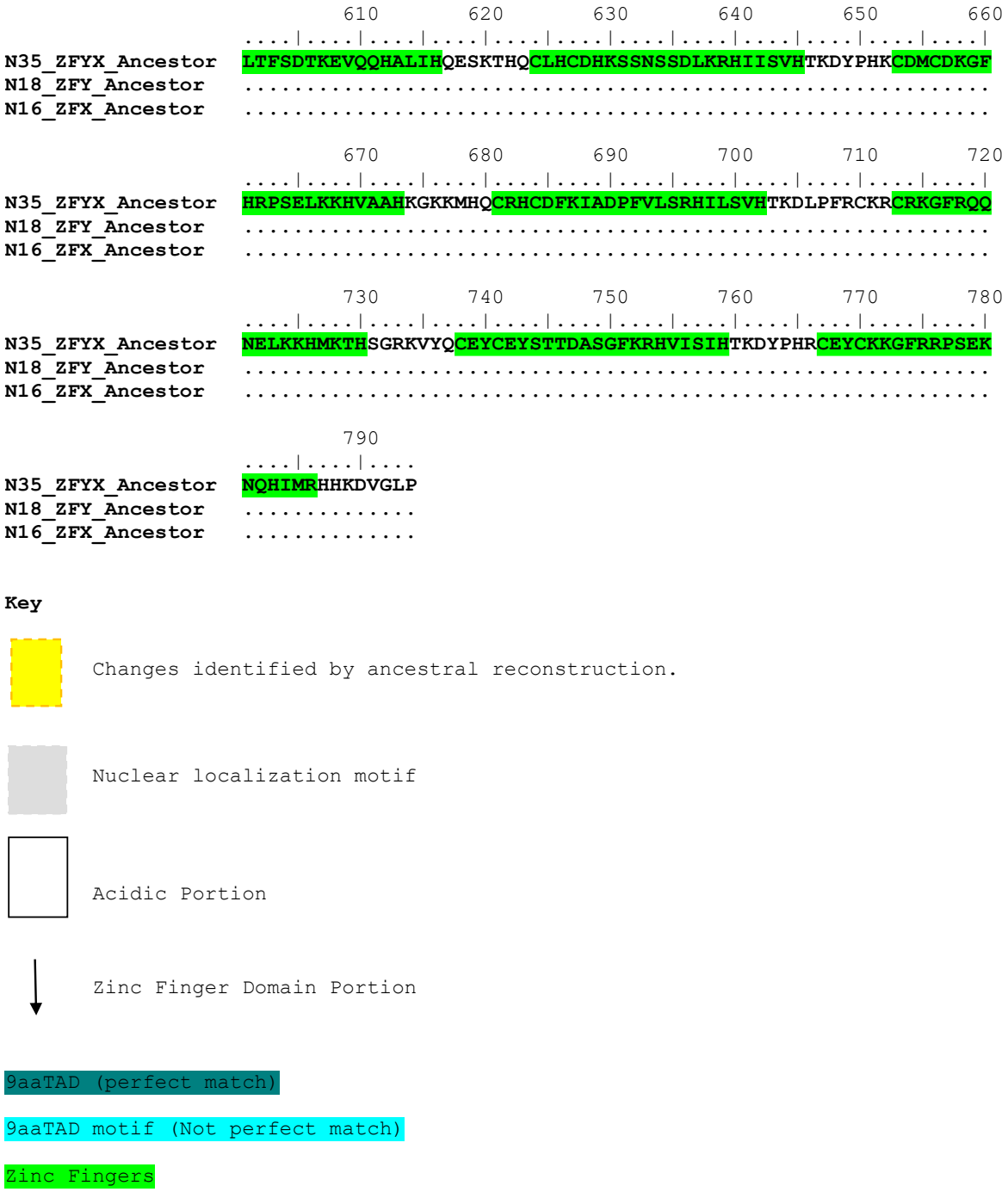


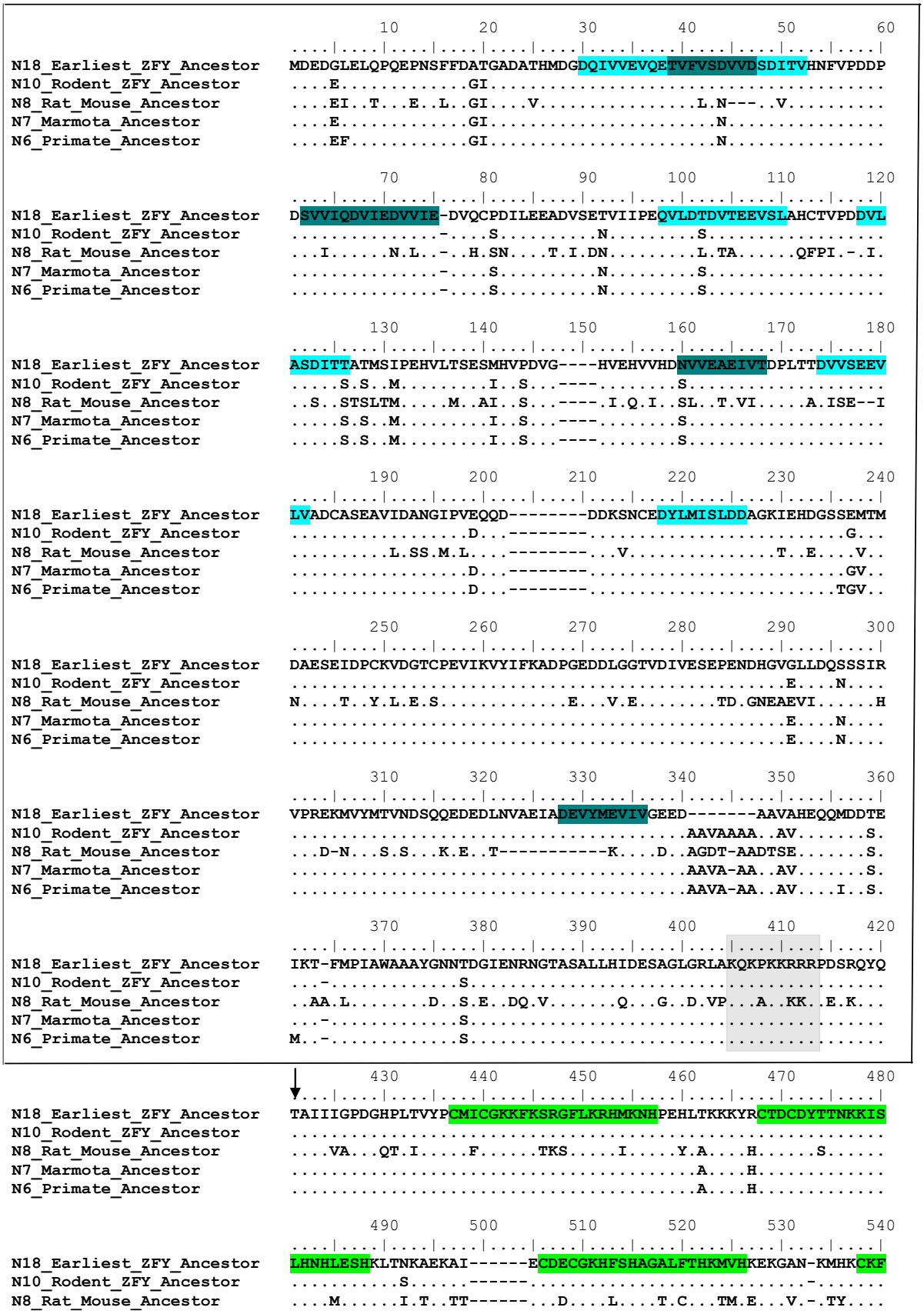
Figure 2.12: Ancestral reconstruction of the first *ZFY*/*ZFX* ancestor following the marsupial/eutherian divergence. N35_*ZFYX*_Ancestor is the predicted progenitor of all eutherian *ZFY* and *ZFX* homologs following the marsupial/eutherian split when looking at a global alignment of *ZFY* and *ZFX* sequences. N18_*ZFY*_Ancestor and N16_*ZFX*_Ancestor describe the theoretical changes resulting in the initial sex chromosome derivation—the last universal *ZFY* and *ZFX* variants prior to Y or X linkage respectively. The phylogenetic tree and node selection can be seen in *Supplementary* **Figure 2.**

The sequence comparison in **Figure 2.12** compares the *ZFYX* ancestor against its N18 and N16 subsequent *ZFY* and *ZFX* divergent nodes. **Figure 2.12** only shows 3

protein substitutions (M>L, V>I, I>M) between the *ZFY*X ancestor and the *ZFY*-specific ancestor. In contrast, no substitutions are seen between the *ZFY*X ancestor and the *ZFX*-specific ancestor. All 3 *ZFY*-specific changes are localised within the more variable AAD of *ZFY* rather than the more conserved DBD region, which persists unchanged. The substitutions although in the AAD do not fall in the predicted more conserved 9aaTAD regions of the domain. Although there are minimal changes, changes in the protein sequence can have major changes in protein structure and function. However, these changes were tolerated when *ZFY* moved to the Y chromosome.

The negligible divergence seen between the reconstructed sequences prompts further examination as more variation between these homologs would be expected during their divergence on non-recombining sex chromosome. Although evidence has suggested *ZFY* is under negative selection further mutations would be expected that would lead to the role divergence of *ZFY* and *ZFX*. Hopefully, by comparing further nodes in the reconstruction will lead to the identification of impactful changes.

Further inspection of *ZFY* sequences at vital nodes of divergence aimed to observe the rapid evolution of *ZFY* within the rodent lineage. Specifically, the reconstructed sequences at these divergent points; the rodent ancestor, the rat/mouse ancestor, the marmot ancestor and the primate ancestor were examined. These nodes were selected for a stepwise analysis of the rodent divergence, as marmot seems to diverge away from the other rodents in the analysis and groups more towards the primate species.


10 20 30 40 50 60
N18_Earliest_ZFY_Ancestor MDEDGLELQPQEPNSFFDATGADATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDDP
N10_Rodent_ZFY_Ancestor ....E.....GI.....
N8_Rat_Mouse_Ancestor ....EI..T...E..L..GI....V.....L.N---..V.....
N7_Marmota_Ancestor ....E.....GI.....N.....
N6_Primate_Ancestor ....EF.....GI.....N.....
70 80 90 100 110 120
N18_Earliest_ZFY_Ancestor DSVVIQDVIEDVVIE-DVQCPDILEEADVSETVIIPEQVLDTDVTEEVSLAHCTVPDDVL
N10_Rodent_ZFY_Ancestor .....-....S.....N.....S.....
N8_Rat_Mouse_Ancestor ...I.....N.L..-..H.SN....T.I.DN.....L.TA.....QFPI.-.I.
N7_Marmota_Ancestor .....-....S.....N.....S.....
N6_Primate_Ancestor .....-....S.....N.....S.....
130 140 150 160 170 180
N18_Earliest_ZFY_Ancestor ASDITTATMSIPEHVLTSESMHVPDVG----HVEHVVHDNVVEAEIVTDPLTTDVVSEEV
N10_Rodent_ZFY_Ancestor .....S.S..M.....I..S...----.....S.....
N8_Rat_Mouse_Ancestor ..S..STSLTM.....M..AI..S...-----.I.Q.I..SL..T.VI.....A.ISE--I
N7_Marmota_Ancestor .....S.S..M.....I..S...----.....S.....
N6_Primate_Ancestor .....S.S..M.....I..S...----.....S.....
190 200 210 220 230 240
N18_Earliest_ZFY_Ancestor LVADCASEAVIDANGIPVEQQD--------DDKSNCEDYLMISLDDAGKIEHDGSSEMTM
N10_Rodent_ZFY_Ancestor .....D...--------.....G...
N8_Rat_Mouse_Ancestor .....L.SS.M.L....--------...V.....T..E....V..
N7_Marmota_Ancestor .....D...--------.....GV..
N6_Primate_Ancestor .....D...--------.....TGV..
250 260 270 280 290 300
N18_Earliest_ZFY_Ancestor DAESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDHGVGLLDQSSSIR
N10_Rodent_ZFY_Ancestor .....E....N....
N8_Rat_Mouse_Ancestor N....T..Y.L.E.S.....E...V.E.....TD.GNEAEVI.....H
N7_Marmota_Ancestor .....E....N....
N6_Primate_Ancestor .....E....N....
310 320 330 340 350 360
N18_Earliest_ZFY_Ancestor VPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEED-------AAVAHEQQMDDTE
N10_Rodent_ZFY_Ancestor .....AAVAAAA..AV.....S.
N8_Rat_Mouse_Ancestor ...D-N...S.S...K.E..T----------K....D..AGDT-AADTSE.....S.
N7_Marmota_Ancestor .....AAVA-AA..AV.....S.
N6_Primate_Ancestor .....AAVA-AA..AV....I..S.
370 380 390 400 410 420
N18_Earliest_ZFY_Ancestor IKT-FMPIAWAAAYGNNTDGIENRNGTASALLHIDESAGLGRLAKQKPKKRRRPDSRQYQ
N10_Rodent_ZFY_Ancestor ...-.....S.....
N8_Rat_Mouse_Ancestor ..AA.L.....D..S.E..DQ.V.....Q...G..D.VP...A..KK..E.K...
N7_Marmota_Ancestor ...-.....S.....
N6_Primate_Ancestor M..-.....S.....
430 440 450 460 470 480
N18_Earliest_ZFY_Ancestor TAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHLTKKKYRCTDCDYTTNKKIS
N10_Rodent_ZFY_Ancestor .....
N8_Rat_Mouse_Ancestor ....VA...QT.I.....F.....TKS.....I.....Y.A....H.....S.....
N7_Marmota_Ancestor .....A....H.....
N6_Primate_Ancestor .....A....H.....
490 500 510 520 530 540
N18_Earliest_ZFY_Ancestor LHNHLESHKLTNKAEKAI------ECDECGKHFSHAGALFTHKMVHKEKGAN-KMHKCKF
N10_Rodent_ZFY_Ancestor .....S.....------.....-.....
N8_Rat_Mouse_Ancestor ....M.....I.T..TT------...D....L....T.C...TM.E...V.-.TY....

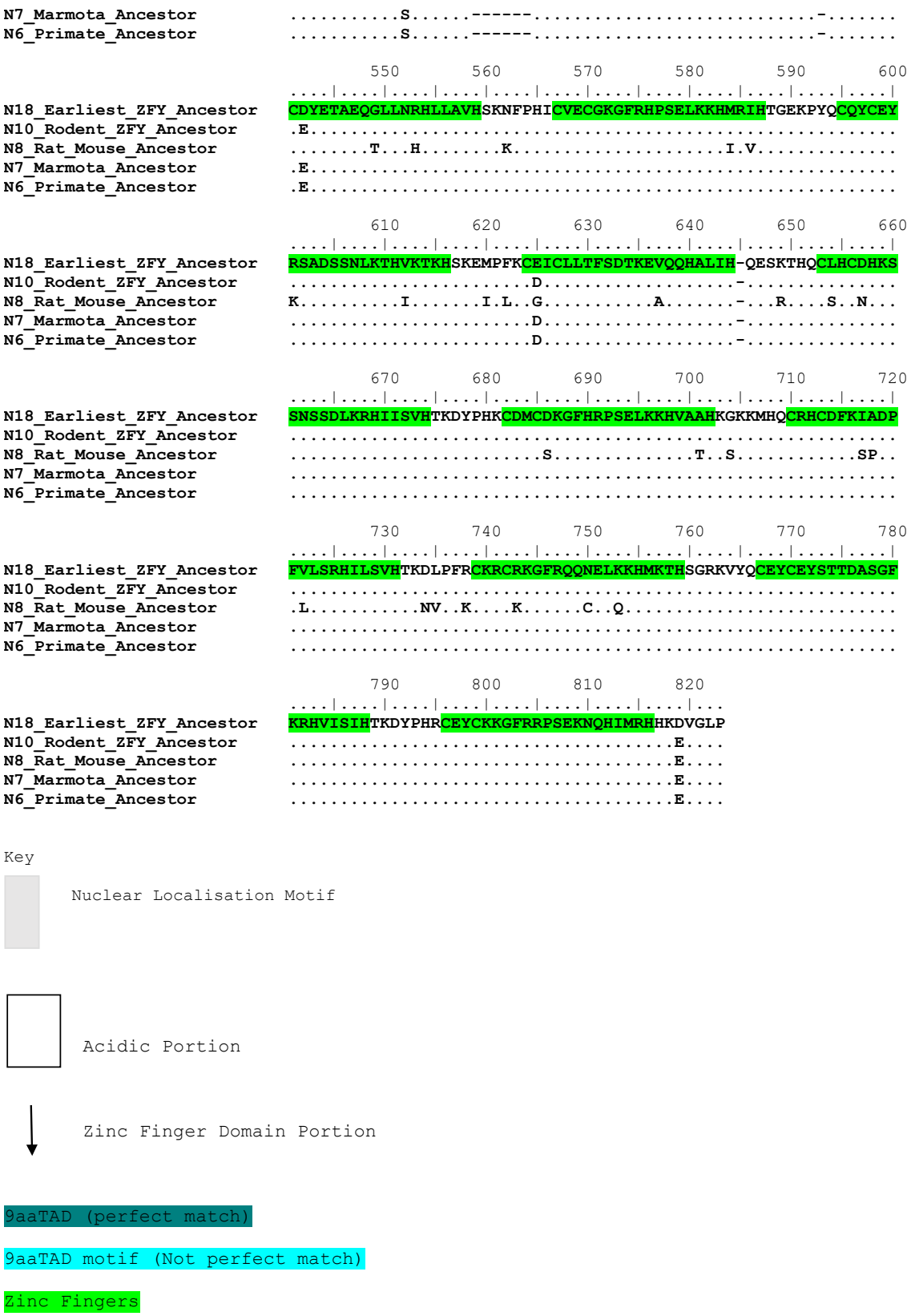
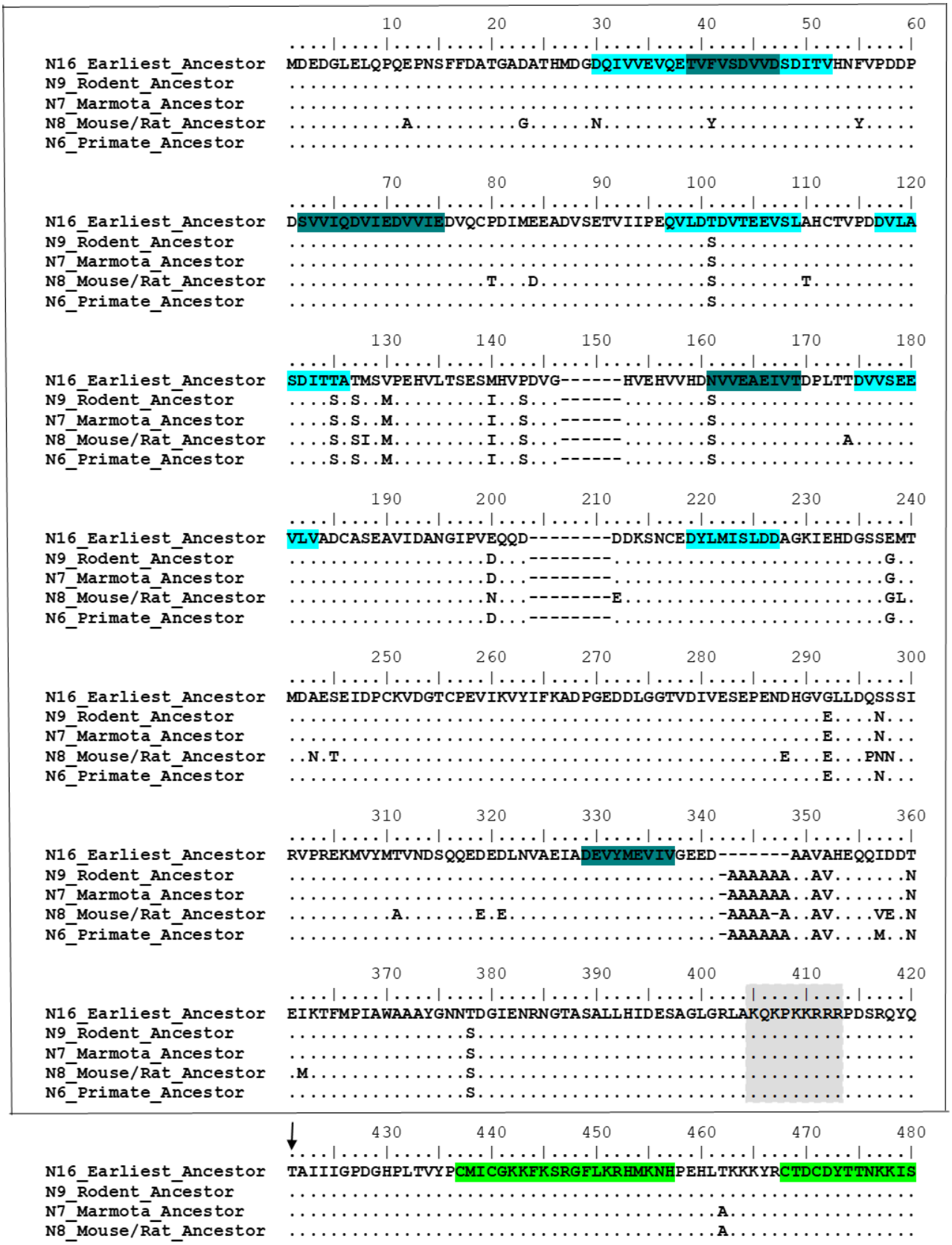


Figure 2.13: Ancestral reconstruction of the *ZFY* ancestors following the move to the Y Chromosome. N10_Rodent_*ZFY*_Ancestor: The last shared predecessor of mouse, rat and marmot *ZFY*. N8_Rat_Mouse_Ancestor: The last common progenitor prior to the divergence of rat and mouse *Zfy*. N7_Marmota_Ancestor: The final marmot *ZFY* variant before becoming an independent lineage. N6_Primate_Ancestor: The most recent evolutionary ancestor linking all analysed

primate *ZFY* sequences. The phylogenetic tree and node selection can be seen in *supplementary* **Figure 3.**

The rapid divergence of the rat and mouse *Zfy* sequences is illustrated in **Figure 2.13**. The first rodent ancestor sequence exhibits minor changes in the protein sequence, with a more noticeable number of substitutions identified in the AAD. However, the divergence of the mouse and rat lineage is evident, as numerous more substitutions are visible within both the AAD and DBD, but more prominently in the AAD. Substitutions are visible in the 9aaTAD and nuclear localisation motif of the AAD which are typically regions of greater conservation. The rat and mouse ancestor also have changes within the zinc fingers, although the zinc finger structures continue to persist due to the consistent presence of C and H residues. These changes suggest potential structural and functional changes of rat/mouse *ZFY* that make it distinct from other species. The marmot ancestor shares similarities with the primate ancestor, providing evidence for the marmot's divergence from the rodent lineage.

This rapid rodent divergence is less evident in the *ZFX* ancestral reconstruction analysis, with fewer substitutions seen between the earliest *ZFX* ancestor and the rat/mouse ancestor.


10 20 30 40 50 60
N16_Earliest_Ancestor MDEDGLELQPQEPNSFFDATGADATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDDP
N9_Rodent_Ancestor
N7_Marmota_Ancestor
N8_Mouse/Rat_Ancestor
N6_Primate_Ancestor
70 80 90 100 110 120
N16_Earliest_Ancestor DSVVIQDVIEDVVIEDVQCPDIMEEADVSETVIIPEQVLDTDVTEEVSLAHCTVPDDVLA
N9_Rodent_Ancestor
N7_Marmota_Ancestor
N8_Mouse/Rat_Ancestor
N6_Primate_Ancestor
130 140 150 160 170 180
N16_Earliest_Ancestor SDITTATMSVPEHVLTSESMHVPDVG------HVEHVVHDNVVEAEIVTDPLTTDVVSEE
N9_Rodent_Ancestor
N7_Marmota_Ancestor
N8_Mouse/Rat_Ancestor
N6_Primate_Ancestor
190 200 210 220 230 240
N16_Earliest_Ancestor VLVADCASEAVIDANGIPVEQQD--------DDKSNCEDYLMISLDDAGKIEHDGSSEMT
N9_Rodent_Ancestor
N7_Marmota_Ancestor
N8_Mouse/Rat_Ancestor
N6_Primate_Ancestor
250 260 270 280 290 300
N16_Earliest_Ancestor MDAESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDHGVGLLDQSSSI
N9_Rodent_Ancestor
N7_Marmota_Ancestor
N8_Mouse/Rat_Ancestor
N6_Primate_Ancestor
310 320 330 340 350 360
N16_Earliest_Ancestor RVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEED-------AAVAHEQQIDDT
N9_Rodent_Ancestor
N7_Marmota_Ancestor
N8_Mouse/Rat_Ancestor
N6_Primate_Ancestor
370 380 390 400 410 420
N16_Earliest_Ancestor EIKTFMPIAWAAAYGNNTDGIENRNGTASALLHIDESAGLGRLAKQKPKKRRRPDSRQYQ
N9_Rodent_Ancestor
N7_Marmota_Ancestor
N8_Mouse/Rat_Ancestor
N6_Primate_Ancestor
430 440 450 460 470 480
N16_Earliest_Ancestor TAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHLTKKKYRCTDCDYTTNKKIS
N9_Rodent_Ancestor
N7_Marmota_Ancestor
N8_Mouse/Rat_Ancestor

```
N6_Primate_Ancestor    ........................................A.................

                                490       500       510       520       530       540
                       ....|....|....|....|....|....|....|....|....|....|....|....|
N16_Earliest_Ancestor  LHNHLESHKLTNKAEKAI------ECDECGKHFSHAGALFTHKMVHKEKGAN-KMHKCKF
N9_Rodent_Ancestor     ..........S......------..............................-.......
N7_Marmota_Ancestor    ..........S......------..............................-.......
N8_Mouse/Rat_Ancestor  ..........S......------..............................-.......
N6_Primate_Ancestor    ..........S......------..............................-.......

                                550       560       570       580       590       600
                       ....|....|....|....|....|....|....|....|....|....|....|....|
N16_Earliest_Ancestor  CDYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHMRIHTGEKPYQCQYCEY
N9_Rodent_Ancestor     .E..................................................
N7_Marmota_Ancestor    .E..................................................
N8_Mouse/Rat_Ancestor  .E.................................................E......
N6_Primate_Ancestor    .E..................................................

                                610       620       630       640       650       660
                       ....|....|....|....|....|....|....|....|....|....|....|....|
N16_Earliest_Ancestor  RSADSSNLKTHVKTKHSKEMPFKCEICLLTFSDTKEVQQHALIH-QESKTHQCLHCDHKS
N9_Rodent_Ancestor     ......................D...................-..............
N7_Marmota_Ancestor    ......................D...................-..............
N8_Mouse/Rat_Ancestor  ......................D...................-..............
N6_Primate_Ancestor    ......................D...................-..............

                                670       680       690       700       710       720
                       ....|....|....|....|....|....|....|....|....|....|....|....|
N16_Earliest_Ancestor  SNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKGKKMHQCRHCDFKIADP
N9_Rodent_Ancestor     ..................................................
N7_Marmota_Ancestor    ..................................................
N8_Mouse/Rat_Ancestor  ..................................................
N6_Primate_Ancestor    ..................................................

                                730       740       750       760       770       780
                       ....|....|....|....|....|....|....|....|....|....|....|....|
N16_Earliest_Ancestor  FVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKVYQCEYCEYSTTDASGF
N9_Rodent_Ancestor     ..................................................
N7_Marmota_Ancestor    ..................................................
N8_Mouse/Rat_Ancestor  ............................S.....................
N6_Primate_Ancestor    ..................................................

                                790       800       810       820
                       ....|....|....|....|....|....|....|....|...
N16_Earliest_Ancestor  KRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKDVGLP
N9_Rodent_Ancestor     ....................................E....
N7_Marmota_Ancestor    ....................................E....
N8_Mouse/Rat_Ancestor  ....................................E....
N6_Primate_Ancestor    ....................................E....
```

Key

Nuclear localization motif

Acidic Portion

Zinc Finger Domain Portion

9aaTAD (perfect match)

9aaTAD motif (Not perfect match)

Zinc Fingers

Figure 2.14: Ancestral reconstruction of the *ZFX* ancestors following the move to the X Chromosome. N9_Rodent_*ZFY*_Ancestor: The last shared predecessor of mouse, rat and marmot *ZFY*. N8_Rat_Mouse_Ancestor: The last common progenitor prior to the divergence of rat and mouse *ZFY*. N7_Marmota_Ancestor: The final marmot *ZFY* variant before becoming an independent lineage. N6_Primate_Ancestor: The most recent evolutionary ancestor linking all analysed primate *ZFY* sequences. The phylogenetic tree and node selection can be seen in *Supplementary* **Figure 4**.

Depicted in **Figure 2.14** is the more conserved evolution of *ZFX* across the taxons compared to *ZFY*. This emphasises that *ZFY* has acquired a unique function different to *ZFX* or that ZFY has lost *ZFX* functions Though the marmot has diverged away from the remaining rodent clade in this analysis, the node is only 36% supported, which indicates that there is weak evidence for this node, but the changes between the marmot and rat/mouse are still dramatic enough to result in the divergence of the order taxon. Similar to *ZFY*, the majority of the substitutions in *ZFX* are within the AAD, with the 9aaTAD and nuclear localisation motifs experiencing fewer alterations. The DBD remains highly conserved across all the ancestors with only 7 protein substitutions identified across the sequences, with only 2 being located within a zinc finger. This further highlights the maintained conservation of the DBD across the *ZFY*/*ZFX* ancestors.

2.4 Discussion

The Y chromosome has an unusual evolutionary history yet continues to persist in the genome harbouring the master-switch gene required for male determination and is abundant in spermatogenesis-specific genes (Bachtrog, 2013);(B. T. Lahn *et al*., 2001). Sex chromosomes have evolved independently from autosomes many times across many lineages, with the therian sex chromosomes evolving around ~180 million years ago following the split from monotremes. Following this split, the sex chromosomes of species like birds and snakes evolved independently, with the original autosomes responsible for the mammalian sex chromosomes still existing in birds today (Bachtrog, 2013);(J. F. Hughes & Page, 2015). Due to their unique function in sex determination, the sex chromosomes are subject to unique evolutionary forces, and they therefore play a role in speciation, adaptation and other evolutionary processes. Although the Y chromosome is known to be important, the chromosome is still targeted for degradation, resulting in most of its original genes being lost over evolutionary time. This is evident in some genetic studies which have shown that some species' Y chromosomes harbour almost no genes and some species' have completely lost their Y chromosome. Although the gene content of the human Y chromosome is consistently targeted by degradation, *ZFY* persists in the human genome emphasising the potentially vital role of this gene (B. T. Lahn *et al*., 2001);(Bachtrog, 2013);(J. F. Hughes & Page, 2015). Since *ZFY* was disproved as the sex-determining gene, its research interest diminished, and further investigation stopped. Therefore, this thesis aimed to investigate *ZFY*'s persistence on the Y chromosome and its importance in male development.

The evolutionary trajectory of *ZFY* over the past 200-300 million years suggests it has been subject to negative selection pressures, especially in regions crucial for its male developmental function, such as the AAD and DBD. Investigations into these region-specific areas confirmed the negative selection pressures on the AAD and DBD showing the importance of maintaining sequence integrity in these regions for DNA-binding targets essential to *ZFY*'s male-determining functions. Within the AAD and DBD, the 9aaTAD and zinc finger domains respectively, are crucial to function and are largely under negative selection to ensure the conservation of these key functioning motifs. Positive selection on these sites would consequently change the coding sequence and, therefore, potentially alter DNA-binding sites necessary for *ZFY*'s male-determining functions. While *ZFY* evolution mostly aligns with expected taxonomic patterns, Rodentia, particularly rat and mice, display accelerated evolution, suggesting potential adaptation to new roles within the rodent lineage. This is evident by the high substitution rates **(Table 2.4)** calculated in the rodent lineage. This

contrasts with *ZFX* evolution, which, while also evolving, shows less pronounced changes, indicating potentially different evolutionary pressures or roles between *ZFY* and *ZFX* in rodents. However, it is well established that rodents have one of the highest mutation rates among mammals, so longer branch lengths for rodent species are typically expected (Nabholz et al., 2008).

The high homology between *ZFX* and *ZFY* raised the possibility of gene conversion events. A gene conversion arises between paralogous genes resulting in the reshuffling and homogenisation of their DNA sequences resulting in the paralogous genes becoming highly similar (Mansai & Innan, 2010). This can cause downstream problems in the inference of both duplicated genes and multigene families' evolutionary history. Geneconv detects recent gene conversions across an alignment of sequences in a pairwise manner, with the significance determined by the random shuffling of variable sites within the alignment. However, it is difficult to robustly detect gene conversions via Geneconv due to the limited power when the gene conversion rate per site is large (Mansai & Innan, 2010). This is suggested to be the result of frequent homogenisation reducing the number of variable sites and the heterogeneity of the configurations at the variable sites (Mansai & Innan, 2010). This means that the detection of few gene conversions could mean two things; (1) gene conversions are not very active in this region or (2) gene conversions are extremely frequent. Constant gene conversion repeats can result in a long stretch of DNA with very few mismatches (Mansai & Innan, 2010). Consequently, gene conversions detected by Geneconv possibly do not reflect real converted regions however, when the data is combined with phylogenetic tree analysis there is very good evidence for potential gene conversions.

While the alignment-wide nucleotide phylogenetic tree didn't reveal many gene conversion events, the clustering of elephant *ZFY* and *ZFX* suggested a possible gene conversion during their evolutionary split. However, upon tree construction of only coding exon 7 which encodes the DBD multiple more possible gene conversion events were identified, whilst this was not evident in coding exons 1-6. Further investigation by Geneconv uncovered three potential conversions in rat, elephant, and stoat, supported by significant permutation p-values, though they fell short of BC-KA p-value thresholds. Although permutation p-values and tract analysis hinted at gene conversion, additional evidence is needed for validation. Notably, global outer-fragment analysis detected a potential gene conversion in cow *ZFY*, backed by strong permutation and BC-KA metrics, possibly originating from outside the Y chromosome. Both phylogenetic tree construction and gene conversion analysis software identified gene conversions across rats, elephants, and stoats. This further reinforces the

argument that the high homology between *ZFY* and *ZFX* could be attributed to gene conversions. It has further been suggested that these gene conversions are unique to the DBD (coding exon 7) potentially due to their high specificity for targets emphasising the continued need for these sites to be under negative selection. Despite these findings, the anticipated number of gene conversions was not as great as expected, indicating that while *ZFY* and *ZFX* share high homology, their divergence might account for their distinct functions. It is also important to consider that the high similarity between these genes could be due to ancestral duplication, meaning the present sequence similarities may be the result of identity by descent not gene conversions.

After aligning ancestral sequences, few changes appeared at the first breakpoint of *ZFX* and *ZFY*, unexpected given their relocation to separate chromosomes. However, further ancestral node analysis revealed more changes in *ZFY* sequences of subsequent lineages, particularly in key functional regions like zinc fingers in the DBD and 9aaTAD in the AAD. Mutations in these regions would alter DNA-binding sites and downstream targets. Tracing back *ZFY* showed far more changes than *ZFX*, potentially explaining the functional divergence between them. This also explains the accelerated Rodentia evolution, evidenced by the greater protein changes in mice *Zfy1* and *Zfy2*, suggesting completely different functionalities. While intriguing, this predictive analysis remains inconclusive without further investigation into the ancestral nodes and sequences.

In conclusion, *ZFY* persistence on the Y chromosome is attributed to negative selection pressure acting on the key functional regions of the gene. This is indicative of a unique functional role differing from *ZFX* on the X chromosome as even with its high homology, *ZFY* has maintained a male-specific function vital to sex determination. Changes throughout the lineage splits have resulted in the uniqueness and functional divergence of the two homologues.

3. Chapter 3: Unravelling *ZFY*'s Splicing Mechanism: Exploring the Role of *RBMY*

3.1 Introduction

3.1.1 Alternative Splicing Increases Genome Complexity

Human *ZFY* exists in two forms regulated by splicing; a canonical full-length isoform, *ZFYL*, and a testis-specific short isoform denoted *ZFYS*. Two types of splicing exist: (1) constitutive splicing is the removal of introns and the ligation of exons in the order in which they appear in the gene and (2) alternative splicing results in the deviation from the preferred sequence by the rearrangement or removal of exons (G. Wang *et al*., 2015);(Stamm *et al*., 2005). Alternative splicing was first described as a concept in 1978, accounting for the discrepancies between the number of protein-coding genes and proteins in humans. This concept disproved the original notion of "one gene - one RNA - one protein," with 95% of human genes identified as undergoing splicing at some point during development (G. Wang *et al*., 2015);(Matlin *et al*., 2005). Alternative splicing plays a major role in cellular differentiation and organism development due to increasing the complexity of gene expression (G. Wang *et al*., 2015). A higher complexity and quantity of alternative splicing has been discovered in higher eukaryotic organisms, with species-specific splicing conservation resulting in species differentiation and genome evolution (G. Wang *et al*., 2015).

Through the emergence of microarray data and expressed sequence-tagged data, analysis has revealed five main types of alternative splicing; mutually exclusive exons, cassette alternative exons, alternative 3' splice site, alternative 5' splice sites and intron retention **(Figure 3.1)** (G. Wang *et al*., 2015);(Jiang & Chen, 2021). In vertebrates and invertebrates, 30% of the alternative splicing patterns were identified as exon-skipping, whilst in lower metazoans the main type is intron retention (G. Wang *et al*., 2015);(E. Kim *et al*., 2006).

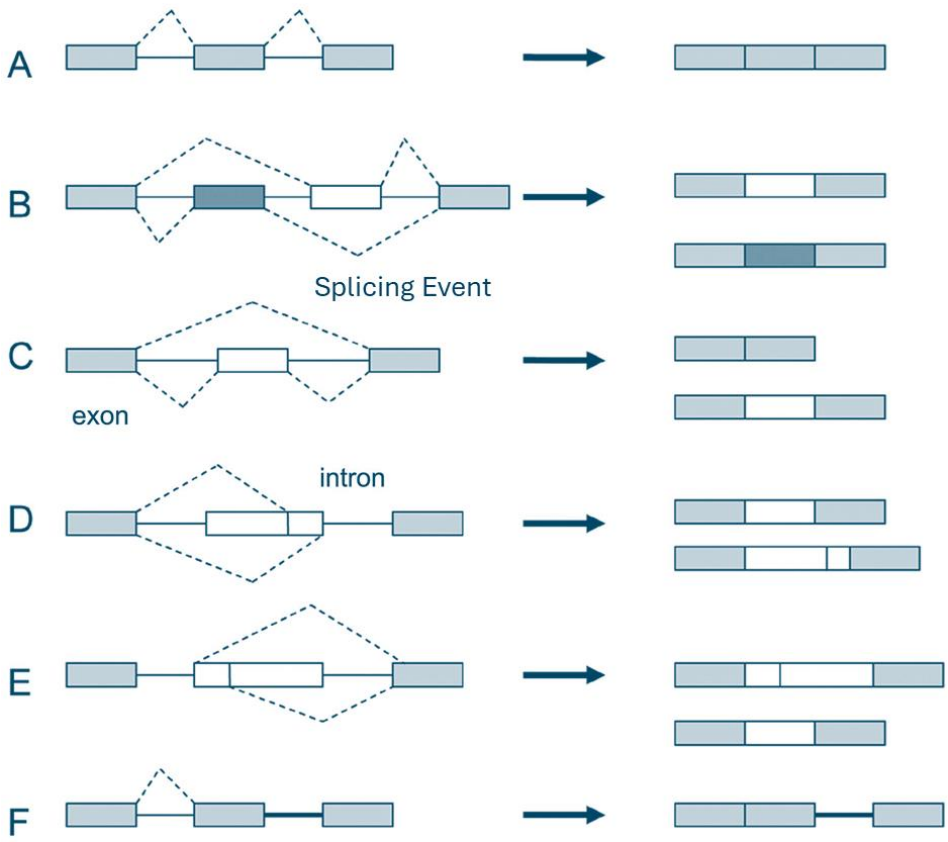


Figure 3.1: The five main types of alternative splicing. A: constitutive splicing, **B:** mutually exclusive exons, **C:** cassette alternative exons, **D:** alternative 3' splice site, **E:** alternative 5' splice site, and **F:** intron retention. Taken directly from (G. Wang *et al*., 2015).

Splicing consists of two major processes: spliceosome assembly and the actual splicing of pre-mRNAs (G. Wang *et al*., 2015). The spliceosome, a large ribonucleoprotein consists of small nuclear ribonucleic proteins (snRNPs) forming a core unit upon splicing signals at the 5' and 3' splice sites. In a stepwise manner, conserved DExD/H-type RNA-dependent ATPases/helicases assemble on the core and execute splicing steps resulting in a variety of exon ligations and intron excisions. Alternative splicing is controlled by a wide range of interactions between cis and trans components (G. Wang *et al*., 2015);(Matlin *et al*., 2005). Alternative splicing is inhibited by negative factors including heterogenous nuclear ribonucleoproteins (hnRNPs) binding to exonic splicing silencers and intronic splicing silencers. However, alternative splicing is activated by positive trans-acting factors binding to exonic splicing enhancers and intronic splicing enhancers (G. Wang *et al*., 2015). This alternative splicing system is a key part of generating complex proteomes supporting embryonic stem and precursor cells during the differentiation of cell lineages, epithelial-mesenchymal transitions, adult organ development and the immune system (Ule & Blencowe, 2019).

3.1.2 Alternative Splicing of *ZFY* Forms a Testis-Specific Short Variant

As discussed in sections 1.7 to 1.7.6.3 of the introduction, the mammalian *ZFY* gene is located on the Y chromosome and up until 2012 only a single ubiquitous *ZFY* protein had been identified. However, in 2012 Decarpentrie *et al* identified a secondary testis-specific *ZFY* transcript encoding a truncated *ZFY* protein with a shorter acidic domain (Decarpentrie *et al*., 2012). Further analysis indicated that this transcript encoded a protein lacking the entire second coding exon because of alternative splicing (cassette splicing event as per **Figure 3.1c**), allowing for one gene to express two splice variants. Moreover, these splice variants are both differentially expressed and functionally distinct, with the short isoform being testis-specific and having greatly reduced transactivation ability compared to the long form (see introduction sections 1.7 to 1.7.3.2 for further detail) (Decarpentrie *et al*., 2012).

Importantly, mammalian *ZFX* does not appear to produce a short splice form equivalent to *ZFYS*. The occurrence of a Y-specific isoform exclusively expressed in the testis poses an evolutionary mystery: which came first? If the ancestral autosomal Zf* gene also expressed a testis-specific splice isoform, then the current Y-specific *ZFYS* represents an ancestral testis-specific expression pattern that has been retained following X/Y divergence and indeed the requirement for a testis-specific splice form could have been a factor driving recruitment of *ZFY* into the non-recombining portion of the eutherian sex chromosomes. Alternatively, if ancestral autosomal Zf* did not produce the short splice form, then the current Y-specific *ZFYS* represents neofunctionalization – a novel expression pattern that has evolved after X/Y divergence. While we cannot directly observe the ancestral expression pattern, we can infer the splice regulation of ancestral autosomal Zf* by looking at the splicing of autosomal *ZFX/Y* relatives in marsupials and birds, where it retains its ancestral autosomal genomic location.

A further unresolved mystery is how *ZFY* splicing is regulated – which gene or genes act to trigger the skipping of the cassette exon during early germ cell development? In this study, based on a range of circumstantial observations set out in sections 1.7.3 to 1.7.3.2 of the introduction, we hypothesised that *RBMY* acts to hinder the incorporation of the *ZFY* core acidic domain exon, leading to the formation of the observed short isoform. Specifically:

• Based on expression data, *ZFYS* is only expressed in cell types that also express *RBMY*.

• The phenotype for human AZFb deficiency (i.e. RMBY deficiency) resembles the mouse phenotype for overexpression of *ZFYL* – this would be predicted if *RBMY* triggers conversion *ZFYL* to *ZFYS*.

• A close relative of *RBMY*, RBMXL2, acts to suppress the use of specific splice acceptor sites, suggesting that the function of *RBMY* is also likely to involve suppressing the inclusion of specific exons in its downstream targets.

This chapter sets out to address these two questions; is the testis-specific form seen in ancestral autosomal species and is *RBMY* causing the skipping of *ZFY* exon 2. To address these core questions this chapter is split into two parts, firstly, published cross-species data will be used to determine whether marsupials and birds also produce a testis-specific short *ZFY* splice form. Secondly, replicate data from the Fenton/Ellis lab demonstrating the expression of both *ZFYS* and *RBMY* in two head and neck squamous cell carcinomas. Finally, use a mammalian splicing reporter system to quantify the effect of *RBMY* expression on exon skipping for the *ZFY* cassette exon that is omitted in *ZFYS*.

3.2 Materials and Methods

3.2.1 RNA-Seq Data Analysis Looking into Splicing Variations

Cardoso-Moreria and team investigated the evolutionary patterns of ZF gene expression in mammalian organs, and they have publicly accessible RNA-Seq data for numerous mammals and organs (Cardoso-Moreira *et al*., 2019). Both the unprocessed and processed RNA-Seq data can be accessed on ArrayExpress. The mammals examined for splicing variation were E-MTAB-6769 (chicken), E-MTAB-6798 (mouse), E-MTAB-6814 (human), and E-MTAB-6833 (opossum). RNA-Seq data was gathered from the testis, brain, cerebellum, liver, kidney, heart, and ovary. Although data from all these organs were collected, the ovaries were excluded from the analysis due to the focus on male-specific aspects. Many time points during development were analysed, but between each species, these time points varied, and the number of replicates also varied, with the most replicates being two. Time points were selected to roughly show “birth”, “mid-meiosis” and “mature adult” as well as possible **(Table 3.1, 3.2, 3.3 & 3.4)**.

Table 3.1: Available RNA-Seq data for chicken at birth, mid-meiosis and adulthood. Six organs have available data each with 2 replicates available. 0dph (days post-hatch), 10wph (10 weeks post-hatch) and adult is when premeiotic, mid-meiotic and mature sperm are the most advanced stages present in the testis.

Organ	0 dph	10 wph	Adult
Testis	2	2	2
Brain	2	2	2
Heart	2	2	2
Kidney	2	2	2
Liver	2	2	2
Cerebellum	2	2	2

Table 3.2: Available RNA-Seq data for mice at birth, mid-meiosis and adulthood. Six organs have available data each with 2 replicates available. 0 days (“birth”), 2 weeks (“mid-meiosis” – stra8 highest) and 9 weeks (“mature adult”).

Organ	0 days	2 weeks	9 weeks
Testis	2	2	2
Brain	2	2	2
Heart	2	2	2
Kidney	2	2	2
Liver	2	2	2
Cerebellum	2	2	2

Table 3.3: Available RNA-Seq data for humans at birth, mid-meiosis and adulthood. Six organs have available data each with varying data availability.

Organ	4-8 months	13-17 years	25-29 years	46 years	54-60 years
Testis	2	2	2	1	1

Brain	2	1	1	0	2
Heart	2	1	1	0	1
Kidney	2	0	0	0	0
Liver	2	1	1	0	2
Cerebellum	2	2	2	0	2

Table 3.4: Available RNA-Seq data for opossum at birth, mid-meiosis and adulthood. Six organs have available data each with varying data availability. 28 days ("birth"), 60 days ("mid-meiosis" – stra8 highest) and 180 days ("mature adult").

Organ	28 days	60 days	180 days
Testis	2	2	2
Brain	2	1	1
Heart	0	1	2
Kidney	1	2	1
Liver	1	2	2
Cerebellum	1	2	0

BAM files were deposited, and these files were downloaded. It was noted that they originally annotated these files using the Ensembl 69 annotation and had only allowed for uniquely mapped reads. Using the wget function the desired bam files were downloaded onto the HPC system. Samtools was used to sort, index and then flank each BAM file with the corresponding coordinates of the ZF gene of interest ~+/-5000bp. These coordinates were identified using the Ensembl 69 GTF files available for each species **(Table 3.5)**. However, for chicken, this was not possible as the Ensembl 69 annotation did not have ZF annotated. This meant that the chicken bam files had to be remapped to the latest Ensembl annotation (109) **(Table 3.5)**, sorted, indexed and then flanked with the corresponding coordinates.

Table 3.5: Genes analysed for each species. The Ensembl annotation used was version 69 except for chicken, where an updated annotation was required to include Zf genes missing from the original version.

Species	Genes Analysed	Ensembl Annotation
Human	ZFY & ZFX	69
Mouse	*Zfy*1, *Zfy*2, *Zfx* & *Zfa*	69
Opossum	Zf	69
Chicken	Zf	109

Using the flanked bam files and the indexed files, sashimi plots could be produced to identify gene expression and splicing events in each organism and each organ. Plots were produced using the pysashimi package (v1.5.0). The -M Min_Coverage function was applied to establish thresholds on the number of reads supporting a junction, with limits set to include junctions supported by a minimum of 10 reads. The complete code can be found in the folder Chapter 3 – Splicing Analysis at https://github.com/IzzyGarcia/Thesis-code.

3.2.2 Cancer Cell Line Maintenance

Two head and neck squamous cell carcinoma cell lines were used to assess endogenous *ZFYS* expression and the influence of *RBMY* on splicing. PCI-30 is a HPV-negative oral tongue squamous carcinoma derived from a 54-year-old male. UM-SCC-104 is a HPV-positive oral cavity squamous carcinoma from a 56-year-old male that uniquely contains HPV-16 and expresses E6/E7 oncoproteins.

Both cell lines were cultured at 37°C under humidified conditions, with 5% CO_2 in Gibco Dulbecco's Modified Eagle Medium (DMEM, Gibco, 11574486) supplemented with 10% Fetal Bovine Serum (FBS, Gibco, 11570516) and 1% L-Glutamine–Penicillin–Streptomycin (Sigma-Aldrich, G6784). For passaging, cells were washed twice with 1x phosphate-buffered saline (PBS) (Oxoid, BR0014G) and incubated with trypsin-EDTA (Gibco, 0.25%, 11560626) at 37°C for 1 minute or until completely detached. The trypsinised cells were collected, centrifuged at 1,200 x g for 4 minutes, and resuspended in fresh media. A logarithmic growth phase of the cells was maintained by passaging the cells.

Trypan blue was used to assess the viability of cells. Trypan blue stain (Logos biosystems, #T13001, 0.4%) was used and does not enter viable cells with an intact cell membrane. The cell mixture was mixed in a 1:1 ratio with the dye (i.e., 10uL to 10uL) and was mixed gently. A haemocytometer was used to count the cells, and a viability percentage and cell count could be calculated.

3.2.3 Reverse Transcription Polymerase Chain Reaction

Reverse transcription polymerase chain reaction (RT-PCR) was performed to assess *ZFY*/*RBMY* expression in the Head and Neck Cancers.

Following the collection of cell pellets, RNA extraction was performed using the Qiagen Rneasy mini kit (QIAGEN, 74104), using the protocol provided. Following RNA extraction, the RNA concentration was quantified using a NanoDrop ND-1000 Spectrophotometer to determine the concentration and purity. Subsequently, the RNA was converted back into complementary DNA (cDNA) using the LunaScript RT Supermix Kit (New England BioLabs, #E3010) which has been optimised for first-strand cDNA synthesis. The protocol followed for cDNA synthesis can be found in **Table 3.6** and **Table 3.7**. Alongside this, a no-template control and no-RT buffer control were set up.

The thermocycler used was the Mastercycler X50a – PCR Thermocycler (Eppendorf, cat no. 6313000042).

Table 3.6: Lunascript RT supermix reaction mixture set up for RNA samples. 20µL total volume reactions were set up for each RNA sample, using 1µg of RNA per sample.

Components	Volume	Final concentration
5x RT supermix	4µl	1x
RNA sample	Variable	1µg
Water	Makeup to 20µl	-

Table 3.7: Thermocycler Conditions for cDNA synthesis. These conditions are the standard stated in the Lunascript protocol. It is expected that 1000ng of RNA is converted to 1000ng of cDNA using this setup.

Cycle step	Temperature (°C)	Time (min)	Cycles
Primer annealing	25	02:00	1
cDNA synthesis	55	10:00	
Heat inactivation	95	01:00	

After cDNA was synthesised from the cancer cell RNA, the 50 ng/µL cDNA reactions were diluted to produce a 5 ng/µL working solution for use in subsequent experiments. Successful cDNA synthesis was validated by PCR amplification of the TBP housekeeping gene **(Table 3.10)** using GoTaq G2 Flexi DNA polymerase (Promega, M7801) in 10µL reactions. The component concentrations used are listed in **Table 3.8**.

Table 3.8: GoTaq® G2 Flexi DNA Polymerase reaction setup. The reaction components and concentrations are stated, and these are standardised across all chapters. The concentrations were optimised based on the kit protocol.

Component	Final concentration
5x GoTaq green Flexi buffer	1x
$MgCl_2$	3mM
dNTP	0.2mM each dNTP (dATP, dCTP,dTTP, dTTP)
GoTaq® G2 Flexi DNA polymerase	1.25u
Forward primer	1uM
Reverse primer	1uM
Water	Makeup to the final volume required

A standardised thermocycler setup shown in **Table 3.9** was used for the GoTaq G2 Flexi polymerase with many of the conditions remaining the same across experiments. The annealing temperature was often adjusted based on the primer pair melting temperature (Tm). Changes to the extension time were also made according to the expected product length, with longer extensions for larger products being required. Finally, the number of PCR cycles was optimised between 25-35 depending on the amplification yield required.

Table 3.9: Standardised GoTAQ G2 Flexi Polymerase Thermocycler conditions. This table states the standard conditions used on the thermocycler, with alterations made where necessary based on the primer pairs being used.

Phase	Temperature (°C)	Time (min)	Cycle Number
Initial denaturation	94	02:00	1
Denaturation	94	00:15	25-35
Annealing	Variable	00:15	
Extension	72	00:30 - 01:00	
Final extension	72	05:00	1
Hold	4	Infinite	1

After PCR, the amplified products were analysed by agarose gel electrophoresis. Agarose gels (2% w/v) were prepared by measuring agarose powder (Melford Biolaboratories, MB1200) and dissolving it in 1X Tris-borate-EDTA (TBE) buffer solution. The 1X TBE was prepared by diluting a 10X stock solution (Fisher Bioreagents, 10x TBE, BP1333-4) with Milli-Q water. This was then microwaved to fully dissolve the agarose. After complete dissolution, SYBR safe (Invitrogen, S33102) was introduced at a 1x concentration and gently mixed. The agarose mixture was then poured into a tank tray and allowed to solidify. Once firm, the comb was gently removed, and 1x TBE was poured to cover the entire gel. A 100bp DNA ladder (Fisher Scientific, Invitrogen #15628019) was loaded on the agarose gels alongside the PCR products. After loading samples, the gels were run at 90V for 45 minutes.

To identify the expression of *ZFY* and *RBMY* in these cancer cells, specific primers were designed. The *ZFY* primers designed in **Table 3.10** will produce two band sizes depending on whether *ZFYL* or *ZFYS* is expressed. The forward primer is located in the first coding exon of *ZFY*, whilst the reverse primer is located in the third coding exon. These primers therefore surround the second coding exon which is spliced out. *RBMY* primers were also designed to identify if it is also mis-expressed in cancers alongside *ZFY*.

Table 3.10: Primer sequences designed for determining *ZFY* expression in Head and Neck cancer. The Human *ZFY* primers were taken directly from Decarpentrie *et al.* (2012). The *RBMY* primers were designed by a previous master's student (Trujillo, 2019). TBP was used as a loading control to ensure even loading of cDNA.

Primer Name	Accession Number	Primer Sequence	Tm (°C)	Product Length (bp)
Human *ZFY*	NM_003411.4	For: GAATTGCAGCCACAAGAGCC Rev: CACCTTGATGACTTCAGGAC	57.2 52.9	*ZFYL*: 729 *ZFYS*: 156
Human *RBMY1A1*	NM_005058.4	For: GAAACCAATGAGAAGATGCTT	51.0 56.3	473

		Rev: TTGCTTCTTGCCACAGCAG		
TBP	NM_003194.5	For: CCCATGACTCCCATGACC Rev: TTTACAACCAAGATTCACTGTGG	55.2 53.4	108

3.2.4 GFP Splicing Construct

A luciferase reporter system from (Younis *et al*., 2010) was adapted to a GFPC1 vector by Dr. Florian Heyd at the Free University of Berlin as stated in (Neumann *et al*., 2020). A cassette containing a chimeric β-globin/immunoglobulin intron was inserted into the middle of the GFP gene, turning it from a single-exon transcript to a two-exon transcript. This change allows the insertion of another exon into the middle of the reporter system **(Figure 3.2)**.

3.2.4.1 Lysogeny Broth Media and Agar

Lysogeny Broth (LB) was prepared using 1% Bacto Tryptone (GIBCO, REF 211705), 0.5% Bacto Yeast Extract (GIBCO, REF 212750) and 1% Sodium Chloride (Fisher Scientific, 10418420). The LB medium was sterilised by autoclaving before use.

LB agar was made with the same composition as the LB medium described above, with the addition of 1.2% agar (GIBCO, REF 214530). The LB agar was also autoclaved before use to sterilise it.

3.2.4.2 DNA Transformation

The reporter construct was generously provided by Dr. Florian Heyd and subsequently transformed into NEB 5-alpha competent *E. coli* cells (NEB, #C2988J) following the manufacturer's protocol and plated on Kanamycin-containing LB agar plates (final concentration 50ug/mL).

3.2.4.3 Plasmid DNA Miniprep & Reporter Cloning

Following the transformation of the reporter construct into *E. coli*, colonies were inoculated in 5mL LB overnight cultures supplemented with a final concentration of 50 µg/mL kanamycin. These were incubated overnight in a shaking incubator at 37°C. The following morning, minipreps using the QIAprep Spin Miniprep Kit (Qiagen, Cat. No./ID: 27104) were performed following the provided kit method using the 5mL overnight cultures. A NanoDrop ND-1000 Spectrophotometer was used for DNA quantification. The DNA concentration, 260/280 and 260/230 ratios were noted down.

Subsequently, the second coding exon of *ZFY*, together with ~500bp flanking intronic sequence was commercially synthesised by GenScript and cloned into the GFP reporter vector using the XmaI/XhoI sites in the intron region. The *ZFY* exon sequences provided for insertion are shown in *supplementary* **Table 6**. In addition to the GFP reporter clones, GenScript also synthesised an *RBMY*-mCherry construct to use for co-transfection experiments alongside the GFP reporters.

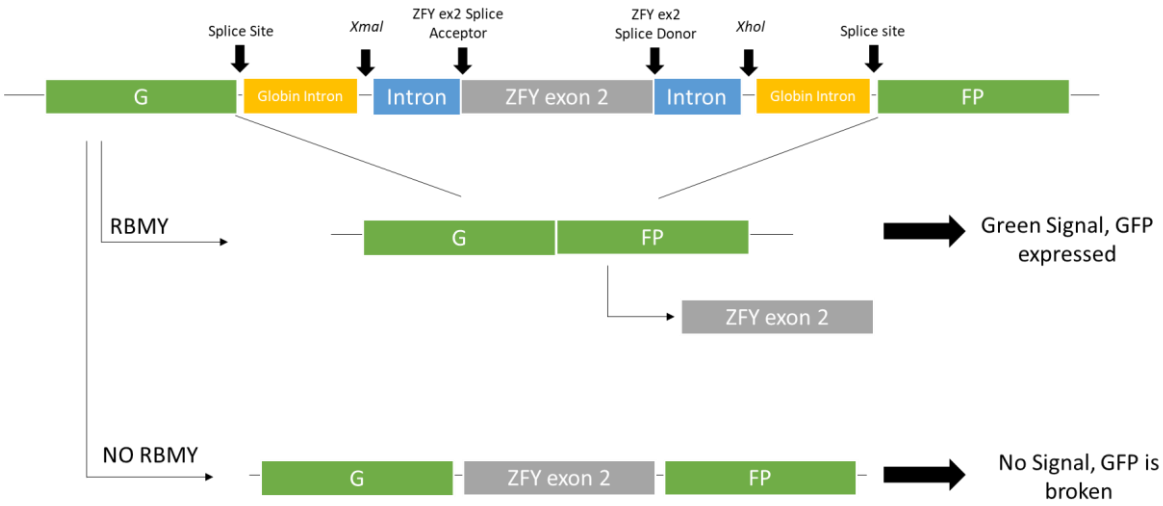


Figure 3.2: Luciferase reporter adapted to a GFPC1 vector schematic demonstrating the addition of the second coding exon of *ZFY*. The second coding exon of *ZFY* is alternatively spliced in males, with the current hypothesis suggesting *RBMY* is the splicing factor. If the additional exon is spliced out a GFP signal will be produced, whilst the inclusion of the additional exon will result in no signal since GFP is broken.

Upon receiving the GFP-Splicing reporter and *RBMY*-mCherry constructs, they were transformed into NEB 5-alpha competent *E. coli* cells (New England BioLabs, C2987H) and plated on LB agar plates with the appropriate antibiotics for selection. The GFP reporter vector is kanamycin resistant, while the *RBMY*-mCherry construct in the pcDNA3.1 backbone is ampicillin resistant (final concentration 100ug/ml). Overnight cultures were prepared, and plasmid DNA was extracted using the QIAprep Spin Miniprep Kit (Qiagen, Cat. No. 27104) as mentioned previously.

3.2.5 GFP-Splicing Reporter Mammalian Cell Transfection

To monitor *ZFY* splicing in a mammalian system, synthetically designed constructs **(Figure 3.2)** were transfected into mammalian cells and fluorescence was used to detect changes in *ZFY* variant expression.

3.2.5.1 HEK293 Cell Maintenance

Human embryonic kidney 293 cells (HEK293), a female hypotriploid human cell line derived from a foetal kidney, was used as the model system to monitor *ZFY* splicing. This adherent cell line was selected based on its reliable growth characteristics and high transfection efficiency. HEK293 cells were provided very generously by the Garrett Lab.

HEK293 cells were cultured at 37°C under humidified conditions, with 5% CO_2 as described in section 3.2.2. A logarithmic growth phase of the HEK293 cells was maintained by passaging the cells every 2-3 days. Due to the cells being adherent, the cells were detached from the flask via trypsinisation. Trypsinisation was performed on these adherent cells as described in section 3.2.2.

3.2.5.2 Lipofectamine 3000 Transfection

Lipofectamine 3000 Transfection (ThermoFisher Scientific, #L3000001) allows for the transfection of nucleic acids into eukaryotic cells at high efficiency.

Experimental conditions included: control (no transfection), *RBMY*-mCherry only (red fluorescence positive control), *ZFY*exon-mutant (green fluorescence positive control), *ZFY*exon only (GFP-Splicing construct), ZFYexon+mCherry (GFP-splicing construct cotransfected with mCherry) and *ZFY*exon+*RBMY* (GFP-Splicing construct cotransfected with *RBMY*-mCherry).

On day zero HEK293 cells were seeded into a 6-well plate with the optimised seeding density (no transfection = 300,000 cells per well & transfection = 450,000 cells per well), topping up to a 6mL total media volume per well. Once seeded, the plates were incubated at 37°C for 24 hours.

On day 1, the cells were transfected with the volumes and quantities stated in **Table 3.11**. A diluted DNA master mix and diluted lipofectamine master mix were produced separately and combined in a 1:1 ratio and then incubated at RT for 10-15 minutes. To note: 2µg of each construct was transfected into the corresponding well, with co-transfections consisting of 2µg of both individual constructs. This DNA: lipid mix was then added to the corresponding wells. The cells were then left for a further 48 hours. Following the incubation period, the cells underwent trypsinisation as previously stated, collected, and then pelleted.

Table 3.11: Lipofectamine 3000 reaction component volumes and protocol. 2µg of the desired DNA construct was added to each well and left for 48 hours.

	Component	**6-well plate volume (per-well)**
	Opti-MEM Medium	125µL

Diluted Lipofectamine Reagent	Lipofectamine 3000 Reagent	5.5µL
Diluted DNA	Opti-MEM Medium	125µL
	DNA	2µg
	P3000 Reagent (2µL/µg of DNA)	4µL
1:1 ratio	Add diluted DNA to diluted lipofectamine 3000 reagents (1:1 ratio)	
	Incubate at RT for 10-15 minutes	
	250µl of DNA-lipid complex added to cells (2µg DNA per well)	

3.2.5.3 Cell Fixing and Harvesting

48 hours post-transfection, the cell media was removed, and the cells were washed twice with ice-cold 1X PBS. The cells were then fixed for 10 minutes at room temperature using 4% paraformaldehyde (PFA) in PBS. After fixation, the cells were again washed with cold 1X PBS, collected in Eppendorf tubes, and stored at 4°C for up to several days.

3.2.5.4 Slide Preparation and Fluorescence Microscopy

Once the cells were collected, slides for fluorescence microscopy were prepared. 5µL of fixed cell suspension was placed onto a Superfrost microscope slide (Fisher Scientific, 11562203) and left to dry on a hotplate (31°C). Once dry, a drop of VECTASHIELD® Antifade Mounting Medium with DAPI (Vector Laboratories, H-1200-10) was placed on top of the cells and a coverslip was gently placed on top with slight pressure.

Prepared slides were then visualised using the Olympus BX61 epifluorescent microscope and SmartCapture 3 software. Images were captured using the Texas Red, FITC, and DAPI filters at a x20 magnification.

3.2.6 Flow Cytometry

Flow cytometry analysis was performed using a BD Accuri C6 Plus system. The FL1 laser (522/30 nm filter) was used to detect GFP fluorescence, while the FL3 laser (670 nm long pass filter) detected mCherry fluorescence. For the flow cytometry setup, the cell count was set to 10,000, and cells were gated based on forward and side scatter to remove any debris. Gating thresholds and alignments were determined using control samples. Using the analysis function on the program software, alignment was set for both GFP and mCherry signal peaks using the control samples **(Figure 3.3)**. Since multiple immunofluorescent labels were used in this experiment, there was a potential for a given fluorochrome to emit signals in the incorrect detector.

Consequently, a compensation factor of 4% was established to account for spectral overlap between fluorochromes.

To prevent sample carryover between runs, the flow cytometry system was flushed with water between samples.

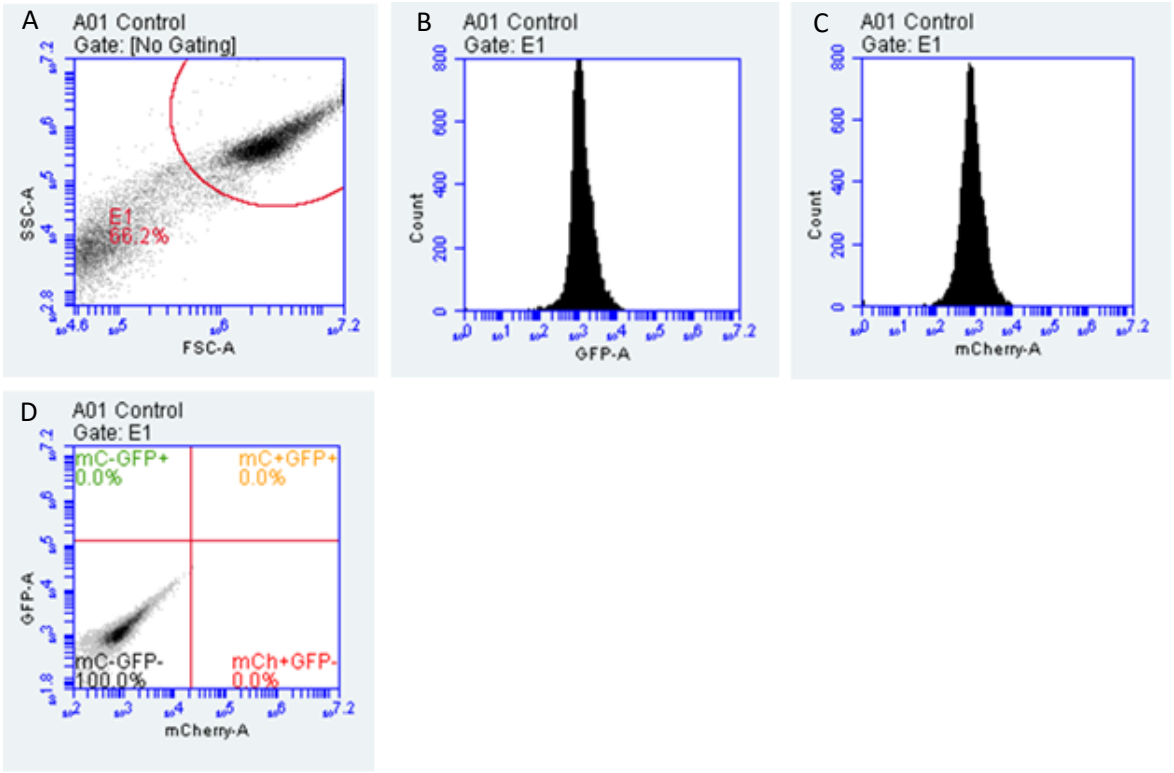


Figure 3.3: Flow cytometry gating strategy using untransfected control HEK293 cells. A: Plot A shows the forward scatter-area (FSC-A) (X-axis) and side scatter-area (SSC-A) (Y-axis) highlighting the gated cell population highlighted by E1. **B:** Gating and alignment of the GFP signal using the negative control. **C:** Gating and alignment of the mCherry signal using the negative control. **D:** Quadrant plot indicating percentages of cells positive for GFP only (Top left), mCherry only (Bottom right), double positive (Top right), or negative (Bottom left).

3.2.7 GFP-Splicing Reporter Polymerase Chain Reaction

RNA extraction and cDNA synthesis were performed as mentioned in section 2.2.3 of this chapter.

To determine if *RBMY* regulates alternative splicing of *ZFY* to generate the *ZFYS* isoform, PCR was performed using the primers listed in **Table 3.13** below and the GoTaq G2 Flexi DNA polymerase kit (Promega, M7801). The PCR reaction setup is described in **Table 3.9** of this chapter. The thermocycler conditions for this PCR can be found in **Table 3.13**. To control for even sample loading, the expression of the housekeeping gene TBP was examined in parallel. The expected band size for *TBP* is 108bp **(Table 3.10)**.

The primers in **Table 3.12** target sequences flanking the alternatively spliced exon and will produce either a short 167 bp amplicon if splicing occurs or a longer 740 bp amplicon containing the exon. Successful splicing of the exon should result in only the short PCR product, while retention of the exon gives the larger band.

Table 3.12: GFP-splicing reporter system primers. These primers were designed and taken directly from (Neumann *et al*., 2020), and work by binding to GFP upstream and downstream of the inserted intron. IDT synthetically designed primers.

Primer Name	Accession Number	Primer Sequence	Tm	Product Length (bp)
GFP Splicing Reporter	-	For: CACATGAAGCAGCACGACTT Rev: TCCTTGAAGTCGATGCCCTT	55.9° C 56.3° C	Long: 740 Short: 167

Table 3.13: Thermocycler conditions for the GFP splicing primers. Since both short and long amplicons were expected in some lanes, a slightly longer extension time was used during PCR to ensure efficient amplification of both products.

Phase	Temperature (°C)	Time (min)	Cycle Number
Initial denaturation	94	02:00	1
Denaturation	94	00:15	30
Annealing	60	00:15	
Extension	72	00:45	
Final extension	72	05:00	1
Hold	4	Infinite	1

3.3 Results

3.3.1 Human and Mouse Splicing Events are not Directly Detected in Short Read RNA-Seq

Using the public organ-specific RNA-Seq data for both humans and mice at varying development stages, reads were flanked to the coordinates of *ZFY* in the Ensembl genome. Reads across the exons could then be plotted graphically using pysashimi, however, a complication in mapping was noted. This meant that very few uniquely mapping intron-spanning reads were identified in both these species. This is due to the presence of both X and Y copies, and specifically the presence of *Zfy1/2* in mice which causes the mapping complication. Due to the high similarity between the X and Y homologues, unique mapping is more difficult. However, the splicing patterns of humans and mice are already known from published data.

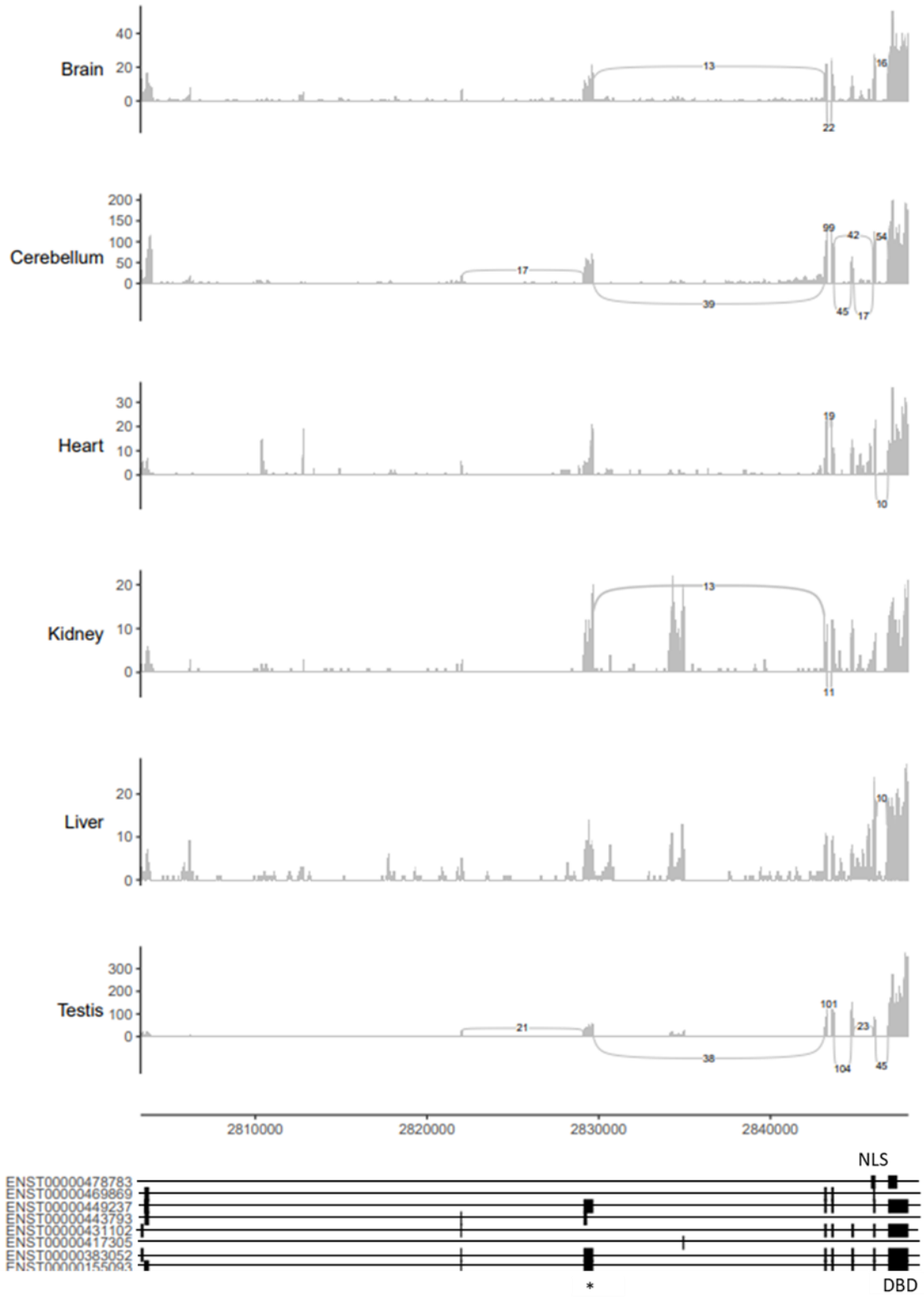


Figure 3.4: Sashimi plot of transcripts mapped to the human genome in the region of ENST00000155093 (zinc finger protein Y-linked). The read count for junction-spanning reads linking specific exons is noted. Forward strand - Exons: 8, Coding exons: 7. DBD: DNA binding domain, NLS: Nuclear localisation signal, * alternatively spliced exon.

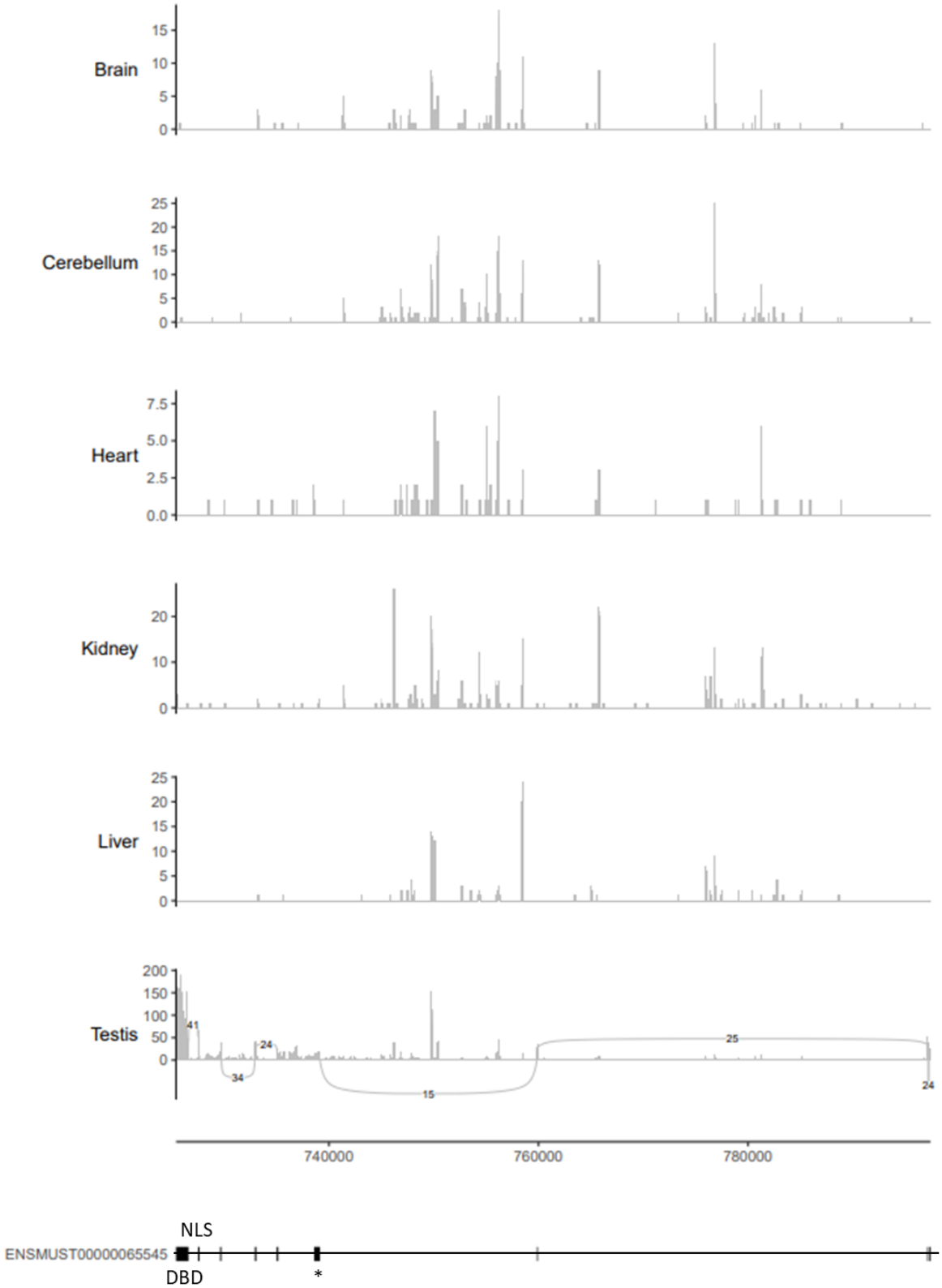


Figure 3.5: Sashimi plot of transcripts mapped to the mouse genome in the region of ENSMUST00000065545 (zinc finger protein Y-linked 1). The read count for junction-spanning reads linking specific exons is noted. Note that the exons run right to left - Reverse Strand - Exons: 9, Coding exons: 7. DBD: DNA binding domain, NLS: Nuclear localisation signal, * alternatively spliced exon.

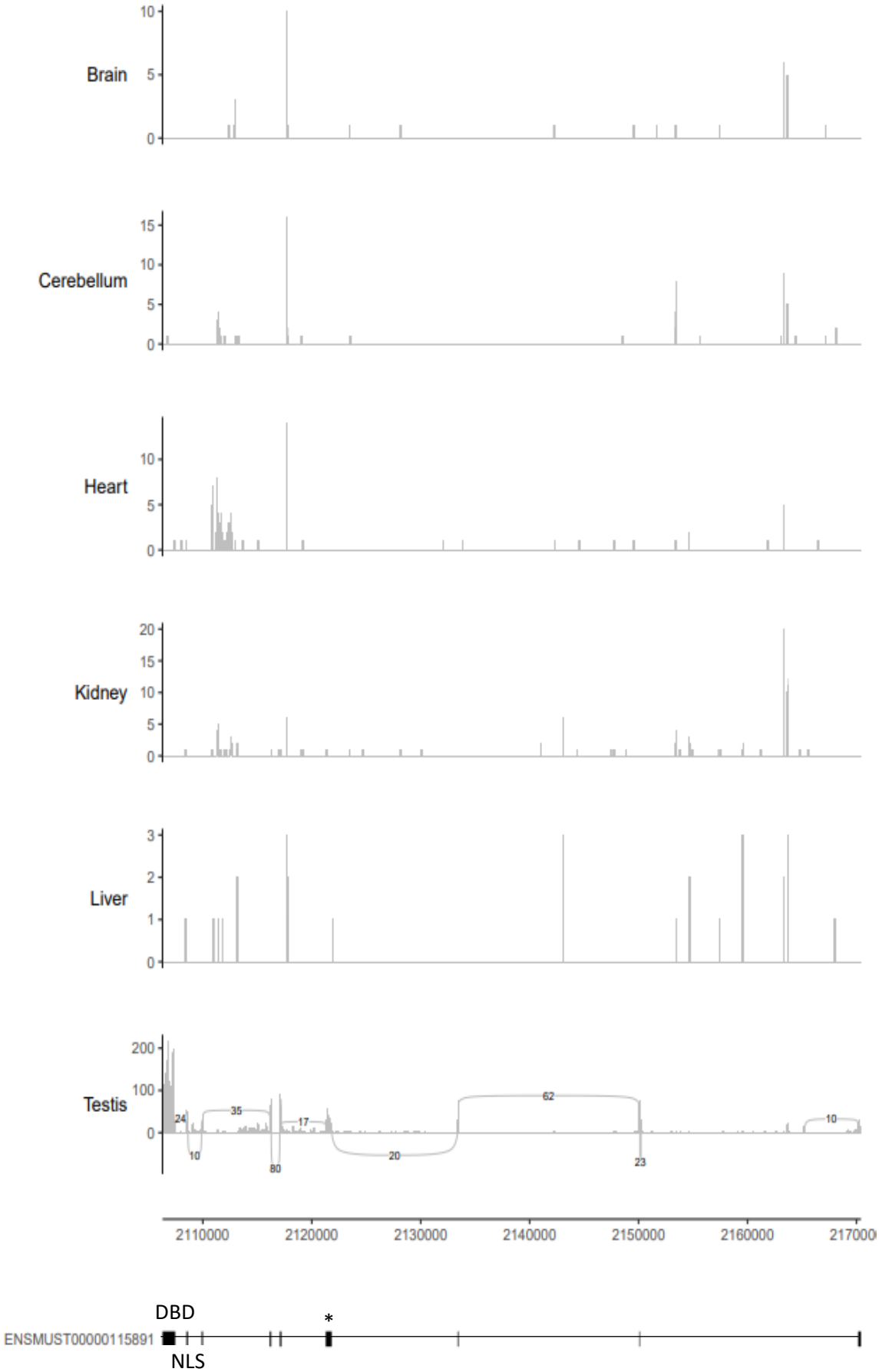


Figure 3.6: Sashimi plot of transcripts mapped to the mouse genome in the region of ENSMUST00000115891 (zinc finger protein Y-linked 2). The read count

for junction-spanning reads linking specific exons is noted. Note that the exons run right to left - Reverse Strand - Exons: 9, Coding exons: 7. DBD: DNA binding domain, NLS: Nuclear localisation signal, * alternatively spliced exon.

As seen in **Figures 3.4, 3.5 & 3.6** there are low read numbers across all organs. In the human plots, reads can be seen in all organs apart from the liver, but even then, the read numbers are too low to identify the testis splicing pattern known to be present in humans and mice. While reads are not expected in the mice organs except for the testis, the read information is still limited. As mentioned before, the presence of X and Y copies complicates mapping, which explains why there are very few uniquely mapping intron-spanning reads in these species. Due to the lack of splice junctions supported at any given age, **Figures 3.4, 3.5 & 3.6** display pooled reads. This decision was made to enhance the number of unique reads mapping to the intron-spanning regions by pooling the data for each species at each age. This would hopefully improve the visibility of any potential splicing pattern, however, for the human and mouse data this still resulted in a low read number as seen.

However, due to Zf* being autosomal in both chicken and opossum there was no additional mapping complication as there are no X and Y copies, meaning there was much better-read mapping across the organs.

3.3.2 The Eutherian Testis-Specific Isoform *ZFYS* is Conserved in Opossum

Using public RNA-Seq data from opossum across a variety of developmental stages, reads were flanked to the coordinates of Zf* in the Ensembl genome. Reads across the exons could then be plotted graphically using pysashimi. **Figure 3.7** summarises mapped transcript reads to the Opossum Zf* gene residing on chromosome 4. **Figure 3.7** clearly shows 12 reads indicative of an exon skipping event resulting in the exclusion of the second coding exon, i.e. conservation of the *ZFYS* splice form in marsupials. Importantly, this exon was never skipped in any other tissue, even the cerebellum where there is a similar absolute level of expression to testis. Therefore, it can be concluded that the generation of a testis-specific short transcriptional isoform likely represents the ancestral state before recruitment of Zf* to the eutherian sex chromosomes, and that subsequently the eutherian Y copy (i.e. *ZFY*) has retained this testis-specific function while the X copy (i.e. *ZFX*) no longer produces the short form.

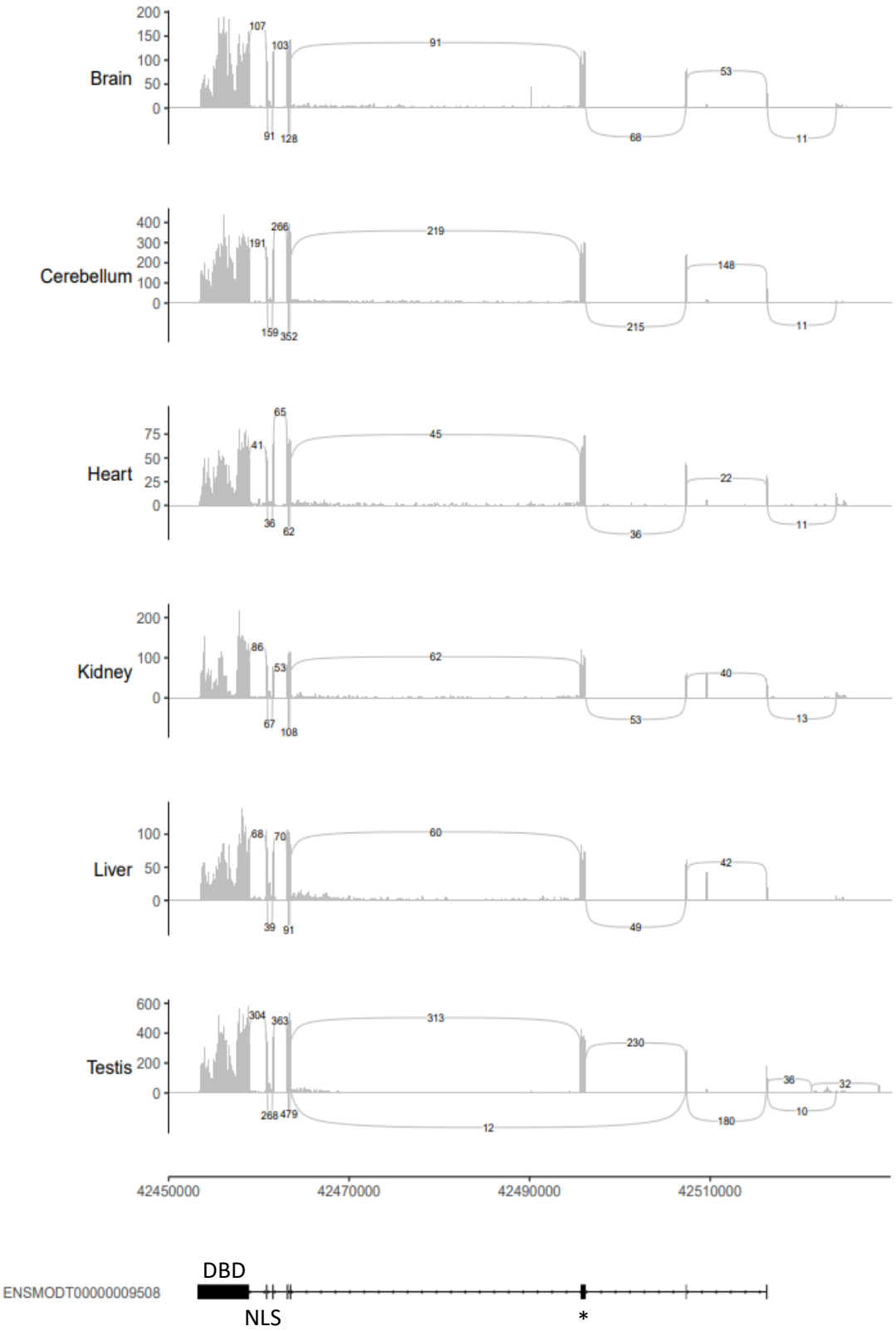


Figure 3.7: Sashimi plot of transcripts mapped to the opossum genome in the region of ENSMODT00000009508 (zinc finger protein X-linked). The read count for junction-spanning reads linking specific exons is noted. Note that the exons run right to left – Reverse Strand. Exons: 9, Coding exons: 7. Only 8 exons are shown here (2-9, missing the first non-coding exon). DBD: DNA binding domain, NLS: Nuclear localisation signal, * alternatively spliced exon.

3.3.3 A Potential Alternative Short Zf* Splice Form is Observed in Chicken Testis

Following the marsupial work, transcript mapping of chicken RNA-Seq data was conducted to investigate whether the observed splicing effect is also present in more distantly related vertebrates. **Figure 3.8** surprisingly does not validate the existence of a testis-specific splicing pattern as seen in eutherian mammals. Instead, it reveals an unexpected finding. In all examined organs, fewer transcript reads align with ENSGALT00010007123 where exon 6 is missing, and notably, exon 6 is generally included. Regarding the spliced second coding exon, it appears to be present in the testis without exclusion. However, numerous additional peaks outside exon regions are discernible in the testis which are not found in other organs and further to this, transcripts are mapped to regions between these sites. This suggests a novel testis-specific splicing event involving Zf* in chickens. Specifically, the splice junctions and read counts shown by the sashimi plot indicate the presence of three novel exons, labelled A, B and C in **Figure 3.8**. These are each linked to each other and to the NLS-containing exon by well-supported splice junctions specifically in testis. There are however no well-supported splice junctions linked to the 5' end of exon A suggesting that this may be an alternative transcriptional start site. Overall, the novel exons imply a testis-specific transcript starting with exon A and including exons A/B/C followed by the NLS and DNA binding domain of *ZFY*. The resulting novel form would therefore entirely omit the acidic domain, and might therefore function equivalently to mammalian *ZFYS*, as a *ZFY* form that retains DNA binding capacity but without transactivation ability. If so, it implies that chickens and mammals both generate "transactivation-dead" *ZFY* variants, but via different mechanisms. Unfortunately, we did not have time (or access to chicken testis tissue) to follow up on these observations in the scope of this PhD.

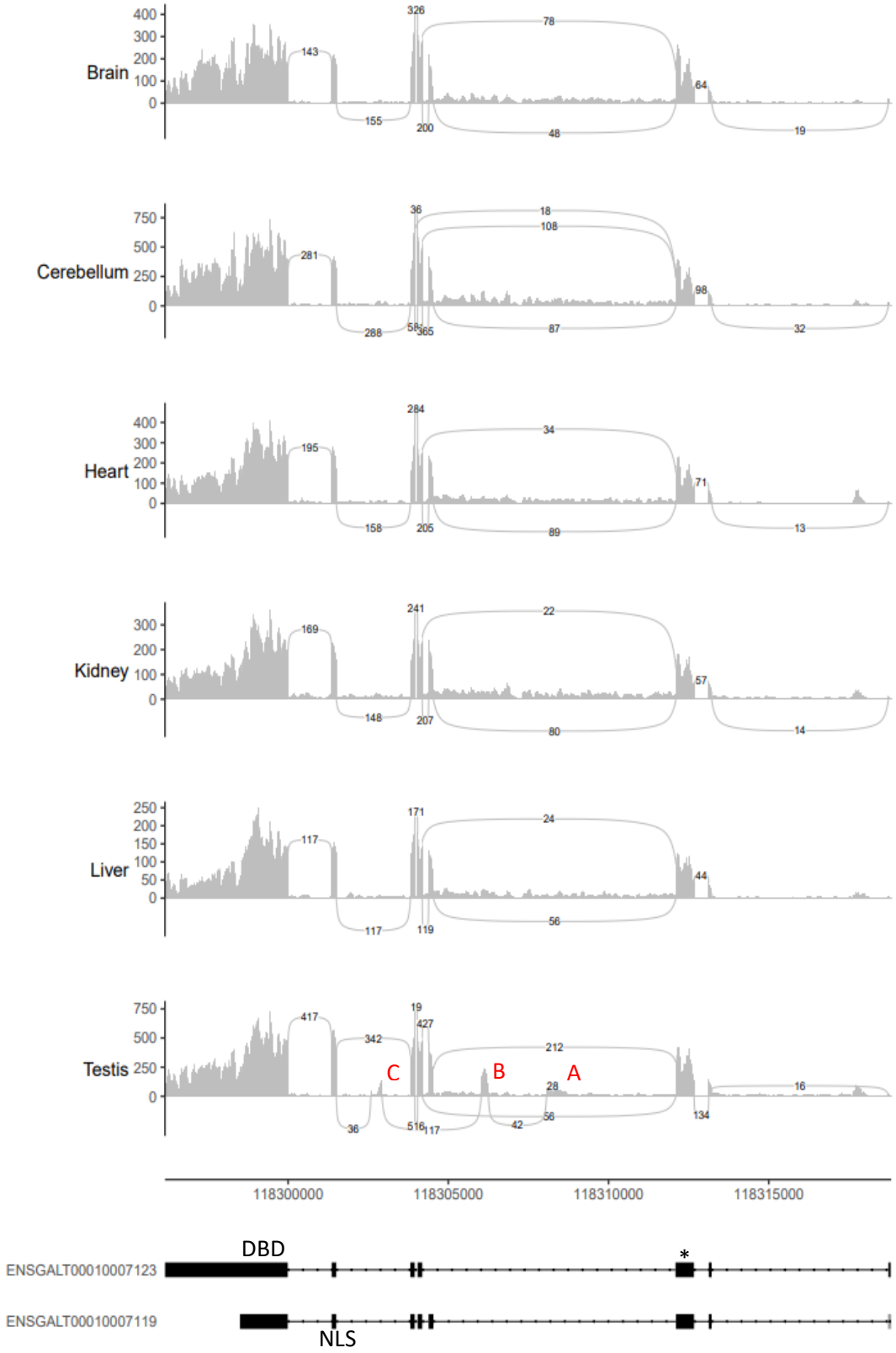


Figure 3.8: Sashimi plot of transcripts mapped to the chicken genome. The read count for junction-spanning reads linking specific exons is noted. ENSGALT00010007123 is the *ZFX*-201 transcript (Note that the exons run right to left – Reverse Strand: 7 exons, 6 coding) annotated on Ensembl and ENSGALT00010007119 is the *ZFX*-202 transcript (Note that the exons run right to left – Reverse Strand: 9 exons, 7 coding) annotated on Ensembl. Note that the exons run

right to left. DBD: DNA binding domain, NLS: Nuclear localisation signal, * alternatively spliced exon.

3.3.4 *ZFYS* and *RBMY* are Expressed in a Head and Neck squamous Cell Carcinoma

Primers for *ZFY* were designed to surround the second coding exon, which undergoes alternative splicing. The forward primer was positioned within the first coding exon of *ZFY*, while the reverse primer was situated in the third coding exon. As *RBMY* is testis-specific in humans (and all mammals), the aim was to also see if it becomes mis-expressed in cancer cells like *ZFY*, primers were therefore designed to determine the level of *RBMY*.

PCR was performed and visualised on a 2% agarose gel to determine the presence of these testis-specific genes in two head and neck squamous cell carcinomas; SCC104 and PCI30.

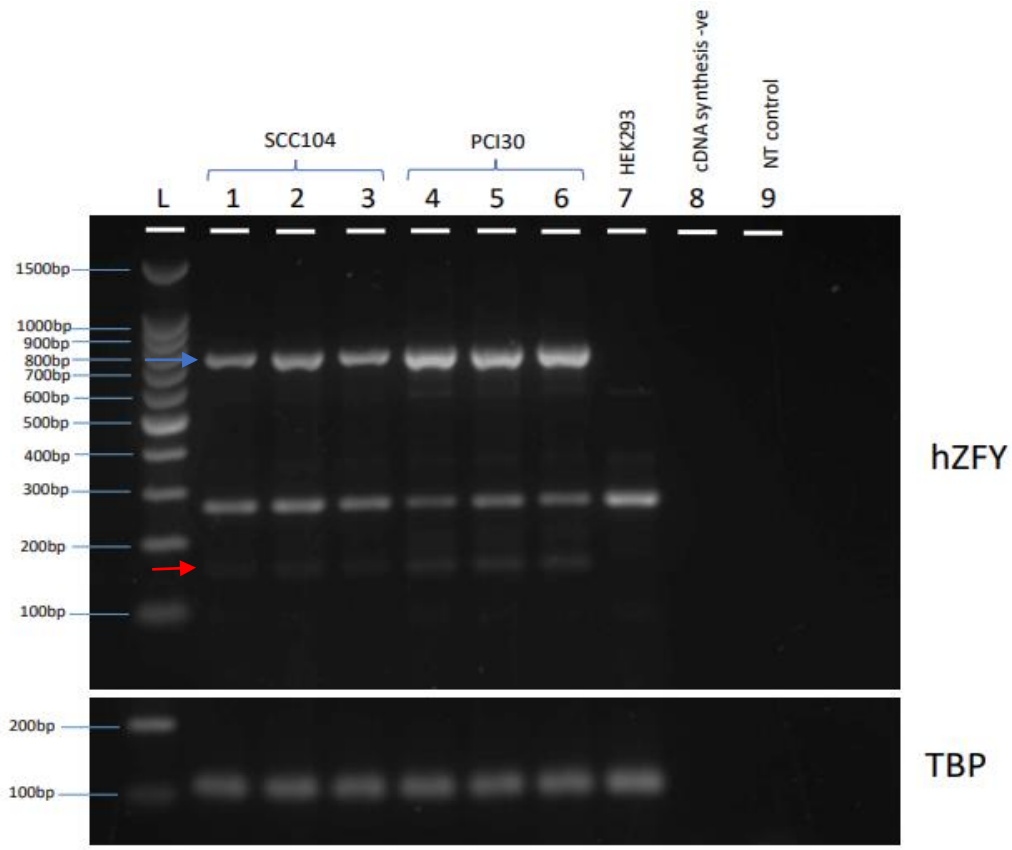


Figure 3.9: 2% agarose gel following *ZFY* PCR amplification in SCC104 and PCI30. Three replicates for SCC104 and PCI30 are shown alongside three control samples. HEK293 negative control: a female cell line with no endogenous *ZFY* expression, cDNA synthesis negative control: to ensure no contamination during cDNA synthesis occurred and a no template control. TBP was used as a confirmation of equal loading. Predicted band sizes: *ZFYL* = 729bp highlighted by the blue arrow and *ZFYS* = 125bp highlighted by the red arrow.

Clear bands in **Figure 3.9** are at the expected size for *ZFYL* (729bp) across both SSC104 and PCI30 which is not seen in the three controls. Furthermore, a faint band at the expected size of *ZFYS* (125bp) is also evident across the SSC104 and PCI30 replicates which is not seen in the control lanes; a greater signal was noted in PCI30. This suggests that these head and neck squamous carcinomas express a low amount of *ZFYS*, a testis-specific gene. The band at ~270bp is present across all lanes including the female control (HEK293) and is, therefore, a non-specific contaminant band. The identity of this band was not further confirmed in this thesis. In parallel,

RBMY PCR amplification was performed in SCC104 and PCI30 to confirm its presence.

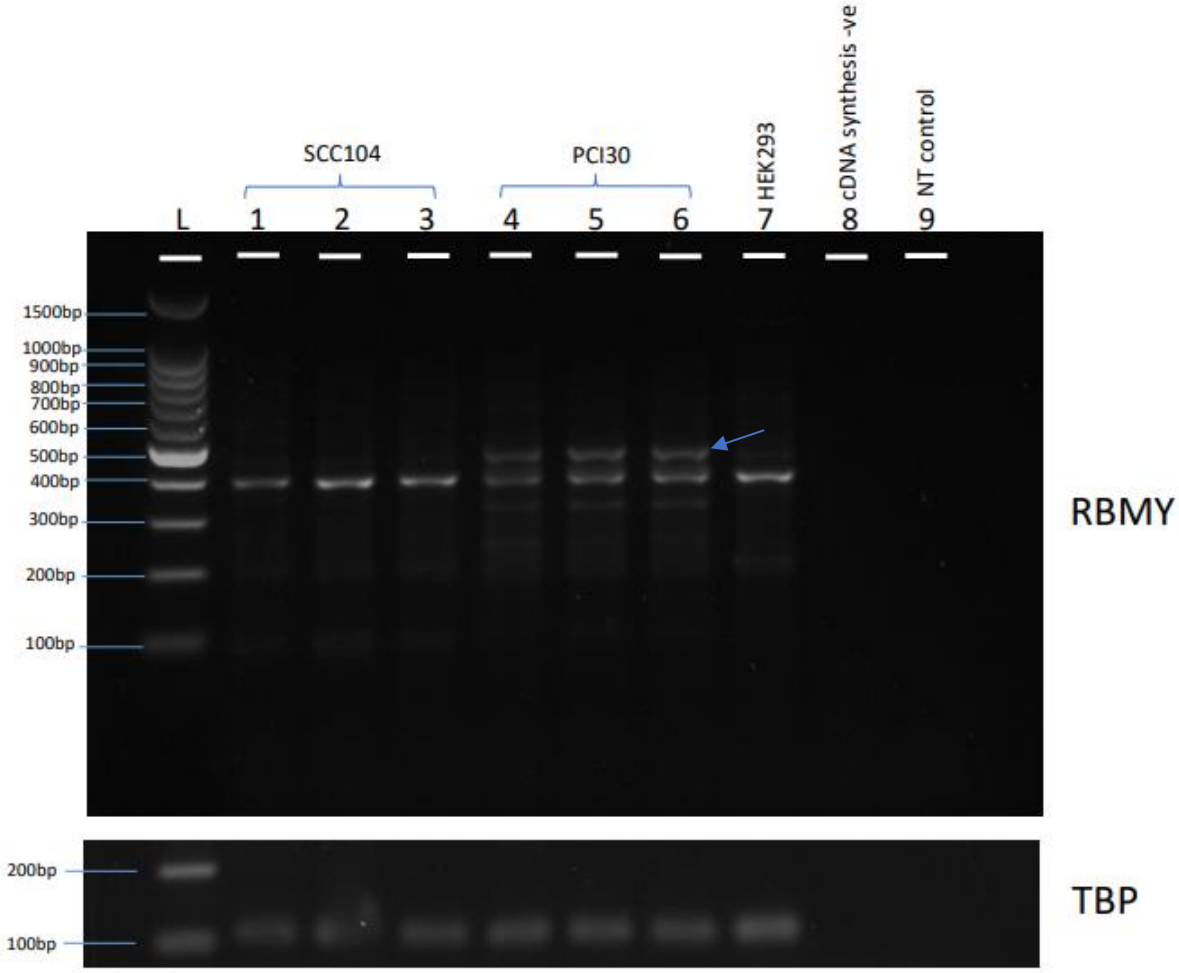


Figure 3.10: 2% agarose gel following *RBMY* PCR amplification in SCC104 and PCI30. Three replicates for SCC104 and PCI30 are shown alongside three control samples. HEK293 negative control: a female cell line with no endogenous *RBMY* expression, cDNA synthesis negative control: to ensure no contamination during cDNA synthesis occurred and a no template control. TBP was used as a confirmation of equal loading. Predicted band sizes: *RBMY* = 473bp - highlighted by the blue arrow.

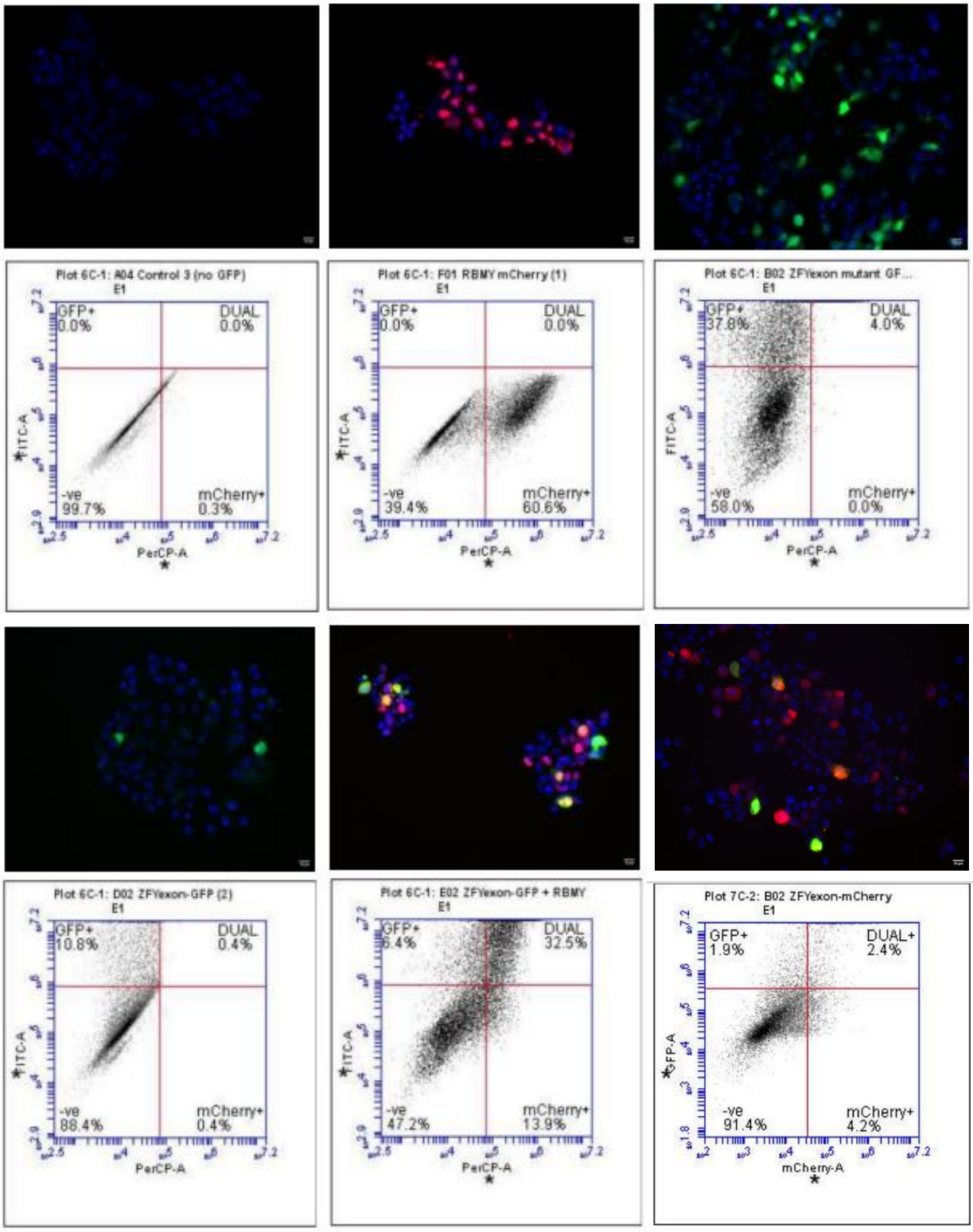
The expression of *RBMY* in PCI30 but not in SSC104 is shown in **Figure 3.10**, by a distinctive band at 473bp. PCI30 exhibits *RBMY* expression and demonstrates a higher abundance of the *ZFYS* band in **Figure 3.9**, supporting the hypothesis that *RBMY* may impede the inclusion of the second coding exon. Notably, these primers generate non-specific bands in lanes corresponding to SCC104, PCI30, and HEK293, prominently at approximately 400bp. The identity of this band is uncertain, but its visibility in the HEK293 lane suggests a potential endogenous gene expressed in females.

These results confirm the co-expression of two testis-specific genes; *ZFYS* and *RBMY* in PCI30, a HPV-negative oral tongue squamous carcinoma. This is less evident in SCC104, but potentially the levels are lower and not so detectable.

3.3.5 Direct Testing of *RBMY* Regulation of *ZFY* Splicing in a Model System

Using a modified GFP splicing reporter system, constructs were synthetically designed to contain the second coding exon and the second coding exon with splice site mutations. The principle of the assay is when the GFP-exon2 reporter is transfected by itself, the *ZFY* exon will be included in the mature transcript, breaking the GFP open reading frame and resulting in no detectable GFP signal. A positive control with splice site mutations was constructed which would prevent the inclusion of the *ZFY* exon, resulting in the GFP open reading frame remaining intact and a GFP signal would be detectable. Finally, when mCherry tagged *RBMY* was cotransfected with the GFP-exon2 reporter, it is predicted that the *RBMY* would trigger the exclusion of the *ZFY* exon, resulting in a detectable GFP signal. By using mCherry and GFP tags changes in fluorescence could be monitored by flow cytometry and microscopy. Subsequent RT-PCRs could also be used to look at the splicing ratio.

A


Plot 6C-1: A04 Control 3 (no GFP)
E1
GFP+ 0.0%
DUAL 0.0%
-ve 99.7%
mCherry+ 0.3%
FITC-A
PerCP-A
Plot 6C-1: F01 RBMY mCherry (1)
E1
GFP+ 0.0%
DUAL 0.0%
-ve 39.4%
mCherry+ 60.6%
FITC-A
PerCP-A
Plot 6C-1: B02 ZFYexon mutant GF...
E1
GFP+ 37.8%
DUAL 4.0%
-ve 58.0%
mCherry+ 0.0%
FITC-A
PerCP-A
Plot 6C-1: D02 ZFYexon-GFP (2)
E1
GFP+ 10.8%
DUAL 0.4%
-ve 88.4%
mCherry+ 0.4%
FITC-A
PerCP-A
Plot 6C-1: E02 ZFYexon-GFP + RBMY
E1
GFP+ 6.4%
DUAL 32.5%
-ve 47.2%
mCherry+ 13.9%
FITC-A
PerCP-A
Plot 7C-2: B02 ZFYexon-mCherry
E1
GFP+ 1.9%
DUAL+ 2.4%
-ve 91.4%
mCherry+ 4.2%
FP-A
mCherry-A

B

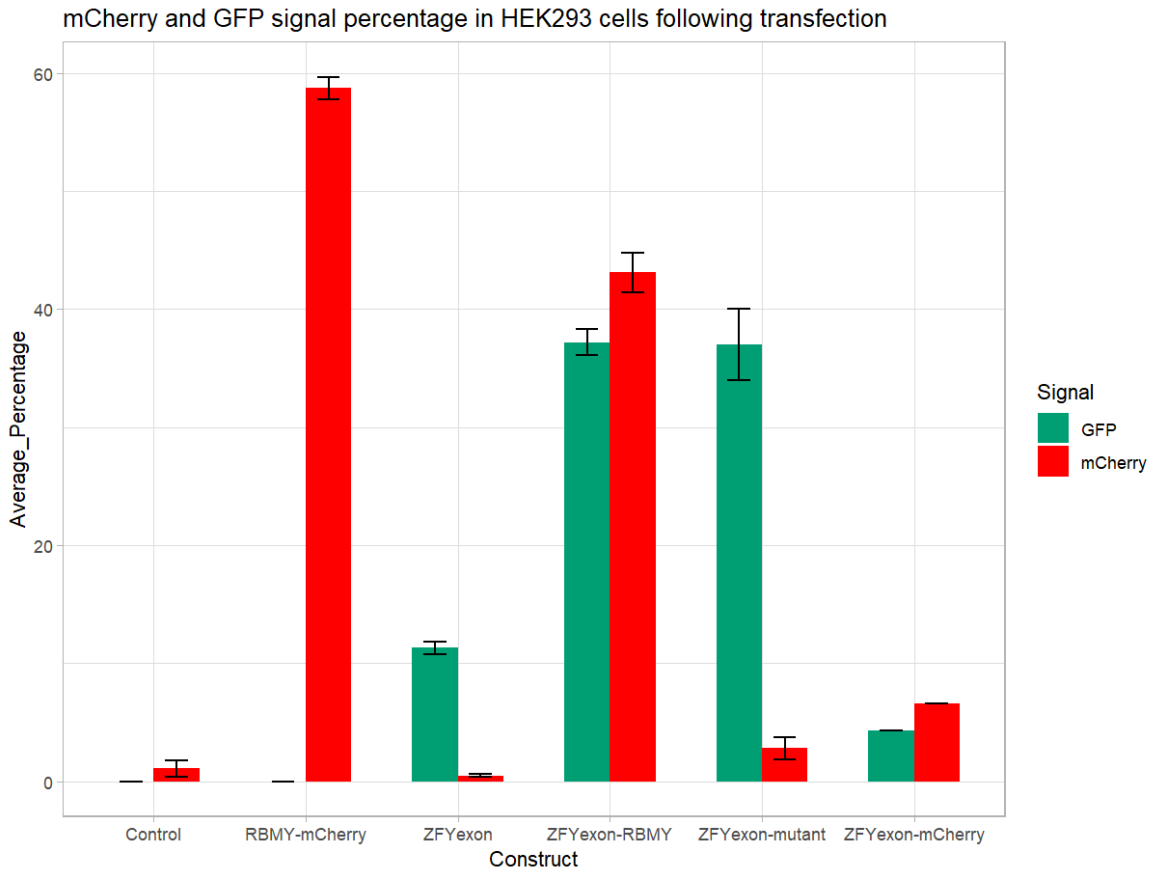


Figure 3.11: Microscopic and flow cytometry analysis of the splicing constructs following transfection into HEK293 cells. A: Each sample was analysed under x20 magnification and flow cytometry. Flow cytometry graphs display four quadrants; -ve = no signal, mCherry+ = Red signal, GFP+ = Green signal and Dual = both red and green signal detected. Samples top panel left to right: control, *RBMY*-mCherry only, *ZFY*exon-mutant (*ZFY* exon with splice site mutations in the reporter system). Samples bottom panel left to right: *ZFY*exon (*ZFY* exon inserted into the reporter construct), *ZFY*exon+*RBMY* (*ZFY*exon cotransfected with *RBMY*), *ZFY*exon-mCherry (*ZFY*exon cotransfected with mCherry-only). **B:** Graph representing the combined repeats data across the transfections n=3 (ZFYexon-mCherry n=1).

Gating was set based on three samples. The control sample with no expected fluorescence was used as a negative control to determine untransfected cells. *RBMY*-mCherry was used as a positive gating for the mCherry filter, and *ZFY*exon-mutant was used as a positive for the GFP filter. By setting alignments based on this, samples were standardised.

In **Figure 3.11A**, untransfected cells are depicted clustering in the -ve quadrant in the bottom left corner, indicating the absence of both green and red signals. Following the transfection with *RBMY*-mCherry, there was a significant shift in cell distribution, with 60.6% now residing in the mCherry +ve bottom right quadrant, signifying that 60.6% of the cells express *RBMY*.The *ZFY*exon-mutant is anticipated to exhibit the highest GFP signal due to splice site mutations causing exon removal even in the

absence of *RBMY*. This restoration of the GFP open reading frame results in a green fluorescence signal. According to **Figure 3.11A**, 37.8% of the cells exhibit a GFP signal, with an additional 4% leaking into the dual quadrant. Despite setting a compensating factor, there is still some evident spill-over into different channels, albeit minor.

The *ZFY*exon without *RBMY* yields only 10.4% of cells fluorescing green compared to the 37.8% observed in the *ZFY*exon mutant. This 10.4% discrepancy may be attributed to a limited level of exon skipping even in the absence of *RBMY*. However, it is evident that when the *ZFY*exon is cotransfected with *RBMY*, the green fluorescence increases. In this cotransfection, the focus is on the dual quadrant, highlighting cells emitting both green and red fluorescence. Notably, 32.5% of the cells exhibit dual fluorescence, meaning that overall, of the cells expressing *RBMY*, 70% also emit a green signal due to the splicing out of the *ZFY* exon from the construct. This shift in signal is also seen in **Figure 3.11B** where the replicates were combined. There is a pattern shift when focusing on *ZFY*exon compared to the *ZFY*exon cotransfected with *RBMY*. When cotransfected, the GFP signal of *ZFY*exon matches that of the *ZFY*exon-mutant with a GFP mean fold change of 3.8. An additional cotransfection control experiment was also conducted using the *ZFY*exon construct alongside a commercially synthesised intermediate mCherry-tag. However, the intermediary construct used in creating the *RBMY*-mCherry construct proved to be unstable, as indicated by the poor transfection efficiency and the weak red signal observed compared to the *RBMY*-mCherry +red signal (**Figure 3.11B** n=1).

Whilst these initial results were promising, PCR confirmation was subsequently performed to determine the ratio of GFP and GFP+exon with primers located in the GFP regions surrounding the inserted exon.

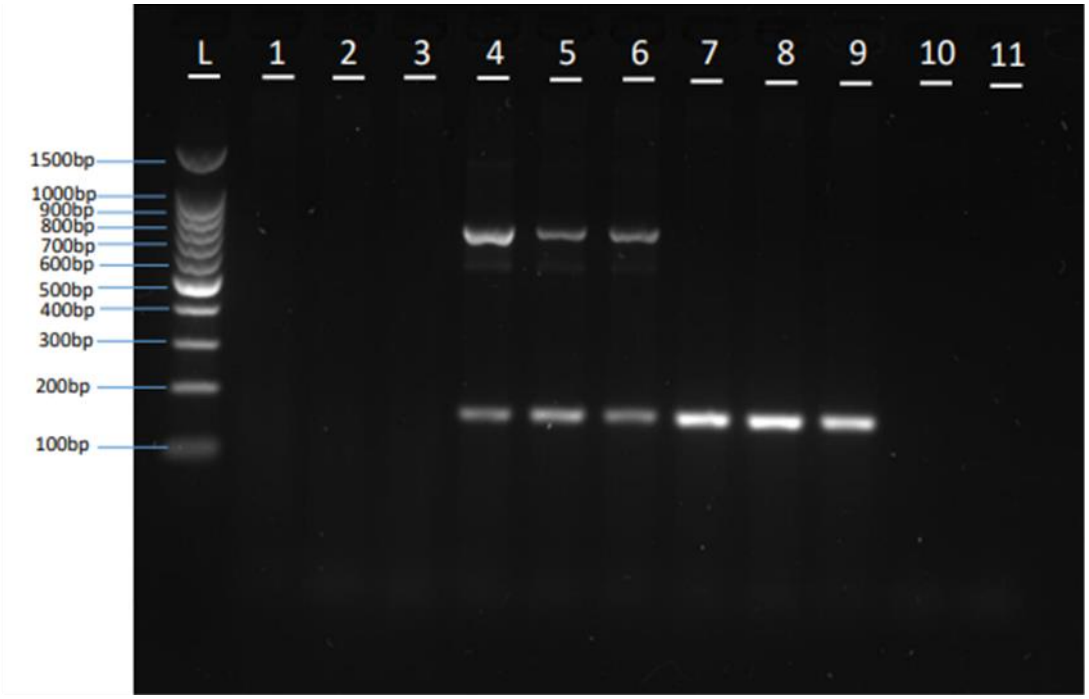


Figure 3.12: 2% PCR agarose gel using GFP splicing reporter primers. L: ladder, **1:** untransfected control, **2:** mCherry only, **3:** *RBMY*-mCherry, **4:** *ZFY*exon, **5:** *ZFY*exon + *RBMY*, **6:** *ZFY*exon + mCherry, **7:** *ZFY*mutant, **8:** *ZFY*mutant + *RBMY*, **9:** *ZFY*mutant + mCherry, **10:** cDNA synthesis negative control, **11:** GoTaq NT control. Expected bands: GFP = 167bp, GFP+exon = 740bp.

In **Figure 3.12**, lanes 7, 8, and 9 show the *ZFY*exon-mutant samples, displaying only a single band at the anticipated GFP size. This implies the removal of the second coding exon with mutated splice sites from the GFP open reading frame, resulting in the production of a full-length, uninterrupted GFP. Negative controls (lanes 1, 2, 3, 10, and 11) show no bands, as both GFP or a GFP reporter system are not expressed. However, lanes 4, 5, and 6, containing the *ZFY*exon, reveal variations in the upper and lower band ratio across the lanes. The relative abundance of the upper (*ZFYL*) and lower (*ZFYS*) bands was quantitated using ImageJ. ImageJ is capable of comparing signal intensity between selected bands. To do this the background was first removed and the mean gray value-only option was selected. Using the analysis tool the background signal was calculated and subtracted from the actual band signal. Each lane was outlined for analysis and the data was plotted for each lane as a peak signal. Finally, using the line tool, the signal peaks were joined, and the wand filter could be used to determine the amount of signal produced from each band under investigation. This data was then used to determine the ratio of *ZFYL* to *ZFYS*.

In lane 4, *ZFY*exon expression alone yields a greater amount of the full-length construct, with a ratio of 72:28 (L:S), represented graphically in **Figure 3.13**. Co-transfection with *RBMY*-mCherry shifts this ratio to 39:61, indicating an increased presence of the expected shorter band after exon removal. Conversely, co-transfection with mCherry alone results in a ratio of 54:46, suggesting a higher splicing-out of the exon in the presence of mCherry alone. This raises concerns and diminishes confidence in the reporter system design, although the most significant change is observed in the presence of *RBMY*. It has also been suggested that the mCherry was produced in an intermediate step of the construct cloning and is potentially less stable.

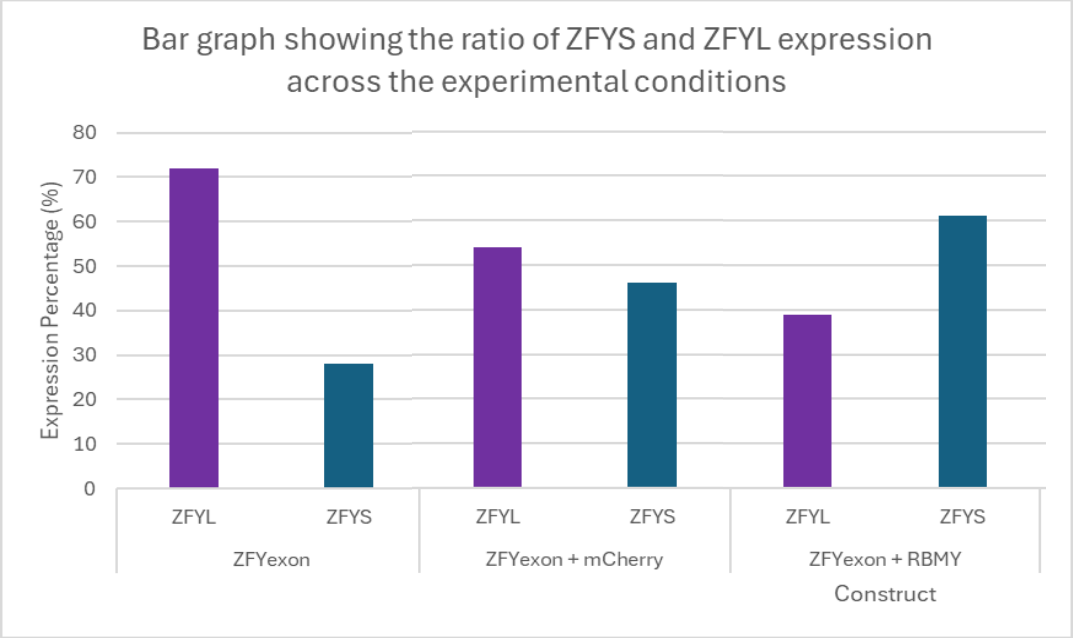


Figure 3.13: A bar graph showing the ratio of *ZFYS* and *ZFYL* expression across the experimental conditions. The X-axis shows the construct and the *ZFYL* and *ZFYS* split, with the expression percentage (%) of the ratio plotted on the Y-axis.

3.4 Discussion

The majority of all protein-coding genes are alternatively spliced (Tazi *et al*., 2009). This plays a crucial role in enhancing the coding potential of eukaryotic genomes by allowing a single gene to produce multiple distinct proteins through the generation of diverse mRNA transcripts (Tazi *et al*., 2009). In 2012, Decarpentrie *et al* identified a short testis-specific *ZFY* variant arising from an alternative splicing event which excluded the second coding exon of the *ZFY* gene (Decarpentrie *et al*., 2012). Despite this finding, the reasoning behind this event and the splicer remains largely unknown as research around *ZFY* dwindled after it was disproved as the sex-determining gene. In this chapter, *RBMY* was explored as a potential splicing factor for *ZFY*, prompted by suggestive evidence hinting at a correlation between these two testis-specific genes.

Following investigation into the possibility of marsupial ZF splicing, it was found that opossum ZF also undergoes a testis-specific splicing event that is not evident in the other organs. Although most transcripts maintain the exon expressing the default full-length form, there is a discernible pattern suggesting the presence of a testis-specific short Zf* isoform pointing to a potentially splicing event before the movement of *ZFY* onto the Y chromosome. Furthermore, when investigating chicken ZF, an unexpected testis-specific phenomenon was identified. In the chicken testis, a distinct pattern of read mapping emerges in contrast to the other organs, yet understanding precisely what is occurring proves challenging. The analysis of mouse and human read mapping data for *ZFY* is limited because the low number of reads mapping to the gene renders the interpretation of the data challenging. The low read count could be due to the multi-mapping reads being discarded. However, RT-PCR evidence has previously demonstrated the existence of a testis-specific short *ZFY* variant in placental mammals.

To validate the previous findings indicating the expression of both *ZFYS* and *RBMY* in head and neck cancer cell lines, PCR analysis was conducted. Some cancer cells have been identified to exhibit unexplained *ZFY* expression, leading to the hypothesis that *ZFY* may possess some indirect oncogenic activity beyond its function in the testis. Due to previous findings in the Ellis-Fenton lab, ectopic expression of the short-spliced variant in HPV- OPSCC cell lines was suggested to explain the male prevalence for HNSCCs. The data presented indicates that *ZFYS* is expressed in both SCC104 (HPV+) and PCI30 (HPV-) at low levels, whereas *RBMY* is only detectable in PCI30. PCI30 exhibits higher *ZFYS* expression and co-expression of *RBMY*, suggesting *RBMY*'s involvement in *ZFY* splicing. However, the absence of *RBMY* expression in SCC104 contradicts this hypothesis. It is possible that the *RBMY*

levels in SCC104 were below detection thresholds, or potentially an alternative *RBMY* variant may be present, facilitating *ZFY* splicing at lower levels. The RT-PCR analysis for *RBMY* revealed an unexpected band, slightly smaller than the expected *RBMY* band, present in all lanes; SCC104, PCI30, and HEK293 cell lines. This observation suggests the expression of an endogenous gene in both male and female cells. A potential candidate for this band could be *RBMX*, the X chromosome homologue of *RBMY*. Given its location on the X chromosome, *RBMX* is expected to be present in both male and female cell lines and could therefore be cross-reacting with the *RBMY* primers. It has been demonstrated that *RBMXL2* suppresses a specific subset of acceptor splice sites, leading to the skipping of certain exons within the testis (Ehrmann *et al*., 2019). However, when looking at the designed primers, the forward primer shows 76% homology to *RBMX* (NM_002139.4). In comparison, the reverse primer is only 26% homologous, suggesting there would be no cross-reaction with the reverse primer. Furthermore, the predicted size of an *RBMX* product is 470bp which is larger than the contaminating band present in **Figure 3.10** and very similar to that of *RBMY*. The potential unintended product could be another variant of *RBMY*, *RBMY1D*, which is predicted to have a product length of 450bp with these designed primers. However, again this seems to be too large for the band visible. This extra band causes more uncertainty in the *ZFYS* oncogenic role.

To monitor the changes in *ZFYS* abundance in the presence of *RBMY*, a GFP-splicing reporter system was utilised. Following HEK293 transfection, the GFP fluorescence was determined to investigate if *RBMY* is splicing the second coding exon of *ZFY*. The results provide some support as an increase in GFP signal was identified following the addition of *RBMY*. However, the *ZFY*exon alone produced a greater background GFP signal than expected (~11%). A potential reason for this is that the “long” transcript which retains the *ZFY* exon has no coding potential and is thus targeted for nonsense-mediated decay. This means that we are preferentially detecting a small fraction of transcripts that do manage to exclude the *ZFY* exon even in the absence of *RBMY*. Although some degree of post-translational modifications is expected to occur in cells, the observed results could potentially indicate an issue with the reporter system itself. However, approximately 70% of the cells expressing *RBMY* also exhibited the spliced GFP version, lending some support to the hypothesis. Nevertheless, due to the higher background signal observed, caution is warranted in interpreting the results. The PCR analysis raised further concerns as a high abundance of the GFP-spliced band was observed in all samples. Although the GFP band signal was more intense in the sample co-expressing *ZFY*exon and *RBMY*

compared to the *ZFY*exon controls, and matched the GFP positive controls, the PCR results alone do not provide conclusive evidence to support the hypothesis.

While the aims of this chapter were largely met and initial evidence supports *RBMY* as a splicer of *ZFY*, further modifications to the reporter system are needed to confidently confirm *RBMY*'s role. Further to this, changes to the cell lines could also be useful. For example, the normal immortal keratinocytes (NIKS) cell line could be used as they express only the long form, alongside PCI30 which contains both the short and long form. Then using these cell lines, *RBMY* transfections could be performed to monitor changes in the ratio of the long and short form in the presence of *RBMY*, the potential splicer of the second coding exon.

Furthermore, a more extensive cancer cell line panel would further aid the analysis of *ZFYS* oncogenic activity. However, the presence of background bands does complicate analysis and subsequent qPCR across the samples is not an option. This means the development of a better assay for *ZFYS* is necessary before any high-throughput analysis of a potential role in cancer is possible.

4. Chapter 4: Overexpressing *ZFY* in HEK293 cells to Conduct Transcriptomic Analysis

4.1 Introduction

4.1.1 The Growing Transcriptomic Field

Transcriptomic technologies play a crucial role in the examination of an organism's transcriptome, which encompasses the total set of RNA transcripts including; messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA) and other non-coding RNAs (ncRNA) (Lowe *et al*., 2017);(Dong & Chen, 2013). This discipline first emerged in the early 1990s when a partial human transcriptome was published with 609 mRNA sequences from the brain. Since then, rapid technological advancements have driven an explosion in transcriptomics making it a global approach. The two main transcriptomic technologies are microarrays and RNA sequencing (RNA-Seq), with RNA-Seq now at the forefront (Lowe *et al*., 2017).

Transcriptomic analysis has facilitated the monitoring of gene expression changes across diverse organisms, enhancing our understanding of human diseases such as cancers (Lowe *et al*., 2017);(Supplitt *et al*., 2021). In the field of oncology, the development and application of transcriptome profiling have been instrumental in the identification of cancer biomarkers, gene signatures, abnormal expression patterns, and targets for anti-cancer therapies. Continued knowledge expansion will further the genomic background and pathogenesis of specific tumours and potentially further enhance personalised medicine (Supplitt *et al*., 2021).

RNA-Seq investigations have been driven by significant enhancements in next-generation sequencing (NGS) capabilities, with increasing knowledge of quantitative and qualitative aspects of the transcriptome of both eukaryotic and prokaryotic organisms (Ozsolak & Milos, 2011). NGS developments have eliminated challenges often seen in microarray and sanger-sequencing-based approaches (Kukurba & Montgomery, 2015). A typical RNA-Seq workflow **(Figure 4.1)** starts with isolating RNA from a biological material of choice (i.e., cells or tissues). The RNA is then reverse transcribed to produce cDNA. Subsequently, the cDNA is either fragmented or amplified through primed cDNA molecules, followed by the ligation of sequencing adaptors. The cDNA is used to prepare a sequencing library. Sequencing is then performed on an NGS platform and the current dominating NGS platform is Illumina. Following NGS, transcriptomic analysis is performed to analyse gene expression to understand the global transcriptional landscape. Conventional RNA-Seq data generates FASTQ-format files containing sequenced reads from the NGS platform

(Lowe *et al.*, 2017). These reads are then aligned to a reference genome and mapped reads can be assembled into transcripts. Gene quantification is then performed using software packages such as DESeq2. Using DESeq2 or other packages, the main objective of many gene expression experiments aims to identify transcripts that exhibit differential expression under different conditions (Kukurba & Montgomery, 2015);(Love *et al.*, 2015).

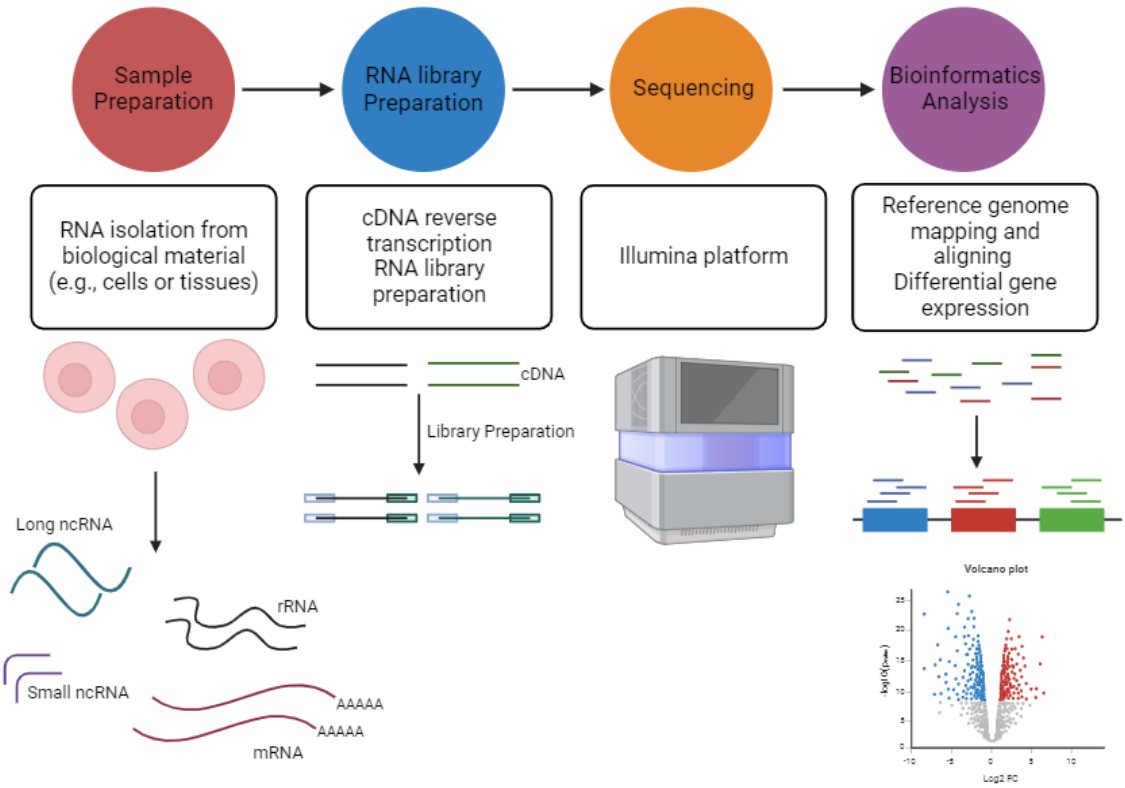


Figure 4.1:Overview of an RNA-Seq workflow. RNA is extracted and isolated from a biological material, for example, cells. The RNA is converted into cDNA by reverse transcription and sequence adaptors are ligated to the cDNA fragment ends. Using these an RNA-Seq library is made, and sequencing can occur. An Illumina platform is normally the NGS technology of choice. Following RNA-Seq, bioinformatics analysis can be performed following the alignment and mapping of transcripts to a reference genome.

4.1.2 Gene Overexpression System

A commonly employed method for investigating the biological pathways of a specific gene of interest is to induce overexpression of the gene in a system (Prelich, 2012). This system has been exploited since the development of yeast transformation techniques and has been evolving since then (Prelich, 2012);(Beggs, 1978). Within this chapter, overexpression of *ZFYS* and *ZFYL* is conducted in a mammalian cell line to identify potential pathway interactions and hint at a more specific function of *ZFY*.

Following overexpression, RNA was isolated and sent for RNA-Seq to subsequently investigate the differential gene expression under the influence of *ZFYS* or *ZFYL*.

For the past 25 years, the HEK cell line has served as a widely employed expression tool (Thomas & Smart, 2005). Derived from the transformation of HEK cells through exposure to sheared fragments of human adenovirus type 5 (Ad5) DNA, this cell line is commonly referred to as the HEK293 cell line (Thomas & Smart, 2005). However, following transcriptomic profiling it is been suggested that HEK293 cells are more likely to be derived from embryonic adrenal gland cells which were potentially inadvertently collected alongside the embryonic kidney cells and cultured (Y.-C. Lin *et al*., 2014). HEK293 cells are extensively utilised due to their rapid growth rate and relative ease of transfection making them a great tool for biological functional analysis (Pulix *et al*., 2021). These reasons justify the selection of this cell line for the experiment, including the additional factor that the cell line is of female origin. *ZFY* is encoded on the Y chromosome and is consequently absent in the female genome. Utilising a female cell line like HEK293 ensures the absence of endogenous *ZFY* expression, with only transfected *ZFY* being present. This means that any changes in gene expression will be a consequence of the transfected *ZFY*.

4.1.3 The Current *ZFY* Functional Knowledge

Following the discovery that *ZFY* is not the sex-determining gene, interest in its function and significance has diminished. Consequently, the complete understanding of *ZFY*'s roles remains elusive, but studies suggest crucial involvement in spermatogenesis. *ZFY* is highly unusual in that it is a core component of the Y chromosomal gene content across all mammals and is therefore one of the very few genes that evade the general processes of gene inactivation and loss on the Y (Waters & Ruiz-Herrera, 2020). The survival of the Y chromosome relies on its functions in spermatogenesis and sex determination, with the remaining Y chromosome genes proving beneficial for male-specific development (Waters & Ruiz-Herrera, 2020). This indicates that *ZFY* must be important for male differentiation as it continues to persist throughout evolution.

Initial studies have suggested that *ZFY* plays a role in spermatogenesis, with roles linked to MSCI, meiosis progression and general spermatogenesis progress (Holmlund *et al*., 2023). Mouse studies have shown that *ZFY*2 functions in sperm development (Vernet, Mahadevaiah, Decarpentrie, *et al*., 2016), and both *ZFY*1 and *ZFY*2 have been shown to be vital for efficient MSCI in spermatocytes (Vernet, Mahadevaiah, Decarpentrie, *et al*., 2016). In *ZFY*-deficient mice testes 17.5 days post-partum, a combination of MSCI leakage in early pachytene cells and pachytene

cell death was noted (Vernet, Mahadevaiah, Decarpentrie, *et al*., 2016). In mice, *ZFY* genes have demonstrated executor functions at MSCI checkpoints, possibly influenced by their genomic positioning on the Y chromosome. Additionally, their location implies a plausible negative feedback loop wherein *ZFY* regulates its expression during the pachytene transition. This proposes that sustained *ZFY* expression advances MSCI to its conclusion, leading to the termination of *ZFY* transcription and allowing prophase to progress. However, if prolonged *ZFY* activity prevents MSCI completion, it can lead to MSCI failure and subsequent cell apoptosis (Vernet, Mahadevaiah, Decarpentrie, *et al*., 2016).

Nakasuji *et al* demonstrated that mice with double knockout of *ZFY*1 and *ZFY*2 exhibited significant sperm abnormalities, encompassing issues with morphology, motility, capacitation, acrosome reaction, oocyte activation, and chromosomal aberrations (Nakasuji *et al*., 2017). This suggests that both *ZFY*1 and *ZFY*2 play a role in many spermatogenesis processes. Sperm head and tail formation has also been linked to *ZFY*1 and *ZFY*2. With previous reports showing that in *ZFY*1-KO, *ZFY*2-KO, and *ZFY*1/2-DKO sperm samples, the occurrence of morphologically abnormal sperm was noted at rates of 4.5%, 82.9%, and 100%, respectively. This further suggests that *ZFY*2 dominates sperm development, but a level of both *ZFY*1 and *ZFY*2 is needed for successful sperm development. This further shows that *ZFY*1 and *ZFY*2 are indispensable for spermiogenesis, but understanding the mechanisms that underlie this is required to shed light on fertilisation and embryonic development failure (Nakasuji *et al*., 2017).

A constraint in numerous present *ZFY* investigations is that studies of *ZFY* function rely on mouse models, while studies of its transactivation ability have only been carried out in yeast. Given that mouse *ZFY* exhibits some homology to human *ZFY* similar functions are expected, but variations do occur particularly since mice have evolved to possess two Y-linked *ZFY* genes: *ZFY*1 and *ZFY*2. This chapter aims to understand the importance of *ZFY* functionally in humans. By overexpressing *ZFYS* and *ZFYL* individually into HEK293 cells the aim is to gain an understanding of potential *ZFY* pathways and functions using RNA-Seq and downstream transcriptomic analysis.

This work will also confirm the relative transactivation power of *ZFYS* and *ZFYL* in a mammalian cell system and see if this recapitulates the yeast finding that *ZFYS* is a far weaker transcription factor. In particular, given the fact that *ZFYS* and *ZFYL* share a common DNA binding domain and thus likely compete for access to their target gene promoters, it is important to understand how their effects differ from each other. From the yeast data two potential hypotheses were identified: (a) that *ZFYS* may be

a transcriptional inhibitor rather than an activator; (b) that *ZFYS* is a weaker transcriptional activator than *ZFYL*. In either case, *ZFYS* would act as a net antagonist to *ZFYL* activity, either directly by counter-regulation in hypothesis (a), or indirectly via competitive inhibition in hypothesis (b).

4.2 Materials and Methods

4.2.1 DNA Constructs and Transformations

To examine the effects of *ZFY* overexpression in mammalian cells, DNA constructs were produced containing the gene of interest in a pcDNA3.1(+) vector backbone with either an N-terminal HA or an N-terminal GFP **(Figure 4.2)**. Both the full-length form and the alternatively short-spliced form of *ZFY* were cloned into the vector with either HA- or GFP-fusion proteins using the XhoI/XbaI sites of the vector **(Table 4.1)** (see *supplementary* data sequence A and Sequence B). These DNA constructs were commercially synthesised by GenScript and provided as lyophilised plasmid DNA.

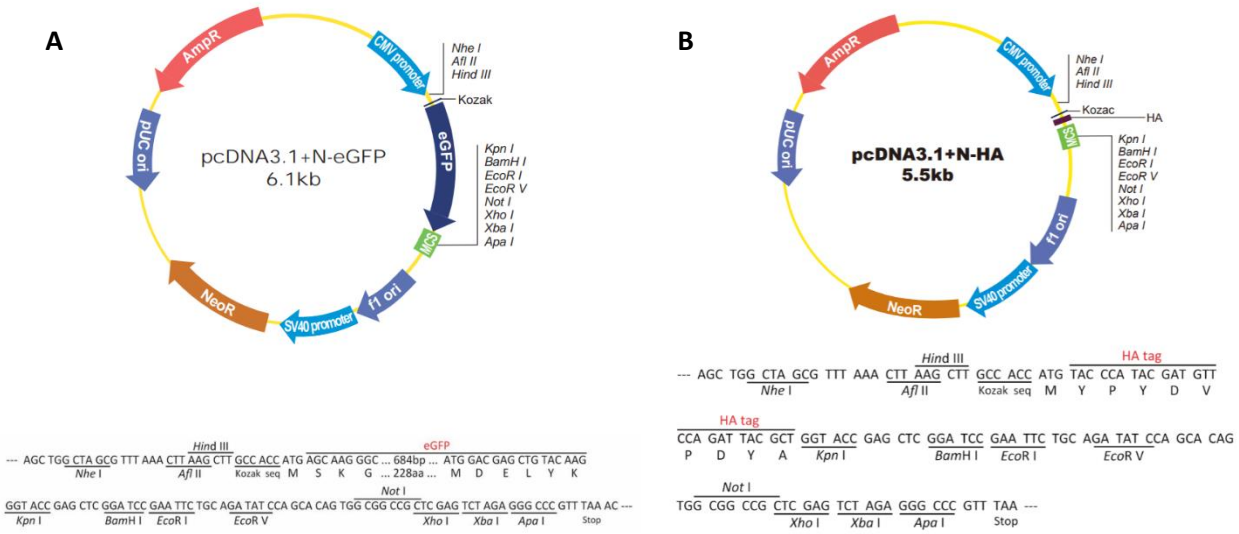


Figure 4.2: pcDNA3.1(+) Plasmid Map from SnapGene. A: pcDNA3.1+N-eGFP plasmid**. B:** pcDNA3.1+N-HA plasmid. pcDNA3.1(+) is a common plasmid type used in mammalian expression, with a CMV promoter and high expression levels. pcDNA3.1 is ampicillin-resistant.

Table 4.1: *ZFY* plasmid constructs produced by GenScript. These DNA constructs were specifically designed for use in our overexpression protocol using the pcDNA3.1(+) plasmid backbone in **Figure 4.2**.

Vector backbone	Insert	Tag	Antibiotic Resistance
pcDNA3.1(+)	h*ZFY*-long (full length)	N-terminal HA-tag	Ampicillin
pcDNA3.1(+)	h*ZFY*-short (isoform)	N-terminal HA-tag	Ampicillin
pcDNA3.1(+)	h*ZFY*-long (full length)	N-terminal eGFP-tag	Ampicillin
pcDNA3.1(+)	h*ZFY*-short (isoform)	N-terminal eGFP-tag	Ampicillin

On arrival of the constructs, the tubes were centrifuged at 6,000 x g for 1 minute at 4°C. 20µL of sterilised water was added to dissolve the DNA with the aid of a vortex.

The plasmids were then transformed into NEB 5-alpha competent *E. coli* cells (NEB, #C2988J) following the manufacturer's protocol, and transformants were selected on ampicillin (Melford, A0104) containing LB agar plates (final concentration 100µg/mL), see 3.2.4.1 to 3.2.4.2 in Chapter 3.

4.2.2 Plasmid DNA Miniprep

Following the transformation of the DNA constructs into *E. coli*, colonies were inoculated in 5mL LB overnight cultures supplemented with a final concentration of 100µg/mL ampicillin. These were incubated overnight in a shaking incubator at 37°C. The following morning, minipreps using the QIAprep Spin Miniprep Kit (Qiagen, Cat. No./ID: 27104) were performed following the provided kit method using the 5mL overnight cultures. A NanoDrop ND-1000 Spectrophotometer was used for DNA quantification. The DNA concentration, 260/280 and 260/230 ratios were noted down.

4.2.3 Mammalian Cell Line

HEK293 cells were used as described in Chapter 3 section 3.2.5.1.

4.2.3.1 Lipofectamine 3000 Transfection

For this experiment, *ZFYS* and *ZFYL* form DNA constructs were transfected into the cells with either an N-terminal GFP tag or an N-terminal HA tag. Two forms of controls were used: a no-transfection control and an empty GFP backbone control pEGFP-N1 (Addgene, 6085-1), as transfection itself may have an impact on the cells. The samples prepared are stated in **Table 4.2**.

Table 4.2: The experimental samples prepared for use in the overexpression experiment. The optimal seeding density was determined to achieve an appropriate confluence at the time of harvest for RNA extraction. It was observed that transfection slightly inhibited growth, requiring a higher seeding density for transfected cells to reach the target confluence. *Control group.

Sample Name	DNA construct transfected	Seeding density (cells/well)
Control 1 *	-	300,000
Control 2 *	-	300,000
Control 3 *	-	300,000
Empty pEGFP-N1 1 *	Empty pEGFP-N1	450,000
Empty pEGFP-N1 2 *	Empty pEGFP-N1	450,000
Empty pEGFP-N1 3 *	Empty pEGFP-N1	450,000
ZFY-Short GFP 1	*ZFY*-Short N-terminal GFP	450,000
ZFY-Short GFP 2	*ZFY*-Short N-terminal GFP	450,000
ZFY-Short GFP 3	*ZFY*-Short N-terminal GFP	450,000
ZFY-Long GFP 1	*ZFY*-Long N-terminal GFP	450,000
ZFY-Long GFP 2	*ZFY*-Long N-terminal GFP	450,000
ZFY-Long GFP 3	*ZFY*-Long N-terminal GFP	450,000

ZFY-Short HA 1	*ZFY*-Short N-terminal HA	450,000
ZFY-Short HA 2	*ZFY*-Short N-terminal HA	450,000
ZFY-Short HA 3	*ZFY*-Short N-terminal HA	450,000
ZFY-Long HA 1	*ZFY*-Long N-terminal HA	450,000
ZFY-Long HA 2	*ZFY*-Long N-terminal HA	450,000
ZFY-Long HA 3	*ZFY*-Long N-terminal HA	450,000

The lipofectamine protocol utilised is outlined in 3.2.5.2 of Chapter 3.

4.2.3.2 Cell Microscopy

The transfection efficiency was assessed to confirm that a satisfactory number of cells (>50%) had incorporated the DNA constructs. Cells were grown on glass coverslips and pre-treated with 1mg/mL Poly-D-Lysine (Sigma-Aldrich, A-003-E) to aid cell adhesion. GFP-tagged DNA constructs were transfected into the cells for microscopy due to the presence of a fluorescent tag. Following transfection, fixation of the cells was performed as described in section 3.2.5.3 of Chapter 3. This was followed by adding a drop of the Antifade Mounting Medium with DAPI (Vector Labs, H-1200) to a Superfrost microscope slide (Fisher Scientific) and gently placing the glass coverslip cell side down onto the Superfrost slide.

These slides were then ready for visualisation using the Olympus BX61 Fluorescence Microscope using the software SmartCapture3. Cells were visualised using the DAPI (385nm) and FITC (475nm) channels, allowing for nuclei detection and subsequent uptake of the GFP-tagged *ZFY* isoforms. Microscope slides were subsequently stored at 4°C in light-protecting boxes.

4.2.3.3 Flow Cytometry

Transfection efficiency was also assessed using flow cytometry. At the 48-hour end time point of transfection, the cells were washed and collected in 1x PBS (Oxoid, BR0014G). Using the BD Accuri C6 Plus, 10,000 cells were counted per sample and the FL1 channel (laser 488nm) was selected to detect the GFP signal. For this particular experiment set, the no transfection control HEK293 cells were used to gate for no GFP signal, whilst the empty-GFP vector was used as a GFP positive signal control. Using these samples, gating was set on the selected cell population to check the transfection efficiency of the constructs. This method gives us a more reliable quantification of transfection efficiency. An example gating plot for the negative control samples is shown in **(Figure 4.3)** – see Results section for further plots.

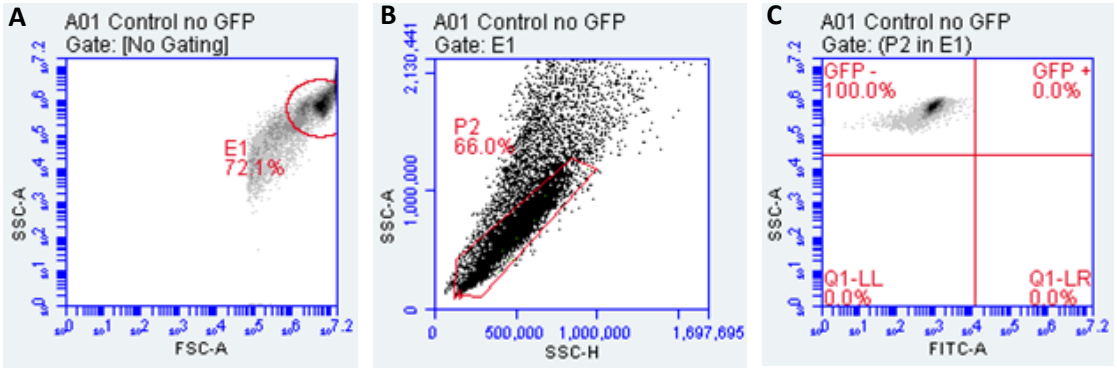


Figure 4.3: Flow cytometry graphs from the negative control sample (see figure 4.7 for results). This sample set was used to set the gating for cells with no GFP signal. **A:** A graph displaying the forward scatter area (FSC-A) on the X-axis and the side scatter area (SSC-A) of the HEK293 cells on the Y-axis. The cell population of interest is highlighted using this plot and is highlighted by the circle called E1. This excludes the debris outside the circle. **B:** This plot is gated based on the E1 population identified in plot A. The side-scatter height (SSC-H) is plotted on the X-axis and the SSC-A is plotted on the Y-axis and allows for the identification and exclusion of signals from cell doublets. P2 is selected as the singlet cells that are used for further gating. **C:** This plot shows the GFP signal in the P2 cell population. FITC-A is plotted on the X-axis and SSC-A is plotted on the Y-axis. The focus in this plot is the top two quadrants, which are labelled GFP- and GFP+. These quadrant positions were gated based on the control populations.

4.2.4 Western Blotting

Following transfection, cells were collected in 20µL of lysis buffer (100Mm Tris-HCL, 600mM NaCl, 4% Sodium deoxycholate, 4% SDS & 4mM EDTA) and incubated on ice for 15-30 minutes. Centrifugation followed at 13,000 x g for 30 minutes at 4°C. The insoluble pellet was then discarded whilst the supernatant was kept. Following this, protein levels were determined using the Bicinchoninic Acid (BCA) protein assay kit (Thermo Scientific, 23235) along with standards containing Bovine Serum Albumin (BSA) at known concentrations (200µg/mL, 40 µg/mL, 20 µg/mL, 10 µg/mL, 5 µg/mL, 2.5 µg/mL, 1 µg/mL, 0.5 µg/mL, 0 µg/mL). The working reagent was prepared by mixing 25 parts micro-BCA reagent A (MA), 24 parts micro-BCA reagent B (MB), and 1-part micro-BCA reagent C (MC). The protein samples and standards were combined 1:1 with the working reagent, mixed thoroughly, and incubated at 60°C for 1 hour. 2µL of each sample was then pipetted onto a µDrop plate. The absorbance was measured at 562nm using a multi-label plate reader. Readings were then blank corrected by subtracting the absorbance readings of the blanks. A standard curve was constructed from the blank-corrected absorbance values of the standards and was then used to estimate the concentration of the unknown samples.

Following the determination of the total protein concentration, 20µg of protein lysate was mixed with Laemmli sample buffer (10% glycerin, 60mM Tris-HCl, 2% SDS, 0.1M DTT, 0.01% bromophenol blue, pH 6.8) and boiled at 95°C for 5 minutes, denaturing the proteins. These samples were then loaded onto a 4-12% Bis-Tris Mini Protein Gel (Invitrogen, NP0321BOX) at 150V for ~1 hour in MOPS SDS running buffer (Invitrogen, NP0001).

Before transfer, the Polyvinylidene Fluoride (PVDF) western blotting membrane (Roche, 03010040001) was activated in 100% methanol and then soaked in Towbin transfer buffer (25mM Tris, 192mM Glycine, 20% (w/v) methanol, pH 8.3) for 10 minutes, alongside the gel and filter paper (BIO-RAD, #1703965). This process allows the equilibration of the components and the removal of any salts and detergents from the gel. Following soaking in the transfer buffer, the proteins were transferred from the gel to the membrane using a semi-dry blotter (Trans-Blot SD Semi-Dry Electrophoretic Transfer Cell, Bio-Rad, 170-3940). The components were layered as follows; filter paper, PVDF membrane, gel, and filter paper. Between the stacking of each layer, they were carefully rolled to ensure no bubbles were present which would disrupt transfer. The transfer occurred at 15V for 20 minutes.

Following the transfer, non-specific binding was prevented by blocking the membrane in 5% BSA (FisherScientific, #BP1600-100) in 1x Tris-buffered saline with tween (TBST) (20mM Tris Base, 150mM NaCl, 0.1% Tween, pH 7.6) for 1 hour at RT, before primary antibody incubation. After blocking, the membrane was incubated in the primary antibody (diluted in blocking buffer, **Table 4.3**) for 1 hour at RT or overnight at 4°C. Unbound or weakly bound primary antibody was removed by washing the membrane 3 x 10 mins in 1x TBST. Following washing, the membrane was incubated in a secondary antibody (diluted in blocking buffer, **Table 4.3**) for 1 hour at RT or overnight at 4°C. Like before, the membrane was washed following incubation 3 x 10 mins in 1x TBST. The membranes were then incubated in Electrochemiluminescence (ECL) Western Blotting Substrate (Thermo Scientific, 10590624) for 5 minutes, before imaging the membrane on the Syngene G:BOX Chemi XX6 using GeneSys image capture software, for chemiluminescent detection of proteins.

Table 4.3: Antibodies used for Western blotting. Associated concentrations and dilutions used for each antibody are included.

Antibody Name	Company	Primary or Secondary	Clonality	Species	Target	Conc.	Dilution
Anti-GFP (B-2)	Santa Cruz Biotechnology [sc-9996]	Primary	Monoclonal	Mouse	GFP-tag	200ug/mL	1:10,000
Anti-HA (F-7)	Santa Cruz Biotechnology	Primary	Monoclonal	Mouse	HA-tag	200ug/mL	1:1,000

	[sc-7392]						
Anti-Beta actin (C-4)	Santa Cruz Biotechnology [sc-47778]	Primary	Monoclonal	Mouse	Beta-Actin	200ug/mL	1:100,000
m-IgG Fc BP-HRP	Santa Cruz Biotechnology [sc-525409]	Secondary	Polyclonal	Mouse	IgG BP HRP Conjugate	100ug/mL	1:10,000

4.2.5 RNA Extraction

Following the collection of the mammalian cell pellets, RNA extraction was performed using the Qiagen Rneasy mini kit (QIAGEN, 74104), using the protocol provided (see 3.2.3 in Chapter 3). After obtaining the RNA concentration, samples were sent to Novogene for sequencing.

4.2.6 RNA-Seq Data Collection and Preparation

RNA sequencing was carried out by Novogene using 150 bp paired-end reads on an Illumina sequencing platform. The following packages were downloaded before analysis; Python (3.6.10), Conda (4.10.3), FastQC (0.11.9), R (3.2.2), HISAT2 (2.2.1), and Samtools (1.14) (Table 4.4). The exact code used for the RNA-Seq analysis in this chapter can be found in Chapter 4 – Transcriptomics at https://github.com/IzzyGarcia/Thesis-code.

Table 4.4: Required Programmes and respective reference and web resources.

Tool and Resources	Reference
Python	https://www.python.org/
Conda	https://docs.conda.io/projects/conda/en/stable/ https://anaconda.org/
FastQC	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
HISAT2	Kim, D., Paggi, J.M., Park, C. *et al*. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37, 907–915 (2019). https://doi.org/10.1038/s41587-019-0201-4
R	R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.
Rstudio	RStudio Team (2019). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL: https://www.rstudio.com/
Samtools	http://www.htslib.org/
FeatureCounts	https://www.rdocumentation.org/packages/Rsubread/versions/1.22.2/topics/featureCounts

The raw high-throughput sequencing data produced by Novogene was provided as FASTQ files containing the sequenced reads. These FASTQ files were downloaded and transferred to the University's high-performance computing cluster using wget and verified using md5sum checksums.

Quality control (QC) analysis of the raw sequencing reads was performed using FastQC software. FastQC calculates a range of quality metrics to assess the quality of raw reads, including evaluations of sequence quality **(Figure 4.4a)**, GC content **(Figure 4.4b)**, and library complexity. A HTML-formatted report is produced for each FASTQ file.

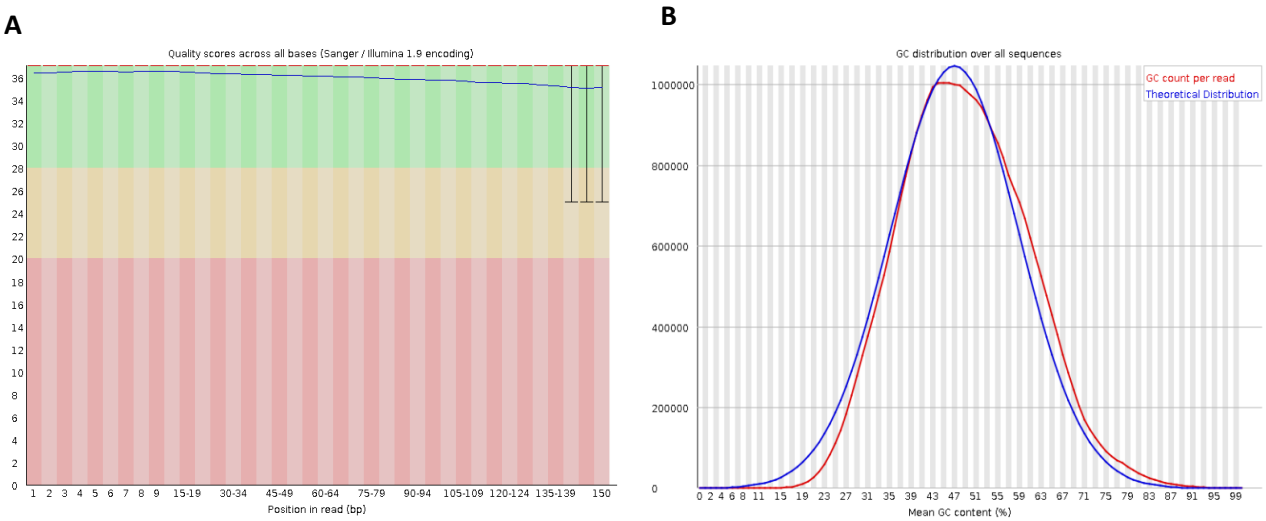


Figure 4.4. An example of FastQC metrics on raw read sequence files. A: Bar chart depicting the Phred quality score distribution across each nucleotide position in the sequenced reads. The Phred quality scores are plotted on the Y-axis and the nucleotide position in the sequenced reads are shown on the X-axis. Green: Good Quality, Orange: reasonable quality, Red: poor quality. **B:** Plot displaying the GC content distribution across all sequenced reads. It is important to note whether the central peak corresponds to the expected GC. The y-axis represents the number of sequences, while the x-axis shows the mean GC content percentage. The blue line represents the theoretical GC content expected and the red line represents the inputted data GC content.

Following the QC checks, the raw reads were aligned to the reference genome using HISAT2. The human reference genome sequence (release 105) and associated chromosome GTF annotation files were obtained from the Ensembl database for use as the reference in the alignment using wget. HISAT2 indexing was performed using the hisat2-build tool (-p 10 added to increase the number of threads) and the toplevel human reference files. The raw reads were then aligned to the indexed reference genome using HISAT2. These sequences are paired-end reads therefore, it is crucial that the input files were not interchanged and were correctly paired. The SAM alignment files produced from HISAT2 were unsorted, Samtools was used to sort the SAM files and convert them to BAM format for subsequent feature counting analysis. Feature counts (2.0.1) summarised the paired-end reads and counted the fragments to produce a count matrix containing all of the read data. The resulting output from feature counts is a tab-delimited text file containing the gene identifier and the corresponding count of reads mapped to that gene for every sample.

4.2.7 DESeq2

The differential gene expression analysis was conducted on the assembled count matrix using the DESeq2 software, and this analysis was carried out in R Studio (4.1.2). Once in R studio the following libraries were installed; DESeq2 (1.34.0), ggplot2 (3.4.0), RColor Brewer (1.1-3), calibrate (6.7-1), enhanced volcano (1.14.0), tidyverse (1.3.2), AnnotationDbi (1.58.0) and org.Hs.eg.db (3.18) **(Table 4.5)**.

Table 4.5: RStudio Packages and respective reference and web resource.

Package Name	Reference
DESeq2	Love, M.I., Huber, W., Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550. URL: https://bioconductor.org/packages/DESeq2
ggplot2	H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. URL: https://ggplot2.tidyverse.org
RColorBrewer	https://rdocumentation.org/packages/RColorBrewer/versions/1.1-3
calibrate	Graffelman, J. and van Eeuwijk, F. (2005). Calibration of multivariate scatter plots for exploratory analysis of relations within and between sets of variables in genomic research. Biometrical Journal, 47(6), 863-879. URL: https://doi.org/10.1002/bimj.200510177
EnhancedVolcano	Kevin Blighe, Sharmila Rana and Myles Lewis (2021). EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. R package version 1.12.0. URL: https://github.com/kevinblighe/EnhancedVolcano
tidyverse	Wickham *et al*., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. URL: https://doi.org/10.21105/joss.01686
AnnotationDbi	Hervé Pagès, Marc Carlson, Seth Falcon and Nianhua Li (2021). AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. R package version 1.56.2. URL: https://bioconductor.org/packages/AnnotationDbi
Org.Hs.eg.db	Marc Carlson (2021). org.Hs.eg.db: Genome wide annotation for Human. R package version 3.14.0.
dplyr	Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.8. URL: https://CRAN.R-project.org/package=dplyr
hmisc	https://hbiostat.org/r/hmisc/

DESeq2 is a widely used algorithm for analysing RNA-Seq data. It assesses the relationship between variance and mean in high throughput count data and identifies differential expression using a negative binomial distribution. The count matrix was imported, and conditions were assigned to produce a coldata metadata, a data frame containing the metadata about each sample. The metadata was refined to include only entries with adjusted p-values (padj) < 0.05, allowing for the extraction of

contrasts between conditions. Subsequently, the data frame was annotated using the org.Hs.eg.db package to incorporate standard gene symbols and their corresponding chromosome locations.

Using DESeq2, QC checks on the samples were performed. Both Principal Component Analysis (PCA) and correlation heatmaps were generated to visualise the sample clustering. The heatmap was generated using the heatmap.2 DESeq2 tool. Verifying that the samples cluster according to their respective conditions is crucial.

After examining sample clustering, specific contrasts were defined, and MA and volcano plots were generated to visualise differences in gene expression profiles between groups. An MA plot compares the gene expression between two genotypes with the log fold-change (LFC) plotted on the Y-axis and the overall expression on the X-axis. The MA plot was produced using the plotMA function from DESeq2, filtered to highlight genes with padj <0.05. A volcano plot also presents and identifies meaningful changes within the dataset. The volcano plot was generated using the enhancedVolcano library, applying filters to highlight genes with padj < 0.05 and absolute log2 fold changes (L2FC) exceeding 1.

The gene lists from DESeq2 were merged with a UniProt annotated CSV file. Duplicate entries and non-coding proteins were then removed, as duplications can occur due to annotation errors on the sex chromosomes. The gene lists were filtered to retain only those genes exhibiting absolute L2FC greater than 1 or less than -1, representing significant differentially expressed genes. Filtering by baseline expression levels was attempted using DESeq2, which normalises count values across samples. However, no clear correlation was identified that would justify setting a rational cutoff for baseline expression filtering. The tidyverse inner_join and anti_join functions were utilised to identify commonalities and differences between the *ZFY* short and long isoform comparisons. This dataset was now ready for further downstream analysis.

4.2.8 *ZFX* Differential Expression Data

ZFX is the X chromosome homologue of the gene of interest *ZFY* and has been implicated in the initiation or progression of a variety of different human cancer types. However, like *ZFY* the underlying mechanism by which *ZFX* influences transcriptional regulation is yet to be determined.

A previous study by Weiya Ni *et al* focused on characterising the transcriptional influence of *ZFX* in HEK293T cells using a CRISPR knockout and add-back approach, with RNA-Seq data available in the supplementary materials (GEO: GSE145160) (Ni *et al*., 2020). This *ZFX* RNA-Seq data was obtained to identify any

commonalities with the *ZFY* dataset, despite being performed in HEK293T rather than HEK293 cells. The HEK293T line is derived from HEK293 cells via transfection with an SV40 origin plasmid and was therefore considered sufficiently comparable for this analysis.

The data from Weiya Ni *et al* was sorted based on p-value <0.05 but not for LFC, this was altered to match the filtering in this thesis' methods (Ni *et al*., 2020). The tidyverse functions anti_join and inner_join were utilised to compare the differentially expressed genes from the *ZFX* dataset of Weiya Ni *et al* to those identified in this thesis, to assess similarities and differences between the two datasets.

4.2.9 Gene Ontology

After differential expression analysis with DESeq2, pathway enrichment was carried out using the Reactome database. Reactome is a freely accessible online resource containing curated biological pathway data for humans that enables visualisation and analysis of pathway interactions. The filtered differential gene lists were inputted into the Reactome analysis tool. Options were selected such as whether to include interactors to expand the analysis background. Reactome then performed enrichment analysis to identify pathways overrepresented among the input genes compared to the genome background. The pathway enrichment results for each gene set were compiled in an Excel document. Filters were applied to highlight pathways with p-values <0.05 regardless of false discovery rate (FDR), as well as more stringently filtered pathways with both p-values <0.05 and FDR <0.05. It is important to note that pathways failing to meet the significance threshold of both p-value <0.05 and FDR <0.05 cannot be justifiably considered enriched.

4.2.10 Primer Design and Selection

Following the downstream analysis of the RNA-Seq data, genes were selected to produce a bioinformatics validation panel. Primer design is a critical step when setting up PCRs for gene expression analysis. PCR primers that anneal poorly or that anneal to more than one sequence during amplification can significantly impact the quality and reliability of the results. The NCBI tool Primer-BLAST is widely used for PCR primer design and is what we used to design the primers.

Using the accession number of the desired gene and selecting the parameters in **Table 4.6**, forward and reverse primers were outputted per gene by the program and subsequently selected for testing. These primers were then ordered and synthesised by Integrated DNA Technologies (IDT) using their standard production type (25nmole purification and desalted) **(Table 4.7)**.

Table 4.6: NCBI tool Primer-blast parameters. Standard parameters were mainly selected as necessary on the NCBI tool.

Product/Amplicon Size	100-150 bp long (for efficient amplification)
Number of primers to return	10 primers
Melting temperature	Minimum of 60°C and a maximum of 63°C; the ideal primer melting temperature is 60°C (with a maximum difference of 3°C in the melting temperatures, Tm, of the two primers).
Exon/intron selection	Primer must span an exon-exon junction
GC content	40-60% to ensure maximum product stability
Advanced settings	Repeat filter - None (only for SNORD3 - because snoRNAs have multiple copies in the genome and get flagged as repeats)

Table 4.7: Primers designed for RNAseq downstream analysis validation. Primers were designed using the NCBI Primer-BLAST tool. https://www.ncbi.nlm.nih.gov/tools/primer-blast/.

Gene name	Accession Number	Primer Sequence	Tm (°C)	Product Length (bp)
WNT7a	NM_004625.4	For: CTCCGGATCGGTGGCTTC Rev: AGGCCCATTTGTGAGCCTTC	59.9 60.6	149
TMPRSS2	NM_005656.4	For: GGGGATACAAGCTGGGGTTC Rev: GATTAGCCGTCTGCCCTCAT	60.1 59.6	113
FGF3	NM_005247.4	For: GGAGAACAGCGCCTACAGTATT Rev: TGCTCCGAAGCATAGAGTCG	60.2 59.6	126
IFIT2	NM_001547.5	For: ACTGCAACCATGAGTGAGAACA Rev: CGATTCTGAAACTCAGTCCGGT	60.2 60.4	149
RBMXL2	NM_014469.5	For: GTTTGGCCAACCAACCACAA Rev: AAGCCATTACGGTCCCCAAG	59.8 60.0	141
ZFX	NM_003410.4	For: TGTTCCCTGAGCTGTGCTTT Rev: TCATCAGTCACAGCTCCTGTC	59.8 59.5	150
FRMPD2	NM_001018071.4	For: CCGCCACATCAGCCCC Rev: GAACCCGACAAGCTTCCAGA	60.2 60.0	118
SNORA7B	NR_002992.2	For: TCCTGGGATCGCATCTGGA Rev: GGAATGGAATGGGTGCCTCT	60.1 59.7	90
SNORD3D (Non-coding, ENSG00000262202)	NR_006882.1	For: TGAACGTGTAGAGCACCGAA Rev: ATCAATGGCTGACGGCAGTT	59.3 60.3	108
CPN2	NM_001080513.4	For: TCGGCCCTCACGAAGATGC Rev: TGGACGAAGCAGTCACAACC	62.4 60.6	107
		For: GCCTTCATGTAGAGGGGACG	57.4	140

ENSG00000289202 (Non-Coding Protein LncRNA)	ENSG00000289202	Rev: GCCCAGCCTTTCAGATCAGT	57.5	

Housekeeping genes are necessary for qPCR normalisation to provide accurate gene expression analysis and primers for these were selected **(Table 4.8)**. The inclusion of these housekeeping genes also known as reference genes serves to correct for sample-to-sample variation therefore improving the reliability of an experiment (Adeola, 2018).

Table 4.8: Housekeeping genes used for qPCR normalisation. Accession number and primer nucleotide sequence included.

Gene name	Accession Number	Primer Sequence	Tm (°C)	Product Length (bp)
GAPDH	NM_001289746.2	For: GTCATCCATGACAACTTTGGTA Rev: GGATGATGTTCTGGAGAGC	52.8 52.7	136
TBP	NM_003194.5	For: CCCATGACTCCCATGACC Rev: TTTACAACCAAGATTCACTGTGG	55.2 53.4	108
ACTB	NM_001101.5	For: CGCCGCCAGCTCACC Rev: CACGATGGAGGGGAAGACG	60.6 57.7	120

4.2.11 Primer Checking and cDNA Synthesis

Following the arrival of the designed primers, 100uM stocks were produced using the corresponding volume of PCR-grade water. From this stock, a 10uM working stock was prepared and then stored at -20°C.

cDNA synthesis was carried out using the remaining RNA samples that were sent for the original sequencing and the protocol followed for cDNA synthesis can be found in 3.2.3 of Chapter 3. A second set of RNA samples from an independent transfection were also subsequently used for further validation.

After cDNA was synthesised from all 18 samples, the 50ng/μL cDNA reactions were diluted to produce a 5ng/μL working solution for use in subsequent experiments. Successful cDNA synthesis was validated by PCR amplification of the GAPDH housekeeping gene using GoTaq G2 Flexi DNA polymerase (Promega, M7801) in 10μL reactions. The component concentrations used are listed in **Table 3.8** in chapter 3.

A standardised thermocycler setup shown in **Table 3.7** (chapter 3) was used for the GoTaq G2 Flexi polymerase with many of the conditions remaining the same across

experiments. After PCR, the amplified products were analysed by agarose gel electrophoresis.

4.2.12 pGEM-T Easy Vector Ligation and Transformation

To clone PCR products the pGEM-T easy vector system (Promega, #A1360) was utilised due to the convenient reduced incubation time as a result of the rapid ligation buffer provided. The ligation was set up as demonstrated in **Table 4.9**.

Table 4.9: pGEM-T easy vector ligation reaction. Ligation reactions were set up following the manufacturer's protocol using the provided positive control insert DNA and a background control without an insert. The positive control contains an insert that will ligate into the pGEM-T vector, while the background control lacks an insert and will not ligate.

Reaction Component	Standard Reaction	Positive Control	Background Control
2x Rapid ligation buffer, T4 DNA ligase	5µl	5µl	5µl
pGEM-T Easy Vector (50ng)	1µl	1µl	1µl
PCR product*	Xµl	-	-
Control Insert DNA	-	2µl	-
T4 DNA Ligase (3 Weiss unit/µl)	1µl	1µl	1µl
Nuclease-free water to a final volume of:	10µl	10µl	10µl

*Molar ratio of PCR product optimised based on size of PCR product

The reactions were mixed by pipetting, and subsequently incubated for 1 hour at RT. Then following ligation, 2µL of the ligation reaction was mixed with 50µL of the NEB 5-alpha competent *E. coli* cells (NEB, #C2988J), and the protocol was completed as mentioned previously using ampicillin as the selection antibiotic. Following this, a dozen colonies were selected for a colony PCR, performed as mentioned above using the GoTaq G2 flexi DNA polymerase. PCR-amplified products were analysed on a 2% w/v agarose gel. From the remaining PCR-amplified product, LB/ampicillin overnight cultures were sent up for DNA extraction to be sent off for sequencing to identify the bands.

4.2.13 Quantitative Reverse-Transcription Polymerase Chain Reaction

Quantitative Reverse-Transcription PCR (RT-qPCR) was performed using the PowerUp SYBR green master mix (Applied Biosystems, ThermoFisher, A25741). Reactions were set up in 10µL volumes for each primer pair and each cDNA sample, as outlined in **Table 4.10**. 10µL/well reactions were prepared in optical 96-well reaction plates (Applied Biosystems, ThermoFisher, N8010560) and the plates were

sealed with adhesive plate seals (Thermo Scientific, AB1170). The components were thoroughly mixed and briefly centrifuged to spin down the contents and remove air bubbles. The PCR instrumentation used was the QuantStudio 3 system, with data analysis performed by the comparative Ct (ΔΔCt) method. The QuantStudio 3 thermocycler uses preset, optimised conditions for SYBR green reactions, so the default cycling program was utilised **(Table 4.11)**. To check for primer contamination, no-template water controls were run for each primer pair on every plate. Expression was normalised to the ACTB endogenous control included on each plate. Reactions were performed in duplicates to check for consistency.

Table 4.10: PowerUp SYBR Green Master Mix. This protocol is the pre-formulated and optimised setup for the 2X master mix designed specifically to amplify targets for accurate gene expression analysis. The total reaction volume should be 10µL, including the desired primer pairs and DNA.

Reaction Component	Final Concentration
PowerUp SYBR green master mix (2x)	1x
Forward Primer	300nM
Reverse Primer	300nM
DNA	10ng
Water	Makeup to 10µL final volume

Table 4.11: qPCR experiment thermocycler cycle. This cycle is the default method for the QuantStudio3 system when using the SYBR green master mix kit.

Cycle Stage	Temperature (°C)	Time (min)	Increment (°C/s)	Number of Cycles
Hold Stage	50	02:00	1.6	1
	95	10:00	1.6	
PCR Stage	95	00:15	1.6	40
	60	01:00	1.6	
Melt Curve Stage	95	00:15	1.6	1
	60	01:00	1.6	
	95	00:01	0.15	

4.2.14 Cancer Dataset and Cancer Correlation Analysis

Correlation analysis is a statistical technique used to determine if two variables are related. In this case, it was applied to assess whether the differentially expressed genes associated with *ZFY* overexpression show any relationship with genes correlated with *ZFY* expression levels across cancer cell lines. The goal was to evaluate if the genes and pathways altered by *ZFYS* and *ZFYL* overexpression in HEK293 cells reflect processes linked to endogenous *ZFY* expression in cancer contexts.

To further analyse potential correlations between *ZFY* expression and cancer, data from cancer cell lines in the Cancer Cell Line Encyclopaedia (CCLE) project of the DEPMAP database was examined. The CCLE contains extensive genetic and pharmacologic profiling data on a large panel of human cancer cell lines generated starting in 2008 **(Figure 4.5)**.

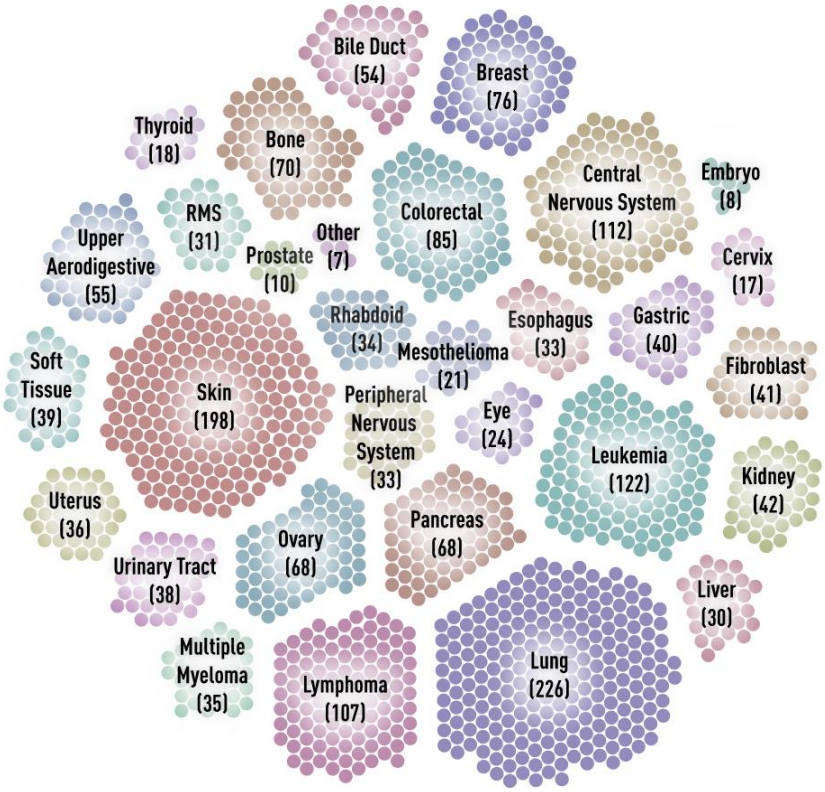


Figure 4.5: CCLE data collection. Consisting of 1829 cancer datasets ranging from lymphoma to breast cancer. Taken from the website DepMap: The Cancer Dependency Map Project at Broad Institute.

The CCLE contains expression data for 1,392 of the total 1,829 cancer cell lines, including 614 females, 770 males, and 445 of unknown gender. Among the cell lines with expression data, 56 were derived from head and neck cancers.

In RStudio the following libraries were loaded; tidyverse (1.3.2), dplyr (1.0.10) and Hmisc (5.1-0) **(Table 4.5)**. The datasets were loaded and prepared. The cell line data was filtered to only include male cell lines since *ZFY* is Y-linked. Of the 770 male lines, 699 had usable expression data. The DepMap_ID column was removed, and the dataset was then filtered for *ZFY*, and the output expression correlation for each cell line was collected. The columns of interest for assessing correlation were the Pearson correlation coefficient (R), p-value, and adjusted p-value for each gene.

The correlation data underwent filtering for significance with a threshold of p-value <0.05, collecting only significant genes. Initially, the Pearson R value was filtered for values ±20%, but this threshold was subsequently adjusted to ±10% due to a

significant lack of overlap between the two datasets. This adjustment to ±10% led to an increase in the detectable overlap between the cancer dataset and the *ZFY* dataset. This weak correlation of ±10% limits the reliability of this analysis and indicates only a modest relationship between the genes differentially expressed with *ZFY* overexpression and endogenous *ZFY* expression levels in cancer cell lines.

Using this correlation data, Leukaemia cell lines were selected for cancer correlation confirmation.

4.2.15 Leukaemia Cell Lines

Three Leukaemia cell lines were selected from experimental data and the available cell lines. MEC-1 (DSMZ.de) is a chronic B cell leukaemia cell line established from the peripheral blood of a 61-year-old. THP-1 (ATCC.org) is an acute monocytic leukaemia cell line derived from the peripheral blood of a 1-year-old male. Finally, U937 (DSMZ.de) is a histiocytic lymphoma established from the pleural effusion of a 37-year-old man. U937 was used as a control, as experimental data has shown no *ZFY* expression in this male cell line.

All Leukaemia cells were cultured at 37°C under humidified conditions, with 5% CO_2. Unlike HEK293 cells, these cell lines were grown in suspension cultures.

THP-1 and UP37 were cultured in an RPMI-1640 media (Sigma, #R0883) with 10% FBS, 1% pen/strep and 1% (2mM) L-glutamine, whilst MEC-1 was cultured in an IMDM media with L-glutamine (Pan-Biotech, #P04-20150) supplemented with 10% FBS and 1% pen/strep. Both of these media are widely used for Leukaemia cell growth. As with HEK293 cells, a logarithmic growth phase was maintained, with the cells being passaged every 2-3 days depending on the cell count. As these cultures are not adherent cells, no trypsin was used. The recommended seeding densities ranged from 100,000 to 150,000 cells/mL. Trypan blue was used to assess the viability of cells as described in 3.2.2.

4.2.16 RNA Extraction Part 2

Cell pellets containing ~$2x10^6$ cells were collected for RNA extraction. For this RNA extraction, the Monarch® Total RNA Miniprep Kit (New England BioLabs, #T2010S) was used, and the additional steps for Leukocyte cells stated in the kit protocol were completed. The concentration of the RNA was determined by nanodrop. Following RNA extraction, the RNA was converted back into cDNA via the LunaScript RT Supermix Kit (New England BioLabs, #E3010). The protocol is stated in **Table 3.6**. The thermocycler conditions for cDNA synthesis are as follows in **Table 3.7**.

4.2.17 Cancer Cell Line Primers

Primers were selected from the correlation analysis. Primers specific to the short and long forms of *ZFY* were also developed, as shown in **Table 4.12**. Primers were purchased from IDT.

Table 4.12: Primers designed from the correlation analysis alongside primers for *ZFYL* and *ZFYS* detection. Primers were designed as previously explained using the NCBI primer tool in **Table 4.6**.

Gene Name	Accession Number	Forward Primer	Tm (°C)	Product Length (bp)
TICAM2	NM_021649.7	For: CGCTCGCCTGCAGATTGAAA Rev: ACACTGTGCCTTTTACCCCAA	58.5 57.0	140
ZBED6	NM_001395895.1	For: GCTGCTGCGAATCACCAAAA Rev: TGGTCTCACCTGAAGCCTCT	56.8 57.9	119
RASAL3	NM_022904.3	For: GCCCCACTGCTTTCAGGTAA Rev: ACGCTCAGCCATGTCTCTTC	57.7 57.3	150
LONRF1	NM_152271.5	For: AGAAGTGGTTTCCGGGCCA Rev: CAAGTCACTGGGTTCTGCTCG	59.4 58.1	131
ZFYS	NM_003411.4	For: GATGGAATAGTGGATGATGC Rev: GTACACCTTGATGACTTCAGGAC	50.7 55.1	126
ZFYL	NM_003411.4	For: AGCAAGATAATGACAAAGCCAG Rev: Same as *ZFYS* reverse primer	53.4 55.1	164
***ZFY* (Both Short and Long)**	NM_003411.4	For: GAATTGCAGCCACAAGAGCC Rev: Same as *ZFYS* reverse primer	57.2 55.1	*ZFYS* = 159 *ZFYL* = 732

4.2.18 Primer Optimisation & RT-qPCR

Using the GoTaq G2 Flexi DNA Polymerase (Promega, M7801) each primer was tested for specificity. The concentration of the kit reagents used was as previously described in **Table 3.8**. The standardised GoTAQ2 thermocycler protocol can be found in **Table 3.9** with slight alterations made to thermocycler times, annealing temperatures, cycle numbers and DNA/template concentration depending on the primer pair used.

PCR products were loaded onto a 2% (w/v) agarose gel (Agarose, Melford Biolaboratories Ltd, MB1200) for 45 minutes at 90V as described in section 3.2.3. A 100bp DNA ladder (Invitrogen, 15628019) was loaded alongside the PCR products.

RT-qPCR was carried out using the QuantStudio3 system (96-Well 0.2mL block) instrument. Comparative Ct was carried out using SYBR Green PCR master mix (Applied Biosystems, 4309155). 10uL reactions were carried out as previously described in **Table 4.10** and the methodology of the thermocycler **(Table 4.11)** remained the same, with exceptions including *ZFYS*, *ZFYL*, *TICAM2* and *ZBED6* where the annealing temperatures were altered based on the optimisation.

Using the data provided by QuantStudio3, the ddCT was calculated and plotted using graph pad. During these calculations, the results were first normalised to the housekeeping gene and then to the female control cell line, HEK293. A 2-way ANOVA test can be carried out on graph pad using the mean of our duplicates, the standard deviation (SD) and the number of repeats. The SD was calculated in Excel using STDV.P. To further identify the significance a multiple comparison could be completed, where within each column the rows were compared.

4.3 Results

4.3.1 Nuclear Localisation of GFP-Tagged Constructs

To determine the localisation of the constructs DAPI was used, a fluorescent stain capable of binding to AT-rich regions of DNA, resulting in the emission of blue light.

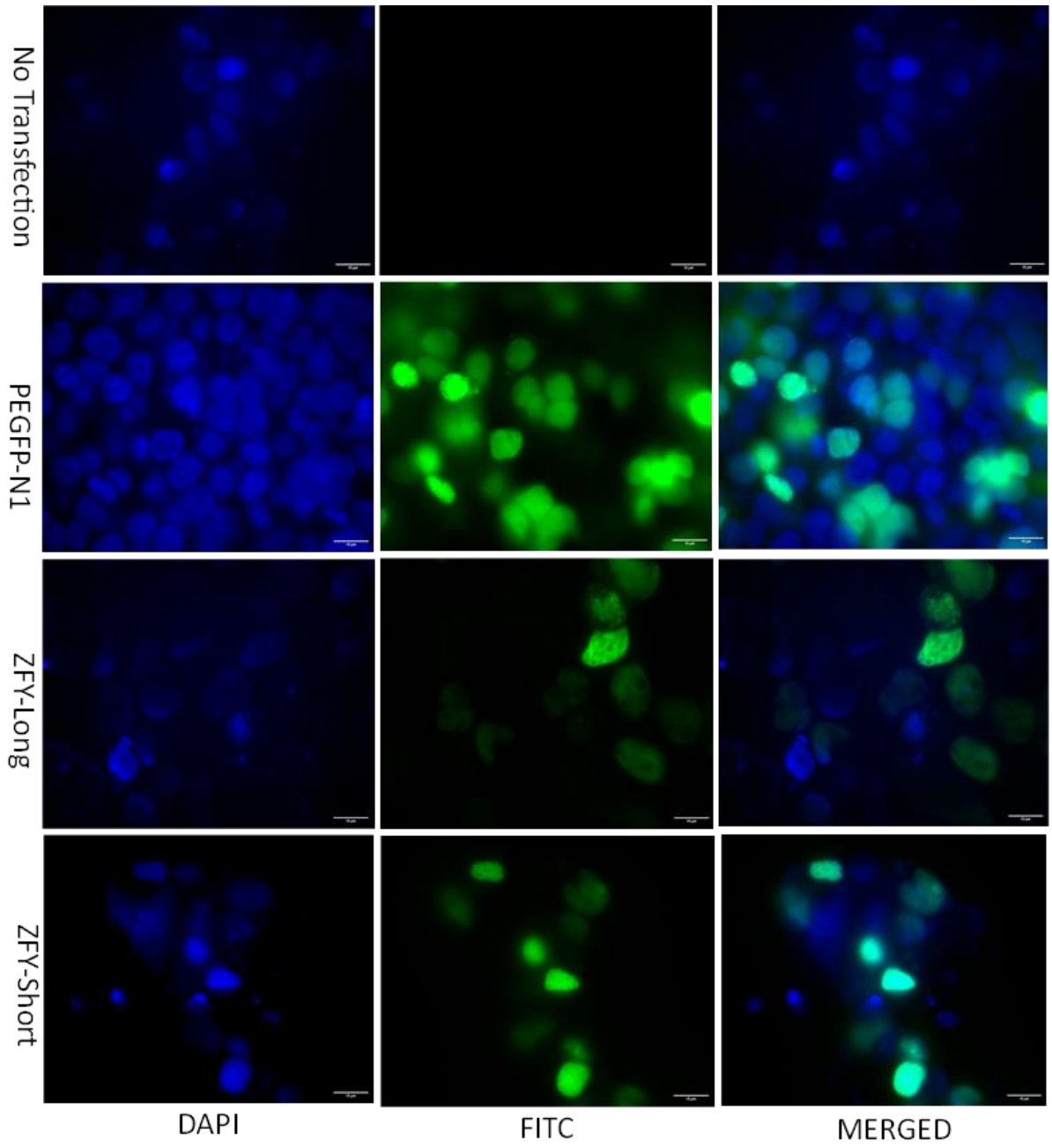


Figure 4.6: DAPI-stained HEK293 cells containing the transfected constructs. DAPI binds to nuclear DNA and is highlighted in this figure as blue fluorescence which is seen in all cell types. The green fluorescence depicts the successful transfection of the GFP constructs and is therefore not identifiable in the no transfection control. Individual channels were taken and the SmartCapture software produced a merged image overlaying the two-coloured signals. Images were captured with an x60 magnification.

DAPI-stained cell nuclei in blue in **Figure 4.6**, is consistent with DAPI's ability to bind A-T-rich double-stranded DNA regions. Green fluorescence is also evident in the transfected HEK293 cells, verifying effective transfection across all three GFP-tagged constructs. As anticipated, both the *ZFY*-Long and *ZFY*-Short constructs localised to the nucleus, aligning with the reported function of *ZFY*. In contrast, free GFP from the pEGFP-N1 control transfection is found in both nucleus and cytoplasm. The absence of green fluorescence in the non-transfected control reflects the lack of GFP expression. Unlike the directly visualisable GFP fusions, the HA-tagged construct does not intrinsically fluoresce, preventing the direct detection of its subcellular localisation and expression via microscopy.

4.3.2 High Transfection Efficiency Achieved

Flow cytometry was used to score the cells based on the presence of GFP to determine the transfection efficiency. The green fluorescence of GFP can be detected by the flow cytometer using the FL1 laser (laser 488nm). The non-transfected HEK293 cells were used as a negative control, whilst the pEGFP-N1 empty vector was used as a transfection-positive control.

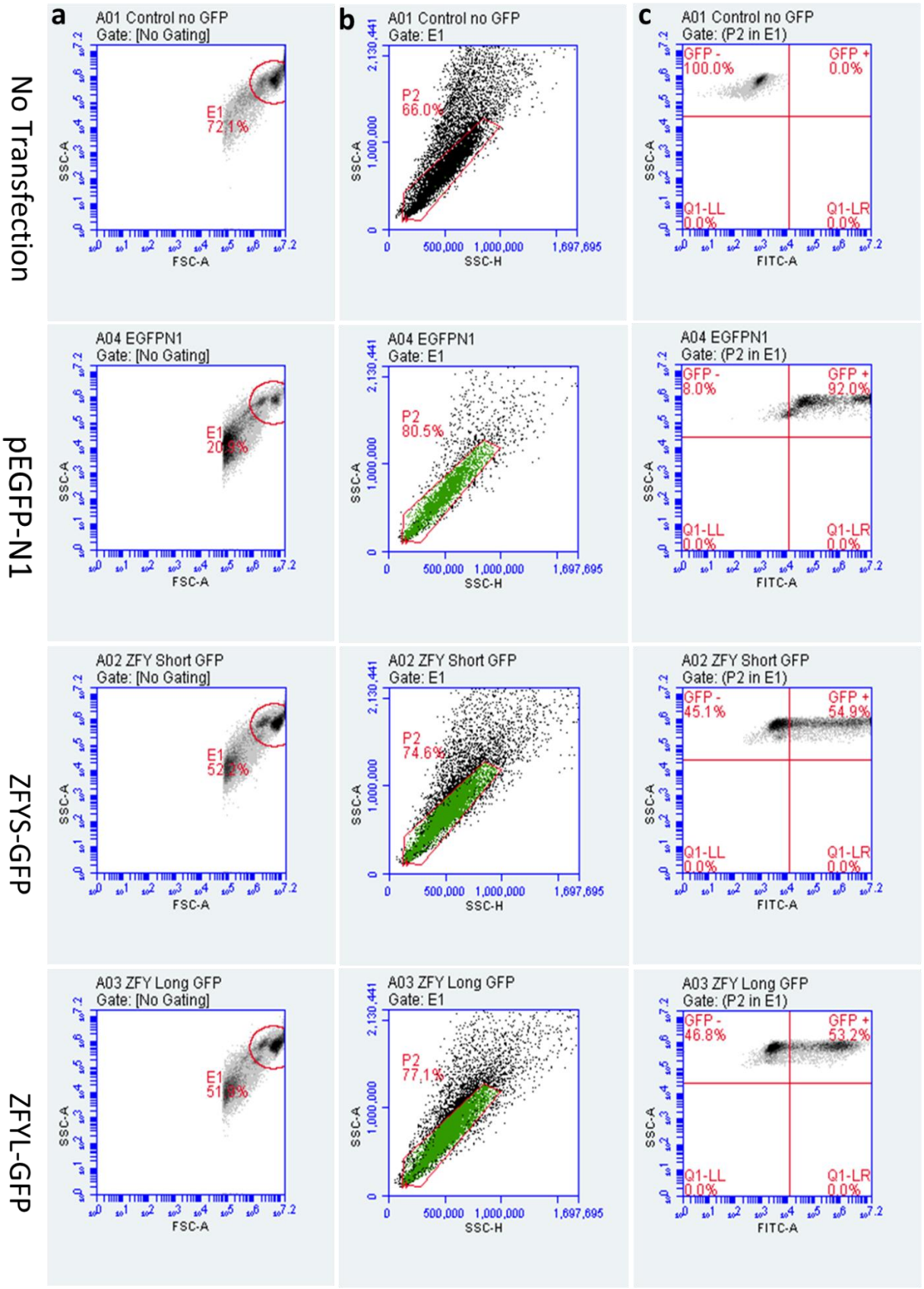


Figure 4.7: Flow cytometry data showing the transfection efficiency percentage in each transfected sample. Gating was set using the non-transfected cells as a

negative control and the pEGFP-N1 transfected cells as a positive control. **A:** This plot illustrates how the FSC-A (X-axis) and SSC-A (y-axis) were used to select the desired cell population and therefore, remove any debris. E1 denotes the selected cell population. **B:** Cell doublets were removed from the selected cell population, P2 is now the wanted cell population. **C:** The plot highlights the cells emitting a GFP signal. The top right quadrant labelled as “GFP+” consists of the cells emitting a green fluorescence signal highlighting that these cells express the transfected constructs. The top left quadrant labelled “GFP-“ contains all the cells with no green signal.

The plots in **Figure 4.7** demonstrate the higher transfection efficiency of the small empty GFP vector (92%) compared to the larger *ZFY*-GFP fusions, as expected. However, both *ZFY*-short and *ZFY*-long achieved transfection efficiencies of >50%, meeting the target aim. Both constructs were equivalently efficient, with only a 1.7% difference noted but this could just be done to noise.

4.3.3 Western Blot Confirmation of Successful Transfection

Protein lysates from transfected mammalian cells were analysed by SDS-PAGE and western blotting. Membranes were probed with anti-GFP and anti-HA antibodies **(Table 4.3)** to validate successful transfection and expression of the tagged *ZFY* constructs at the expected sizes.

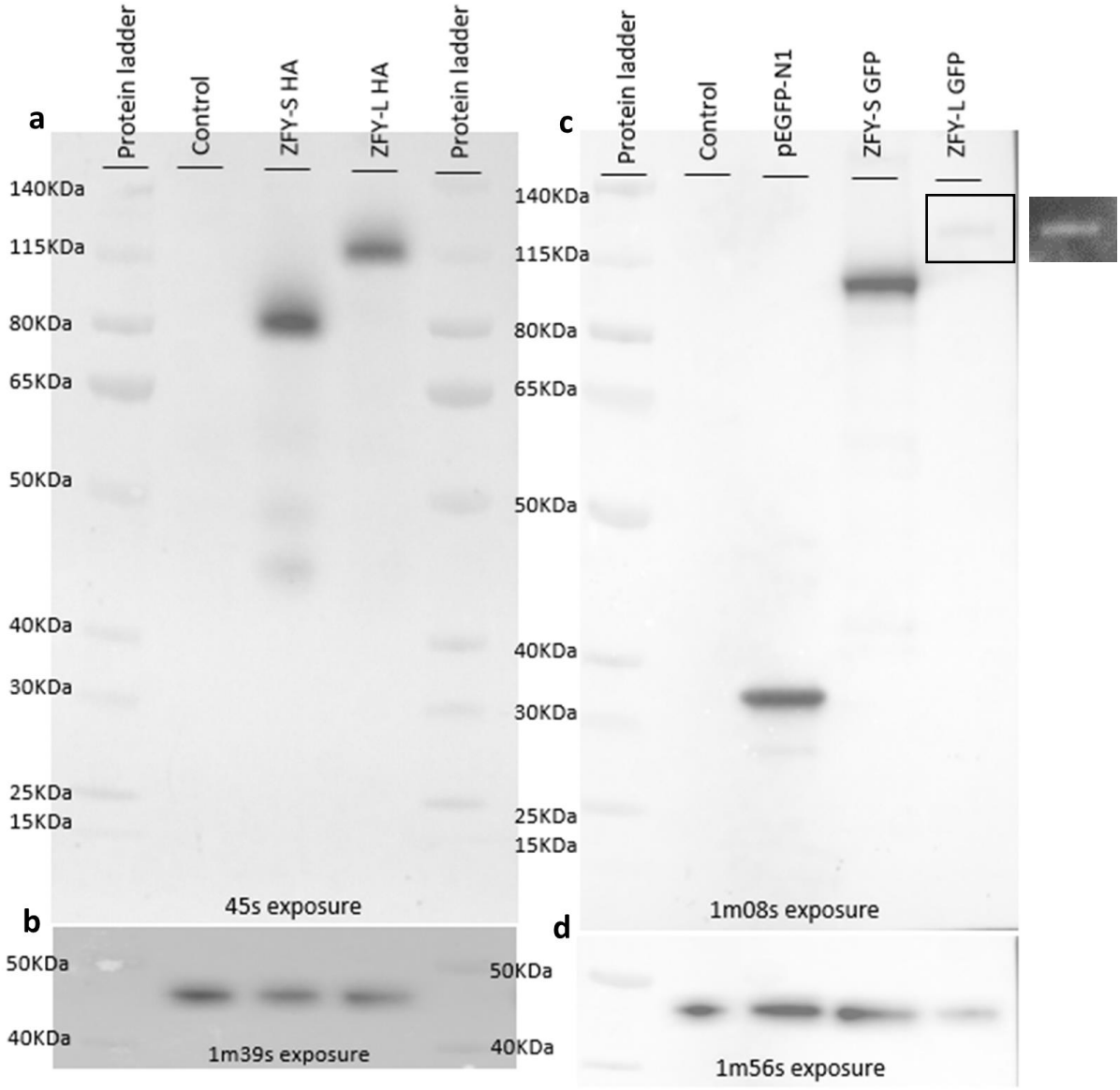


Figure 4.8: Western blots of the transfected constructs. A: Western blot of the HA-tagged construct lysates. Control = non-transfected HEK293 cell lysate, *ZFY*-S HA = *ZFY*-S HA-tagged transfected HEK293 cells & *ZFY*-L HA = *ZFY*-L HA-tagged transfected HEK293 cells. **B:** Beta-actin antibody used as a loading control. **C:** Western blot of the GFP-tagged construct lysates. Control = non-transfected HEK293 cell lysate, pEGFP-N1 = empty GFP transfected cells, *ZFY*-S GFP = *ZFY*-S GFP-tagged transfected HEK293 cells & *ZFY*-L GFP = *ZFY*-L GFP-tagged transfected HEK293 cells. **D:** Beta-actin antibody used as a loading control. Expected molecular weights; pEGFP-N1 = 27-30KDa, *ZFYS*-GFP = ~96.3KDa, *ZFYL*-GFP ~117.5KDa, *ZFYS*-HA ~70.4KDa, *ZFYL*-HA ~91.6KDa & Beta-Actin = 42KDa.

Shown in **Figure 4.8** is the detection of the transfected constructs in protein lysates at the expected molecular weights by western blotting. Strong band intensity demonstrates efficient transfection and expression of the HA- and GFP-tagged *ZFY* isoforms in the mammalian cells. The protein bands run slightly higher than the

expected molecular weight, which is likely to be due to *ZFY* containing highly negative regions which could prevent binding of SDS. Notably, the *ZFYL*-GFP transformation consistently gave lower signal by Western blot, potentially due to reduced expression or translation efficiency of the longer construct. This is consistent with the slightly weaker GFP fluorescence seen for this construct in **Figure 4.7**. Following multiple confirmations of successful transformation, RNA was isolated and sent for RNA sequencing.

4.3.4 Good QC Analysis and Alignment Rates

Following RNA sequencing data collection, quality control analysis was performed. Reports were produced by FastQC (version 0.11.9). The main focus was the alignment rate and the number of QC-failed reads. It was also noted if there was any adapter contamination.

Table 4.13: QC analysis of the FASTQ files provided by Novogene. Sample read number, alignment rate and QC-results from the FastQC analysis are provided in the table below.

Sample Name	Number of Reads	Alignment Rate	QC-failed Reads
R1_Control_1	71723038	97.27%	0
R2_Control_2	82017584	97.25%	0
R3_Control_3	67042865	97.29%	0
R4_1_pEGFP_N1_1	89520020	86.54%	0
R5_pEGFP_N1_2	71873235	86.57%	0
R6_pEGFP_N1_3	63953343	87.60%	0
R7_1_*ZFYS*_GFP_1	59298530	95.28%	0
R8_*ZFYS*_GFP_2	76739990	95.27%	0
R9_*ZFYS*_GFP_3	49949802	95.46%	0
R10_1_*ZFYL*_GFP_1	54178329	96.70%	0
R11_1_*ZFYL*_GFP_2	76377424	96.45%	0
R12_*ZFYL*_GFP_3	56767674	96.49%	0
R13_1_*ZFYS*_HA_1	71574319	95.68%	0
R14_*ZFYS*_HA_2	57479841	96.08%	0
R15_1_*ZFYS*_HA_3	63091116	95.99%	0
R16_*ZFYL*_HA_1	65338642	96.44%	0
R17_*ZFYL*_HA_2	65031044	96.62%	0
R18_*ZFYL*_HA_3	65074987	96.54%	0

Ideal alignment rates are >80% but generally expected to exceed 50%. As shown in **Table 4.13**, all samples displayed good alignment rates above 80%, with no reads failing QC checks. The total read counts were lower in the *ZFY* transfected samples compared to control samples. No adapter contamination was present so read trimming was not necessary.

4.3.5 *ZFY* Over-expression does Not Silence the X Chromosome

Chromosome counting was performed by quantifying mapped reads per chromosome to evaluate the potential triggering of chromosome-wide silencing by *ZFY* overexpression. X chromosome inactivation during spermatogenesis reduces X-linked expression to <1%. While chromosome-wide effects were not expected in HEK cells, *ZFY*'s influence on X-silencing in meiotic contexts provided a rationale for this analysis. A substantial decrease in X-linked reads in *ZFY* transfected cells could indicate ectopic induction of X-inactivation-like effects.

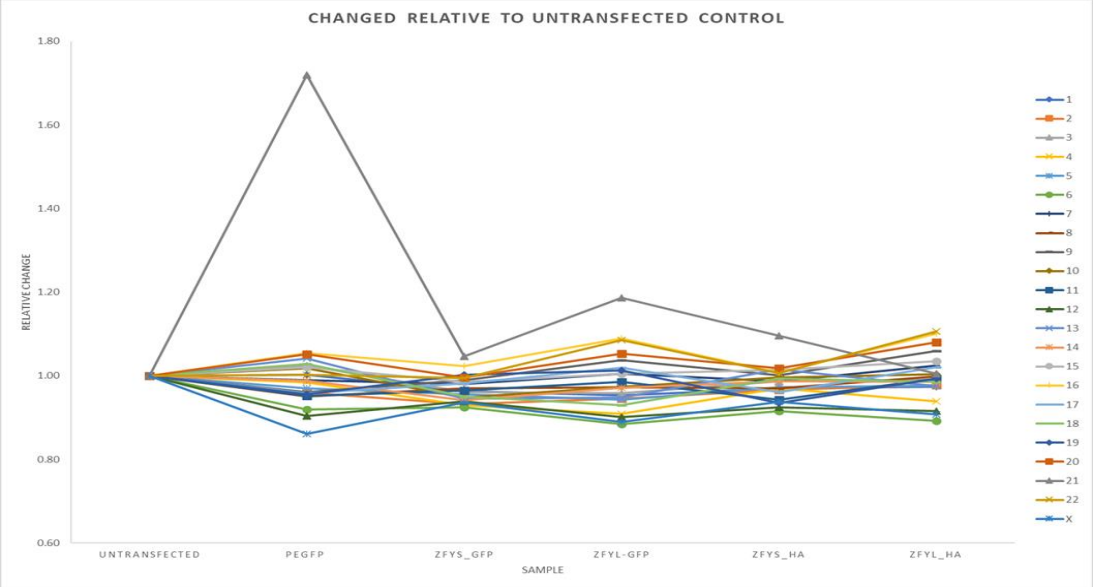


Figure 4.9: Graph showing the relative change in chromosome count across the different samples. For each construct a chromosome count was calculated, and the relative change was plotted on the Y-axis. The chromosome number and its corresponding colour are noted as a key in the above graph.

In **Figure 4.9** there is no evidence of X chromosome silencing resulting from *ZFY* overexpression, as the relative proportion of X-linked reads is consistent across samples. The most pronounced change is an increase in chromosome 21 representation in the GFP empty vector control cells, with a minor elevation also visible for *ZFYL*-GFP. The reason for this is unknown, and potential could be down to one specific gene giving a false read. Therefore, there is no decrease indicative of ectopic X-inactivation triggered by *ZFY* overexpression.

4.3.6 DESeq2 Differential Gene Expression Analysis

DESeq2 takes count data from high-throughput sequencing to test for differential gene expression across samples. DESeq2 stands as a prevalent tool in RNA-Seq analysis, offering an extensive perspective on gene expression alterations across various conditions, distinguishing itself from alternative packages (Love *et al*., 2014)(Love *et al*., 2015). By using negative binomial distribution DESeq2 can make it account for the dispersion across the entire dataset providing more accurate p-values and FDR estimates (Love *et al*., 2014)(Love *et al*., 2015).

4.3.6.1 Dataset Dispersion

To ensure that the sequencing data is accurately modelled, dispersion is calculated for each gene. DESeq2 calculates the variation by using the mean gene expression level via the “shrinkage” model. The dispersion can then be modelled based on the expression level using estimated maximum likelihood estimations and the dispersion value for each gene is then plotted.

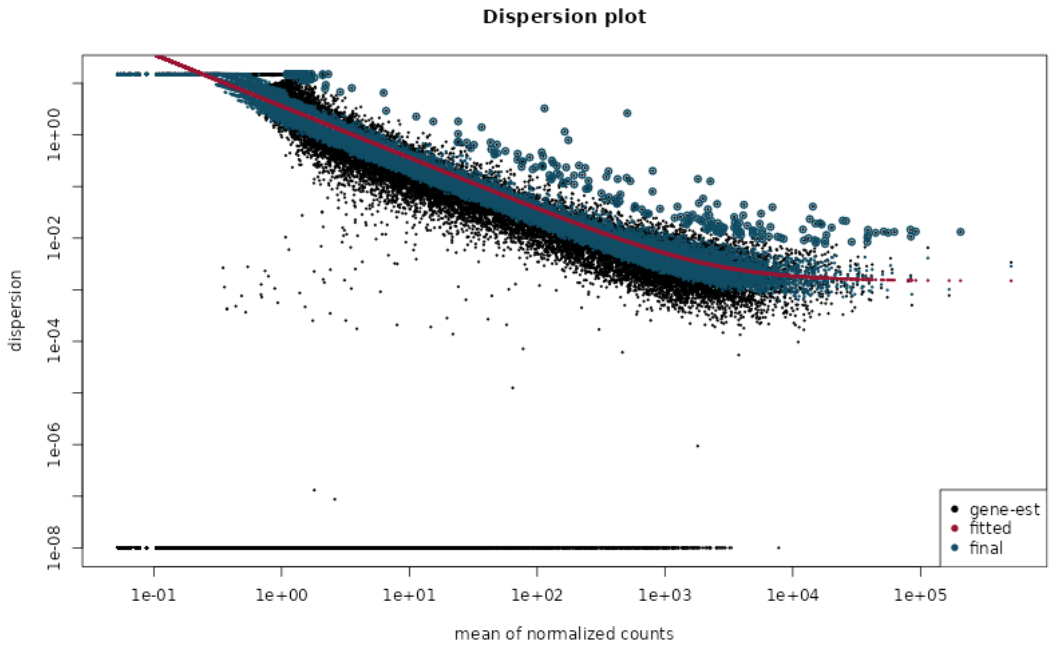


Figure 4.10: Dispersion Plot. A dispersion plot shows the gene variance on the Y-axis and the mean expression on the X-axis. Empty GFP was used as the control sample. Created using the bfigplotDispEsts package readily available in DESeq2.

The desired decreasing dispersion at higher mean counts is seen in **Figure 4.10**, enabled by the multiple biological replicates per condition. It is important that the data follows the red fitted line which plots the expected dispersion value for genes at a given expression strength. Greater numbers of replicates provide stronger shrinkage and power for estimating variation, facilitating the identification of differential expression. The inclusion of three biological replicates improved the detection of gene

expression changes across conditions and leads to the data following the “fitted” data line. It is vital that the dispersion is accurate as correctly estimating these parameters is vital for detecting differential expression. Underestimation can lead to false discovery.

4.3.6.2 Sample Clustering

To explore the similarities and differences across the samples, sample clustering QC was performed using DESeq2. Sample clustering indicates how well the replicates cluster together and will show any major sources of variation within the data. Sample clustering was accessed using Principal Component Analysis (PCA) **(Figure 4.11)** and Hierarchical clustering Heatmap **(Figure 4.12)**. PCA highlights the variation in the dataset in a two-dimensional way. The greatest variation is called the first principal component, PC1. A heatmap works similarly to the PCA but shows the gene expression correlation across the samples in the dataset. Initially, the analysis was conducted by comparing the transfected cells with the non-transfected control cells. However, subsequent examination suggested that transfection itself induced gene

variation. Consequently, the analysis shown proceeded using the empty-GFP vector as the control group.

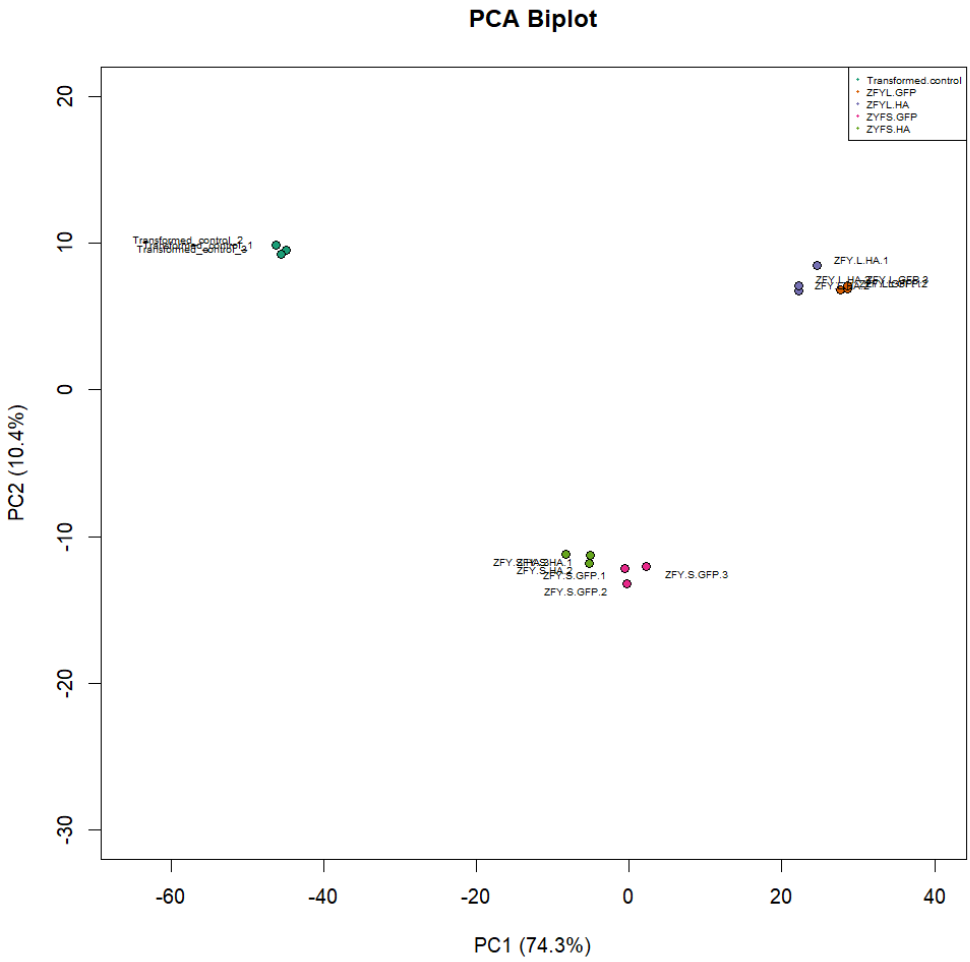


Figure 4.11: Principal Component Analysis Plot. Focusing here on the major variances noted as PC1 (X-axis) and PC2 (Y-axis). This was made in DESeq2 using the empty vector as the control group.

In the depicted PCA plot in **Figure 4.11**, it is evident that the primary source of variance is attributed to the introduction of *ZFY* via transfection into the mammalian cells. This transfection effect accounts for a substantial 74.3% of the observed changes, as indicated by PC1. *ZFYS* seems to be positioned intermediately on the axis, indicating that the *ZFYS* transfection produces a similar change as the *ZFYL* transfection but to a lesser degree. Additionally, about 10.4% of the variations (PC2) appear to distinguish *ZFYS* from both *ZFYL* and the empty-eGFP control capturing *ZFYS*-specific biology. This can be attributed to the introduction of foreign DNA, potentially influencing biological processes linked to genes responsible for immune responses to viral infections. Consequently, these alterations may induce minor changes in gene expression, contributing to the overall dataset variance, albeit at a relatively modest percentage. Furthermore, *ZFY*-L constructs seemed to show similar

effects to that of pEGFP-N1 but the PC2 effect could be due to the larger size of the *ZFYL* constructs which are harder to incorporate into the genome.

Another graphical representation of sample clustering is a heatmap.

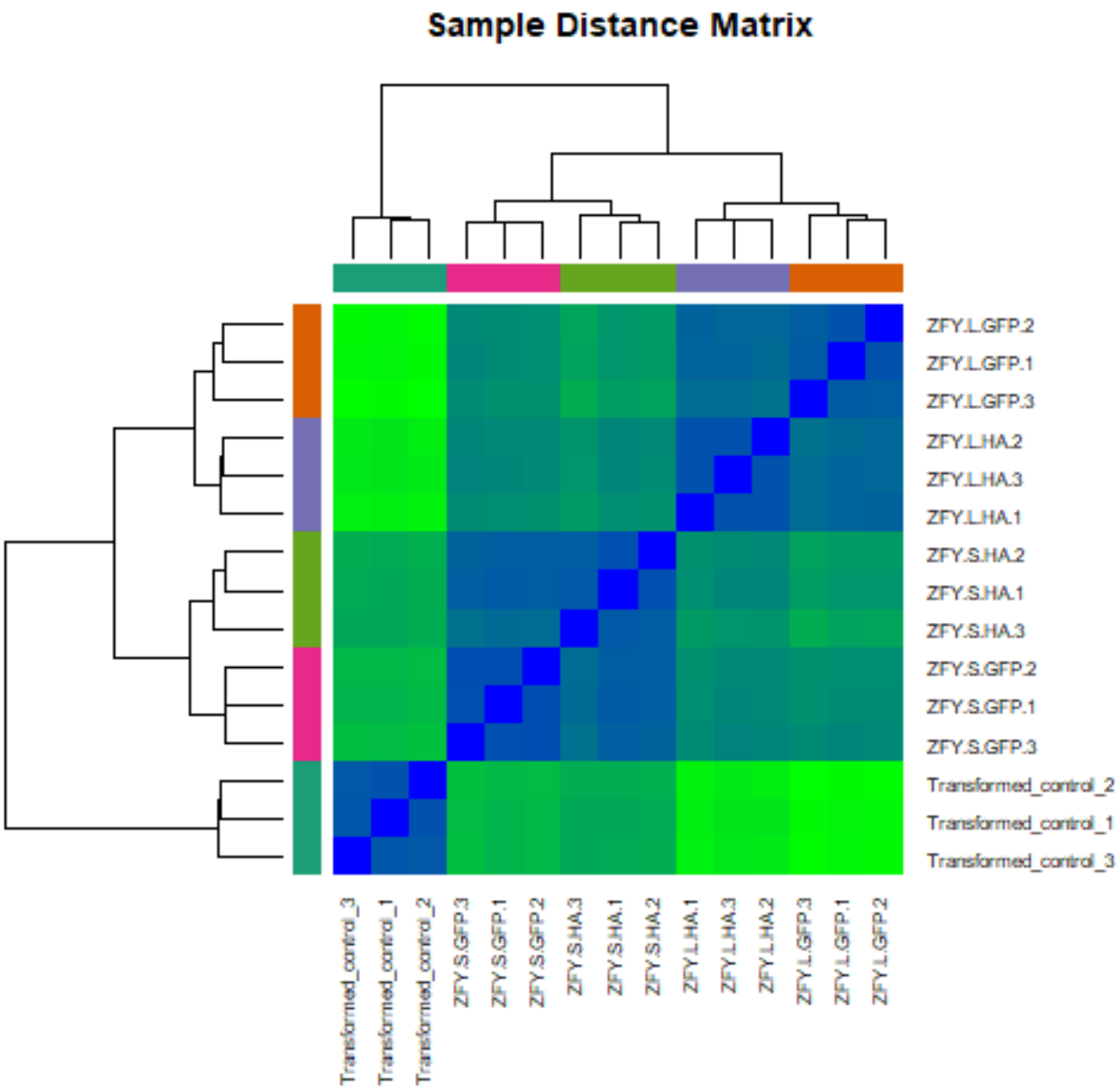


Figure 4.12: A sample distance matrix otherwise known as a heatmap. As with the previous plots the empty GFP was used as the control group. In this plot, blue denotes similarity and green denotes dissimilarity between sample groups, with the branches showing the grouping of the samples. The lighter the green the greater the difference in the transcriptome. The heatmap.2 package was used in DESEQ2 to produce the plot.

Based on the presented plots in **Figure 4.11** and **Figure 4.12**, it can be asserted with confidence that the replicates exhibit cohesive grouping, demonstrating consistency in both experimental and informatics methods. Notably, both the constructs and tags

exhibit clear grouping in both the heatmap and PCA plots. Specifically, all six *ZFYS* samples cluster together distinctively, while the six *ZFYL* samples also exhibit cohesive grouping. We can confirm that since the HA- and GFP-tagged versions of each construct cluster together the tag is not distorting them greatly. Furthermore, from this we can answer one of the hypotheses, as *ZFYS* and *ZFYL* cluster together, they must regulate the same genes in the same direction, ruling out the hypothesis that *ZFYS* has a direct antagonistic effect to *ZFYL*. However, *ZFYS* still may compete for binding and could be a weaker activator. These observation supports the conclusion that *ZFY* significantly influences gene expression levels in these mammalian cells.

Subsequently, using the unfiltered data, the L2FC was plotted with a significant p-value set to <0.05. This looked at the entire set of genes including protein coding and non-protein coding to compare the two *ZFY* variants and their effect on the cell's gene expression.

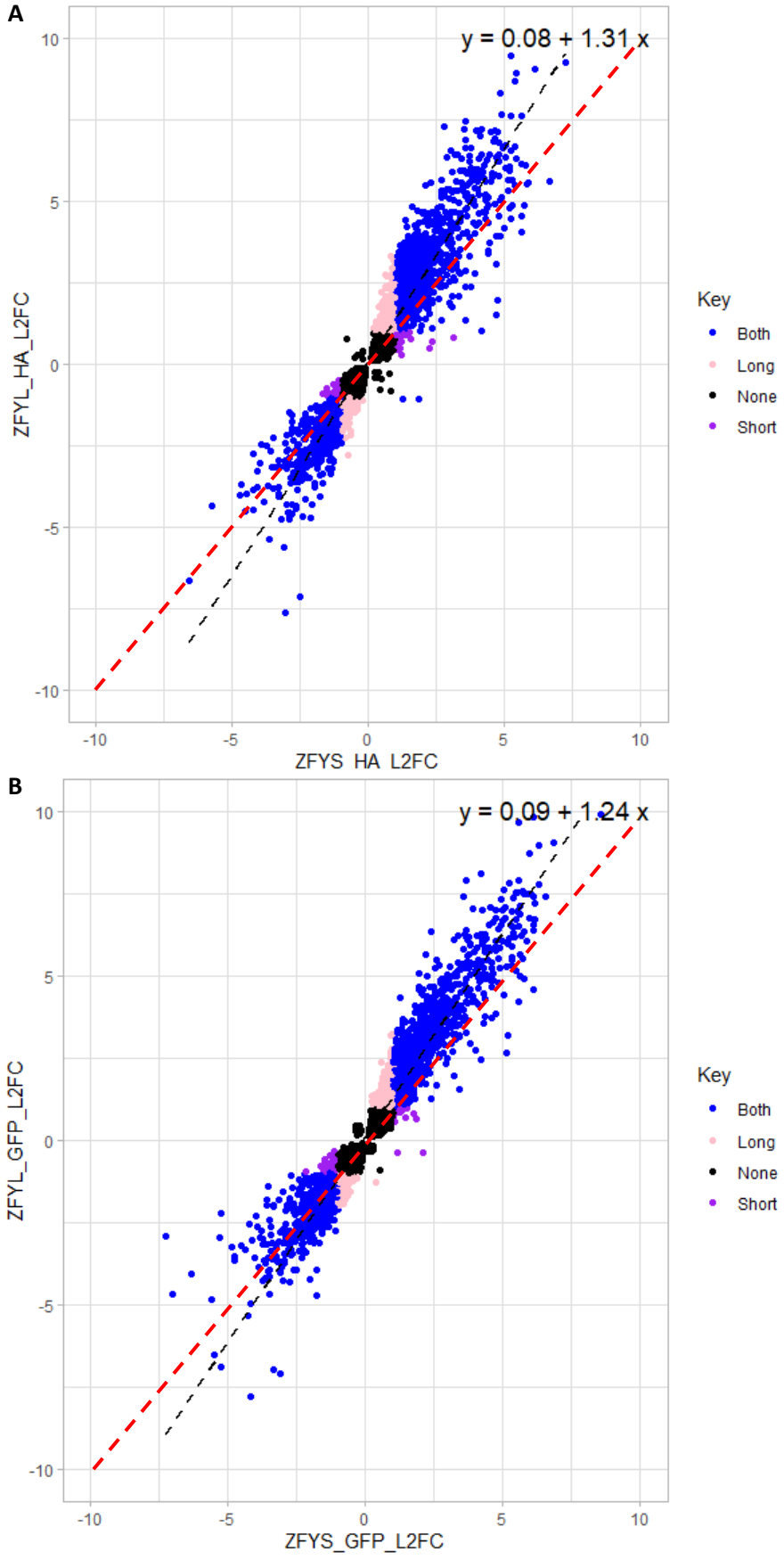

A
y = 0.08 + 1.31 x
ZFYL_HA_L2FC
ZFYS_HA_L2FC
Key
Both
Long
None
Short
B
y = 0.09 + 1.24 x
ZFYL_GFP_L2FC
ZFYS_GFP_L2FC
Key
Both
Long
None
Short

Figure 4.13: L2FC plots comparing the transcriptomic effect of both *ZFY* variants and the different tagged versions. A: HA-tagged *ZFY* construct comparison, with the log_2 fold change in *ZFYS* relative to pEGFP-N1 control plotted on the X-axis and log_2 fold change in *ZFYL* relative to pEGFP-N1 control plotted on the Y-axis. Only genes found to be significant at an adjusted p-value <0.05 are plotted. **B:** GFP-tagged *ZFY* construct comparison, with the log_2 fold change in *ZFYS* relative to pEGFP-N1 control plotted on the X-axis and log_2 fold change in *ZFYL* relative to pEGFP-N1 control plotted on the Y-axis. For both graphs, the colours indicate if the gene expression was changed by greater that 2-fold (= 1 log_2 unit) in response to both constructs, to one individual construct or neither. The black dashed line indicates the linear regression of the *ZFYL* response vs the *ZFYS* response for all points shown. The red dashed line indicates a 1:1 relationship, i.e. the expected value if the fold change were the same in both the *ZFYL* and *ZFYS* experiments.

Figure 4.13 demonstrates that both *ZFYS* and *ZFYL* regulate many of the same genes in the same direction. However, is it notable that positively regulated genes (right hand side of the chart) largely fall above the 1:1 line, indicating a stronger upregulation by *ZFYL* than by *ZFYS*. Conversely, negatively regulated genes fall below the 1:1 line, indicating stronger downregulation by *ZFYL* than *ZFYS*. There seem to be very few genes targeted solely by *ZFYS*, with potentially a few targeted solely by *ZFYL*. However, given how close this graph falls to a straight line, this suggests that there are no genes that selectively regulate by only one isoform: i.e. every gene regulated by *ZFYL* is also regulated to some degree by *ZFYS* and vice versa. The only things that fall outside this trend are a few points located close to the origin that are located within the “noise” scatter of the linear relationship. Again, there is very little difference between the GFP and HA-tagged constructs indicating that the tag is not introducing appreciable bias to the study.

While it is clear from the overall scatter plot that *ZFYL* has a stronger effect on all its downstream targets than *ZFYS*, the magnitude of this difference is hard to estimate directly from the graph and will be further understated since the Western blot experiment (**Figure 4.8**) showed that for both the HA- and GFP-tag experiments, the *ZFYL* construct was expressed more poorly than the *ZFYS* construct. Thus, in this RNA-Seq expression dataset, *ZFYL* is exerting a larger downstream effect despite being present in fewer copy numbers per cell.

To attempt to quantify this, we examined the read counts mapping specifically to the GFP portion of each construct, in order to compare the true relative expression level of the transfected construct in each case. This gives a measure of the construct expression that is not biased by the different length of the transgene in each experiment.

Table 4.14: GFP-Count for each transfection to determine the number of transgene transcripts in each transfection.

Sample Name	Transfection Name	GFP-Hit
R1	Control	0
R2	Control	0
R3	Control	0
R4_1	pEGFP-N1	5966700
R5	pEGFP-N1	4548180
R6	pEGFP-N1	3754494
R7_1	*ZFYS* GFP	425511
R8	*ZFYS* GFP	584082
R9	*ZFYS* GFP	408822
R10_1	*ZFYL* GFP	166502
R11_1	*ZFYL* GFP	262579
R12	*ZFYL* GFP	132234
R13_1	*ZFYS* HA	27
R14	*ZFYS* HA	23
R15_1	*ZFYS* HA	27
R16	*ZFYL* HA	0
R17	*ZFYL* HA	0
R18	*ZFYL* HA	0

In **Table 4.14** a larger number of GFP transgene transcripts are present in the *ZFYS*-GFP samples compared to the *ZFYL*-GFP samples, with the average GFP read level in the *ZFYS* experiment being 2.53 times higher than the read level in the *ZFYL* experiment. It was also noted that pEGFP-N1 transgene expression level is much higher (as also seen in **Figure 4.8**) most likely due to it being a much smaller construct. The *ZFYS*-HA experiment did have a few reads mapping to GFP which is not expected. However, the numbers are so small (about 10,000 times lower) that it is most likely due to be a result of index hopping during the RNA sequencing process. This low-level cross-contamination of the sequencing dataset was too small to pose a problem in downstream analysis.

Returning now to **Figure 4.13**, we note that the absolute magnitude of the regulatory changes is compressed by the log2 transformation. On converting these numbers back to absolute fold changes, the most strongly upregulated genes showed approximately 2^{6}-fold upregulation by *ZFYS* (a 64-fold increase relative to the pEGFP-N1 control) and a $2^{7.5}$-fold upregulation by *ZFYL* (a 181-fold increase relative to the pEGFP-N1 control). Thus, although the *ZFYL* construct is only expressed at

approximately one third the level of *ZFYS* construct **(Table 4.14)**, it nevertheless produced an approximately ~3x higher effect at the most strongly regulated target genes, indicating that for these genes *ZFYL* is almost 10-fold as efficient a transcriptional activator on a per-molecule basis.

Overall, we conclude that while *ZFYS* does not directly antagonise *ZFYL* via transcriptional repression, it may function *in vivo* as a competitive inhibitor by competing for target binding sites and having a weaker downstream effect.

4.3.6.3 Differential Gene Expression Contrasts

To understand the variations in gene expression profiles among various genotypes, utilising MA plots and volcano plots is an excellent initial step.

MA plots graphically present the average of normalised counts against the log2 fold change for all genes. MA plots colour the significant differentially expressed genes making it a very efficient way to illustrate the shift in gene expression and the LFC "shrinkage". While a volcano plot plots and annotates the differentially expressed genes identified in the contrast.

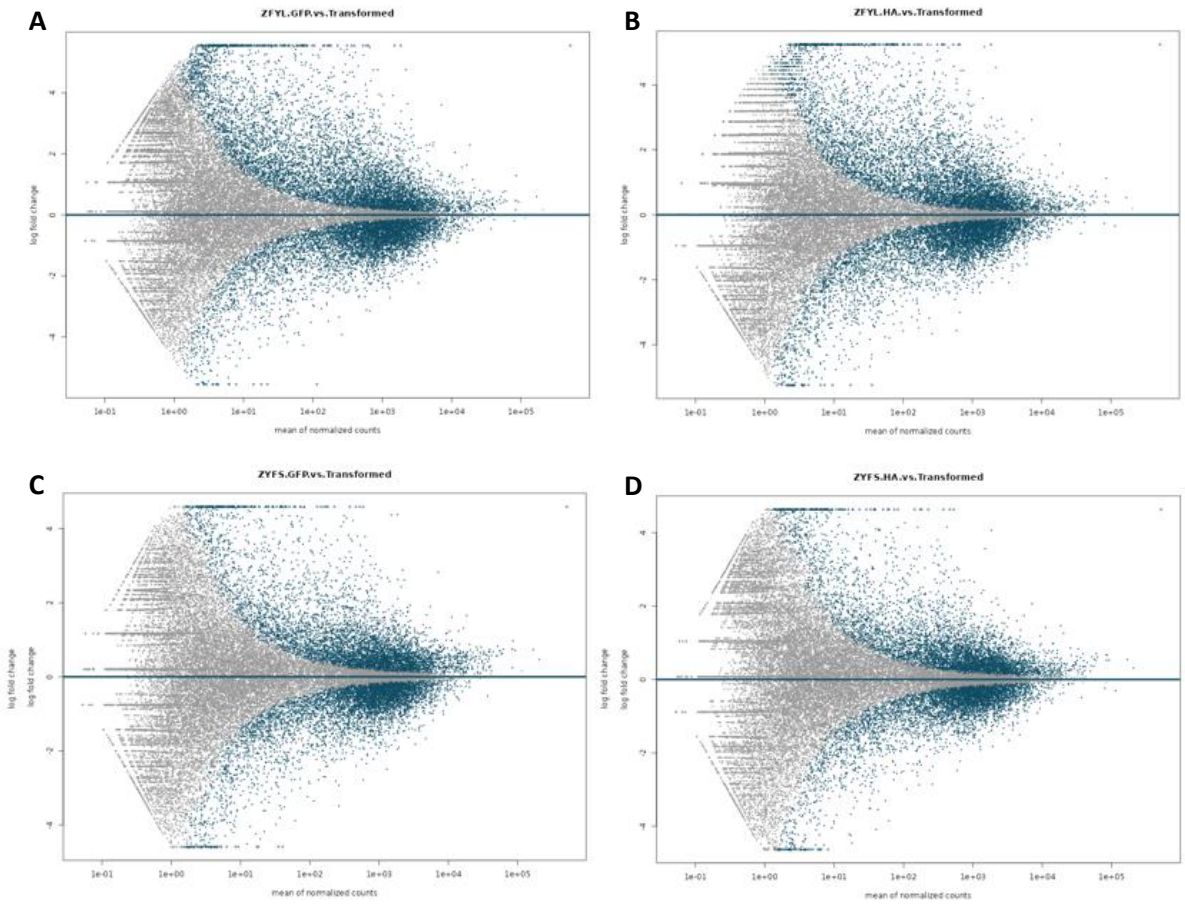


Figure 4.14: MA plots showing the log fold-change (LFC) plotted on the Y-axis and the overall expression on the X-axis. A: *ZFYL*-GFP vs Transformed, **B:** *ZFYL*-

HA vs Transformed, **C:** *ZFYS*-GFP vs Transformed and **D:** *ZFYS*-HA vs Transformed. Blue dots denote significant changes.

Highlighted in **Figure 4.14** is the significant number of differentially expressed genes across all samples in comparison to the empty GFP vector control samples shown by the blue colouration. Further interpretation is difficult from the MA plot.

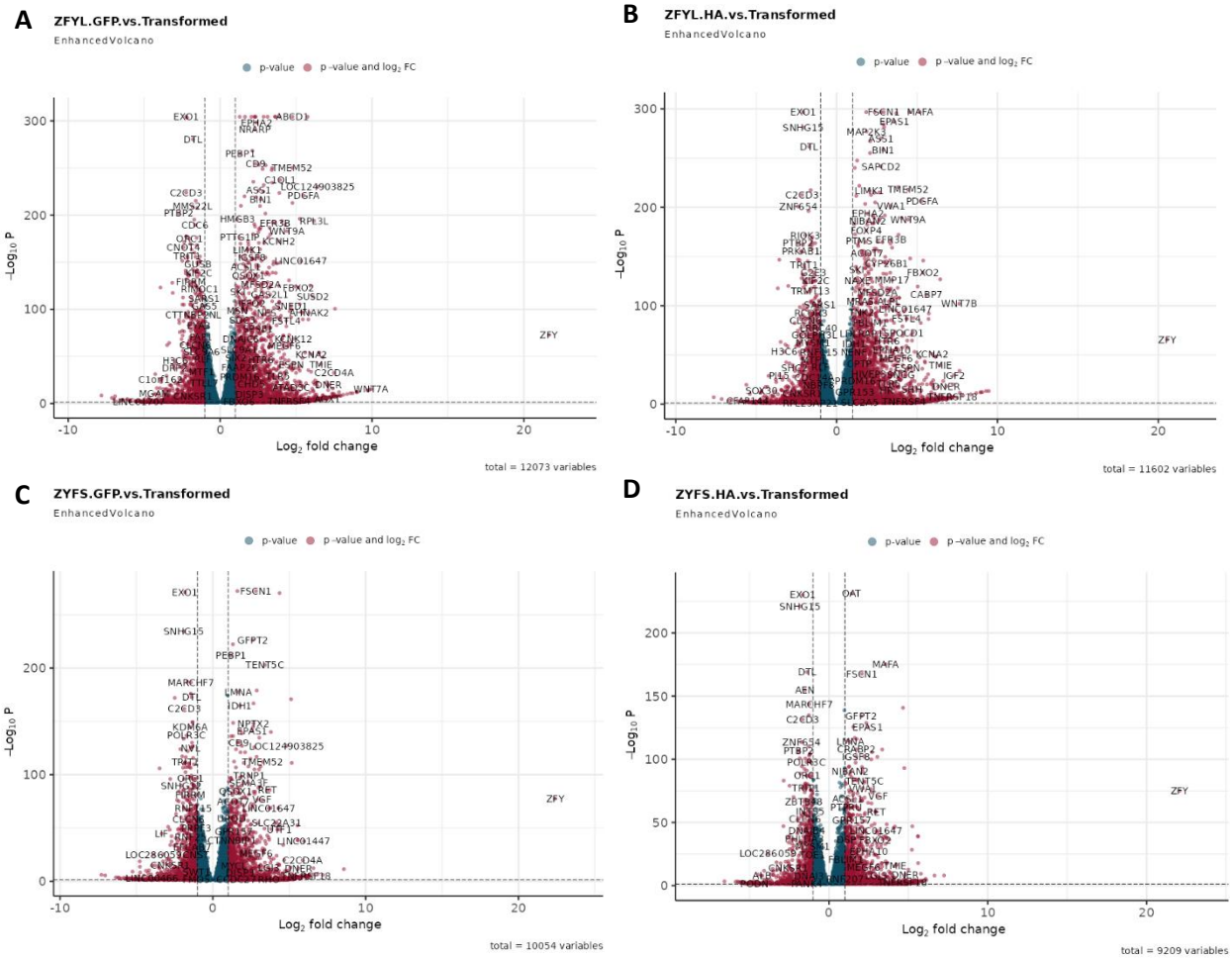


Figure 4.15: Volcano Plots showing the differential gene expression across the constructs. The Y-axis shows the negative base-10 log of the p-value and the X-axis depicts the logarithmic fold change, limited to values greater than 1 or less than -1 highlighted in red. **A:** *ZFYL*-GFP vs Transformed, **B:** *ZFYL*-HA vs Transformed, **C:** *ZFYS*-GFP vs Transformed and **D:** *ZFYS*-HA vs Transformed. Genes significant for p-value and L2FC are highlighted in red.

Based on these results, further investigation into the differentially expressed genes proceeded. Subsequent analysis opted to focus solely on the *ZFY*-GFP samples in comparison to the empty control GFP as transfection alone seems to cause some changes in the genome. Moreover, when looking at the sample clustering there does

not seem to be a significant tag effect suggesting that the HA- and GFP-tagged constructs are showing similar gene expression profiles.

4.3.6.4 Differential Gene Lists

After opting to focus solely on the *ZFY*-GFP samples in comparison to the empty control GFP, analysis of the raw DESeq2 dataset revealed a substantial number of differentially expressed genes, numbering in the thousands. Notably, the disparity was more pronounced in *ZFYL*-GFP, with 12,057 differentially expressed genes compared to 10,027 in *ZFYS*-GFP. However, these genes were initially filtered based solely on a p-value < 0.05 in DESeq2. To refine the dataset for significant analysis, additional filtering was applied by introducing L2FC criteria (>1 or <-1). Consequently, this filtering reduced the gene count to 5,706 for *ZFYL* and 3,293 for *ZFYS*.

The subsequent steps involved distinguishing between protein-coding and non-protein-coding genes, given the focus on protein-coding genes. Through annotation with UniProt IDs and identification of protein-coding genes, 4,063 *ZFYL* protein-coding genes and 2,377 *ZFYS* protein-coding genes remained. Before commencing any analysis, a final data sorting step was implemented to eliminate duplicates, resulting in 4,040 *ZFYL* and 2,362 *ZFYS* differentially expressed genes.

After completing the data sorting process, both inner and anti-join functions were employed to identify shared genes that might be regulated by both *ZFYS* and *ZFYL*, as well as to identify unique genes potentially regulated exclusively by one of the variants. It was observed that *ZFYS* and *ZFYL* exhibited 2,090 differentially expressed coding genes in common, indicating the regulation of a substantial number of shared genes. Additionally, *ZFYL* regulated an extra 1,950 coding genes not regulated by *ZFYS*. In contrast, *ZFYS* had only 272 uniquely differentially expressed genes, suggesting a comparatively lower regulatory impact. This implies that *ZFYL* functions as the more active transcription factor.

Subsequently, eliminating any *ZFX* differentially expressed genes identified in Weiya Ni *et al* enabled an examination focused on genes exclusively influenced by *ZFY* and not *ZFX*, given the assumption of their binding to similar targets/sites. Following this filtering process, out of the 2,090 genes regulated by both *ZFYS* and *ZFYL*, 1,866 were not concurrently regulated by *ZFX*. Likewise, for the *ZFYL*-exclusive genes, the count decreased from 1,950 to 1,879, and for *ZFYS*, it reduced from 272 to 268. The observed pattern suggests a limited overlap in targets between *ZFY* and *ZFX*, as a considerable number of the differentially expressed genes identified did not exhibit differential expression by *ZFX*. This might elucidate the reason behind the divergence of *ZFX* and *ZFY*, as their functions appear to significantly differ from each other. Refer to **Figure 4.16** for the presented data.

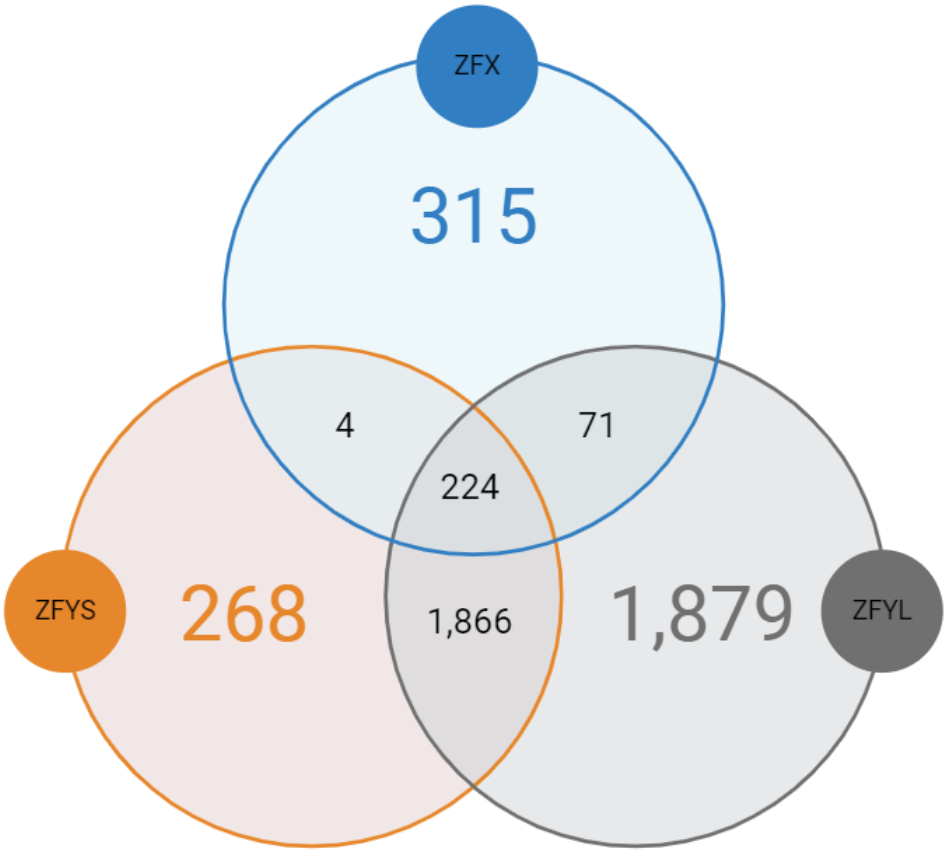


Figure 4.16: A Venn diagram showing the differentially expressed genes. These genes were identified in the *ZFY* DESeq2 analysis alongside the *ZFX* data from Weiya Ni *et al* (Ni *et al*., 2020).

It was observed that a significant majority of the differentially expressed genes were predominantly upregulated, indicating a L2FC greater than 1. Among the 1,866 genes influenced by both *ZFYS* and *ZFYL* but not *ZFX*, 58.4% (1,090) exhibited an L2FC exceeding 1. Furthermore, of the 1,879 genes exclusively regulated by *ZFYL* (excluding *ZFX*-regulated genes), 64.6% (1,213) were found to be upregulated. In

contrast, when considering *ZFYS*-exclusive genes (excluding *ZFX*-regulated genes), the majority displayed downregulation, with only 28.4% (76 genes) of them being upregulated. In Weiya Ni *et al*, a similar observation was made as out of the 614 genes identified to be differentially expressed, 68.1% (418) of the genes were upregulated (Ni *et al*., 2020). A gene validation panel was created utilising the gene lists.

4.3.7 Bioinformatics Validation

To validate the bioinformatic analysis employed, a panel of genes identified as differentially expressed were selected for primer design for subsequent RT-qPCR analysis to confirm whether they are in fact up or downregulated in the cells. Nine genes **(Table 4.15)** were used in this confirmatory experiment, originally eleven gene primers were designed but, after primer testing and optimisation two of the primers were proven to not work specifically. Alongside the differentially expressed genes selected, primers were also designed from housekeeping genes for the normalisation.

Table 4.15: Gene name of the selected genes with their corresponding log2 fold changes (L2FC) in the *ZFYS* and *ZFYL* transfections respectively.

Gene Name	*ZFYS* L2FC	*ZFYL* L2FC
*WNT7*A	8.58	9.93
TMPRSS2	6.30	8.99
FGF3	5.98	8.72
IFIT2	-5.26	-6.92
RBMXL2	-3.76	-4.27
ZFX	-1.13	-1.47
SNORA7B	5.39	0
ENSG00000289202	3.74	0
SNORD3D	0	0

Listed in **Table 4.15** are the L2FC for both *ZFYS* and *ZFYL* from the transfection datasets, representing the expected RT-qPCR results if properly validated. The selected upregulated genes exhibit greater L2FC in the *ZFYL* data, further implying that *ZFYL* is a more active transcription factor despite lower TPM compared to *ZFYS*. This pattern holds for the downregulated genes as well. Numerous snoRNAs were differentially expressed, especially in the *ZFY* transfected lines, alluding to an unknown regulatory role. Notably, snoRNAs were excluded from downstream analyses as non-coding proteins were removed. While the selected snoRNAs showed

no changes with *ZFYL*, two of the snoRNAs increased with *ZFYS*. *SNORD3D* served as a control since both datasets showed no expression. Validation used two distinct RNA sets: Transfection Set 1 was submitted for sequencing, while the independent Transfection Set 2 was not sequenced.

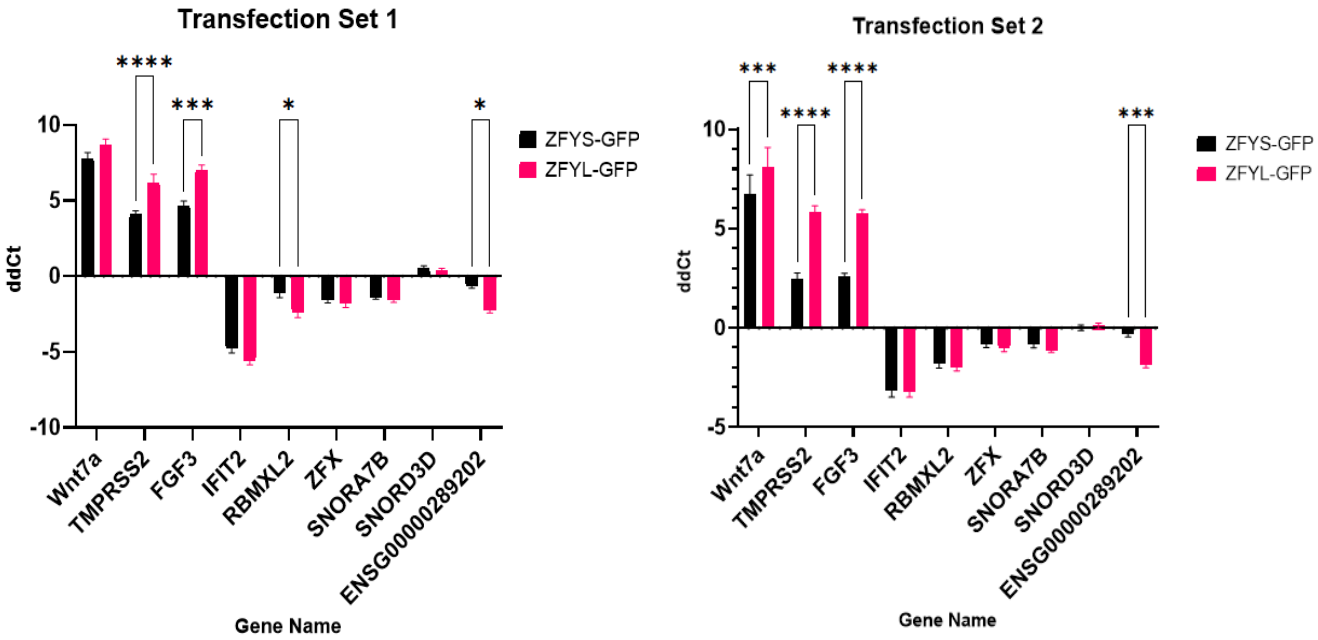


Figure 4.17: RT-qPCR data using the designed primers for the genes of interest selected from the RNA-Seq analysis datasets. ddCT shows the L2FC of the genes when normalised to the housekeeping gene, ACTB. * P ≤ 0.05, ** P ≤ 0.01, *** P ≤ 0.001, **** P ≤ 0.0001.

Highlighted in **Figure 4.17** is that both sets of transfections corroborate with each other and show very similar results. All the genes expected to be upregulated (not including the snoRNAs), and more so by *ZFYL* are showing this and these include *WNT7*A, *TMPRSS2*, and *FGF3*. Again, the downregulated genes, which are expected to be more greatly downregulated by *ZFYL* are confirming the *in-silico* analysis. With regards to the snoRNAs, these validate the *in-silico* analysis less. *SNORD3D* is not expressed in either the transfection line and this is confirmed by RT-qPCR, but the remaining snoRNAs do not corroborate the analysis. *SNORA7B* and ENSG00000289202 unfortunately are showing a downregulation by both forms of *ZFY* which is unexpected but the same is seen across both transfection sets. Genomically, most SNORA and SNORD genes are located within the introns of other protein-coding genes (Zimta *et al.*, 2020). This means they usually don't have their own poly-A tail and are instead released from the mRNA of the "host" gene during splicing. So, the signal seen for any given snoRNA is a complex mix of fully processed snoRNAs, potentially also a contribution from partially processed "host" mRNAs and

may vary depending on how the RNA and cDNA are prepared, whether there is selection for poly-A RNA or whether the cDNA synthesis is primed with oligo-dT or random hexamers (Zimta *et al*., 2020). This could provide some explanation as to why there is a lack of confirmation through PCR.

4.3.8 Gene Ontology and Enrichment Analysis

Enrichment analysis was performed using Reactome, and many data set variations were used in this analysis. See *supplementary* **Table 7** for the genes associated with the highlighted enriched pathways mentioned below.

The first analysis looked at the upregulated pathways affected by both *ZFYS* and *ZFYL* but not *ZFX*. The *ZFX* differentially expressed data was excluded from this analysis. This means genes with a L2FC >1 were selected to identify potential activated pathways.

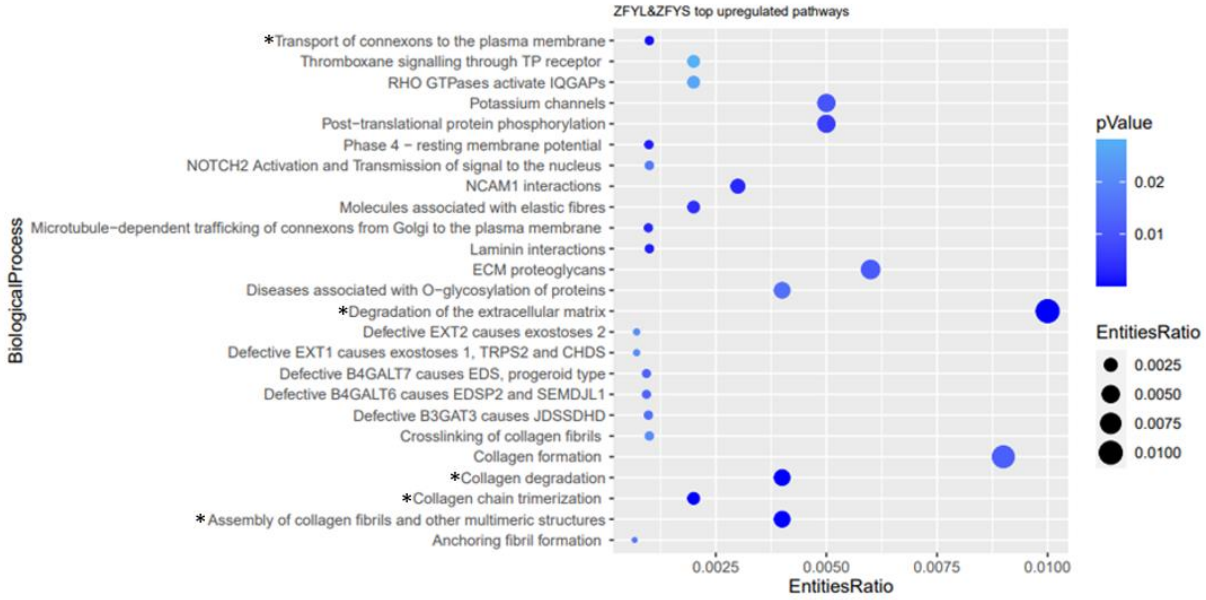


Figure 4.18: Enrichment analysis plot showing the pathways controlled by *ZFYS* and *ZFYL* upregulated genes. This plot only shows the pathways identified with a p-value <0.05, the * represents the pathways which also had an FDR < 0.05.

Figure 4.18 shows that although many of the genes upregulated by both *ZFYS* and *ZFYL* act on many biological pathways with a significant p-value, they do not pass the FDR significance. This means they cannot necessarily be trusted as target pathways of *ZFY*. The significant pathways include many collagen-associated biological processes and extracellular matrix processes. Many of the genes falling in these pathways are collagen type proteins, matrix metallopeptidases and other matrix related genes such as ADAM15 which is an enzyme involved in the sperm epididymal

maturation acrosome reaction (Pastén *et al*., 2014). Collagen serves as a scaffolding protein within the extracellular matrix (Siu & Cheng, 2004);(Siu & Yan Cheng, 2008). In adult mammalian testes, the Sertoli and germ cells are closely associated with a modified form of extracellular matrix known as the basement membrane. Research has shown the vital importance of the extracellular matrix in providing support to Sertoli and germ cells within the seminiferous epithelium, particularly in relation to spermatogenesis (Siu & Yan Cheng, 2008). This suggests a potential explanation as to why collagen-related pathways are being activated.

Following selecting just *ZFYL*-affected genes, the enrichment analysis was performed again.

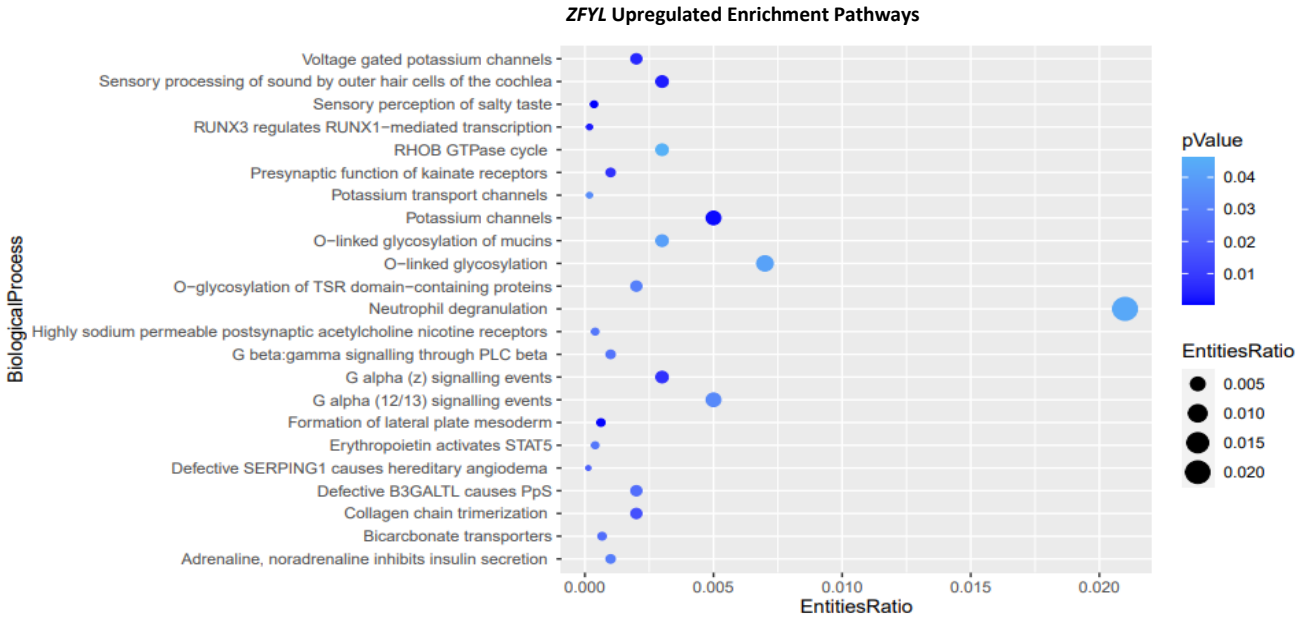


Figure 4.19: Enrichment analysis plot showing the pathways controlled by just *ZFYL* upregulated genes. This plot only shows the pathways identified with a p-value <0.05, the lack of * demonstrates that none of the highlighted pathways passed FDR < 0.05.

As seen above, **Figure 4.19** shows that many pathways are targeted by the *ZFYL* differentially expressed genes with a significant p-value, but none of these pathways pass the FDR significance. Many of the pathways seem to be linked to ion channels, post-translation modification & G alpha signalling to state a few. Ion channels are vital to sperm physiology due to their roles in sperm-cell differentiation and maturation, motility activation, chemotaxis towards the oocyte, and fertilisation to name a few (Pinart, 2022). Potassium regulation has been linked to spermatozoa volume regulation which fertility depends on greatly (Barfield *et al*., 2005). Additionally, research has demonstrated the involvement of the G protein-coupled receptor (GPCR) superfamily in regulating ion-water balance in the epididymis, facilitating the development of efferent ductules, forming the blood-epididymal barrier, and

promoting sperm maturation (D. Zhang *et al*., 2020). Initially, these pathways seemed surprising; however, upon closer investigation, their relevance to spermatogenesis and sperm function becomes somewhat evident.

However, when looking specifically at *ZFYS* upregulated pathways a different story is seen. Using Reactome, pathways of significance were identified for the *ZFYS*-identified genes.

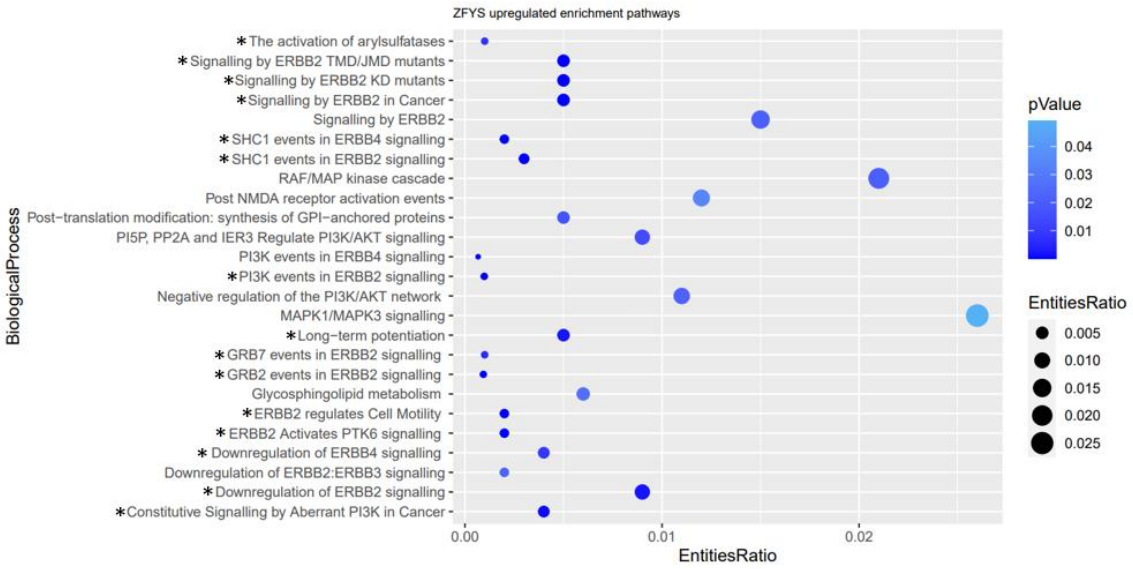


Figure 4.20: Enrichment analysis plot showing the pathways controlled by just *ZFYS* upregulated genes. This plot only shows the pathways identified with a p-value <0.05, the * represents the pathways which also had an FDR < 0.05.

Unlike the *ZFYL* Reactome pathways, many of the pathways identified in this *ZFYS* enrichment analysis pass FDR significance thresholds, as shown in **Figure 4.20**. These more robust p-values and FDRs indicate true biological pathways impacted by *ZFYS*. Numerous significant pathways centre on ERBB2 signalling, well-established in cancer pathology and drug targeting. Intriguingly, ERBB2 itself does not emerge among the differentially expressed genes, implying potential upstream or downstream modulation. PI3K represents another enriched cancer-associated cascade, critical for cell cycle control and thus proliferation (Rascio *et al*., 2021). Though not directly dysregulated, the enrichment of these canonical oncogenic programs suggests *ZFYS* may broadly influence tumorigenesis by altering key signalling nodes.

Finally, of the 206 genes identified to be upregulated by both *ZFYS* and *ZFYL* as well as *ZFX*, pathway analysis was performed to identify any major pathways affected by both isoforms and homologues.

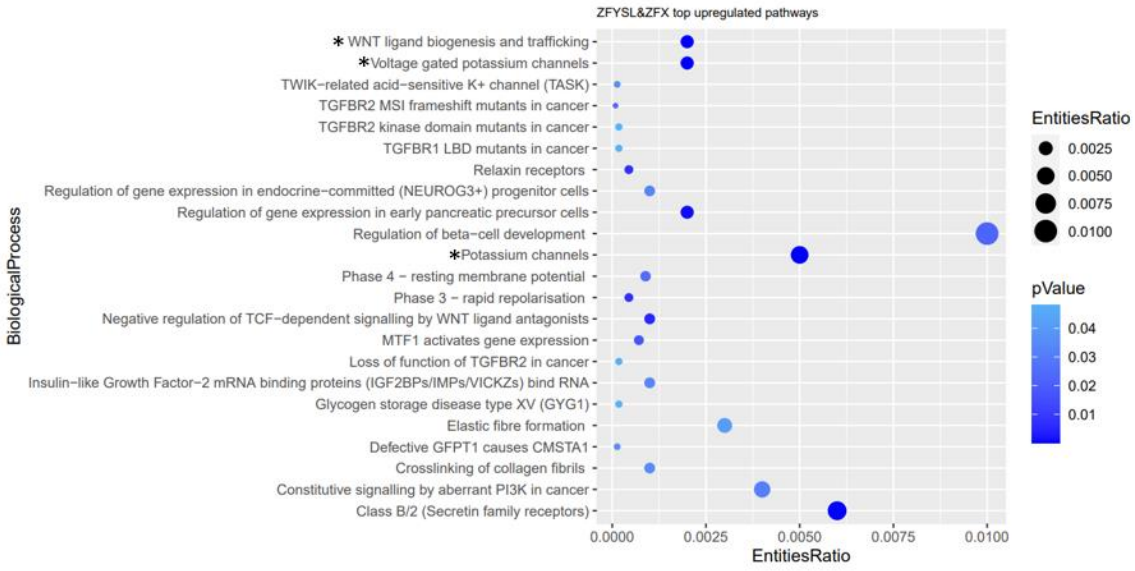


Figure 4.21: Enrichment analysis plot showing the pathways controlled by just *ZFYL*, *ZFYS* and *ZFX* upregulated genes. This plot only shows the pathways identified with a p-value <0.05, the * represents the pathways which also had an FDR < 0.05.

From **Figure 4.21** above, only three of the pathways are truly significant, and one is particularly interesting. All three ZF* seem to be targeting *WNT* signalling, with *ZFYS* and *ZFYL* upregulating these specific *WNT* proteins; *WNT*7A, *WNT*7B, *WNT*9A, *WNT*4, *WNT*11, *WNT*3A, and *WNT*5B. *WNT* proteins have been shown to be expressed within the testis and have been linked to spermatogenic roles. *WNT*7A for example has been found to be expressed in both early and late spermatids in humans, with *WNT*7A also being seen in pachytene spermatocytes and late spermatids in mice (The Human Protein Atlas, 2024a)(Takase & Nusse, 2016). Another example is that *WNT*3A has also been shown to be expressed in human testes in spermatogonia and spermatocytes (Young *et al*., 2020). *ZFX* upregulated all the same *WNT* proteins found in the *ZFY* data with the addition of *WNT*3. *WNT* signalling plays a major role in adult tissue homeostasis and therefore plays a role in many diseases including cancer (J. Liu *et al*., 2022). The other two pathways noted to be significant are related to potassium channels which participate in many critical biological functions such as fertility and play important roles in disease (Tian *et al*., 2014).

4.3.9 Cancer Correlation Analysis

Using the CCLE cancer expression data, the next analysis looked at *ZFY* in relation to cancer, since it is believed there is possibly some form of relationship. This data set contains 1,829 different cancer cell lines, however, only 1,393 have expression

data available. These cell lines cover a range of cancer types ranging from brain cancer to prostate cancer. Of the 1,829 cancer cell lines, 614 are female, 770 are male and 445 are unknown. Initially, the project had a particular interest in *ZFY* expression in Head and Neck cancer since a previous student showed possible *ZFYS* expression in a head and neck cancer cell. Looking more specifically at Head and Neck Cancer, the following data in **Table 4.16** was available.

Table 4.16: Head and Neck cancer cell line data available from CCLE for informatics analysis.

HEAD AND NECK CANCER	Total	Female	Male	Unknown
	79	13	49	17
Average Age (yr.)	59	58	59	NA
Number of Primary cancers	40	10	29	1
Number of Metastatic cancers	12	2	9	1

As *ZFY* is located on the Y chromosome, the focus was on the male cancer cell lines. Out of the 49 male HNC cell lines, only 44 of them have available expression data. A majority of the cancers were primary cancer cell lines, and the average age was calculated to be 59 yr. It was also noted that all of the 79 HNC cell lines, only 56 had expression data, but all 56 cell lines expressed *ZFX*. After extracting the *ZFY* expression data from each cell line, it was noted that only 22 of the male head and neck cancers expressed some form of *ZFY*. However, this data is not specific for long or short.

To explore the expression of *ZFY* across various cancer cell types, TPM data was gathered to observe the diversity present among them. TPM is a way of normalising gene expression, where for each transcript the number of reads mapped is divided by the transcript's length, giving a normalised transcript-level expression. Looking at the RNA-Seq TPM expression, *ZFYS* has a TPM of 30708.76 whilst *ZFYL* has a TPM of 17355.82. These TPMs are extremely high, with *ZFYS* being even higher due to its shorter gene length compared to *ZFYL*. So, that begs the question is the large number of differentially expressed genes seen in the RNA-Seq data due to the very high expression of *ZFY*? Are all these differentially expressed genes being acted on by *ZFY* or is it as a result of downstream regulation? Does it matter how much *ZFY* is present, is it just a case of "on" or "off" if any *ZFY* is present?

As seen in **Figure 4.22**, *ZFY* is being expressed in some cancer cell lines, however, from this data we do not know if the cancers are expressing *ZFYS* or *ZFYL* but based on the RT-PCR in Chapter 3 and that *ZFYS* is mostly testis specific, it is presumed that they are likely expressing *ZFYL*. The absolute levels of *ZFY* expression are low and are much lower than the overexpression data results obtained in our experiments. Thus, the potential *ZFY* target gene expression in these cancers may also be much lower compared to any expression data we have.

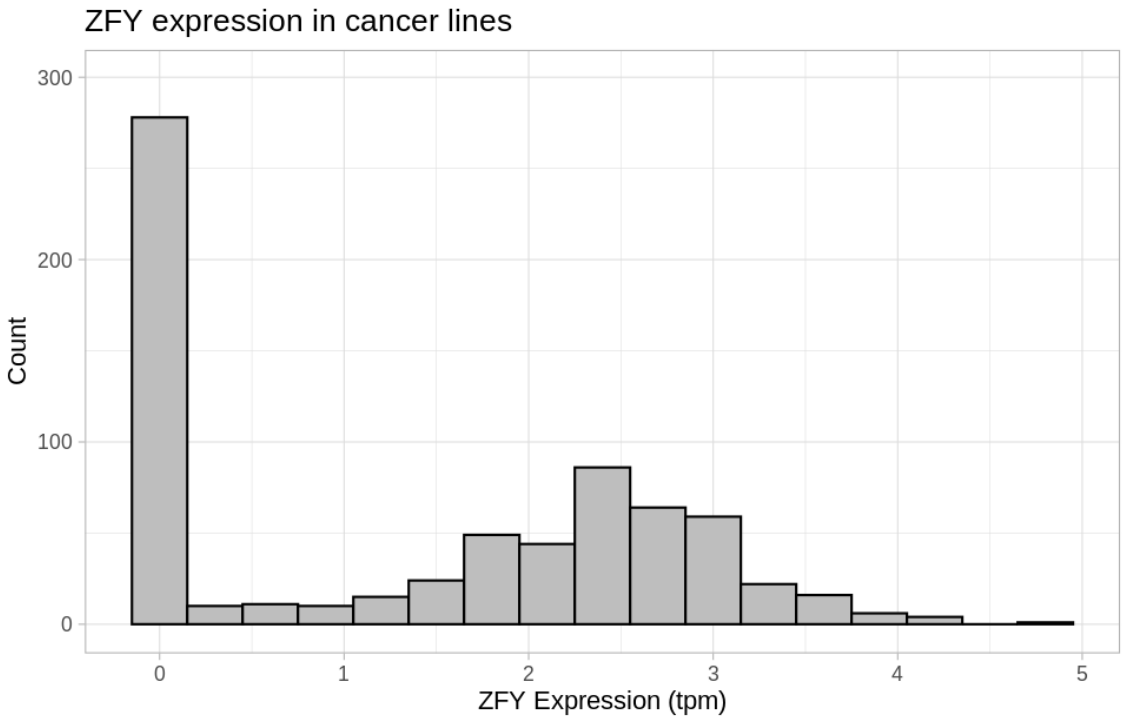


Figure 4.22: TPM expression for *ZFY* in the male cancer cell lines in the CCLE database. Male cancer cell lines only. The transcript per million is plotted on the X-axis against the number of cancer cell lines on the Y-axis.

When looking at *ZFX* expression **(Figure 4.23)** in the cancer cell lines, most of the cancers both male and female express *ZFX*, however, the range is similar to that of *ZFY*, ranging between 1 TPM and 6 TPM. However, the expression is much more consistent compared to that of *ZFY*.

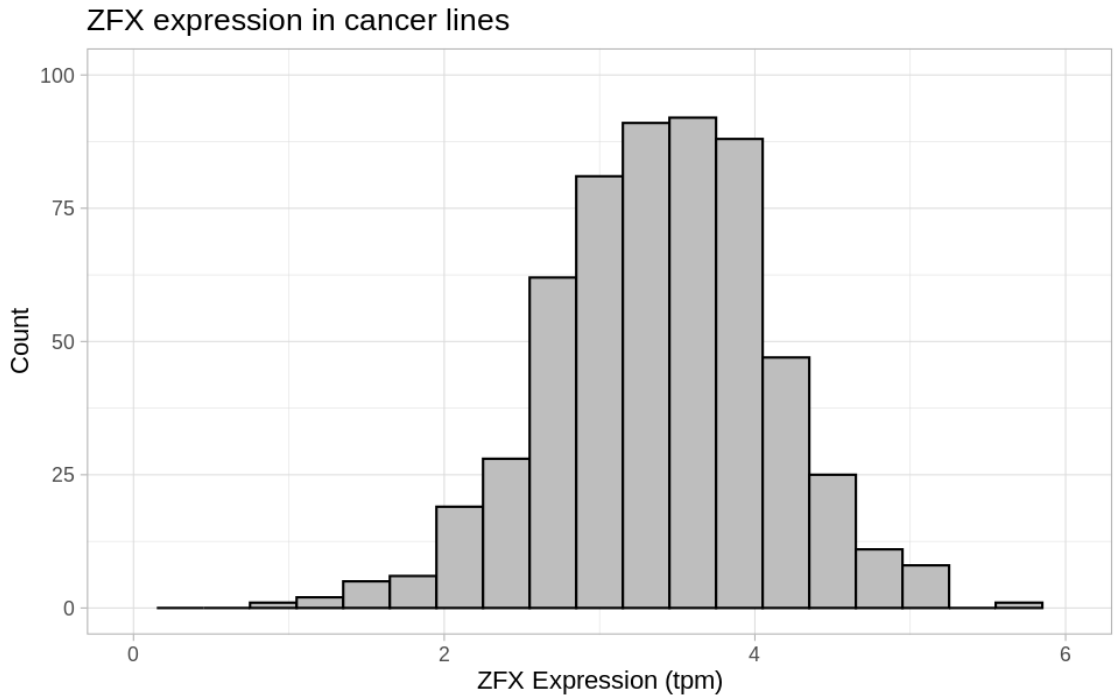


Figure 4.23: TPM expression of *ZFX* in both male and female cancer cell lines available from the CCLE database. The transcript per million is plotted on the X-axis against the number of cancer cell lines on the Y-axis.

In the RNA-Seq data presented in this thesis, it was noted that *ZFX* was downregulated in the presence of both *ZFYS* and *ZFYL*, however, the TPM level is still higher than what is seen in the above cancer cell lines (CCLE data). The RNA-Seq data indicated that the TPM of *ZFX* in *ZFYS* and *ZFYL*-expressing cells was 20.08 and 16.84, respectively. While this expression is significantly lower than the *ZFY* TPM observed after ZFY transfection into the cells, it remains higher than the levels detected in the cancer cell line data.

GEPIA2 was used to look at the gene expression profile across all their tumour samples and paired normal tissues. GEPIA2 uses a different tumour data set compared to the CCLE used above, but it can give a general idea of which tumour types seem to be expressing *ZFY*.

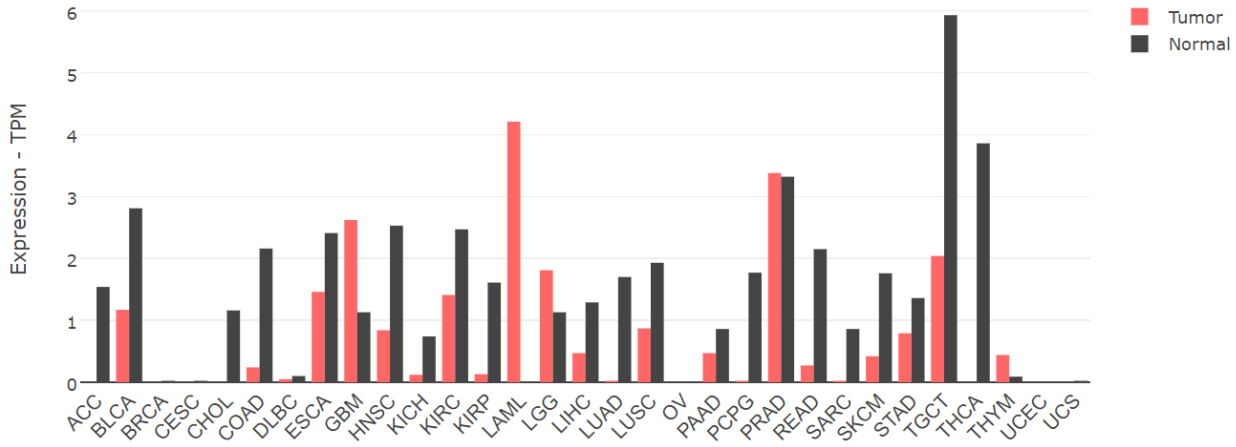


Figure 4.24: GEPIA2 *ZFY* expression data in both tumour and normal tissues ranging across a variety of tissue types. The Y-axis plots the TPM value for *ZFY* expression in each tissue.

The highest *ZFY* expression in a tumour sample is found in LAML (acute myeloid leukaemia) **(Figure 4.24)**, a type of blood cancer where there is the presence of excess immature white blood cells or myeloid cells. The TPM is still relatively low at ~4 but is greater than that of other tumour cells. However, the greatest *ZFY* expression in normal tissues is in TGCT (Testicular Germ cell tumours) and this TPM (~6) is the greatest compared to all other tissue types both normal and cancerous.

Knowing this, a correlation analysis was performed to identify if any of the genes identified from the RNA-Seq data are also present in the cancer cell lines which also express some form of *ZFY*. The *ZFY* expression data was extracted including the p-values and p.adj values. The data collected was then filtered to a p.adj value of < 0.05. Originally, the Pearson R value was filtered at ±15%.

Originally, the entire database was included in the analysis, investigating the male, female and unknown gender cancer cell lines as well as looking at the genes identified to be differentially expressed by both *ZFYS* and *ZFYL*. And as mentioned above Person R was filtered to ±15%.

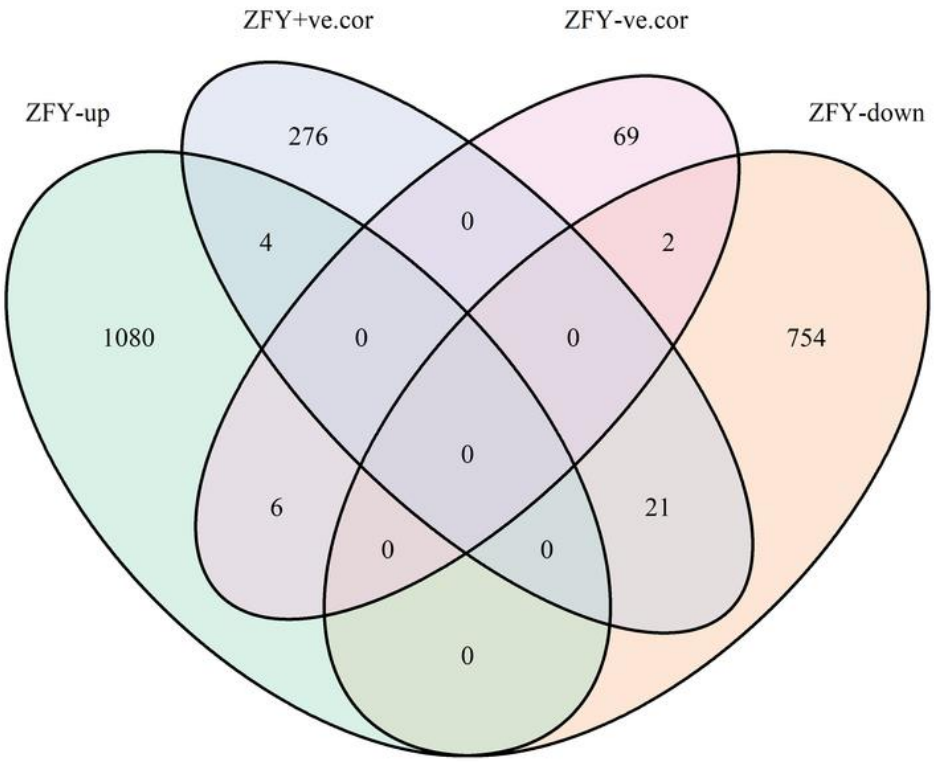


Figure 4.25: A Venn diagram showing the correlation between the RNA-Seq differentiated genes and the genes identified to be expressed in the cancer cell data possibly related to *ZFY* expression. Venn diagram consists of all the available cell lines with the Pearson R correlation set at ±15% and a p.adj filtered to <0.05.

As seen in **Figure 4.25**, it is clear that there is no correlation between the RNA-Seq dataset and the cancer cell line dataset. Only 4 genes were identified to be correlating in an upregulated manner, and only 2 correlating in a downregulated manner. This led to changing the filtering parameters.

After the data was filtered for male-only cell lines, the Pearson R filter was changed to ±10% (>0.1 and <-0.1). This left a total of 6,078 genes from the correlation analysis. During this analysis, groups were assigned as follows; *ZFYS* and *ZFYL* shared genes, *ZFYL* only and *ZFYS* only genes.

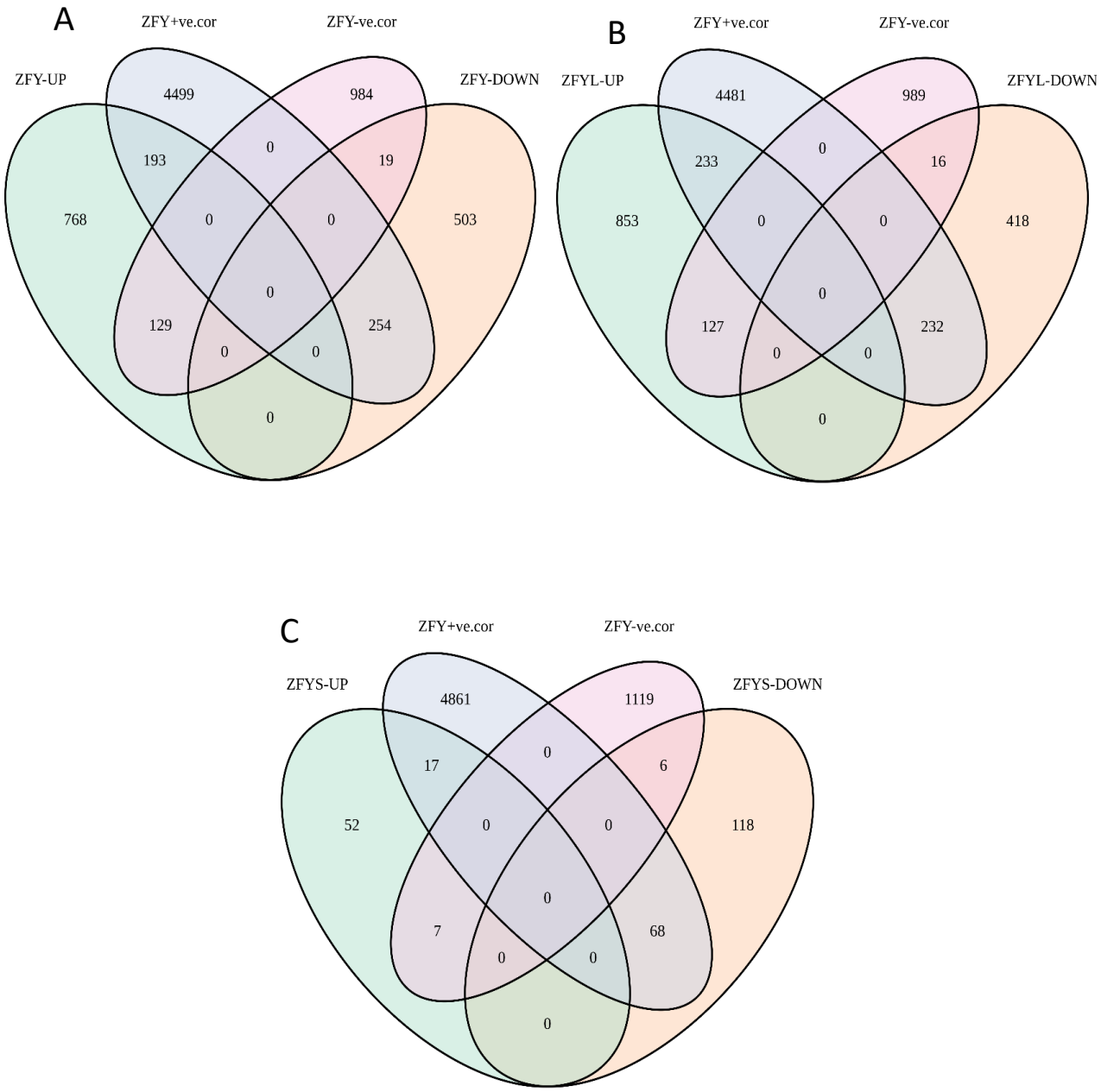


Figure 4:26: Venn Diagrams demonstrating the correlation between the three RNA-Seq datasets to the cancer correlation dataset at a ±10% filter. A: *ZFYS* and *ZFYL* common genes, **B:** *ZFYL* only genes & **C:** *ZFYS* only genes.

As seen in **Figure 4.26**, there are more correlating genes between datasets when applying a lower Pearson R threshold, indicating weak concordance. Specifically, 193 common upregulated genes emerged in both *ZFYS* and *ZFYL* analyses **(Figure 4.26A).** However, the *ZFYL*-restricted search yielded substantially more overlaps (233 genes) with RNA-Seq, highlighting stronger *ZFYL*-cancer data correlations. Far fewer correlating genes surfaced with *ZFYS* (17), though the smaller *ZFYS* dataset likely impacts this. Despite low-level *ZFY* expression (1-5 TPM) in all the male cancer

lines, a negligible correlation exists between the cancer profiles and *ZFYS* RNA-Seq. By loosening the correlation stringency, the gene overlaps increased but this also risks including falsely correlating genes lacking biological relevance.

Looking at the correlating genes of *ZFY* in the cancer cell line data it was found that a high number of the +ve correlating genes are located on the Y chromosome. After taking a closer look at the top 20 +ve genes identified as correlating to *ZFY* in the cancer dataset, 18 are located on the Y chromosome **(Table 4.17)**. Therefore, the genes correlating with *ZFY* in the cancer cell line data might be the result of Y gene expression rather than *ZFY* specifically. This could explain why the DE genes and correlation genes do not overlap as much as one would have thought. So, even though there are some genes whose expression level directly correlates with *ZFY* expression level, the effect is small (10%). Other genes controlled by *ZFY* (RNAseq) may be "on" or "off" but the levels do not correlate with the level of *ZFY*. Any Y chromosome gene directly controlled by *ZFY* would not appear in this data since HEK293 cells are female and do not have a Y chromosome, therefore are not represented in the correlation analysis.

Table 4.17: The top 20 highest correlating genes. The table highlights their corresponding Pearson R-value, and the yellow highlighted rows represent the genes not found on the Y chromosome.

Gene Name	Pearson R	Y chromosome (Y/N)
RPS4Y1	0.891685436	Y
DDX3Y	0.83286187	Y
PRKY	0.815367812	Y
UTY	0.801096988	Y
USP9Y	0.785331051	Y
KDM5D	0.767756627	Y
EIF1AY	0.762723158	Y
NLGN4Y	0.588353735	Y
TMSB4Y	0.555203993	Y
DHRSX	0.517796892	Y
ZBED1	0.51560238	Y
AKAP17A	0.491856345	Y
SLC25A6	0.483214945	Y
GTPBP6	0.450333614	Y
PPP2R3B	0.36609227	Y
ASMTL	0.348864219	Y
CD99	0.347906748	Y
LRRC41	0.330371183	N
PLCXD1	0.325485188	Y
MTMR9	0.313868379	N

As seen in **Table 4.17** many of the genes with high correlating values are located on the Y chromosome, but many of the genes are also genes located on both the X

chromosome as well. Furthermore, the top 20 genes correlate >30% which is much higher and stronger than the values collected from the correlation analysis with this thesis dataset. *LRRC41* and *MTMR9* are the only genes identified to not be expressed on the Y (X) chromosome. *LRRC41* is located on chromosome 1, whilst MTMR9 is located on chromosome 8.

As mentioned above 17 upregulated genes were identified to be correlating between the cancer dataset and the *ZFYS* dataset. The top 4 highest correlating genes of the 17 were identified and are stated in **Table 4.18** below. It is noted that these genes still have a relatively low correlation ranging between 15% and 17%.

Table 4.18: Pearson R data for the top four highest *ZFYS* correlating genes.

Gene Name	L2FC	Pearson R
TICAM2	4.50	0.17
ZBED6	2.99	0.16
RASAL3	3.03	0.15
LONRF1	1.10	0.15

TICAM2 is a TIR domain-containing adaptor molecule which facilitates both the inflammasome response and IFN-inducible genes (Funami *et al*., 2015). Mice lacking *TICAM2* demonstrated heightened protection against severe systemic inflammation and multi-organ injury due to mucosal damage (R. Lin *et al*., 2020). Zinc finger, BED-type containing 6, *ZBED6*, is a transcription factor that represses IGF2 which impacts development, cell proliferation and growth (Akhtar Ali *et al*., 2015). Mouse *ZBED6* targets have been correlated to developmental disorders and cancers with human *ZBED6* binding linked to genes controlling transcription, macromolecule synthesis and apoptosis (Akhtar Ali *et al*., 2015);(X. Wang *et al*., 2013). *RASAL3* is a critical member of the GAP family predominantly expressed on T-lineage cells (Muro *et al*., 2018). They function as a negative regulator of TCR-induced MAPK activation in a T cell line with further roles in neutrophil responses (Muro *et al*., 2018);(Saito *et al*., 2021). Finally, *LONRF1* has been shown to have roles in oxidative damage response and tissue remodelling during wound healing (D. Li *et al*., 2023). These four genes have strong roles in the immune response which is particularly heightened in cancer during the early stages of tumour initiation. This activation aims to mount a protective effector immune response, with the goal of eliminating immunogenic cancer cells (H. Gonzalez *et al*., 2018). There is no evidence suggesting that any of these correlating genes have a role during spermatogenesis. Whether these proteins are potential biomarkers of *ZFYS* requires further investigation especially due to the low correlation.

Using these four genes as possible *ZFYS* biomarkers, further analysis was performed to identify potential *ZFYS* expression in the cancer cell line data. When looking at which cancer cell lines express *ZFY* and the identified four possible biomarkers the expression of these was set to >=1 cutoff.

Table 4.19: This table presents the cancer cell types and the percentage that expresses both *ZFY* and each biomarker above the >= 1 cutoff filter. The number in the bracket is the total number of samples in the dataset.

Cancer Type	*TICAM2 + ZFY*	*ZBED6 + ZFY*	*RASAL3 + ZFY*	*LONRF1 + ZFY*
Bile Duct Cancer (10)	20%	30%	0%	50%
Bladder Cancer (20)	35%	50%	0%	45%
Bone Cancer (16)	31%	81%	38%	81%
Brain Cancer (50)	68%	68%	0%	73%
Colon/Colorectal Cancer (40)	13%	30%	0%	35%
Embryonal Cancer (1)	0%	0%	0%	0%
Endometrial Cancer (2)	50%	50%	0%	100%
Engineered (1)	0%	0%	0%	0%
Oesophageal Cancer (23)	4%	9%	0%	13%
Eye Cancer (4)	0%	0%	0%	25%
Fibroblast Cancer (18)	83%	83%	0%	83%
Gallbladder Cancer (2)	0%	0%	0%	0%
Gastric Cancer (23)	9%	30%	0%	35%
Head and Neck Cancer (44)	**27%**	**32%**	**0%**	**50%**
Kidney Cancer (21)	48%	43%	0%	48%
Leukaemia (63)	**27%**	**63%**	**70%**	**73%**
Liver Cancer (21)	24%	48%	0%	52%
Lung Cancer (148)	22%	51%	0.7%	51%
Lymphoma (50)	36%	60%	68%	70%
Myeloma (13)	23%	38%	38%	38%
Neuroblastoma (17)	0%	59%	0%	76%
Pancreatic Cancer (28)	18%	14%	0%	14%
Prostate Cancer (9)	11%	56%	22%	56%
Rhabdoid Cancer (7)	14%	57%	0%	71%
Sarcoma (10)	30%	80%	0%	80%
Skin Cancer (51)	47%	65%	0%	78%
Teratoma (1)	0%	0%	0%	100%
Thyroid Cancer (6)	33%	33%	0%	33%

Highlighted in **Table 4.19** are numerous cancer cell lines of the bladder, bone, brain, fibroblast, kidney, leukaemia, lymphoma and sarcoma origin co-expressing *ZFY* alongside one of the four selected biomarkers potentially hinting at *ZFYS* expression. In head and neck cancers, the fraction of cell lines with concurrent *ZFY* and biomarker expression ranges from 0-50%, hinting at variable coordinate regulation. However, the low correlation value warrants cautious interpretation. Intriguingly, no head and neck cancer line displays joint *ZFY* and *RASAL3* positivity, potentially attributable to *RASAL3*'s integral immune function. While the trends imply possible *ZFYS* interconnectivity with oncogenic drivers, the variability across cancer types highlights context specificity.

Upon further examination of the Head and Neck cell lines, a varying number of cell lines displaying *ZFY*, and the potential biomarkers were noted. Breaking down the data for Head and Neck Cancer, 12 of the 44 lines with expression data were found to express both *ZFY* and *TICAM2* simultaneously.

Table 4.20: The head and neck cancer cells identified to express both *ZFY* and *TICAM2*.

Head and Neck Cancer	**12**
Squamous Cell Carcinoma 3x primary 3x metastatic	6
Squamous Cell Carcinoma, buccal mucosa 1x primary	1
Squamous Cell Carcinoma, laryngeal 1x metastatic	1
Squamous Cell Carcinoma, oral 3x primary	3
Squamous Cell Carcinoma, tongue 1x metastatic	1

TICAM2 was identified as the highest correlating gene in the analysis with a Pearson R-value of 0.17 (17%), and 12 cell lines (27%) have been identified to express *TICAM2* alongside *ZFY* as shown in **Table 4.20** above. The cell lines are broken down into subsections based on their location i.e., tongue or laryngeal. Of the 12 cell lines, 7 were identified as primary cancer and 5 metastatic. The average age of men was calculated as 55 years.

The same analysis was performed to look at cell lines looking at both *ZFY* expression and *ZBED6*, the second highest correlating gene identified.

Table 4.21: The head and neck cancer cells identified to express both *ZFY* and *ZBED6*.

Head and Neck Cancer	14
Squamous Cell Carcinoma 1x primary 3x metastatic	4
Squamous Cell Carcinoma, buccal mucosa 1x primary	1
Squamous Cell Carcinoma, laryngeal 2x metastatic	2
Squamous Cell Carcinoma, oral 5x primary 1x metastatic	6
Squamous Cell Carcinoma, tongue 1x metastatic	1

Though *ZBED6* had a slightly lower correlation R-value, 14 (32%) of the head and neck cancer cell lines were identified as potentially expressing *ZFYS*. Of the 14 cell lines in **Table 4.21**, half were primary, and half were metastatic, with the average age of men being 52 years. The majority of cases were specifically oral head and neck cancer.

RASAL3 was a correlating gene identified, however, no Head and Neck cancers were identified to express both *ZFY* and *RASAL3*. But this could be because of the low correlation between the datasets.

However, LONRF1 a gene identified to have the same correlation R as *RASAL3* was found to be expressed alongside *ZFY* in 50% of the head and neck cancers male cancer cell lines.

Table 4.22: The head and neck cancer cells identified to express both *ZFY* and *LONRF1*.

Head and Neck Cancer	22
Squamous Cell Carcinoma 3x primary 5x metastatic	8
Squamous Cell Carcinoma, buccal mucosa 2x primary	2
Squamous Cell Carcinoma, laryngeal 2x metastatic	2
Squamous Cell Carcinoma, oral 6x primary 1x metastatic	7

Squamous Cell Carcinoma, sinus 1x metastatic	1
Squamous Cell carcinoma, tongue 1x Primary 1x Metastatic	2

Out of the 22 cancer cell lines identified in **Table 4.22** with *ZFY* and *LONRF1* expression, 13 are primary and 9 are metastatic and the average age was identified to be 55 yrs. Like with *ZBED6* many of the cell lines are oral in origin with 7 lines identified.

Filtering for cell lines that express *ZFY* as well as, *TICAM2*, *ZBED6* and *LONRF1* then the number of head and neck cancers is reduced to 7 out of the 44 male cell lines with expression date, equating to 16% of the head and neck cancers.

Based on this analysis, potential biomarkers specific to *ZFYS* expression have been pinpointed. Consequently, it is anticipated that these cancer cell lines exhibit the short form. Nonetheless, considering the low correlation, it would not be unexpected if this assumption proves false.

Since there were other cancer types in the dataset with a greater number of cell lines expressing *ZFY* and one of the potential biomarkers, it was decided to focus on another specific cancer group. This cancer of interest was leukaemia, due to three main reasons. One reason is that globally, the leukaemia disease burden is higher in males than females, with mortality also being greater in males (Cancer Research, 2024). In the UK, 40% of the leukaemia cases are female while 60% are males (Cancer Research, 2024). Nevertheless, it remains unclear whether the surge in prevalence could be attributed to Y chromosome loss. One of the most prevalent alterations observed in adult male blood cells is the mosaic loss of chromosome Y, closely linked to various hematopoietic and non-hematopoietic human diseases like leukaemia (Q. Zhang *et al*., 2022). Studies indicate that up to 60% of acute myeloid leukaemia cases exhibit mosaic loss of chromosome Y (Q. Zhang *et al*., 2022). Moreover, as depicted in **Figure 4.24**, the greatest expression of *ZFY* in cancerous tissues was observed in acute myeloid leukaemia, aligning with the extensive presence of *ZFY* expression in leukaemia cell lines documented in the CCLE database. Finally, the four prospective *ZFYS* biomarkers exhibit some form of immune system-related functionalities, which constitute a significant component of leukaemia.

Following this interest, enquiries about cell lines were made. Dr Giorgia Chiodin, a colleague of Dr Tim Fenton who was able to provide the following Leukaemia Cell lines stated in **Table 4.23** which were suitable for the RT-qPCR experiments to further look into the possible *ZFYS* expression in Leukaemia. A total of three Leukaemia cell lines were used with varying expressions of *ZFY* and the identified possible

biomarkers of *ZFYS*. As with the previously collected expression data collected for *ZFY* and the biomarkers, the expression data for the Leukaemia cell lines were obtained.

Table 4.23: Leukaemia Cell Lines collected for experimentation. Including information regarding the age (Yr), subtype and expression levels (transcript per million, TPM) of the genes of interest.

Cell Line	Age (Yr)	Disease Subtype	ZFY (TPM)	*TICAM2* (TPM)	*ZBED6* (TPM)	*LONRF1* (TPM)
THP-1	1	Acute Myelogenous Leukaemia (AML), M5 (Eosinophilic/Monocytic)	2.695994	0.879706	2.250962	2.794936
MEC-1	61	Chronic Lymphoblastic Leukaemia (CLL), B-cell	2.367371	1.361768	1.367371	3.152183
U937	37	Acute Myelogenous Leukaemia (AML)	0	1.014355	1.95977	3.294253

U937 was used as a male negative control since there is no *ZFY* expression noted in this specific cell line as shown in **Table 4.23**, however, there is still an apparent expression of the selected biomarkers. But in U937 the highest expression is seen to be LONRF1, the least correlating of the chosen biomarkers.

THP-1 and MEC-1 have very similar *ZFY* expression, but their expression of the biomarkers varies. This could be due to the lack of correlation or suggests that *ZFYS* is not expressed. But like with U937, LONRF1 is the most expressed out of the four genes and is even expressed higher than *ZFY* itself.

Following, thawing and growth of the cells, RNA was extracted, and primers were synthesised. The primers were checked for specificity since cross-reactions between *ZFYS*, *ZFYL* and *ZFX* are all highly likely. After altering the annealing temperature, extension time and template concentration the best conditions for each primer to get single bands were identified. *LONRF1* was found to not be specific enough after multiple attempts of optimising and it was decided that these primers would not be included.

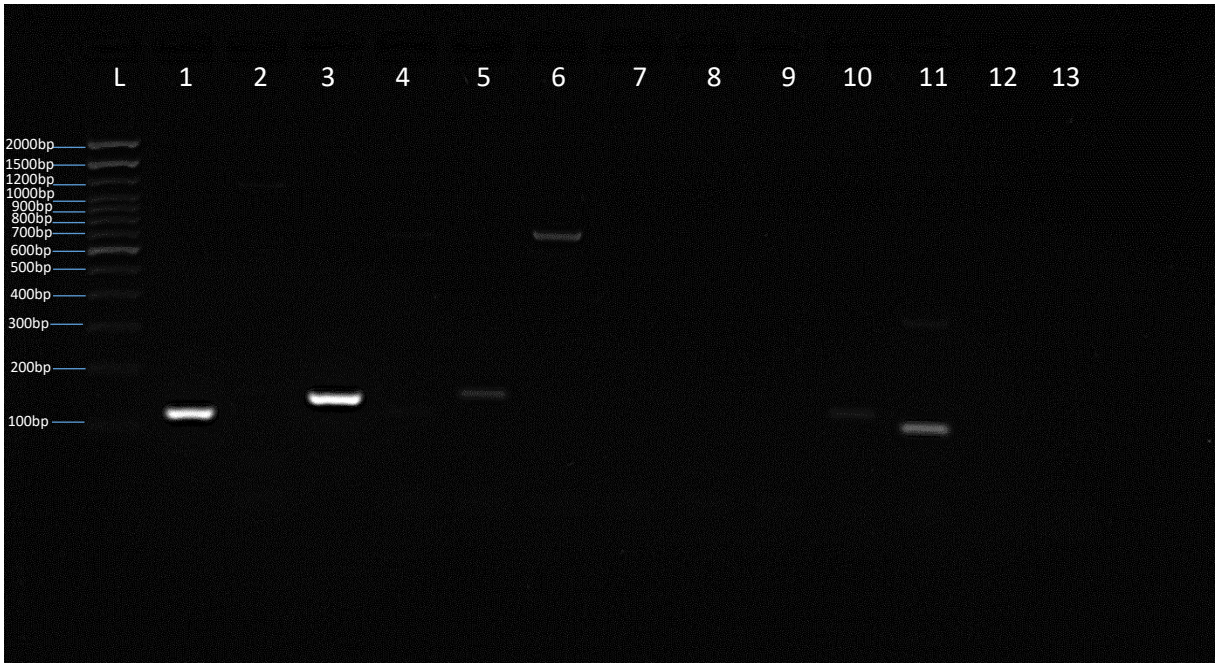


Figure 4.27: 2% agarose gel showing the final optimised PCR conditions for each primer pair designed. **L:** 100bp DNA Ladder **1:** *ZFYS* construct with *ZFYS* primers 126bp (52°), **2:** *ZFYS* construct with *ZFYL* primers 164bp (57°), **3:** *ZFYS* construct with *ZFY* both primers (62°), **4:** *ZFYL* construct with *ZFYS* primers (52°), **5:** *ZFYL* construct with *ZFYL* primers (57°), **6:** *ZFYL* construct with *ZFY* both primers (62°), **7:** *ZFYS* primers No-Template control (52°), **8:** *ZFYL* primers No-Template control (57°), **9:** *ZFY* both primers No template control (62°), **10:** *TICAM2* primers 140bp (56°), **11:** *ZBED6* primers 119bp (56°), **12:** *TICAM2* No-template control (56°), **13:** *ZBED6* No-template control (56°). For *ZFY* both primers the expected bp for *ZFYL* is 732bp and 159bp for *ZFYS*.

Lane 1 in **Figure 4.27** shows a strong band at the expected size (126bp) for *ZFYS* when the template DNA was a designed *ZFYS* construct, but when the *ZFYS* primers are used alongside a *ZFYL* DNA construct no band is present, as seen in lane 4. Lane 2 shows that the *ZFYL* primers are not producing a band in the presence of the *ZFYS* DNA construct, whilst in lane 5 a band at the expected size of 164bp is seen when a *ZFYL* DNA construct is used. The *ZFY* primers designed across the splice site to detect both *ZFYS* and *ZFYL*, produce 159bp and 732bp size bands respectively. These expected band sizes are seen in lane 3 and lane 6 with very little if any cross-reactivity at these conditions. The no-template controls for these three sets of primer pairs produce no bands, so no primer contamination is present. With regards to *TICAM2* and *ZBED6*, a clear band at the expected size of 140bp for *TICAM2* is seen in lane 10, with *ZBED6* also producing the correct size band. However, as seen in lane 11, a second very faint band is visible for the *ZBED6* primers, however, this was deemed to be very minimal. The no-template controls for both *TICAM2* and *ZBED6* also showed no contamination as no bands are identifiable.

Following primer optimisation, the *ZFY* splice site spanning primers were used to detect both *ZFYS* and *ZFYL* with the newly prepared Leukaemia cDNA.

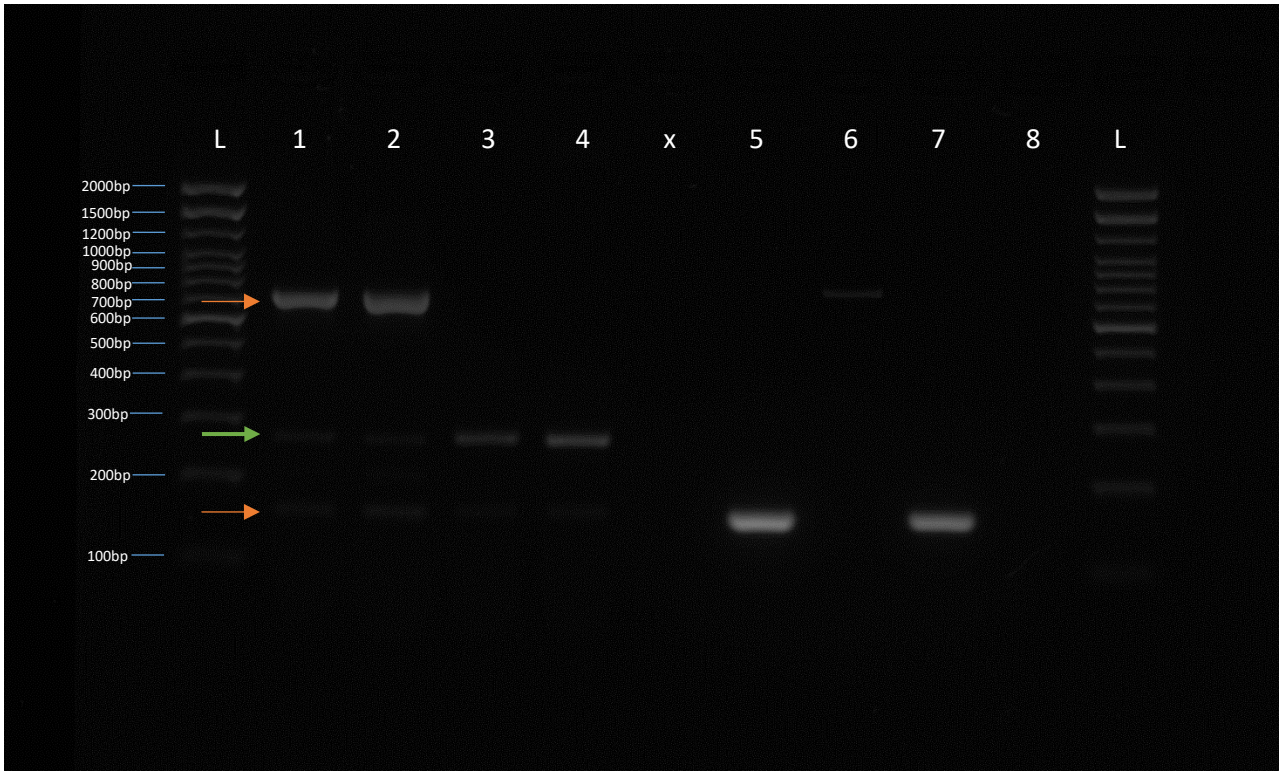


Figure 4.28: 2% agarose gel showing the PCR reactions performed using the *ZFY* primers able to detect both *ZFYS* and *ZFYL*. Leukaemia cDNA was used as the template DNA. **L:** 100bp DNA ladder, **1:** MEC-1 cDNA, **2:** THP-1 cDNA, **3:** U937 cDNA, **4:** HEK293 cDNA, **5:** *ZFYS* +ve control **6:** *ZFYL* +ve control, **7:** Mixture of *ZFYS* & *ZFYL* +ve control, **8:** No-template control.

The primers designed to span the splice junction are not functioning as expected as seen in **Figure 4.28**. Although MEC-1 (Lane 1) and THP-1 (Lane 2) show bands at the expected sizes of 732bp and 159bp (orange arrows), absent in the negative controls, an extra ~280bp (green arrow) band appears in both the male and female negative controls. This 280bp product is likely not contamination since the no-template control is clean but could represent primer cross-reactivity. The *ZFYS* and *ZFYL* positive controls display anticipated banding, however, the *ZFYS*/*ZFYL* mixed positive only exhibits *ZFYS*, probably because shorter amplicons outcompete longer ones during PCR even when the longer target (*ZFYL*) concentration exceeds the short target (*ZFYS*).

Unfortunately, the lack of *ZFYS* expression in these cell lines suggests unreliable Pearson R values that were filtered too leniently. Although some *ZFYS* are expressed, their expression levels are notably lower than those of *ZFYL*. The data instead indicates *ZFYL* expression, contrary to the initial hypothesis disproving the potential *ZFYS* biomarkers identified.

Following this, RT-qPCR was then performed with the above primers as well as previously designed primers for genes identified to be differentially expressed by *ZFYS* and *ZFYL*.

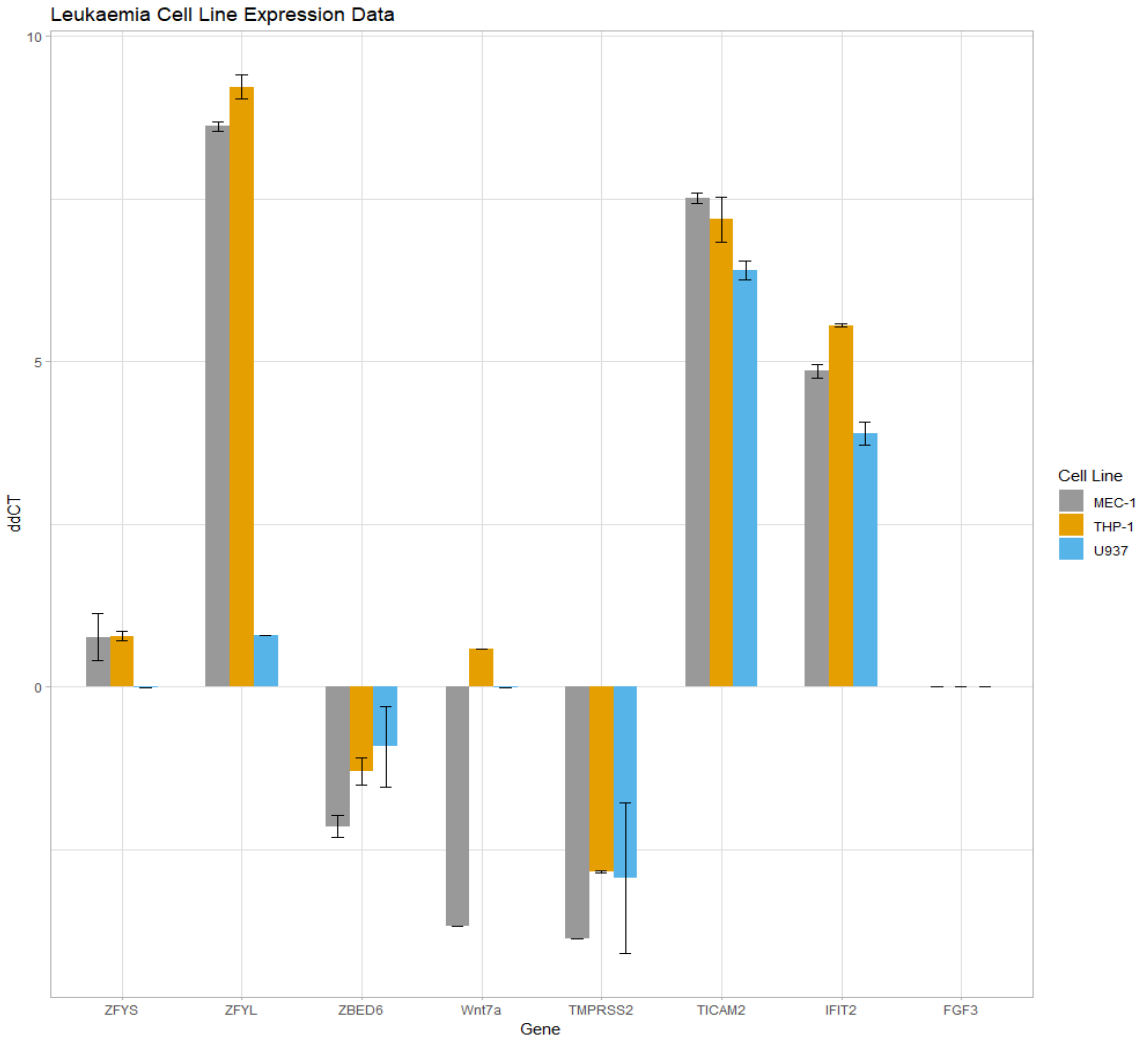


Figure 4.29: RT-qPCR graph produced using Graph-pads 2way-anova analysis. The ddCT was calculated and plotted for each gene in each Leukaemia cell line. The results were first normalised to the housekeeping gene ACTB and then normalised again to a female control cell line, HEK293. Error bars show standard error.

Figure 4.29 correlates to the *ZFY* expression previously seen in **Figure 4.28**. It is clear that these Leukaemia cell lines are expressing very small amounts of *ZFYS* or are possibly expressing *ZFX* which is being picked up non-specifically by the RT-qPCR, but this data falls in the cycle number range where the amount is so low that it is unreliable. However, *ZFYL* seems to be expressed in both MEC-1 and THP-1, with THP-1 expressing *ZFYL* at a slightly higher level, corresponding to the expression data from CCLE. *TICAM2* was identified to be correlating with *ZFYS*, but even with no *ZFYS* expression an upregulation of *TICAM2* is seen in all three leukaemia cell lines which does correspond to the expression data. However, this does suggest that *TICAM2* is not a reliable biomarker of *ZFYS*. On the other hand,

ZBED6 was expected to be upregulated in all three of the cell lines, yet there is a downregulation of *ZBED6*. It can also be seen in **Figure 4.29** that the genes *WNT7*A, *TMPRSS2* and *FGF3* identified as being upregulated in HEK293 cells in the presence of *ZFYS* and *ZFYL* are not following the same expression pattern in these leukaemia cells, suggesting a different mechanism of action in cancerous cells. The same can be said for *IFIT2* which was found to be downregulated in the RNA-Seq dataset but seems to be upregulated in leukaemia cells.

Overall, this indicates a lack of correlation between the cancer database and the RNA-Seq dataset, possibly ruling out *ZFY* as a potential cancer-testis antigen. Nevertheless, additional investigations would necessitate the use of a larger number of cancer cell lines to confirm experimentally.

4.4 Discussion

The function and significance of *ZFY* are not fully understood, with research interest declining due to it not being the sex-determining gene. In this chapter, we aimed to explore its potential roles and its possible role as a cancer-testis gene. Cancer-testis antigens are a subset of tumour antigens typically expressed exclusively in male germ cells within the testis under normal circumstances (Scanlan *et al*., 2002). However, under cancerous conditions, they are also detected in somatic tissues. *ZFYS* has been proposed as a cancer-testis gene because of its restricted expression in the testis, alongside potential aberrant expression in head and neck cancer. Identifying potential tumour antigens is pivotal for advancing immunogenic cancer therapies. The earliest recognised cancer-testis antigen, MAGE-A1, was discovered in the early 1990s, isolated from a panel of melanoma cell lines (van der Bruggen *et al*., 1991);(Scanlan *et al*., 2002). As of 2023, a total of 730 cancer-testis antigens, spanning 100 distinct gene families, have been identified across various cancers with further research underway (Ai *et al*., 2023).

To investigate *ZFY*'s role and *ZFYS* as a potential cancer-testis gene, RNA-Seq was performed to identify potential changes to the transcriptome as a consequence of *ZFY* overexpression. The initial analysis of differential gene expression revealed a substantial alteration in the transcriptome after the integration of both *ZFYS* and *ZFYL* into the genome, with *ZFYL* exhibiting an even more pronounced change. The evidence presented here indicates that *ZFYS* and *ZFYL* regulate all the same genes, but *ZFYL* is more active on a per-molecule basis. This could hint towards understanding how the alternative splicing regulatory system works during spermatogenesis. This initially suggested that *ZFYL* acts as a more potent transcription factor in comparison to *ZFYS*, displaying a significantly greater number of potential unique interactors. However, a limitation of this experiment lies in the exceptionally high expression of *ZFY* observed post-transfection. The TPM levels observed in this experiment are markedly higher compared to the expression levels seen in cancer and the endogenous levels seen in cancer cells (7.0 nTPM) (The Human Protein Atlas, 2024b). This heightened expression may be a contributing factor to the extensive number of differentially expressed genes detected.

Subsequent gene ontology analysis examined the potential pathways activated by *ZFY*. Reactome pathway analysis employs p-value and FDR criteria to assess the significance of the potential targeted pathway. Failure to meet both criteria undermines the reliability of the identified pathway, necessitating caution when analysing the ontology, especially for pathways meeting only p-value thresholds.

Many of the identified target pathways identified by ontology enrichment analysis meet p-value significance but fail to meet the FDR criteria.

4.4.1 A Potential Extracellular Matrix Function

After compiling the most significant pathways for the upregulated genes shared between *ZFYS* and *ZFYL* (excluding *ZFX*), five pathways were found to meet both significance criteria. These pathways were; collagen chain trimerization, degradation of the extracellular matrix, collagen degradation, assembly of collagen fibrils and other multimeric structures and transport of connexons to the plasma membrane. Although the remaining pathways did not meet the FDR criteria, they exhibited connections to the significant pathways, with many associated with the extracellular matrix and ion channels, such as extracellular matrix proteoglycans, laminin interactions and potassium channels. This initially led to confusion regarding why multiple extracellular matrix pathways were upregulated by *ZFY*. However, upon closer examination, some connections could be established. A paper published in 2014 conducted a literature review of the human sperm proteome and performed Reactome pathway analysis (Amaral *et al*., 2014). In their analysis, they also identified the upregulation of extracellular matrix organisation pathways seen in **Table 4.24**.

Table 4.24: Extracellular Matrix Organisation pathways likely active in human sperm. These pathways were identified by Reactome using proteins identified across multiple sperm proteomic studies from (Amaral *et al*., 2014).

Extracellular Matrix Organisation (REACT_118770; P = 1.7 × 10^{-3}; ratio = 0.48)	
Collagen Formation (P = 2.6 × 10^{-6})	Collagen biosynthesis and modifying enzymes (P = 3.2 × 10^{-5}) Assembly of collagen fibrils and other multimeric structures (P = 3.0 × 10^{-4})
Degradation of the extracellular matrix (P = 4.9 × 10^{-3})	Degradation of collagen (P = 2.3 × 10^{-3})

The extracellular matrix is expected to have roles in regulating spermatogenesis, Sertoli cells and the blood-testis barrier with collagen forming a large part of the extracellular matrix (Siu & Yan Cheng, 2008). Type IV collagen and laminins both forms building blocks of the basement membrane in the testis, a specialised form of extracellular matrix (Siu & Yan Cheng, 2008). The basement membrane in the seminiferous epithelium houses the spermatogonia, which are the progenitors of male germ cells (O'Donnell *et al*., 2017). When spermatogonia detach from the basement membrane, meiosis begins, and they transform into preleptotene primary spermatocytes starting the process of spermatogenesis (O'Donnell *et al*., 2017). The degradation of the extracellular matrix and the remodelling of connective tissues are essential for facilitating crucial changes during germ-cell migration (Asgari *et al*.,

2021). Any abnormalities in germ cell migration pathways can result in the disconnection of the cellular matrix, halting of development, and apoptosis in male germ cells (Asgari *et al*., 2021). The extracellular matrix is essential not only for spermatogenesis but also for sperm cells to adhere to the extracellular matrix of the egg, the zona pellucida, during fertilisation (Bi *et al*., 2002). The acrosome reacts to penetrate the zona pellucida, allowing for the final step of plasma membrane fusion between the sperm and egg (Talbot *et al*., 2003).

The sperm plasma membrane is a crucial structure responsible for protecting sperm from extracellular damage and adapting to physiological damage through modifications in membrane fluidity, activation of ion channels, reorganisation of surface proteins, and calcium-induced acrosomal exocytosis (Amaral *et al*., 2014);(Tapia *et al*., 2012). The membrane comprises approximately 70% phospholipids, 25% neutral lipids (cholesterol), and 5% glycoproteins, with its composition finely regulated within the male reproductive tract (Puga Molina *et al*., 2018).

For successful reproduction of the sperm cell and oocyte, complex changes in the plasma membrane of the sperm are vital to produce the diploid zygote (Flesch & Gadella, 2000). Upon sperm capacitation, changes in the sperm plasma membrane result in increased affinity for the zona pellucida due to physiological and biochemical changes (Flesch & Gadella, 2000). The zona pellucida, the extracellular matrix protecting the plasma membrane, primes sperm cells to trigger the acrosome reaction necessary for penetrating through the zona pellucida, ultimately leading to the fusion of the sperm plasma membrane with the egg oolemma, resulting in the incorporation of the sperm cell into the oocyte.

Capacitation is associated with the loss of membrane cholesterol and modification of other membrane lipids; activation of the cAMP/PKA pathway increases protein tyrosine phosphorylation and intracellular pH (Pinto *et al*., 2023). These modifications are orchestrated by decreasing calcium permeability and increasing potassium permeability, resulting in hyperpolarisation of the sperm membrane potential. The Na+/K+ ATPase electrogenic pump plays a pivotal role in regulating sperm function in this process (Pinto *et al*., 2023). While, failing FDR, potassium channel pathways were seen to be potentially upregulated by *ZFY*. Potassium channels are crucial to sperm motility and capacitation through increases in K^+ permeability (Shukla *et al*., 2013). Multiple types of motility-related potassium channels in sperm cells have been reported; inwardly rectifying K+ channels, voltage-gated potassium channels, SLO K+ channels, and cyclic nucleotide-gated channels (Nowicka-Bauer & Szymczak-Cendlak, 2021).

In summary, the activation of extracellular matrix organisation pathways by *ZFY* may underscore its pivotal role in the progression of spermatogenesis. Moreover, given its potential association with extracellular matrix degradation, *ZFY* may be linked to fertilisation, facilitating the penetration of sperm cells through the zona pellucida, and possibly connecting a role to potassium channels as well. This emphasises the potentially pivotal role of *ZFY* in male fertility, further highlighting the importance of its retention on the Y chromosome.

4.4.2 *ZFYL* has a Potential Neuronal-Like Function in Sperm

Analysis of *ZFYL* did not uncover any pathways that met both p-value and FDR significance criteria; only their p-value was deemed significant. Once more, pathways associated with the extracellular matrix and potassium ion channels were identified. Surprisingly, pathways associated with neuronal function were identified. It has been discovered that mammalian sperm express numerous "neuronal" and classical neurotransmitter receptors that play essential roles in sperm function particularly exocytosis during the acrosome reaction (Pierce *et al*., 2009);(Ramırez-Reveco *et al*., 2017). The acrosome reaction, a pivotal process in sperm function, mirrors several aspects of presynaptic secretion. Sperm functions such as capacitation are regulated by second messengers similar to those observed in neuronal exocytosis. These include ion fluxes, sterol oxidation, activation of protein kinase A, and calcium signalling (Ramırez-Reveco *et al*., 2017). This suggests that mammalian sperm and neurons share similar mechanisms and “neuronal” receptors (Meizel, 2004). Although the functions of sperm and neurons differ significantly, this could elucidate why synaptic pathways are identified by Reactome, as there is potential overlap between these similar pathways.

4.4.3 Gene Ontology Potentially Confirms *ZFYS* as a Potential Cancer-Testis Gene

Gene ontology analysis suggests that *ZFYS* potentially interacts with the ErbB2 signalling pathway, with less direct connections to other signalling pathways such as the PI3K/AKT and RAF/MAP kinase pathways, which do not meet FDR significance. These pathways have all exhibited aberrations in various cancers, implying a potential role of *ZFYS* as a cancer-testis gene.

ErbBs are receptor tyrosine kinases that are essential for normal physiology but have also been implicated in cancers as early as the 1980s, with ErbB2 being mutated in multiple epithelial tumours (Yarden & Sliwkowski, 2001);(Hynes & MacDonald, 2009). ErbB receptors function in multiple cellular processes including proliferation, cell migration, metabolism and survival, which means their aberrant expression can be

detrimental. Tumours that exhibit constitutive activation of ErbB2 and EGFR trigger the activation of numerous intracellular signalling proteins and pathways similar to those activated by wild-type receptors. These pathways include the MAPK, PI3K/Akt, and mTOR pathways, as well as Src kinase and STAT transcription factors. It has been shown that breast cancers positive for ErbB2 also maintain high PI3K activity, further proving a link between these pathways (Hynes & MacDonald, 2009).

In the RNA-Seq data lists, *ErbB2* is not differentially expressed however, other genes including *ErbB4, NRG-1 and NRG-2* were identified as being upregulated. Both *NRG-1* and *NRG-2* bind to *ErbB4* (Hynes & MacDonald, 2009);(Veikkolainen *et al*., 2011). *ErbB4* differs from *ErbB2* and *ErbB3* as it is capable of acting as a fully functional receptor tyrosine kinase both as a homo- as well as a heterodimer. Upon activation by neuregulin, such as *NRG-1* or *NRG-2*, *ErbB4* can form homodimers or heterodimers with other ErbB family members, leading to the activation of kinase and autophosphorylation functions. This phosphorylation event subsequently initiates intracellular pathways, including the PI3K/Akt and Ras/MAPK cascades (Veikkolainen *et al*., 2011). This indicates that *ErbB4* mutations play a significant role in the development of various cancers, as evidenced by notable occurrences in colorectal, lung, gastric, prostate, hepatocellular carcinoma, and breast cancers (El-Gamal *et al*., 2021).

ErbB signalling was identified as an activated pathway associated with the integration of *ZFYS* into the genome. This might further elucidate the possible activation of the PI3K/AKT and Ras/MAPK pathways in the Reactome gene ontology analysis, as they serve as downstream targets of ErbB signalling. However, they do not meet FDR criteria due to the weaker association.

4.4.4 The ZF Family Appears to Target the *WNT*-Signalling Family

Analysis of the upregulated genes across *ZFX* and both *ZFY* variants suggests a potential target within the *WNT*-signalling pathway. Several members of the *WNT* family were found to be upregulated in the RNA-Seq analysis, with *WNT7*A being among the most differentially expressed genes across all three ZF* experiments.

The *WNT* signalling pathway, with its ancient origins and remarkable conservation across metazoan animals, plays essential roles in embryonic development and the upkeep of adult homeostasis (Komiya & Habas, 2008);(J. Liu *et al*., 2022). It influences various cellular processes including proliferation, differentiation, migration, genetic stability, apoptosis, and the renewal of stem cells (Pai *et al*., 2017). Disruption of *WNT* signalling has been associated with significant diseases, spanning from non-cancerous to cancerous conditions. The pathway was originally identified in 1982,

with its research interest continuing to increase due to being an attractive target pathway for disease treatment (J. Liu *et al*., 2022). *WNT* signalling is subdivided into two branches; canonical and non-canonical (Komiya & Habas, 2008). The canonical *WNT* pathway primarily oversees cell proliferation, whereas the noncanonical *WNT* pathway governs cell polarity and migration (J. Liu *et al*., 2022). Despite their distinct roles, both pathways operate under mutual regulation (J. Liu *et al*., 2022). In this data we found that the *ZFY* mostly targeted the noncanonical pathway, which could show ties to spermatid elongation, a profound example of cell polarisation and the migration of cells into and out of Sertoli cell crypts during cytoplasm shedding (L. Li *et al*., 2017). Both developing spermatids and Sertoli cells have been defined as highly polarised cells in the testis, with animal study models showing the changes in spermatid polarity are linked to subsequent disruption of spermatid adhesion because of changes in the cytoskeletal organisation within the seminiferous epithelium (L. Li *et al*., 2017). This then further links to roles later on in cancer such as metastasis rather than early during cell proliferation. It has been shown that the non-canonical *WNT* pathway can mediate cancer cell migration and motility which are key for metastasis, a specific example of a *WNT* protein involved in this is *WNT*5A (Y. Chen *et al*., 2021).

*WNT*s are glycoproteins rich in cysteines that interact with the N-terminal extracellular cysteine-rich domains of Frizzled family receptors, which are G protein-coupled receptors (GPCRs) (Pai *et al*., 2017). The *WNT*s have been divided based on canonical or non-canonical classification; canonical (*WNT*1, 2, 3, 3a, 8a, 8b, 10a and 10b) and non-canonical (*WNT*4, 5a, 5b, 6, 7a, 7b and 11) **(Figure 4.30)** (Ackers & Malgor, 2018);(Chae & Bothwell, 2018). Following receptor binding, the canonical pathway is characterised by the accumulation of β-catenin, which subsequently translocates into the nucleus to regulate gene expression (Ackers & Malgor, 2018). In contrast, the non-canonical pathway operates independently of β-catenin, engaging in diverse intracellular signalling and target gene expression mechanisms. This pathway can be further subdivided into the *WNT*/Ca2+ pathway, which modulates gene expression through NFAT and also inhibits β-catenin signalling via NLK, and the planar cell polarity pathway regulates cytoskeletal rearrangements and cell survival through RhoA (Ackers & Malgor, 2018).

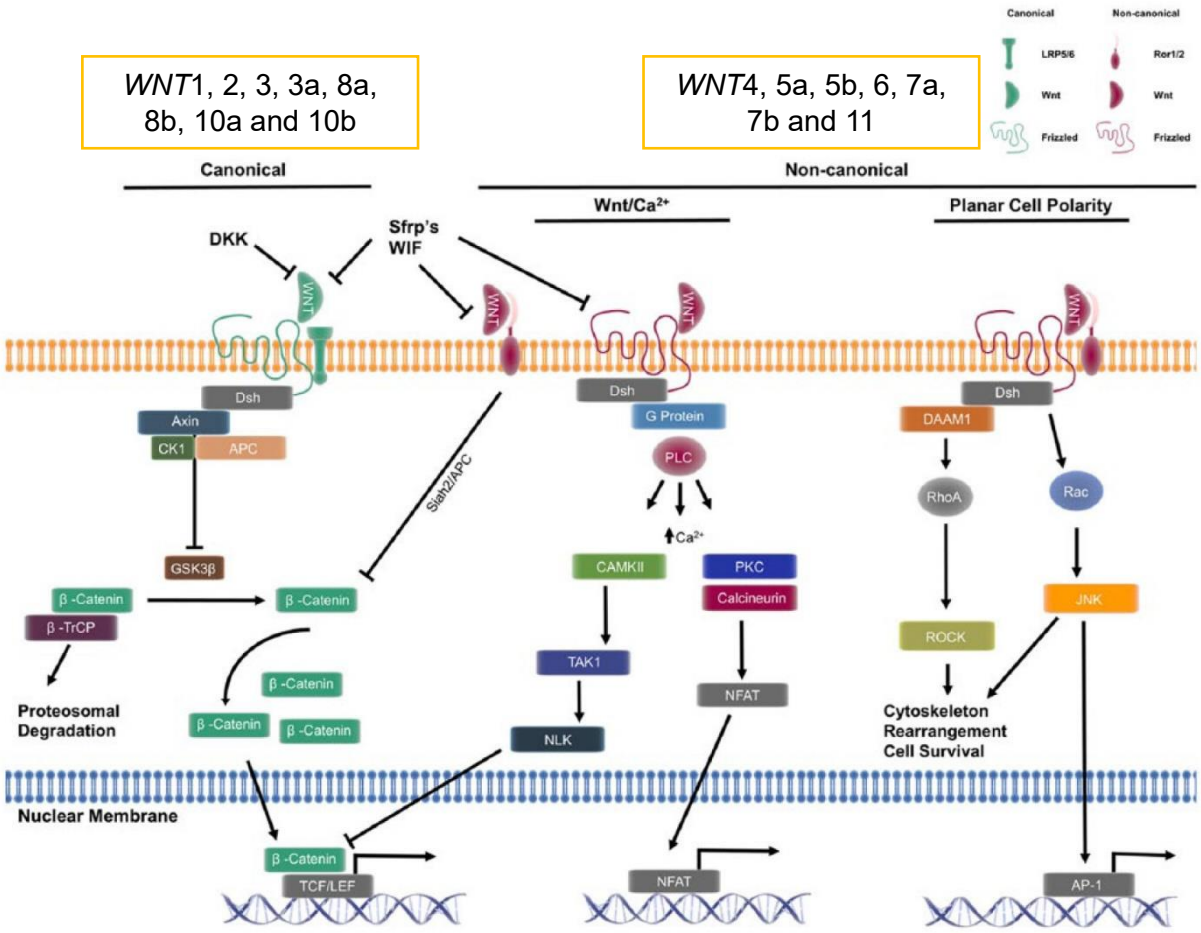


Figure 4.30: *WNT* signalling pathways (Ackers & Malgor, 2018). The canonical pathway is associated with the accumulation of β-catenin. The non-canonical pathway is further subdivided into *WNT*/Ca2+ and Planar cell polarity signalling.

All three ZF* seem to be targeting *WNT* signalling, with *ZFYS* and *ZFYL* upregulating these specific *WNT* proteins; *WNT*7A, *WNT*7B, *WNT*9A, *WNT*4, *WNT*11, *WNT*3A, and *WNT*5B. *ZFX* upregulated all the same *WNT* proteins found in the *ZFY* data with the addition of *WNT*3. Supporting this discovery was the identification of the activation of signalling by *WNT* (P = 4.8 × 10−17) in the sperm proteome project (Amaral *et al*., 2014). Many of the identified *WNT*s in the RNA-Seq have been identified as non-canonical. *WNT* signalling as mentioned plays a major role in cell growth and development with studies showing it is fundamental to mammalian spermatogenesis with documented studies showing it also governs mouse sperm maturation (Koch *et al*., 2015);(Zeng *et al*., 2023). More specifically they showed that multiple *WNT*/STOP target proteins are expressed within sperm and are vital for the promotion of sperm proteome stability and motility. High expression of *WNT*10a, *WNT*2b, *WNT*1 ligands were seen in the caput epididymis, with *WNT*1-null mutants dying perinatally. It was therefore, suggested that multiple *WNT* ligands more have roles in sperm maturation (Koch *et al*., 2015);(Zeng *et al*., 2023). A study revealed the significance of *WNT*

signalling in controlling mammalian spermatogenesis via mrhl RNA (lncRNA) (Akhade *et al*., 2016). This RNA negatively regulates canonical *WNT* signalling but is downregulated when *WNT* signalling is activated in mouse spermatogonial cells. Consequently, the reduction of mrhl RNA triggers the activation of several meiotic differentiation marker genes that play a role in spermatogenesis as well as *WNT* signalling (Akhade *et al*., 2016). An additional study showcased that spermatozoa, which are transcriptionally silenced, respond to *WNT* signals arising from the epididymis (Koch *et al*., 2015). Mice with a mutation in Cyclin Y-like 1, a *WNT* regulator, were found to be sterile due to spermatozoa immobility and malformation. The researchers also inferred that *WNT* signalling primarily coordinates post-transcriptional sperm maturation via GSK3 (Koch *et al*., 2015). Lastly, while not highlighted as upregulated in this dataset, *WNT*5a, a constituent of the non-canonical *WNT* signalling pathway, has been evidenced to significantly influence Sertoli cell junctions via the planar cell polarity signalling pathway (Fu *et al*., 2021);(Rey, 2021). *WNT*5a also balances *mTORC1* and *mTORC2* in actin-dependent processes which are vital for maintaining the blood-testis barrier. Their study revealed that when *WNT*5a was knocked down, there were changes in the expression levels and distribution patterns of blood-testis barrier-associated proteins, along with other actin-binding proteins and F-actin. Consequently, elongated spermatids became embedded within the seminiferous epithelium as a result of polarity loss. Hence, the absence of *WNT* signalling might lead to male infertility as a consequence of Sertoli cell junction dysfunction (Fu *et al*., 2021).

Since *WNT* signalling plays a crucial role in cell growth and development, it is logical that it would have an involvement in spermatogenesis and the development of both male and female reproductive systems, thus explaining the potential upregulation of this pathway by ZF*. Nevertheless, this pathway is closely intertwined with cancer. Dysregulated *WNT* signalling has been associated with cancer stem cell renewal, proliferation, and differentiation, all of which are interconnected with tumorigenesis (Y. Zhang & Wang, 2020);(Corda & Sala, 2017). This makes the *WNT* signalling pathway a very attractive target for cancer intervention. Research indicates that cancer cells can hijack the non-canonical *WNT* signalling pathway for migration and metastasis. This phenomenon is observed, for instance, in melanoma where there is an overexpression of *WNT*5a (Corda & Sala, 2017).

The significance of *WNT* signalling in developmental processes and its correlation with cancer render this potential *ZFY*-targeted pathway an intriguing finding, suggesting that *ZFY* could be a potential cancer-testis antigen.

Furthermore, several snoRNAs were identified to be regulated by both *ZFYS* and *ZFYL* within the data set, but we were not able to confirm this by qPCR. However, due to time constraints these results were not further pursued, this does not mean that the lack of confirmation should be taken as a definitive since snoRNAs are processed differently and they would therefore require a dedicated experiment focused on the small RNA content to formally study this.

4.4.5 *ZFY* has a Weak Cancer Correlation

Subsequent bioinformatics analysis, utilising a vast publicly available cancer database, revealed a weak correlation between the expression of *ZFY* target genes and cancer. Candidate biomarkers for *ZFYS* were discovered because the cancer datasets failed to differentiate between *ZFYS* and *ZFYL*. These biomarkers were subsequently tested in leukaemia cell lines, given the large number of potentially *ZFY*-positive leukaemia cell lines.

Initially, head and neck cancer was the primary focus, but due to insufficient evidence, the experimental validation of the biomarkers shifted to leukaemia. Ageing men especially those over the age of 70 years show an increased prevalence of mosaic loss of chromosome Y in peripheral leukocytes (Ljungström *et al*., 2022). A considerable number of male leukaemia patients exhibit partial or complete loss of their Y chromosome, a condition linked to a more aggressive clinical course and an intermediate prognosis (Holmes *et al*., 1985). Despite this, *ZFYL* expression was verified in MEC-1 and THP-1 cells, whereas *ZFYS* expression was not, even though it was anticipated to be expressed due to the presence of potentially correlated genes. However, “this is consistent with the low correlation values making the biomarkers less reliable.

It was observed that despite the presence of *ZFYL* expression in leukaemia cells, there were differing levels of expression in the genes previously utilised for RNA-Seq validation. For instance, *TMPRSS2*, which was previously shown to be upregulated by both *ZFYL* and *ZFYS*, was found to be downregulated in both *ZFY*-positive leukaemia cell lines. This suggests that *ZFY* may function somewhat differently in leukaemia.

Although we couldn't verify the expression of *ZFYS* in the leukaemia cells, gene ontology analysis revealed potential cancer-related pathways that could be influenced by *ZFYS*, such as ERRB2 and *WNT* signalling pathways. Further cancer studies would be necessary to delve into *ZFYS*'s role as a potential cancer-testis antigen.

Overall, while we have suggested many potential *ZFY* specific functions outside of spermatogenesis, in this thesis we have not been able to full confirm *ZFY*, specifically the short variant as a cancer-testis gene.

4.4.6 A Potential Feedback Mechanism with *RBMY*

A major part of this thesis looks at *RBMY* as the splicing regulator of *ZFY*. An interesting finding from this RNA-Seq data is the fact that RBMXL2 is downregulated by both *ZFYL* and *ZFYS* and this was more pronounced for *ZFYL* compared to *ZFYS*. RBMXL2 is a close relative of both RBMX and *RBMY* making it an intriguing bit of data. Due to the limitation of HEK293 cells being female, this thesis cannot confirm whether *ZFY* is altering *RBMY* transcription. However, if *ZFYL* downregulates *RBMY*, a feedback mechanism could be enforced resulting in further *ZFYL* expression and subsequently a reduction in *ZFYS* expression post-meiosis.

This alongside the data represented in chapter 3 is very promising, but further investigations would be required to pinpoint an exact mechanism of action.

5. Chapter 5: Utilising Proteomics to Understand the Role of *ZFY* Through the Identification of its Interacting Partners.

5.1 Introduction

The objective of this chapter was to express and purify the acidic domains of ZFY-long and ZFY-short using an *E. coli* expression system. Chromatography techniques were employed, leveraging the presence of a His-tag and the protein's high negative charge for purification.

5.1.1 The lac Promoter as a Useful Target

Protein expression and purification are widely utilised methods in biochemical studies to answer fundamental questions regarding protein structure and protein interactions (Bornhorst & Falke, 2000);(Wingfield, 2015). Protein expression is commonly associated with a recombinant expression system such as *E. coli* (Wingfield, 2015), due to its well-established use, fast growth kinetics with a ~20-minute doubling time and ease of achieving high-cell density cultures (Rosano & Ceccarelli, 2014). The lac promoter in *E. coli* **(Figure 5.1)** is commonly targeted to induce protein expression using Isopropyl-β-D-thiogalactoside (IPTG), a common inducer of the promoter's transcriptional activity (Briand *et al*., 2016). In its inactive state, the lac repressor protein (lacI) blocks the lac promoter preventing RNA polymerase from binding and therefore prevents the initiation of transcription. However, in the presence of lactose, the lac promoter becomes active as allolactose binds lacI inducing a conformational change and interrupting its binding capability. This then allows RNA polymerase to bind, and transcription commences. IPTG is an analogue of allolactose and functions to hinder the binding of lacI to the lac promoter thereby triggering the initiation of transcription. This makes the lac promoter a commonly used promoter for protein expression in *E. coli* (Briand *et al*., 2016).

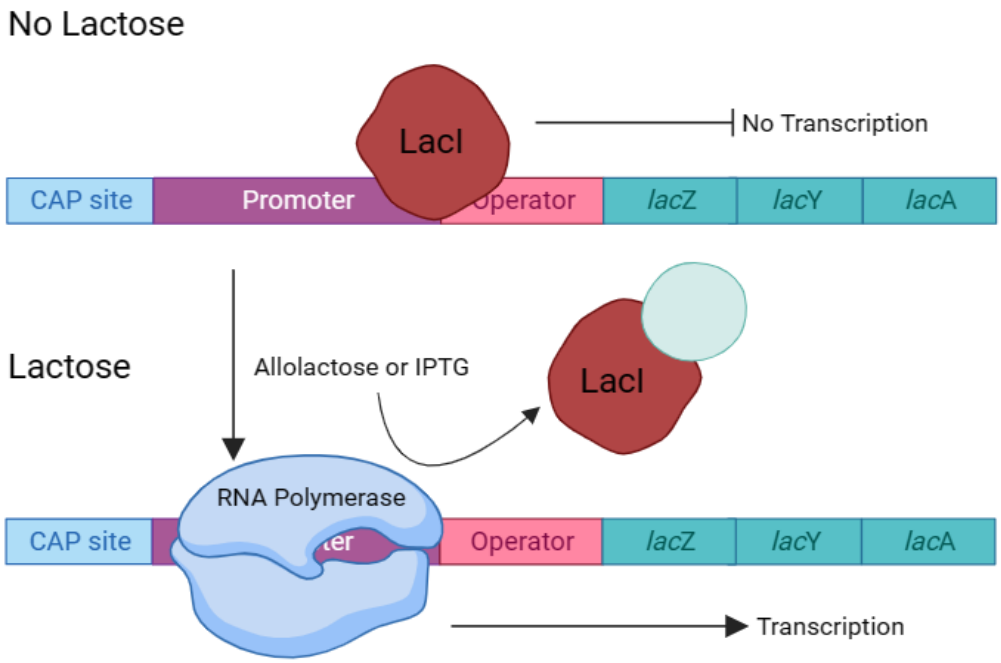


Figure 5.1: *E. coli* Lac operon system.

Following protein expression, the isolation of a protein usually employs chromatographic purification for subsequent downstream use (Wingfield, 2015). Numerous chromatographic techniques have been adopted for protein separation (Coskun, 2016), such as column and ion-exchange chromatography which are both used in this chapter due to *ZFY*'s protein structure and charge.

Polyhistidine-tag affinity chromatography, known as nickel column purification **(Figure 5.2)**, allows the isolation of recombinant proteins engineered to contain a polyhistidine tag (His_6 tag) (Carter & Outten, 2021);(Bornhorst & Falke, 2000). Using a Ni-NTA resin as bait, the proteins with a His_6 tag bind to the resin, whilst the proteins with no His_6 tag flow through without binding to the resin resulting in the separation of tagged and untagged proteins with wash steps aiding in the purification and isolation of the His_6 tagged proteins (Carter & Outten, 2021);(Bornhorst & Falke, 2000). Ion-exchange chromatography is a useful and widely used technique for protein separation as it relies on interactions between charged molecules **(Figure 5.2)** (Selkirk, 2004);(Fekete *et al*., 2015). The net charge of a protein is determined by the amino acids encoded. Lysine, arginine and histidine have a positive charge while aspartic acid and glutamic acid have a negative charge at physiological pH. Shifts in buffer salt concentration can be used to generate a stepwise salt elution gradient to elute proteins in increments based on their charge. Tightly adhered proteins require higher ionic strength to be eluted upon applying a salt gradient (Selkirk, 2004);(Fekete *et al*., 2015).

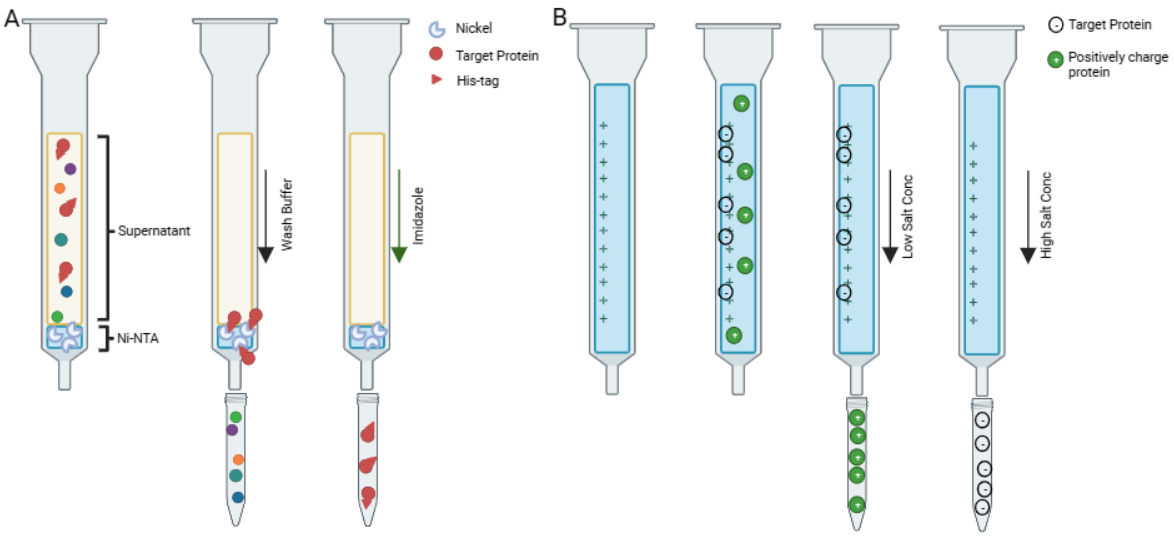


Figure 5.2: A: Nickel Column Chromatography. A modified protein with a His_6-tag can be captured and purified using a Ni-NTA resin in a gravity flow column system. The His_6-tag binds to the nickel resin, allowing for unbound proteins to be washed through the column. To elute the tagged protein, imidazole must be added to replace the His-tag bound to the Nickel. **B: Anion exchange Chromatography.** Ion exchange works on the principle of ionic interactions, anion exchange uses a positively charged resin with affinity for negative surface charges. By increasing the salt gradient, proteins bound by weak forces are eluted at low salt concentrations, with stronger interactions requiring a higher salt concentration to be eluted.

Downstream of protein expression and purification, many methods now exist to investigate protein-protein interactions and investigate potentially relevant biological pathways. Pull-down assays are a valuable technique to detect physical interactions between proteins (Louche *et al*., 2017). This assay follows similar principles to the affinity purification "bait" system and utilises washing and elution steps. Using this tag method, a tagged protein can be targeted and pulled down alongside any proteins it may be interacting with in the system (Louche *et al*., 2017). Mass-spectrometry-based proteomics enables complex profiling of protein-protein interactions using peptide matching (Brymora *et al*., 2004). Advancements in mass spectrometry have led to increased capacity and capability in protein identification and analysis.

5.1.2 *ZFYS*' Structure in Relation to its Role as a Transcription Factor

Transcription factors control gene expression through the binding of coactivators to their acidic activator domains (Staller *et al*., 2022). Whilst these regions are poorly conserved and intrinsically disordered their function remains (Staller *et al*., 2022). It has been noted that a common feature of these acidic activator domains is their high acidity (net negative charge) due to their richness in aspartic acid and glutamic acid amino acids (Staller *et al*., 2022);(Ferreira *et al*., 2005). For transcription of genes by RNA polymerase II to start, the assembly of general transcription factors at the gene's

promoter is vital (Felinski *et al*., 2001). Due to their acidic residues, the acidic activator regions have been suggested to interact with target proteins through electrostatic interactions (Ferreira *et al*., 2005). Within these regions 9aaTAD's have been deemed as critical hydrophobic regions essential for binding target factors for transcription (Ferreira *et al*., 2005). *ZFY* is a transcription factor and consists of a highly negative acidic activating domain at its N-terminus. Due to this *ZFY* has a predicted negative charge of -16 and an isoelectric point between 5.65 and 5.99. The isoelectric point of a protein is the pH at which the net charge of the protein molecule is 0 (Tokmakov *et al*., 2021). This means that at a pH ranging between 5.65 and 5.99, *ZFY* has a net charge of 0, however, a pH below this means *ZFY* would be positively charged and negatively charged if the pH rises above the isoelectric point. These isoelectric point figures are of use when carrying out ion exchange chromatography. Although identified as a transcription factor, *ZFY*'s binding partners and targets are widely unknown.

To elucidate the functions of *ZFY*, this chapter aimed to express and purify the acidic activating domain of *ZFY* to identify interactions that differ between *ZFYS* and *ZFYL*. The original aim was to purify the protein's acidic domain and use a pull-down system in testis tissues to determine interacting coactivators. However, due to difficulties experienced when trying to express and purify *ZFY* in an *E. coli* system, the method of target identification was altered but the end result was still successful.

5.2 Materials and Methods

5.2.1 DNA Construct Design

DNA constructs consisting of a pET-15b backbone **(Figure 5.3)** and the acidic domain of either *ZFYL* or *ZFYS* were synthetically produced by Genscript. pET-15b contains a His_6-Tag which can be used for nickel column purification. These DNA constructs aimed to express and purify the two forms of the acidic domain to identify specific binding partners through a pulldown system (for sequence information see *supplementary* data Sequence C and D).

Following the arrival of the lyophilised plasmid DNA from Genscript containing our insert, the vials were briefly centrifuged at 6,000 x g for 1 minute at 4°C, and 20µL of sterilised water was added to dissolve the plasmids.

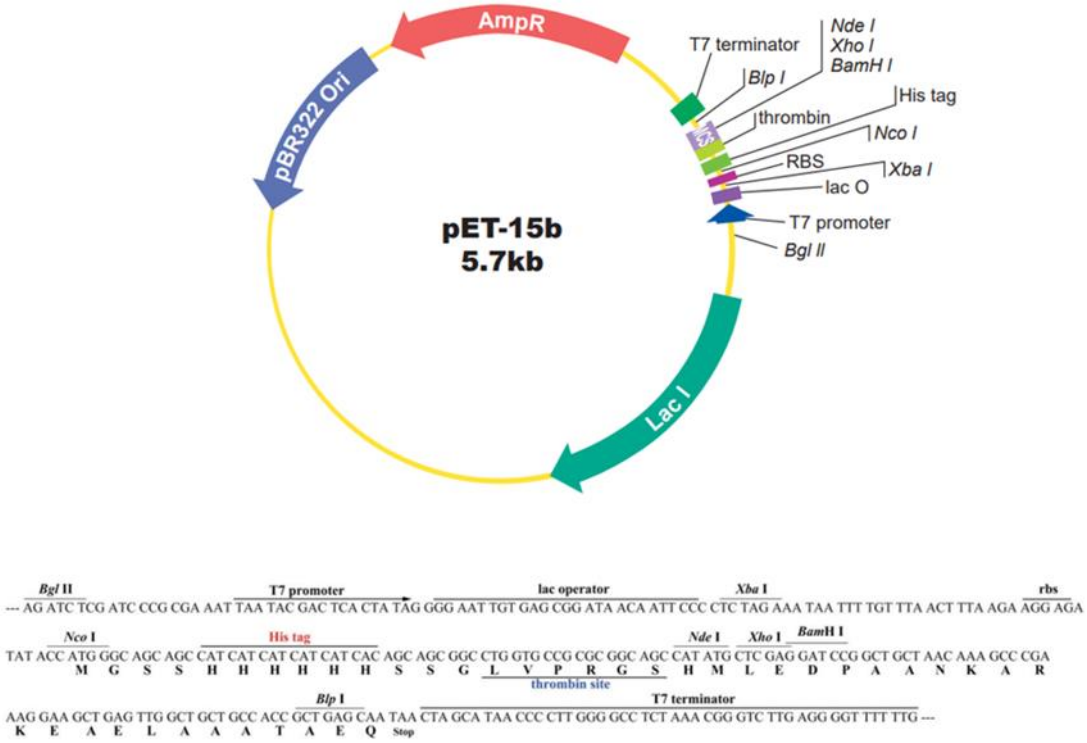


Figure 5.3: pET-15b vector map (Source: Addgene). The 5.7Kb plasmid contains a histidine tag and a thrombin cleavage site. pET-15b confers ampicillin resistance in bacteria.

The *ZFYS* acidic domain construct has an expected molecular weight of 26.3KDa and an isoelectric point of 4.39. The *ZFYL* acidic domain construct has an expected molecular weight of 46.9KDa and an isoelectric point of 4.03.

5.2.2 Restriction Digest

Restriction digests using BgI II (Promega, R6081) and Nco I (Promega, R6513) confirmed the size of the plasmids (*ZFYS*: 6,576bp, *ZFYL*: 7,089bp).

In sterile tubes, the components were assembled in a 20µL reaction **(Table 5.1)** and subsequently mixed and incubated for 1-4 hours at 37°C. The samples were combined with loading dye (Promega, # G190A) at a 1x final concentration and visualised on a 2% agarose gel (see chapter 3 section 3.2.3).

Table 5.1: Setting up restriction enzyme digest protocol. This protocol taken directly from Promega can be used for both BgI II and Nco I. Reactions are performed in a volume of 20µl and incubated at the enzyme's optimum temperature.

Component	Volume (uL)
dH_2O	16.3
Restriction Enzyme 10X buffer	2
Acetylated BSA, 10µg/µl	0.2
DNA, 1µg/µl	1.0
Restriction Enzyme, 10µ/µl	0.5

5.2.3 *E. coli* Competent Cells

During the optimisation of protein expression and purification, a variety of competent cells were trialled and are stated in **Table 5.2** below.

Table 5.2: A table with the competent cell systems chosen for protein expression. The corresponding antibiotic necessary for successful expression and purification is noted along with the final working concentration used.

Competent Cell Name	Company	Catalogue Number	Antibiotic Resistance	Antibiotic Concentration
BL21(DE3) Competent Cells	Sigma-Aldrich	CMC0014-440UL	Ampicillin	Amp = 100ug/mL
BL21(DE3) pLysS Competent Cells	Promega	L1195	Ampicillin Chloramphenicol	Amp = 100ug/mL Cam = 50ug/mL
Rosetta (DE3) Competent cells	Novagen (Sigma-Aldrich)	70952	Ampicillin Chloramphenicol	Amp = 100ug/mL Cam = 25ug/mL

5.2.4 Transformation of Plasmids into Competent Cells

The first bacterial system used was the BL21(DE3) competent cells containing the phase T7 RNA polymerase gene linked to the IPTG-inducible promoter, for use with any expression plasmid containing the T7 promoter. This strain was selected as a good starting point since it is an all-purpose strain used for high-level protein expression and easy induction. However, the transformation process remains consistent between the different competent cells.

The transformation protocol was as follows; thaw the competent cells on ice for 10-15 minutes, once thawed add 1pg-100ng of DNA to the chilled competent cell aliquot. The cells were then incubated on ice for 30 minutes, and then heat shocked at 42°C for 45 seconds. After heat shocking the cells, they were placed back on ice for a further 5 minutes. 950µL LB media was then added to the cell mixture and incubated at 37°C for 1 hour while shaking. After incubation, the cells were spread onto LB agar plates containing the correct antibiotic **(Table 5.2)** and incubated at 37°C overnight. The following morning the plates were checked for colony growth.

5.2.5 Plating Assay in BL21(DE3) Competent Cells

After transforming BL21(DE3) cells with *ZFYS* and *ZFYL* constructs, colonies were inoculated in 5mL LB cultures, and the growth of the colonies was monitored until an OD600 reading of 0.5 was reached (~$1x10^8$cells/mL). Subsequently, a 10-fold serial dilution was performed to obtain approximate cell concentrations of 10^4, 10^3,10^2, and 10^1 cells/mL. 100µL of the diluted samples were spread onto ampicillin (100ug/mL) agar plates and IPTG (0.1mM) + ampicilin (100ug/mL) agar plates, which were incubated overnight at 37°C. This was done to determine the desired concentration for colony counting and examination.

Based on the dilution series, it was determined that cell concentrations of 10^3 and 10^2 cells/mL produced a suitable number of colonies. The diluted samples were prepared and plated as before. Plates were incubated overnight at 37°C and then checked the following morning. For each plate, the colonies were counted, and the size of the colonies was noted.

5.2.6 Protein Growth and Induction

5mL LB overnight starter cultures in a 50mL falcon containing the appropriate antibiotic **(Table 5.2)** were inoculated from colonies and grown overnight shaking at 37°C. The following day, a 1% volume of the starter culture was added to the larger culture volume. These cultures were incubated at 37°C whilst shaking and the OD600 reading was monitored until the cultures had an absorbance reading between 0.4 and 0.6.

At the target OD600, a pre-induction sample was collected and spun down in a bench-top centrifuge (FisherScientific, accuSpin Micro17, #13-100-675) at full speed for 3 minutes and subsequently resuspended in 1x Sodium Sodecyl Sulfate (SDS) sample buffer (10% glycerin, 60mM Tris-HCl, 2% SDS, 0.1M DTT, 0.01% bromophenol blue, pH 6.8). To the remaining culture, IPTG was added at varying final concentrations (0.4mM – 0.8mM). The post-induction incubation time and temperature were also

optimised, ranging from 3 hours at 37°C to overnight at 30°C. After induction, post-induction samples were processed identically to the pre-induction sample.

The remaining growth culture was centrifuged at 6,000 RPM (6,353.5 x g) for 20 minutes at 4°C to obtain the cell pellet.

5.2.7 Cell Lysis

This protocol is for a 50mL culture, and this was altered depending on the culture size. Cell pellets were gently resuspended in 5mL lysis buffer (25mM Sodium Phosphate, 100mM sodium chloride, pH 7.3). Following this, 100µg/mL lysozyme and 0.1% Triton X-100 were added. The culture was mixed and incubated at RT for 30 minutes. Following incubation,10mM $MgCl_2$ and 2ug/mL of DNaseI were added and again the culture was mixed. Lysozyme digest was performed at RT for 20 minutes, but this incubation was lengthened to 30 minutes to further reduce viscosity. Cultures were subsequently cooled on ice for 5 minutes before continuing. Cells were lysed on ice by sonication at 20 amps, with cycles of 10/20 seconds on and 10/20 seconds off. Sonication time was increased to improve lysis, especially for larger culture volumes. The lysates were then centrifuged at 4,000 RPM (3,584 x g) at 4°C for 30 minutes, a clear supernatant was wanted. For larger cultures, the ultra-centrifuge was used. The clear supernatant was saved for downstream purification.

5.2.8 Freezing-Thawing Protocol

In BL21(DE3) PlysS cells there is an overexpression of lysozyme which means freeze-thawing can be used as an alternative lysis method. Proceeding up to the lysis process, all the methods remained the same as previously mentioned. The pellets were resuspended in the lysis buffer as previously stated, other buffers were also tested (10mM Tris-HCl lysis buffer). Here the difference is no lysozyme, Triton X-100, and $MgCl_2$ need to be added. However, protease-phosphatase inhibitors (Pierce Protease and Phosphatase Inhibitor Mini Tablets, EDTA-Free, A32961) and 20µg/mL DNaseI were added per sample.

Three rounds of freeze-thawing were completed, with freezing at -20°C and thawing at 4°C. It was ensured that the samples were completely frozen and completely thawed between each round. Following freeze-thawing the samples were centrifuged at 4000 RPM (3,584 x g) at 4°C for 30 minutes. The supernatant was collected, and the pellets were resuspended in the lysis buffer for visual analysis (see section 5.2.12).

5.2.9 Nickel Column Chromatography

Following cell lysis and the collection of the supernatant, nickel-chelating affinity purification was performed using a BIO-RAD gravity flow column. A HisPur Ni-NTA Resin (ThermoScientific, #88221) was carefully added to the column and was washed with 10mL of Milli-Q water. The resin was then equilibrated with 10mL wash buffer (20mM sodium phosphate, 200mM Sodium Chloride, pH 7.3), during this optimisation period 50mM imidazole was later added to the wash buffer to aid in removing loosely bound proteins. Once, the column was cleaned and equilibrated, the supernatant was loaded and the flow-through was collected. The column was washed with 10mL of wash buffer and fractions of the flowthrough were collected. Elution was performed with 10mL of elution buffer (20mM sodium phosphate, 200mM sodium chloride, 250mM imidazole, pH 7.3) in ~2/5mL fractions. A high imidazole clean buffer (20mM sodium phosphate, 200mM sodium chloride, 500mM imidazole, pH 7.3) was finally passed through to remove any remaining bound proteins. The collected fractions were visualised (see section 5.2.12).

5.2.10 Dialysis

Following nickel affinity purification, the desired protein is in a high imidazole concentrated buffer, this is not suitable for long-term protein storage, nor is it suitable for the next step of protein purification using the AKTA. This purification step was added later after discovering that nickel column purification was not effective enough. Dialysis was therefore performed on the eluted protein fractions. Initially, dialysis was performed overnight against 4L of 25mM piperazine, pH 5.4 at 4°C using Snakeskin dialysis tubing (ThermoScientific, 3.5K MWCO, 22 mm, #10005743). However, significant precipitation occurred during dialysis, resulting in protein loss. The dialysis buffer was altered later to 25mM Bis-Tris at a pH of 6.4. By shifting the pH unit by 1, the hope is the protein will tolerate the buffer salt better. Switching to a 25mM Bis-Tris dialysis buffer improved solubility and prevented precipitation while still maintaining pH control. The use of this buffer increased protein yields for *ZFYS* purification by preventing losses from precipitation.

5.2.11 Anion Exchange Chromatography

For further purification, anion exchange chromatography was performed using a HiTrap Q HP 5mL column on an AKTA Start system, taking advantage of the high negative charge of the protein. Buffers for ion exchange require a pH at least 0.5 unit above the pI. The AKTA Start system works on a salt gradient using two buffers; Buffer A (0% salt) and Buffer B (100% salt), increasing the salt concentration leads to the

elution of proteins based on their charge. Originally, Buffer A was 25mM piperazine, pH 5.4 and Buffer B was 25mM piperazine with 500mM NaCl. After optimising dialysis to Bis-Tris, the ion exchange buffers were changed to 25mM Bis-Tris pH 6.4 for Buffer A and 25mM Bis-Tris, 500mM NaCl pH 6.4 for Buffer B to match.

The protocol followed on the AKTA system is noted below in **Table 5.3.** The column (5mL) is first primed and equilibrated using 1 column volume of Buffer A and 5 column volumes of Buffer B. The protein elute volume is then added to the column. The AKTA follows a similar washing and elution process to the Nickel column. 10 column volumes of wash buffer were used to clear away any unbound proteins. Then when the elution stage started Buffer A and Buffer B were mixed in an increasing gradient and protein elution was monitored by UV. This elution stage produced around 20 fractions, which were then loaded onto a 12% SDS gel for Coomassie staining to identify where our protein of interest was being eluted (see section 5.2.12).

Table 5.3: Anion exchange on the AKTA start purification system using the Hi-trap Q HP sample protocol. This method was adapted from the default Hi-Trap Q HP column to suit anion exchange for both *ZFYS* and *ZFYL*. CV: Column Volume.

Phase	Variable	Value
Methods Setting	Column Volume	5ml
	Column	Hi-trap Q HP
	Flow rate	5ml/min
Wash sample value	Sample volume	5ml of buffer A (1 CV)
Prime and Equilibration	Equilibration volume	25ml of buffer (5 CV)
Sample Application	Sample volume	8.5ml of protein sample
Wash out unbound with buffer A	Wash buffer volume (Buffer A)	50ml (10 CV)
Elution	Starting Buffer B Conc	0%
	Fraction volume	5ml
	Linear gradient B conc	100%
	Elution linear gradient volume	20ml
Prime and Equilibration end	Wash volume	25ml (5 CV)

5.2.12 Sodium Sodecyl-Sulfate-Polyacrylamide Gel Electrophoresis

For protein analysis, SDS-PAGE was performed using either 12% gradient gels prepared in-house **(Table 5.4)** or precast 4-12% Bis-Tris gels (ThermoFisher Invitrogen NuPAGE, #10472322). The precast gels used a premade 20x buffer (ThermoFisher Invitrogen NuPage NP0001) diluted to a 1x concentration using Milli-Q water. A different running buffer was required for the 12% gradient gels and a 10x stock was produced (30g/L Tris, 140g/L glycine & 10g/L SDS), this was diluted down to 1x when necessary.

Table 5.4: 12% gradient SDS gel resolving and stacking solutions. The resolving gel solution is prepared first, poured into the gel cast, and allowed to polymerise

before adding the stacking gel on top. *The APS and TEMED are added right before pouring since they initiate polymerization.

Stock solution	Resolving	Stacking
MilliQ	5.6mL	8.23mL
30% Acrylamide	16mL	2.7mL
1.5M Tris pH 8.8	7.8mL	-
0.5M Tris pH 6.8	-	3.75mL
10% SDS	300uL	150uL
APS*	300uL	150uL
TEMED*	30uL	15uL

All gels were run in an Invitrogen™ XCell SureLock ™ Mini-Cell (Product Code. 10093492) at 150V for ~80 minutes. Gels were either stained with Coomassie blue (0.1% Coomassie in 40% ethanol, 10% acetic acid) or were transferred to PVDF membranes for western blots (see below).

For Coomassie staining, the gels were submerged in Coomassie blue solution for 1 hour whilst rocking, then washed in destain (10% acetic acid, 20% methanol).

5.2.13 Western Blotting

Western blotting was performed to check which protein bands contained the His_6-tag and therefore, *ZFY*. This was checked at various stages of the process to identify protein expression and loss throughout the purification process. Western blots were carried out as previously described in section 4.2.4.

A *ZFY*-antibody was later trialled to determine its specificity due to the high sequence similarity between *ZFY* and *ZFX*. The antibodies used in this chapter can be found in **Table 5.5** below.

Table 5.5: Antibodies used for Western blot. Associated concentration and dilution used for each antibody are included.

Antibody Name	Company	Primary or Secondary	Clonality	Species	Target	Conc	Dilution
Anti-polyHistidine-Peroxidase antibody	Sigma-Aldrich [A7058]	Primary	Monoclonal	Mouse	His-HRP Conjugate	5.0 - 11.0 mg/ml	1:2,000
Anti-his-tag antibody (H-3)	Santa Cruz [SC8036]	Primary	Monoclonal	Mouse	His-tag	200µg/ml	1:500
Recombinant Anti-*ZFY* antibody	Abcam [AB185541]	Primary	Monoclonal	Rabbit	*ZFY*	NA	1:1,000
M-IgG Fe BP-HRP	Santa Cruz [SC525409]	Secondary	Monoclonal	Mouse	HRP	100µg/mL	1:10,000
Goat Anti-Rabbit IgG	Bio-Rad [170-6515]	Secondary	Polyclonal	Rabbit	IgG (H + L)-HRP Conjugate	1.0mg/ml	1:10,000

5.2.14 Mass Spectrometry

5.2.14.1 Protein Identification

After possible identification of desired proteins via Coomassie staining and western blotting samples were sent for analysis via mass spectrometry (MS).

The actual use of the mass spectrometry machines, and training was conducted by the previous School of Biosciences MS manager, Dr Kevin Howard. Preparation of samples was performed and then handed off to Dr Howard for analysis.

Matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) MS was performed for protein identification of in-gel digested protein bands stained with Coomassie Blue.

In-gel protein digestion is a two-day process. On day 1 after cutting bands out of a gel, reduction and alkylation were performed. The gel particles were washed with 100µL of 50 mM NH_4HCO_3:acetonitrile (1:1) for 15 min. This was spun down (1 min at 5000 RPM/2,236 x g), and the liquid was removed. 100µL of acetonitrile was then added and this was left for a further 15 minutes. The samples were spun again, and the liquid was removed. The gel pieces were swelled by the addition of 10mM DTT in 50mM NH_4HCO_3, adding enough liquid to cover the gel (~75µL). Incubated for 30 minutes at 56°C. The samples were spun down again, and excess liquid was removed. The gel pieces were shrunk by adding 100µL of acetonitrile for 1 minute and then the liquid was removed. 75µL of 55mM chloroacetamide in 50mM NH_4HCO_3 was added to the gel and then incubated for 20 minutes at RT, in the dark. The gel pieces were spun down, and the chloroacetamide solution was removed. The gel pieces were washed with 100µL of 50mM NH_4HCO_3:acetonitrile (1:1) for 15 min again. The gel pieces were spun down, and all liquid was removed. This washing step was repeated with 50mM NH_4HCO_3. Then the gel pieces were shrunk again by the addition of 150µL of acetonitrile for 15 minutes. Again, these were spun down, and all liquid was removed ensuring that all liquid had been successfully removed.

The gel pieces were then rehydrated in (25mM NH_4HCO_3, 10% acetonitrile) containing 10ng/µL of trypsin on ice for 30 minutes. After 15 minutes the samples were checked to ensure the liquid hadn't been absorbed by the gel pieces. This trypsin solution was then removed and 15uL of digestion buffer (25mM NH_4HCO_3, 10% acetonitrile) was added again to cover the gel and these samples were left overnight at RT.

On day 2, the peptides were extracted, 15µL of acetonitrile was added and the samples were sonicated in an ultrasound bath for 15 minutes. The samples were spun down, and the supernatant was collected in a microcentrifuge tube. Following this 10µL of 50% acetonitrile with 5% formic acid was added and this was placed back

into the ultrasound bath for 15 minutes. Again, the samples were spun down, and the supernatant was collected. Both supernatants were combined, and these could then be analysed by MALDI-TOF (Ultraflextreme, Bruker) without further treatment.

Following this in-gel digest, 1µL of each sample was plated onto a MTP anchorchip 384 plate. The samples were left to dry and once dried 1µL of matrix solution was added and then left to dry. A 1µL calibration buffer was spotted between each four samples and again this was left to dry. The protocol used RP700_3500Da, as this was ideal for our protein size.

5.2.14.2 Intact Mass Spectrometry

To determine intact protein mass, electrospray LC-MS was performed using a Bruker micrOTOF-Q II mass spectrometer on desalted protein samples. Desalting was done by reverse-phase HPLC on a Phenomenex Jupiter C4 column (5µm, 300Å, 2.0mm x 50mm) running at 0.2 ml/min on an Agilent 1100 HPLC system, using a water/acetonitrile/0.05% TFA gradient. This elution process was monitored at 280nm & 214nm and then directed into the electrospray source, operating in positive ion mode, at 4.5 kV, and mass spectra were recorded from 500-3000m/z. The data was then analysed and deconvoluted to give uncharged protein masses using Bruker's Compass Data Analysis software.

5.2.14.3 Top-Down Sequencing

Top-down sequencing is a method for identifying proteins by comparing the sequences from the N-terminal to the C-terminal against our construct. The intact proteins (from the above samples 5.2.14.2) were ionised by electrospray ionisation and trapper. Fragmentation for tandem mass spectrometry is accomplished by electron-capture dissociation or electron-transfer dissociation. Top-down MS interrogates protein structure through the measurement of an intact mass followed by direct ion dissociation in the gas phase.

5.2.15 GFP Pull-Down

Due to being unable to express and purify both forms of *ZFY* from *E. coli* culture, an alternative method of protein expression was performed in a mammalian cell system. This method would still lead to the identification of protein-binding partners of *ZFY* via MS.

5.2.15.1 Mammalian Cell Line

The cell line used for the experiment was HEK293 cells, as previously used for the splice regulation analysis (Chapter 3) and RNA-Seq analysis (Chapter 4). The cell maintenance and culturing was performed as previously mentioned (section 3.2.5.1) ensuring that a logarithmic growth phase was maintained to keep the cells viable.

5.2.15.2 Lipofectamine 3000 Transfection

After growing up the cells until an adequate number of cells were present, the constructs **(Table 5.6)** were transfected into the cells using Lipofectamine 3000 using the same method as stated previously in section 3.2.5.2. The seeding densities, volumes, and quantities were kept the same as in previous transfections to ensure consistency between the experiments.

For this GFP-pull-down experiment, the *ZFY* constructs with the GFP tag were used alongside a pEGFP-N1 vector as a control.

Table 5.6: Plasmid constructs transfected into HEK293 cells. Both versions of *ZFY* with a GFP-tag were inserted into a pcDNA3.1(+) backbone containing an ampicillin-resistant gene. pEGFP-N1 was used as an empty GFP control.

Vector backbone	Insert	Tag	Antibiotic Resistance
pcDNA3.1(+)	h*ZFYL* (full length)	N-terminal eGFP-tag	Ampicillin
pcDNA3.1(+)	h*ZFYS* (isoform)	N-terminal eGFP-tag	Ampicillin
pEGFP-N1	NA	eGFP	Kanamycin

5.2.15.3 Cell Harvesting and Lysate Preparation

Following the 48-hour transfection period, the cells were harvested and pelleted using the following method. Harvesting of cells and cell lysis was performed with ice-cold buffers and at 4°C as much as possible to prevent protein degradation. For each 6-plate well, the cell media was collected and discarded. Then the wells were washed with ice-cold 1x PBS (Oxoid, BR0014G) twice gently to ensure the cells did not detach from the plate. The cells were then collected in 1mL of the cell media (composition as previously mentioned) and pelleted at 1,200 x g for 5 minutes at 4°C.

After obtaining the cell pellet, the GFP-Trap Agarose Kit (Chromotek, gtak-20) protocol was followed. This kit uses GFP nanobodies coupled to agarose beads to immunoprecipitated GFP-tagged proteins together with their binding partners. Cell pellets were resuspended in 200uL of ice-cold RIPA buffer (10mM Tris/Cl pH 7.5, 150mM NaCl, 0.5mM EDTA, 0.1 % SDS, 1 % Triton™ X-100, 1 % deoxycholate, 0.09 % sodium azide) supplemented with 100Kunitz U/mL DNase I (Merck, DN25-10MG),

2.5mM $MgCl_2$ (Fisher Scientific, BP214-500), Pierce protease and phosphatase inhibitor mini tablets (EDTA-free) (ThermoFisher Scientific, A32961) and Phenylmethylsulfonyl (PMSF) (Roche, 10837091001). The tubes were then placed on ice for 30 minutes and were extensively pipetted every 10 minutes to lyse the cells. Following incubation, the lysates were centrifuged at 17,000 x g for 10 minutes at 4°C. The clear lysate (supernatant) was transferred to a pre-chilled Eppendorf. At this stage, a 10µL aliquot of the lysate was taken for western blot analysis. The lysates were then diluted in 300µL of dilution buffer (10mM Tris/Cl pH 7.5, 150 mM NaCl, 0.5mM EDTA, 0.018 % sodium azide) supplemented with 1mM PMSF and mini tablet inhibitor cocktail as above.

5.2.15.4 GFP-Trap Agarose Protocol

Following lysate preparation, the agarose beads were equilibrated. The beads were gently resuspended by pipetting up and down, but not vortexed. After fully resuspending, 25µL of the bead slurry was transferred into a 1.5mL reaction tube. Then 500µL of dilution buffer was added and the beads were sedimented by centrifugation at 2,500 x g for 5 minutes at 4°C. The supernatant was then discarded. Following bead equilibration, the diluted lysate was then added to the beads and rotated end-over-end for 1 hour at 4°C. Washing the beads then followed. The beads were again sedimented by centrifuging at 2,500 x g for 5 minutes at 4°C. An aliquot of the supernatant was saved for analysis (flow-through/non-bound sample) but the remaining supernatant was discarded. The beads were then resuspended in 500µL of wash buffer (10mM Tris/Cl pH 7.5, 150mM NaCl, 0.05 % Nonidet™ P40 Substitute, 0.5mM EDTA, 0.018 % sodium azide). Again, the beads were centrifuged at 2,500 x g for 5 minutes at 4°C, the supernatant was then discarded. This washing step was repeated twice. Following the last wash step, the beads were then transferred to a clean tube.

The samples were then eluted with 2x SDS-Laemmli sample buffer (120mM Tris/Cl pH 6.8, 20% glycerol, 4% SDS, 0.04% bromophenol blue, 10% β-mercaptoethanol) as followed. The beads were resuspended in 80µl of 2x SDS-Laemmli sample buffer and then boiled for 5 minutes at 95°C to allow for the dissociation of the immunocomplexes from the beads. The samples were then centrifuged at 2,500 x g for 2 minutes at 4°C and the supernatant was taken for analysis.

SDS-page analysis was carried out, loading the input diluted lysate, the flow-through, and the elution sample (section 5.2.12). SDS-gels were Coomassie stained and transferred to PVDF membranes for western blotting as previously described, using the anti-GFP antibody in **Table 5.5**.

5.2.15.5 Proteomics via Mass Spectrometry

Following analysis of the pulldown samples by SDS-Page/Western blots, the samples were then sent for mass spectrometry and proteomics analysis. The pulldown was loaded onto an SDS-page gel and run for roughly 1cm to concentrate the sample. The band was then cut out from the bottom of the loading dye to the bottom of the well. The bands were cut into smaller pieces (~8) to make digestion easier. This method was suggested as it cleans up and simplifies the samples before tryptic digestion.

Once the band was cut, the gel pieces were digested as described in section 5.2.14.1. At each step, 50uL more of reagent was added compared to the amount described in 5.2.14.1. However, changes in the protocol occur during the overnight digest. The gel pieces were rehydrated in 50uL of digestion buffer (12.5 mM NH_4HCO_3, 10% acetonitrile) with 5ng/uL of trypsin and this was left overnight at RT.

For analysis by the nanoLCMS, the samples were vacuum-dried and resuspended in 20uL of 5% acetonitrile, and 0.1% TFA and were then cleaned up using a Pierce C18 spin tip. This step, and the use of the nanoLCMS and the proteomics analysis was carried out by Dr Kevin Howard.

Analysed data was collected and processed further. Data cleaning to reduce contamination was performed using Table 1 made by Hodge and colleagues (Hodge *et al*., 2013).

5.3 Results

5.3.1 BL21 Expression System 0.4mM IPTG 3-hour Induction

Initial protein expression and purification methods involved inducing the protein in a BL21 *E. coli* system by adding 0.4mM IPTG and incubating for 3 hours at 37°C in a shaking incubator. Before induction with IPTG, pre-induction samples were taken when the desired optical density was reached, and this was also done after induction to retrieve a post-induction sample.

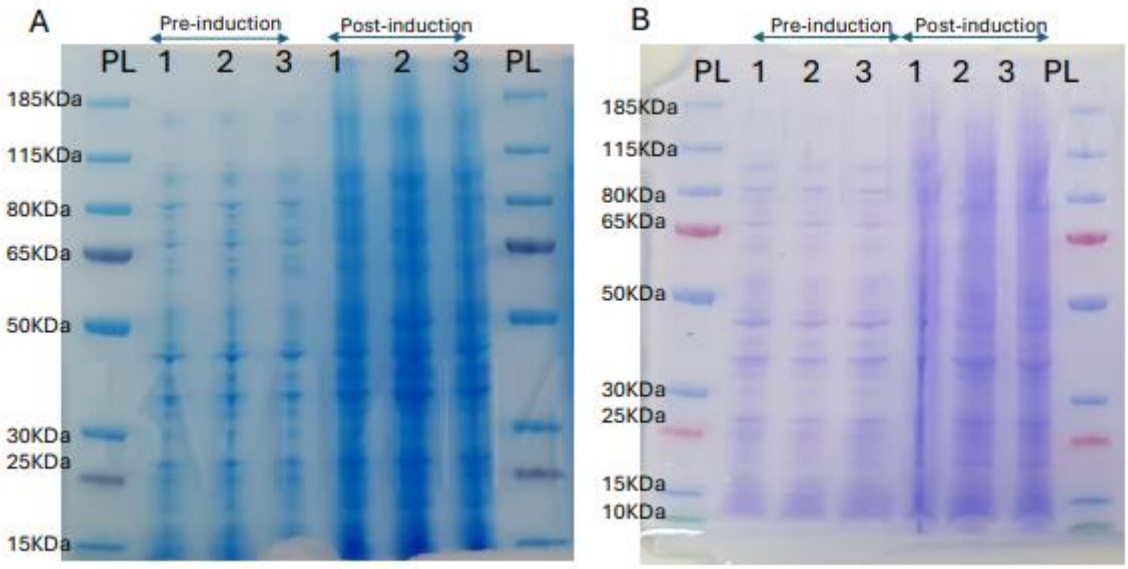


Figure 5.4: 4-12% Bis-Tris SDS gels loaded with the pre-and post-induction samples for both *ZFYS* and *ZFYL* constructs. A: *ZFYS* construct, **B:** *ZFYL* construct. **PL:** Protein ladder, **1,2 & 3** represent the 3 replicates for both the pre-and post-induction samples for each construct.

Displayed in **Figure 5.4A** are the BL21 *E. coli* colonies transformed with *ZFYS* before and after induction with IPTG which initiates the transcription of *ZFYS*. The *ZFYS* acidic domain construct had a predicted molecular weight of 26.3KDa inclusive of the His-tag, thrombin site, cloning site and *ZFYS*. Therefore, a band at this site would be expected post-IPTG induction across all repeats, with little to none expected in the pre-induction sample as transcription has not been induced. At the size-specific site of *ZFYS*, a potential band is visible across all replicates post-induction at low levels, however, this band is also visible pre-induction. This could be due to the leaky expression of the lac operon or could potentially be a native *E. coli* protein with the same molecular weight as *ZFYS*. Other bands can be seen across the pre- and post-induction samples at varying molecular sizes, with banding seen in the 30KDa-50KDa range. **Figure 5.4B** displays the BL21 *E. coli* colonies transformed with *ZFYL* before and after induction. The *ZFYL* acidic domain construct had a predicted molecular weight of 48.9KDa inclusive of the His-tag, thrombin site, cloning site and *ZFYL*. At

the size-specific site of *ZFYL*, a potential protein band is visible, but it seems as if the levels are even across the pre- and post-induction samples, which would indicate that this band could be an *E. coli* native protein. The banding pattern at this size site is also seen in **Figure 5.4A** for the *ZFYS*-transformed BL21 cells, again indicating this protein band might not be *ZFYL*. This potentially indicates that *ZFYL* is not expressing well in these conditions. It was also noted that streaky banding patterns are visible in **Figure 5.4** suggesting overloading of the wells or potential DNA contamination.

Following IPTG induction, the samples were lysed by sonication. This sonication protocol was as follows, intermediate micro-probe at 20 amps for a total of 2 minutes (cycles of 10s on and 10s) on ice.

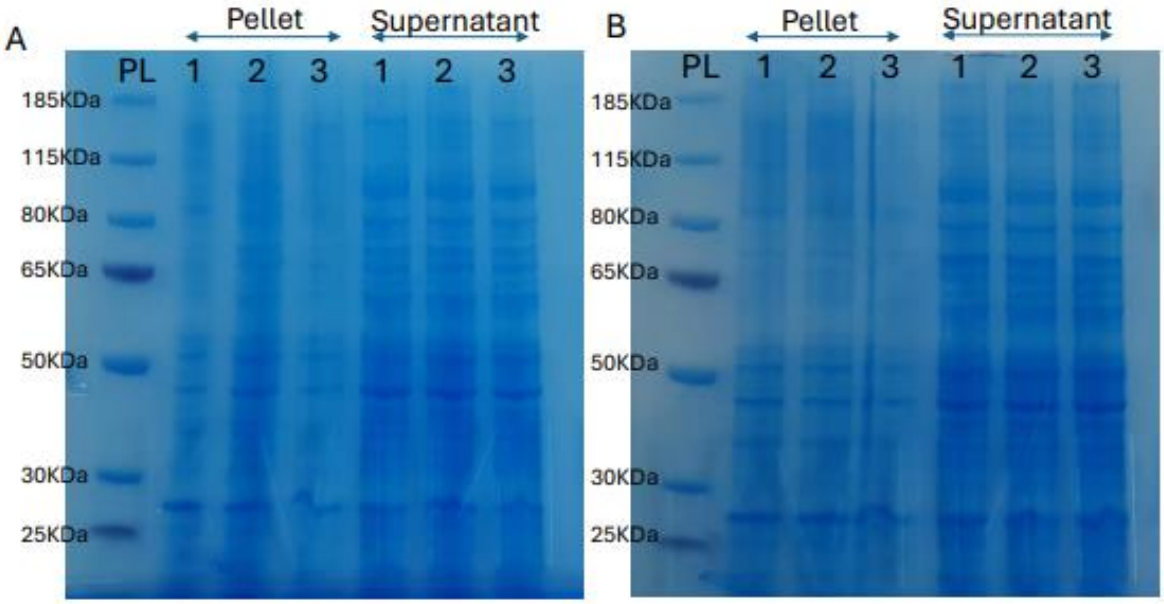


Figure 5.5: 4-12% Bis-Tris SDS gels to confirm successful lysis of the *E. coli* cells. A: *ZFYS* construct, **B:** *ZFYL* construct. **PL:** Protein ladder, **1,2 & 3** represent the 3 replicates for both the pellet and supernatant samples for each construct post-sonication.

Successful lysis would be indicated by the presence of the desired protein bands in the supernatant, as the pellet contains all the insoluble material of the *E. coli* cells. **Figure 5.5A/B** indicates that this lysis method requires amendment as a large amount of protein visible in the supernatant is also seen in the pellet. The bands at the expected weights for *ZFYS* and *ZFYL* are present in the supernatant suggesting that these proteins are soluble if they are confirmed, but bands are still visible in the pellet. However, it should be noted that the pellet is 5x the concentration of the supernatant and therefore, the amount of protein is not as high as presumed by **Figure 5.5.** The major issue though is that these protein bands are evident in both the *ZFYS* and *ZFYL* samples, which leads to further uncertainty about their identity. *ZFYL* may be cleaved

in the *E. coli* system resulting in a *ZFYS* size band, but the reverse of this is not possible. At this point, it is hard to confirm whether *ZFYS* and *ZFYL* are being expressed. Further improvements to the lysis method are also required to improve the supernatant protein yield.

Nickel column purification followed cell lysis. The supernatant collected from sonication was taken and loaded onto the column.

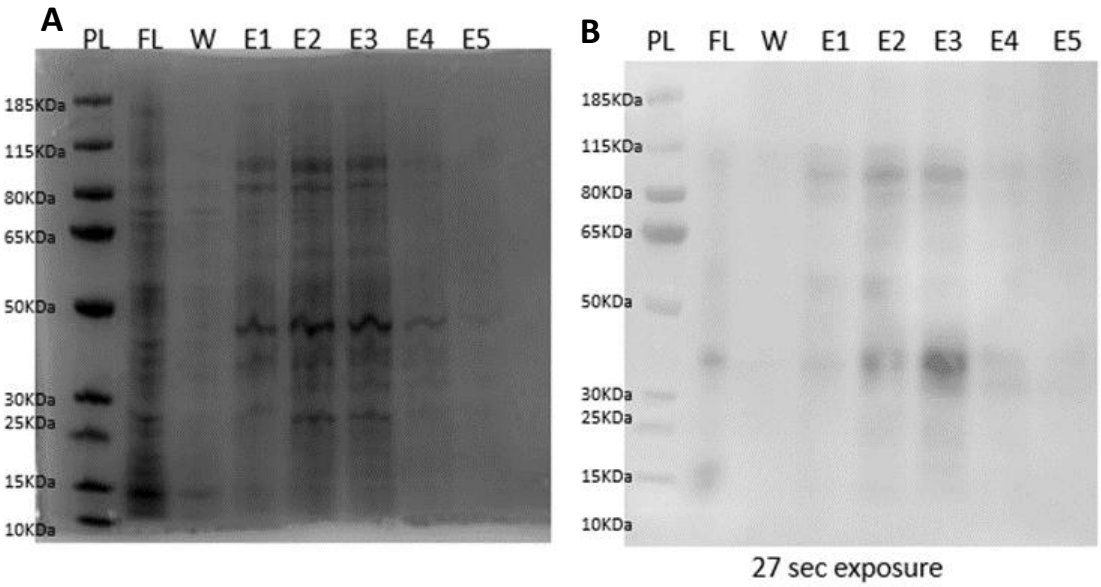


Figure 5.6: Nickel column purification results for *ZFYS*. A: Coomassie-stained SDS-page gel & **B:** Western blot using an anti-His antibody. Both images were taken on the ChemiDoc Imaging system. **PL:** Protein ladder, **FL:** supernatant flow through, **W:** wash through, **E1-E5:** Elution fractions collected. 4-12% Bis-Tris SDS-Page gel.

Figure 5.6 shows the post nickel chromatography purification factions collected and subsequently analysed by SDS-Page and western blotting for *ZFYS*. It is clear that the washing step is not as effective as hoped, which is made evident in **Figure 5.6A** as many proteins are present in the elution fractions, which is when an imidazole-based buffer is added to dislodge the his-tagged protein from the nickel in the resin. A band at ~26KDa is strongly present in elution fractions E2 and E3, but these bands do not produce a signal in **Figure 5.6B** indicating they do not have a His-tag present suggesting that this is not *ZFYS*. A further strong signal is present at ~48KDa in the elution fractions, however, this protein doesn't produce an antibody signal. A His-antibody signal is seen ~35KDa **(Figure 5.6B)**, with faint Coomassie-stained bands visible corresponding to this weight **(Figure 5.6A)**. This signal is higher than expected which raises questions. This larger band size can be explained by the large number of negatively charged residues in the protein. An equation has been derived to account for negatively charged residues by Guan and colleagues which is $y = 276.5x$

− 31.33 (where x is the percentage of negatively charged amino acids and y is the average MW per amino acid) (Guan *et al*., 2015). Using this equation, the *ZFYS* construct is predicted to run at 33.14kDa, very similar to the band observed by the western. It is therefore possible that this is *ZFYS* being expressed although at very low levels. It is also noted that this protein band is visible in the flow through, which suggests the His-tag is binding weakly, resulting in protein loss. Further His-antibody signals are seen at higher molecular weights, but this is suspected to be cross-reactivity.

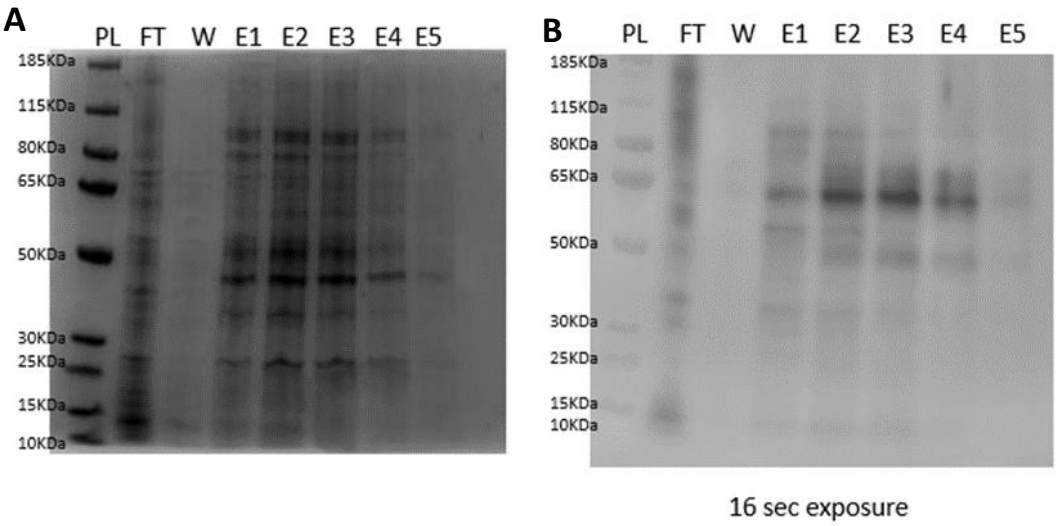


Figure 5.7: Nickel column purification results for *ZFYL*. A: Coomassie-stained SDS-page gel & **B:** Western blot using an anti-His antibody. Both images were taken on the ChemiDoc Imaging system. **PL:** Protein ladder, **FL:** supernatant flow through, **W:** wash through, **E1-E5:** Elution fractions collected. 4-12% Bis-Tris SDS-Page gel.

Shown in **Figure 5.7** is the post nickel chromatography purification factions collected and subsequently analysed by SDS-Page and western blotting for *ZFYL*. The wash buffer is not sufficient for removing loosely bound proteins from the column, as many proteins subsequently appear in the elution fractions, making the fractions very messy which is evident in **Figure 5.7A**. Many of the same banding patterns as seen in the *ZFYS* fractions are visible in **Figure 5.7A** and suggest a lot of *E. coli* protein contaminants which are not being removed by the wash buffer. **Figure 5.7B** shows a strong His-antibody signal within the elution fractions E1, E2, E3, and E4 running at a molecular weight of ~60KDa but this band is not as visible by Coomassie staining solely. This indicates the protein is expressed at low yields as immunoblot sensitivity is much higher. This protein runs higher than *ZFYL* is expected to, however, when considering the high negative charge of *ZFY* and using the formula to account for this, a predicted running molecular weight of 61.23KDa is expected for this construct. Therefore, it can be presumed that this signal is the result of *ZFYL* expression. Again,

this antibody seems to have a high cross-reactivity level, indicated by the presence of other bands even at a short exposure time.

From **Figure 5.6** and **Figure 5.7,** it is clear that *ZFYS* and *ZFYL* yield is low, as bands are only detectable by western blot, which is 100-1000x more sensitive than Coomassie staining. This indicates that this induction methodology is not inducing enough *ZFY* expression and requires alteration. Furthermore, it is evident that both the lysis and nickel chromatography methods need optimisation to ensure cell lysis is fully complete and improve the yield/purification of *ZFY* from the column (see section 5.3.6). In an effort to enhance yields, the lysis buffer was supplemented with proteinase and phosphatase inhibitors to protect against potential protein degradation.

5.3.2 BL21 Expression System 0.8mM IPTG Overnight Induction

In efforts to enhance the production of both *ZFY* isoforms, the IPTG concentration was elevated from 0.4mM to 0.8mM to more potently drive induction. Simultaneously, prolonged overnight incubation at 30°C aimed to permit adequate time for protein production while balancing potential protein degradation during prolonged incubation periods.

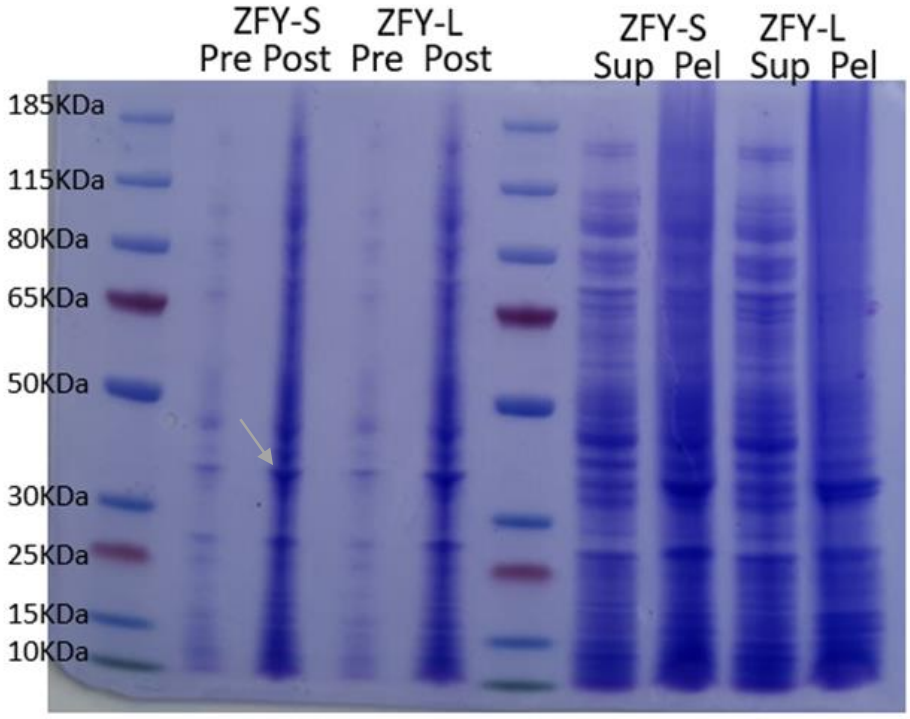


Figure 5.8: 4-12% Bis-Tris SDS gels loaded with the pre- and post-induction sample for both the *ZFYS* and *ZFYL* constructs, as well as the lysed samples following sonication. Pre: pre-induction, **Post:** post-induction, **Sup:** supernatant, **Pel:** pellet.

Displayed in **Figure 5.8** is the protein expression for *ZFYS* and *ZFYL* pre- and post-induction after implementing optimised growth parameters. Following the results shown in section 5.3.1 it is known that *ZFYS* and *ZFYL* run at higher molecular weights than expected at ~33kDa and ~61KDa respectively. Compared to previous results, *ZFYS* yields noticeably increased relative to earlier attempts, with a pronounced ~33kDa band visible post-induction (grey arrow). A band at the same weight is also present in the *ZFYL* construct across all lanes, this could be the result of protein degradation or premature translation termination. Furthermore, leaky expression is still apparent with the same band being visible in the pre-induction lane. In contrast, the expected ~61kDa *ZFYL* product remains difficult to identify post-induction, suggestive of inadequate full-length construct expression despite adjustments. Following induction, the cell pellets were lysed by sonication using a lengthened more aggressive method. The sonication time was increased to a total of 5 minutes with 20-second cycles, compared to the previous 2 minutes with 10-second cycles to hopefully improve cell lysis. Even with this increase in sonication, a high yield of *ZFY* protein exists in the pellet. This could indicate that the protein is both soluble and insoluble, but this is not likely. Regardless, *ZFYS* is more easily detectable post-lysis while negligible *ZFYL* appears in the lysate supernatant even after enhanced disruption attempts.

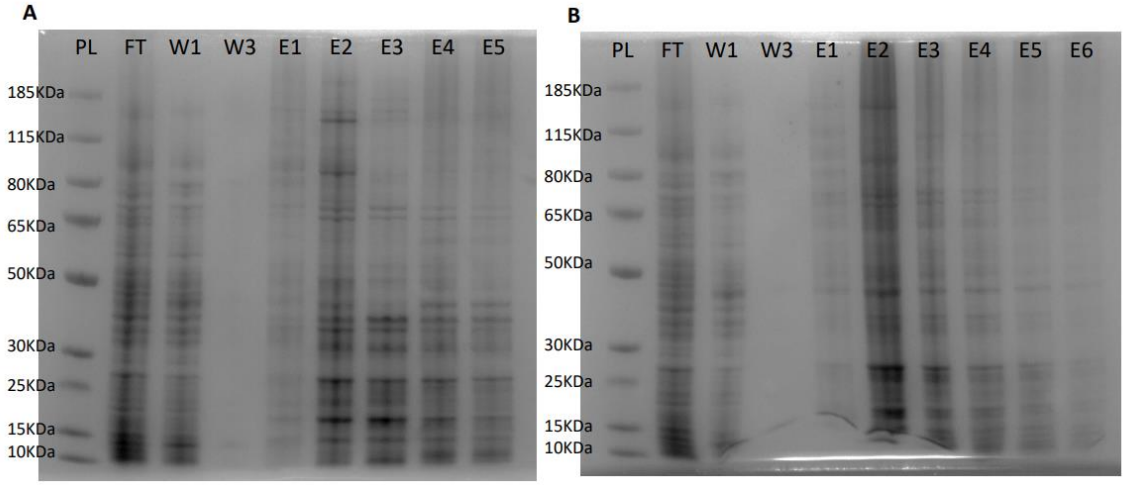


Figure 5.9: 4-12% Bis-Tris SDS-Page gel for the fractions collected from the nickel column purification. A: *ZFYS* construct nickel column purification & **B**: *ZFYL* construct nickel column purification. **PL:** Protein ladder, **FT:** Supernatant flow through, **W1 & W3:** wash through fractions, **E1-E6:** Elution fractions.

The nickel column purification profile in **Figure 5.9** is after implementing modified sonication times during the lysis process. In both the *ZFYS* and *ZFYL* fractions, initial flow-through (FT) and wash (W1, W3) fractions are depleted of numerous loosely adhered proteins. Subsequent elutions (E1-E5/6) following the addition of high imidazole concentrations the his-tagged constructs were displaced. In **Figure 5.9A** a

band is present at the predicted ~33KDa size however, as there is no clear dominant band visible in **Figure 5.9A** at this time it is hard to conclusively say if there is successful *ZFYS* production. But it is reassuring that **Figure 5.9B** does not show as much signal at this size. **Figure 5.9B** demonstrates the continued low yields of *ZFYL*, with a potential band at the expected size for *ZFYL*, but again confirmation is difficult due to low signal and messy background due to other protein contaminants. Moreover, shown in **Figure 5.9A** significant potential *ZFYS* loss is seen in the FT and wash fractions, indicating that *ZFYS* may not be binding sufficiently to the column. New resin was subsequently used to ensure that nickel binding was sufficient.

Following this result, expression and purification were trialled again using the same induction method but altering the lysis method further to aid in improving protein yield. Subsequent improvements to the lysis method included increasing the extension of the incubation times in the cell lysis protocol to lengthen lysozyme, triton, and DNase I activity in the lysis process. Furthermore, the sonication time was further increased to a total of 7 minutes with 20-second cycles. An ultracentrifuge was then utilised to spin down the lysates at a much higher speed of 40,000 RPM (165,000 x g).

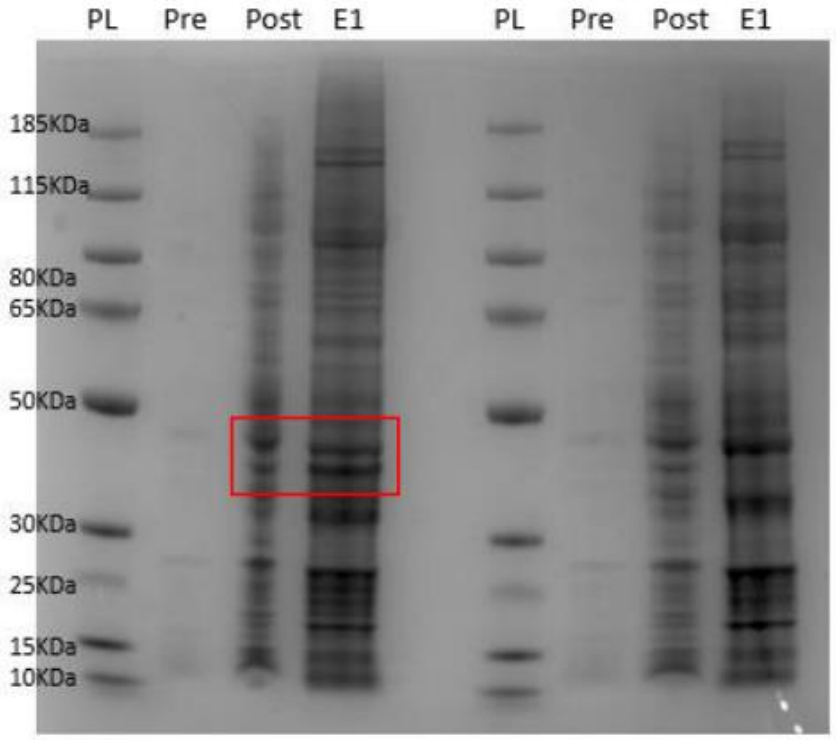


Figure 5.10: 4-12% Bis-Tris SDS-Page gel with the pre- and post-induction samples for both *ZFYS* (left) and *ZFYL* (right) alongside the first elution fraction collected from the Nickel gravity column. PL: Protein Ladder, **Pre:** Pre-induction sample, **Post:** Post-induction sample and **E1:** elution fraction 1.

Yield improvements (*ZFYS* more than *ZFYL*) were noted when lysis method alterations were implemented however, it was noted that *ZFYS* purification often

displayed a characteristic doublet banding pattern as highlighted by the red box in **Figure 5.10**. This recurring two-band profile prompts questions regarding their respective identities and relative quantities. If the upper band indeed represents full-length *ZFYS* while the lower constitutes a processed or degraded *ZFYS* protein, overall *ZFYS* yields may exceed original estimates. However, such truncation would also indicate the short isoform is vulnerable to proteolytic cleavage or instability. A similar faint upper band sometimes visible for *ZFYL* could similarly suggest splicing/degradation phenomena may affect both isoforms when expressed in bacteria. However, due to consistently low *ZFYL* outputs, duplicate bands are more difficult to conclusively discern. Further scrutiny of the two visible *ZFYS* products is warranted to determine if changing the protocol conditions will increase stability.

Due to the low yields, a western blot was used to confirm the double banding pattern seen in **Figure 5.10**. Is *ZFY* both the short and long isoforms being targeted for degradation in this BL21 expression system?

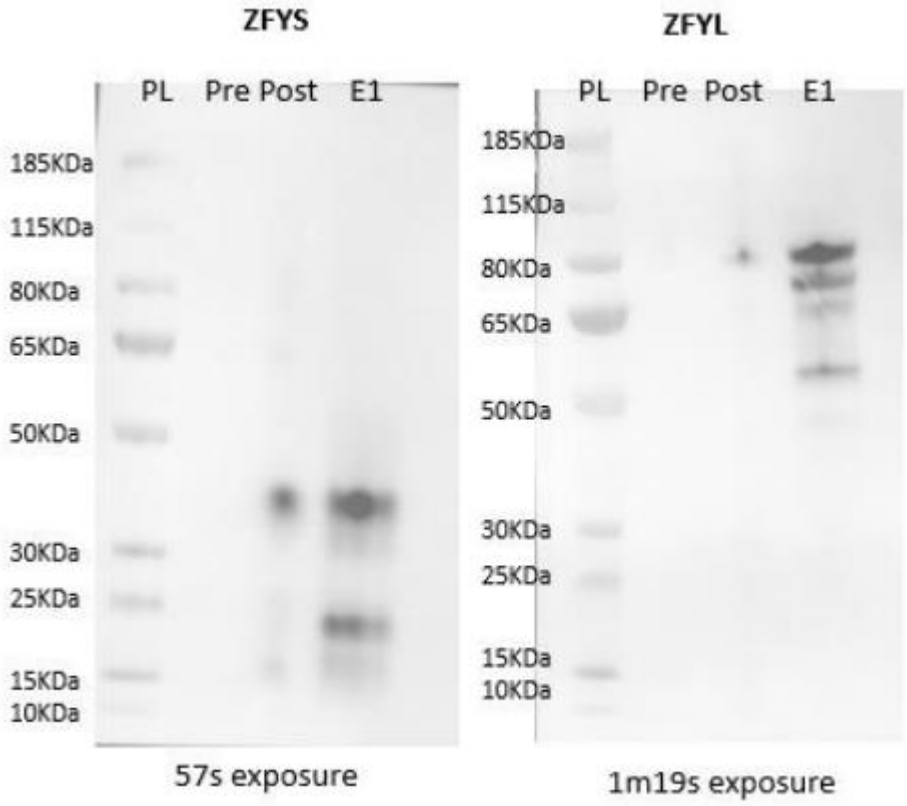


Figure 5.11: Western blots using the anti-His antibody for both the *ZFYS* and *ZFYL* constructs. PL: Protein Ladder, **Pre:** pre-induction sample, **Post:** post-induction sample, **E1:** Elution fraction 1.

Suspicions of possible degradation events during the lysis and purification of *ZFYL* and *ZFYS* are confirmed in **Figure 5.11**. Using the His-antibody, clear additional

bands are seen post-purification which are not as evident before lysis in the post-induction samples, for both *ZFYS* and *ZFYL*. In **Figure 5.11**, *ZFYS* shows a strong signal for the lower band seen in the red box in **Figure 5.10** in E1, with a potential merging of the upper and lower band signal due to the closeness of the bands. This suggests that the possible truncated *ZFYS* protein is favoured by *E. coli*. However, strong His signal is also seen at lower weights between 15KDa and 25KDa indicating large degradation events of the C-terminus end, leaving the His-tag untouched allowing for binding and antibody detection. These bands are not present in the post-induction sample, which indicates that these degradation events are occurring during the lysis process.

The double banding pattern is evident in **Figure 5.11** *ZFYL* samples, with two strong band signals seen at ~80KDa which is higher than expected. Subsequent lower bands are also present indicating protein degradation events. However, this western shows that the Coomassie stained band in *ZFYL* similar to *ZFYS* is not a spliced version resembling *ZFYS* since no His-signal is detectable. This indicates an *E. coli* protein is expressed with a similar molecular weight.

Despite the potential degradation of the proteins, the yield seems substantially higher once the expression, lysis and chromatography methods have been adapted and optimised.

5.3.3 Mass Spectrometry Confirmation

As the only confirmation so far for successful protein expression is via western blot further confirmation of *ZFYS* and *ZFYL* was important before continuing with optimisation steps. Mass spectrometry was performed to determine the intact mass of the proteins. The elutions were desalted by dialysis. Electrospray mass spectra were recorded on a Bruker micrOTOF-Q II mass spectrometer by taking an aliquot of the sample and further desalting it on line by reverse-phase HPLC on a Phenomenex Jupiter C4 column. Note that *ZFYL* had an insufficient protein yield, so intact mass spectrometry at this time could not be completed.

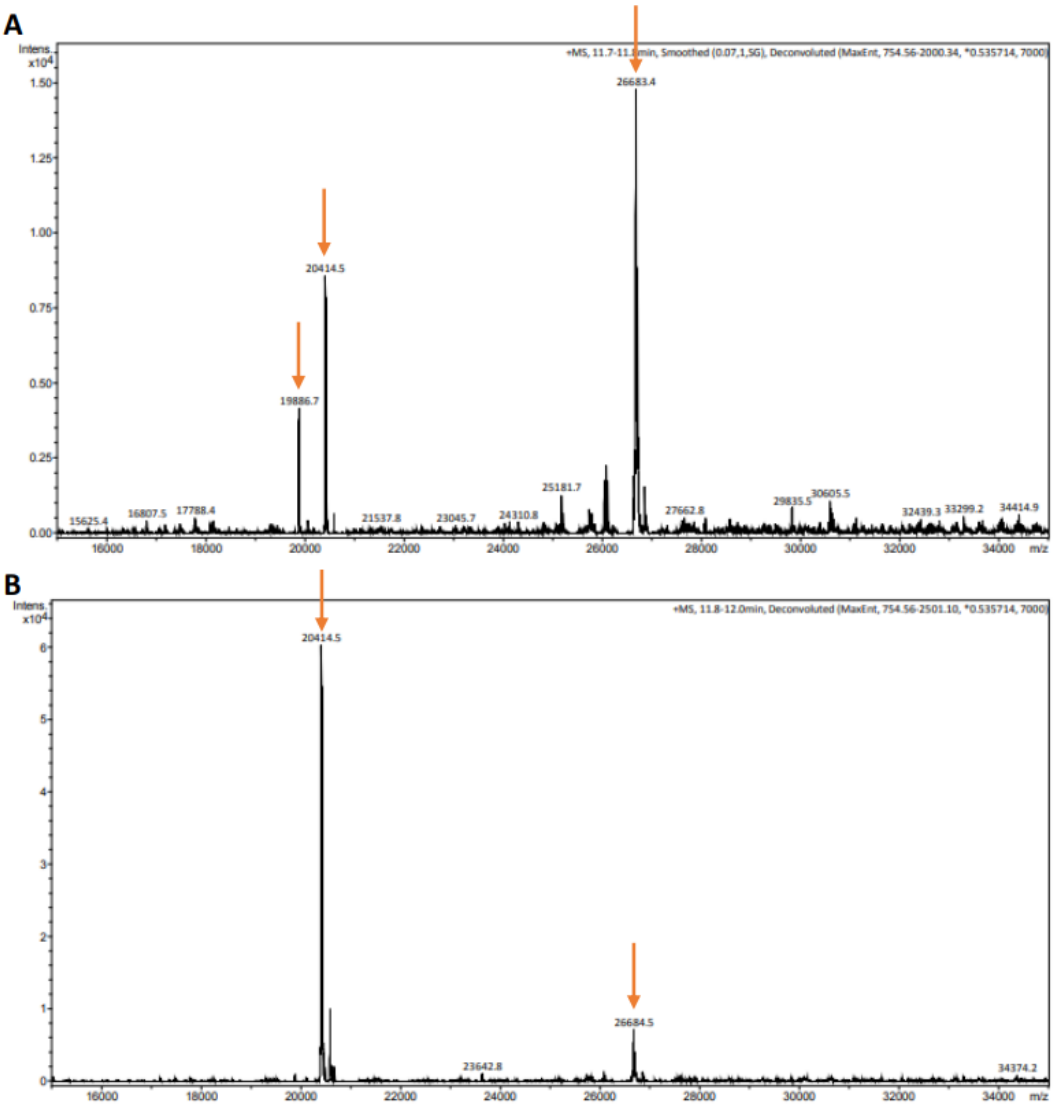


Figure 5.12: Electrospray LC-MS intact protein mass for *ZFYS* showing the desalting chromatograms at 214nm. A: Shows the raw mass charge envelope produced when analysing a protein by mass spectrometry and **B:** shows the final deconvoluted data.

ZFYS predicted weight including the His-tag is 26270.69 Daltons (26.3KDa). **Figure 5.12** showed two major peaks in the sample; one at 26683.4 Daltons (26.7KDa) and a second at 20414.5 Daltons (20.4KDa), with the 20.4KDa peak being the most abundant **(Figure 5.12A and 5.12B)**. The 26.6KDa peak eluted earlier on the water, acetonitrile, 0.05% trifluoroacetic acid gradient which suggests that this protein is slightly more hydrophilic than the 20.4KDa peak. **Figure 5.12A** also shows trace amounts of a third peak at 19886.6 Daltons (19.9KDa) which eluted alongside the 26.6KDa peak on the gradient. This peak is absent in **Figure 5.12B** following deconvolution of the data.

A 413 Dalton difference is therefore seen between the *ZFYS* expected molecular weight and the weight of the upper band. This could be the result of modifications or adducts but the cause is generally unknown. However, it should be noted that the initial methionine would be cleaved as it is common in *E. coli* expressed proteins to be removed when the second residue is a glycine, which is the case here. This would result in an expected weight of 26123.3Da (26.1KDa) meaning the difference is actually 147.39Da. The mass difference between the upper and lower band is 6268.9 Daltons (6.27KDa), which corresponds roughly to a loss of the final 57 amino acids. This could be the result of a cleavage event or a premature termination of translation. Since intact mass analysis yielded unclear size discrepancies, the identity of the purified proteins was subsequently verified by mass spectrometry-based peptide mapping. The *ZFYS* and *ZFYL* bands were carefully cut out from the SDS-Page gel and an in-gel digest was performed to extract the peptides.

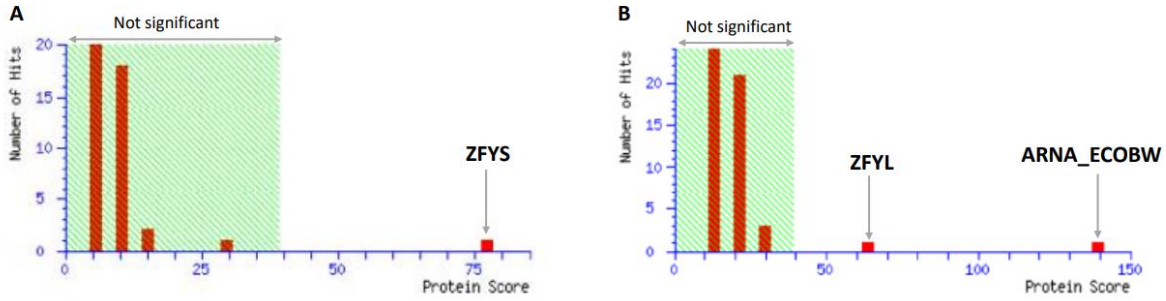
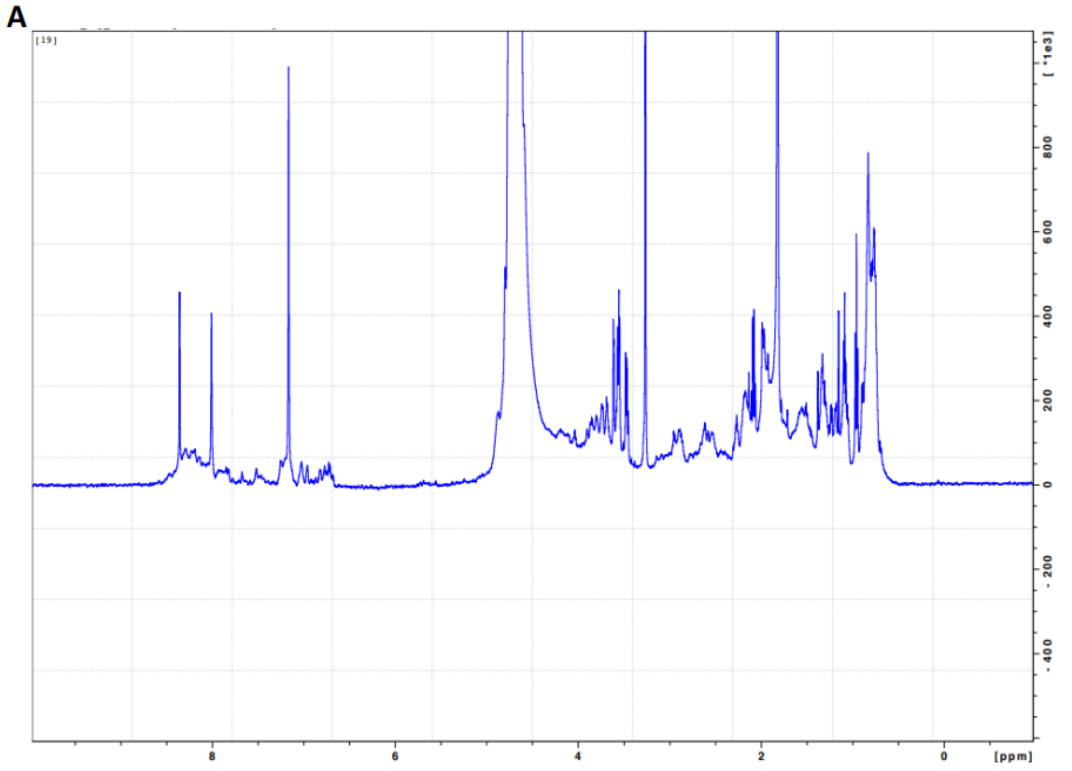


Figure 5.13: MS/MS protein identification. A: *ZFYS* predicted digested band & **B:** *ZFYL* predicted digested band. The green area highlights proteins with a protein score deemed to not be significant by the analysis platform.

Figure 5.13A and **Figure 5.13B** show significant hits for *ZFYS* (~76) and *ZFYL* (~62) following the digestion of gel protein bands. Although *ZFYL* is a significant protein hit in **Figure 5.13B** a second protein ARNA_ECOBW also has a significant protein hit with a higher protein score of ~140. This protein encodes for Biofunctional polymyxin resistance protein ArnA, an enzyme responsible for catalysing the oxidative decarboxylation of UDP-glucuronic acid. ARNA_ECOBW has a molecular weight of 74,289Da (74.3KDa) which is very similar to the expected weight of *ZFYL*. It can be presumed that ARNA_ECOBW is expressed by *E. coli* in the background and then was subsequently cut out of the SDS-Page gel alongside *ZFYL*. These *ZFYS* and *ZFYL* peptide hits were promising and confirmed that the induction and purification protocols did indeed yield detectable amounts of both *ZFYL* and *ZFYS* proteins, albeit in low yield and at this stage still relatively impure.

Following peptide fingerprinting mass spectrometry, the samples were subsequently sent for 1D NMR to identify the potential structural arrangement of *ZFY*. NMR was carried out by Dr Gary Thompson at the University of Kent. By looking at the NMR spectra the presence of different chemical environments can hopefully be identified. Identification of the folding pattern is also possible using NMR. Due to the low concentration of the samples, they were run for a prolonged period.



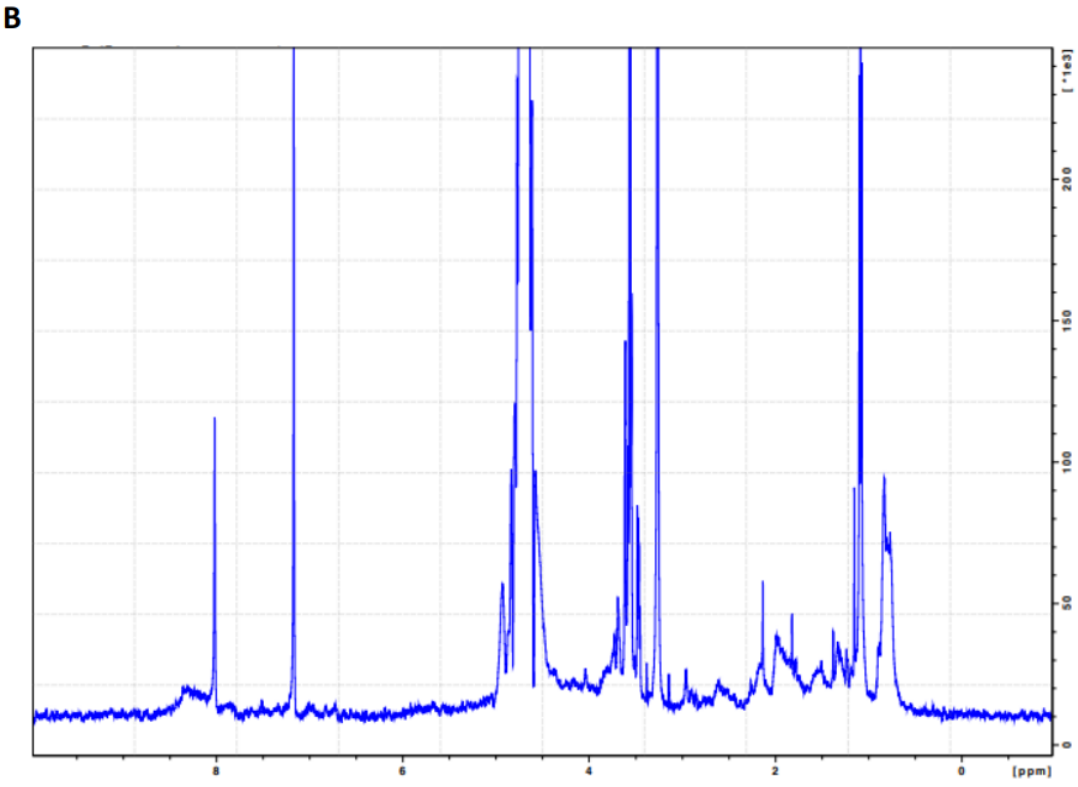


Figure 5.14: 1D NMR spectra A: *ZFYS* & **B:** *ZFYL*. Spectra is plotted with frequency (parts per million, ppm) on the horizontal axis and intensity on the vertical axis.

Whilst the protein concentration was low, some details regarding the structure of *ZFYS* and *ZFYL* were identifiable. Both spectra in **Figure 5.14A** and **Figure 5.14B** show significant contaminant peaks including imidazole and glycerol. These contaminants have arisen despite the samples being purified on the nickel column and dialysed not once but twice. The peaks in the region of ~8ppm are likely to be unfolded protein, whilst the peaks in the region 0.5-4ppm are consistent with unfolded peptides as they are sharp with little dispersion. Moreover, there is an absence of methyl peaks between 0.5-1.5ppm which further suggests an unfolded protein state. Spectra's A and B show similar peak patterns, suggesting high similarity between the two variants. Overall, it can be inferred that both variants are mostly unfolded proteins, but potentially not completely unfolded. This is consistent with the expected structural arrangements of proteins with hydrophobic 9aaTAD patches within the acidic domain as they are more likely to fold into short helices due to electrostatic charges. Seeing a similar pattern for both the *ZFYS* and *ZFYL* proteins is a good indicator that both versions are expressing as expected. This alongside the peptide matching mass spectrometry is hinting towards successful *ZFY* expression.

However, this current method of expression and purification in combination are producing low yields as well as truncated variants, with *ZFYL* expression being so poor that intact mass spectrometry could not be performed. It has previously been thought that *ZFYS* could be toxic to the cell, yet *ZFYL* seems to be expressed more poorly. Therefore, a plating assay was performed to see how the colonies change physiologically following the induction of *ZFYS*/L.

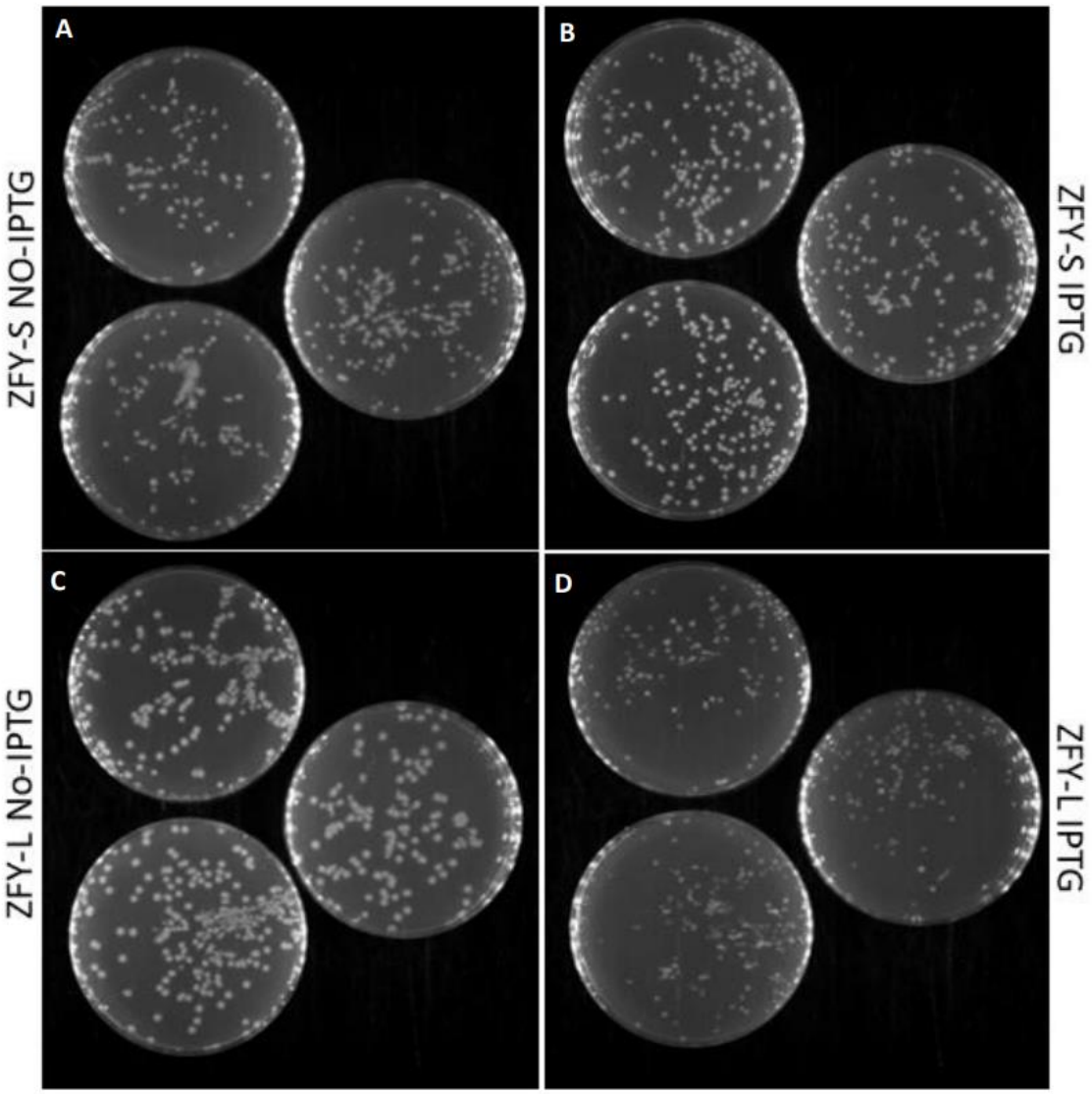


Figure 5.15: Plating assay LB plates. A: *ZFYS* No-IPTG colonies, **B:** *ZFYS* IPTG induced colonies, **C:** *ZFYL* No-IPTG colonies, **D:** *ZFYL* IPTG induced colonies. 10^3 cell dilution was prepared and spread onto the selection plates and incubated overnight at 37°C. The colonies were subsequently counted.

Table 5.7: Colony count from plating assay shown in Figure 5.15. The number of colonies from each plate, with the calculated average for each. All conditions were done in triplicates.

	ZFYS	1	2	3	Average	*ZFYL*	1	2	3	Average
No-IPTG	10^3	195	159	183	179	10^3	142	242	203	197
IPTG	10^3	108	150	100	119,3	10^3	178	137	105	140

Using **Figure 5.15** and **Table 5.7** from the plating assay notes on colony size and numbers could be made. **Figure 5.15** shows that BL21 colonies remain similar in size pre- and post-induction when transformed with the *ZFYS* construct, whilst for *ZFYL* transformed cells the size of the colonies seems to be smaller when induced with

IPTG activating *ZFYL* transcription. The reasoning behind this size change is unknown, but this could hint towards *ZFYL* having a toxic trait rather than *ZFYS* and could explain the lower protein yield of *ZFYL* seen this far. *ZFYL* although smaller in size post-induction, the number of colonies is greater compared to *ZFYS* and this is also evident in the pre-induction count as well **(Table 5.7)**. It was noted that for both variants, that a reduction in colony number was seen post-induction. If induction is indeed toxic, then fewer bacteria colonies would be expected to survive and grow. This could hint that the induction of both *ZFY* variants acidic domain is having a non-specific toxic effect.

So far, *ZFYS* and *ZFYL* expression has been confirmed, but yield and purity is low. Furthermore, protein degradation is very evident and *ZFYL* has been potentially identified as having a toxic effect on the cells based on the low yields and reduced colony size and number post-induction. Further optimisation is required in order to yield sufficient, pure protein usable for downstream applications.

5.3.4 Rosetta Cell Expression System

In an attempt to increase protein yield and stop protein cleavage, we considered whether codon usage might be a factor. Mammalian and bacterial cells typically have different favoured codons for each amino acid, and thus the native human sequence in our constructs might be poorly translated in *E. coli*. To test this hypothesis, we repeated the expression experiments using Rosetta cells. These are a derivative of BL21 that carries additional tRNAs for AGG, AGA, AUA, CUA, CCC, GGA codons on a compatible chloramphenicol-resistant plasmid, enabling them to efficiently translate these specific codons which are prevalent in mammalian sequences but suboptimal for protein translation in *E. coli* (Tegel *et al*., 2010). All other protocol steps remained the same using the optimised conditions determined previously.

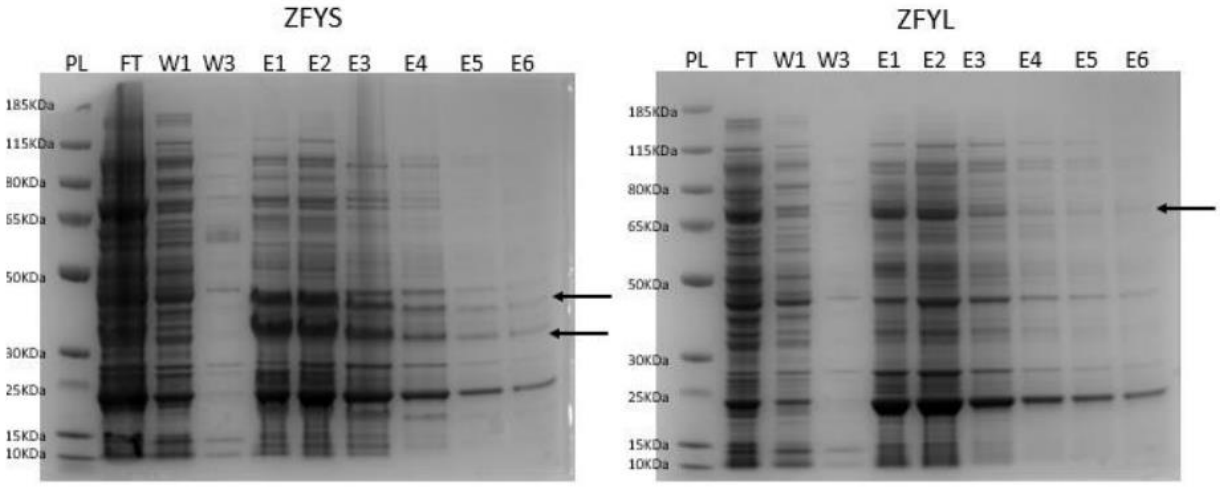


Figure 5.16: Nickel column purification fractions collected following Rosetta cell protein induction and lysis. PL: Protein Ladder, **FT:** flow through, **W1-W3:** Wash through, **E1-E6:** Elution fractions. 4-12% Bis-Tris SDS-Page gel.

As indicated by the arrows in **Figure 5.16**, transitioning protein production to Rosetta *E. coli* cells somewhat boosted yields for both *ZFY* isoforms following equivalent nickel chromatography compared to the poor BL21 outputs (refer to **Figure 5.9)**. The optimised Rosetta system produced a stronger expression of the ~33kDa potentially truncated *ZFYS* protein across all elutions, with the previously noted ~45kDa upper *ZFYS* band also being highly visible. Despite an earlier lack of western signal, its clear presence and nickel enrichment indicate abundant production of both protein bands from this construct worthy of further study. This could hint that a cleavage (or premature termination) event is still occurring in this expression system resulting in the band pairs. A major contrast to the BL21 expression is seen for *ZFYL* in **Figure 5.16** as the previously weak *ZFYL* signal emerges more strongly and is now clearly identifiable via Coomassie staining which was not previously possible. There is still a strong signal at about ~23KDa for both constructs, this is possibly a cleavage event occurring in both constructs. This result indicates promising yield improvements following the move to the Rosetta cells, but there is still a lower abundance of *ZFYL* compared to *ZFYS* even in this optimised system. Additionally, notable *ZFY* abundance in the flow-through fractions implies a portion fails to bind the nickel resin as intended. His-tag inaccessibility due to misfolding could prevent interaction, though imidazole carryover from incomplete prior column washing provides an alternative potential explanation. If the low binding affinity is due to protein structural issues, improper folding is likely contributing to the continued lower recovery of full-length *ZFYL* since solubility and intractability depend on appropriate conformational states.

ZFY abundance for both isoforms is more evident however purity is still low. To further increase the purity of both protein isoforms, anion exchange was performed to exploit *ZFY*'s highly negative charge.

5.3.5 Anion Exchange Chromatography

Anion exchange is a form of ion exchange chromatography, which uses the net surface charge of proteins for separation. Anion exchange uses a positively charged ion exchange resin with an affinity for negatively charged surface areas. Proteins applied to an ion exchange column will elute at different times within a salt gradient based on their charge.

Anion exchange was performed on the AKTA start system. Nickel column elution fractions were dialysed in preparation for anion exchange and were added to the column to monitor elution following the addition of a salt gradient.

This first attempt of anion exchange used the following buffer composition; buffer A was 25mM piperazine, pH 5.4 (0% salt) and buffer B was 25mM piperazine with 500mM NaCl, pH 5.4 (100% salt).

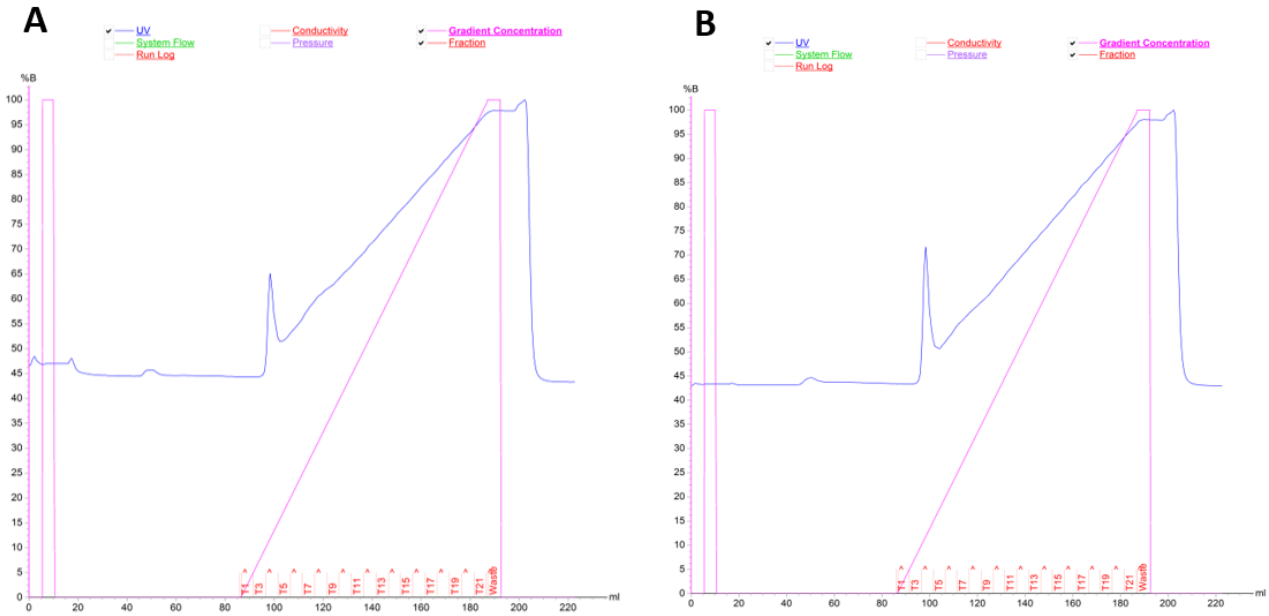


Figure 5.17: AKTA-Start anion exchange graphs. A: *ZFYS* elutions & **B:** *ZFYL* elutions. %B plotted on the Y-axis shows the percentage of salt added to the column, with the X-axis representing the fraction collected.

Initial anion exchange chromatography profiles reveal sharp elution peaks within the T5 fraction for both *ZFY* isoforms (refer to **Figure 5.17**). A salt concentration of ~75mM (15%) resulted in the elutions of both protein peaks in **Figure 5.17A** and **Figure 5.17B**. This similar protein behaviour is expected between the constructs due to their similar charge properties and subsequent ionic interactions. Such aligned

behaviours on the column provides encouraging analytical evidence that *ZFY* is being purified rather than anomalous contaminants. However, subsequent analysis of the peak contents remains imperative to conclusively confirm whether the prominent signals indeed correspond to selective elution of *ZFYS* and *ZFYL* specifically versus potential co-eluting background proteins.

The T5 fraction, along with the adjacent fractions around the identified peak, was then visualised through SDS-page electrophoresis.

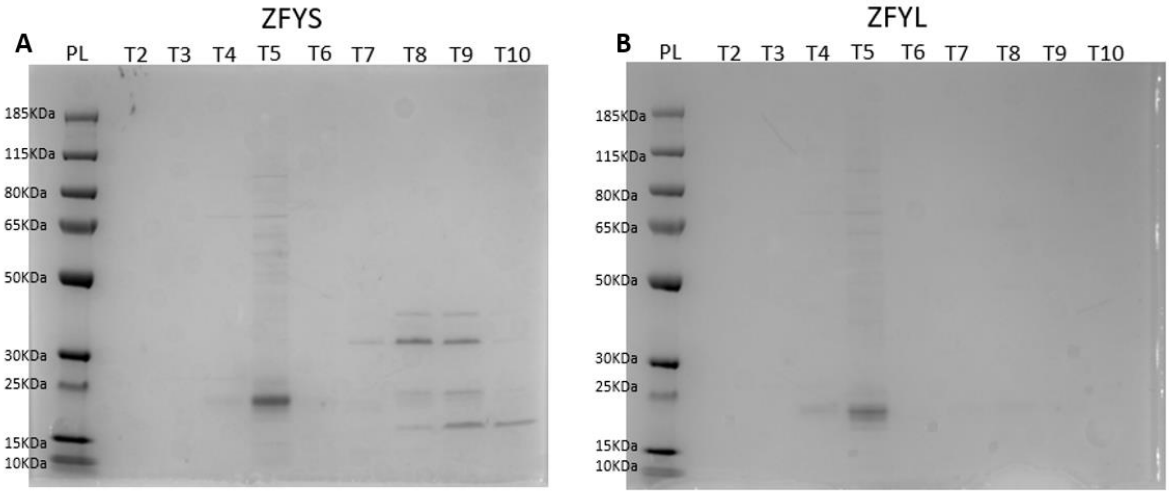


Figure 5.18: 4-12% Bis-Tris SDS-Page gel of AKTA anion exchange fractions collected from both A: *ZFYS* and **B:** *ZFYL* runs. **PL:** Protein ladder, **T2-T10:** Elution fractions from the ion exchange.

Upon analysing the contents of fraction T5 for both samples shown in **Figure 5.18**, the prominent peak species reflects a protein with a molecular weight of ~23KDa, with this protein band being very abundant throughout the optimisation process. This band has consistently been highly abundant following the expression and purification of both constructs and even produced a His-antibody signal in the *ZFYS* western blot (refer to **Figure 5.11**). This aligns with this protein being a potential cleavage product rather than an intended full-length protein. Other background proteins are also evident in the T5 fraction, suggesting that this salt concentration of 75mM is removing loosely bound proteins with a higher charge to *ZFY*. **Figure 5.18B** shows no eluted protein bands in the surrounding fractions matching the previously identified weight of *ZFYL*. However, **Figure 5.18A** has bands matching previous *ZFYS* results in T7, T8 and T9, with doublet bands being visible across the fractions. Looking back at **Figure 5.17** there was a potential bump noted in the UV line which could correspond to the elution of *ZFYS* at a salt concentration of 150mM. Smaller contaminants accompany the *ZFYS* target bands in fractions T7-T9, highlighting shared surface charge properties resulting in co-elution. The absence of *ZFYL* versus the persistence of potential *ZFYS*

hints at better expression of the spliced form, though purity remains hampered by continued co-purification of bacterial proteins with similar charge affinities.

Despite the lack of visible *ZFYL* bands, fractions T7-T9 were then pooled across both *ZFYS* and *ZFYL* ion exchange fractions under the assumption that the previously low *ZFYL* expression levels may fall below staining detection limits though still recoverable via pooling. Moreover, assuming that *ZFYL* should elute at a salt concentration similar to *ZFYS*, the corresponding fractions were combined. The combined samples of *ZFYS* and *ZFYL* were dialysed to eliminate salt, and spin concentration was performed to assess its potential effectiveness in facilitating *ZFYL* identification despite its low yield.

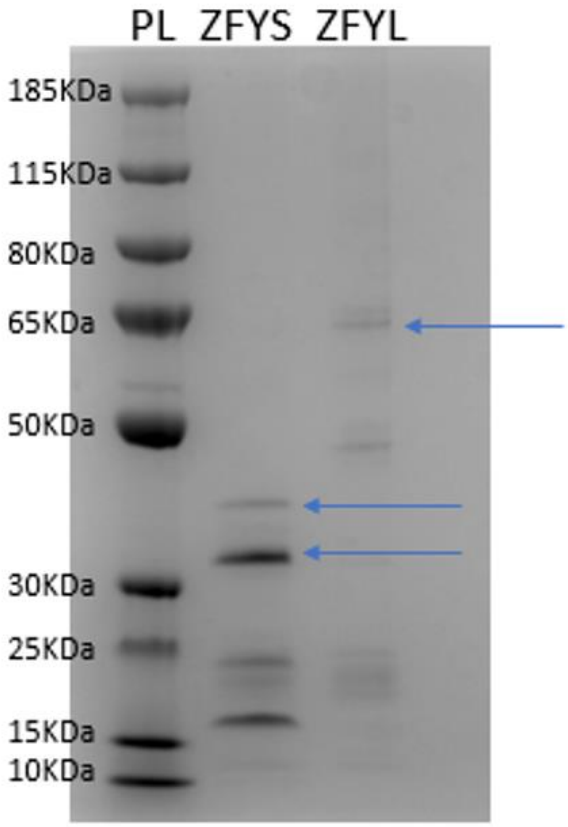


Figure 5.19: 4-12% Bis-Tris SDS-Page gel showing the *ZFYS* and *ZFYL* protein samples following dialyse and spin concentration. PL: protein ladder. ***ZFYS*:** Concentrated *ZFYS* sample following anion exchange. ***ZFYL*:** Concentrated *ZFYL* sample following anion exchange. Arrows indicating the *ZFYS* and *ZFYL* proteins.

Gratifyingly, concentrating the samples post-ion exchange now renders a *ZFYL* band visible by staining (**Figure 5.19)**, verifying low-level yields remaining below the detection limits of Coomassie staining. This supports that *ZFYL* rather than *ZFYS* could be toxic to the cells. Significant cumulative losses during purification likely contribute to negligible visualisation. In contrast, both *ZFYS* bands stain strongly but

the suspected truncate consistently dominates, implying instability of the full short isoform. Moreover, the bands highlighted by the arrows in **Figure 5.19** were confirmed to have a His-tag via western blotting providing further confirmation.

5.3.6 Further Buffer Optimisation

To further clean the purification of both *ZFYS* and *ZFYL*, further optimisation steps were performed. These included the addition of 50mM imidazole to the wash buffer to aid in the removal of loosely bound contaminant proteins which are subsequently eluting with the *ZFY* constructs. In addition to incremental losses during purification processes, visible precipitate formation during dialysis indicated additional product instability and aggregation likely exacerbates unsatisfactory yields, this was more evident for *ZFYS*. Alterations to the buffer compositions were therefore made to enhance protein stability. The 25mM piperazine (pH 5.4) buffer was changed to a 25mM Bis-Tris (pH 6.4) buffer. This pH shift by 1 will hopefully mean the *ZFY* will tolerate the salt better.

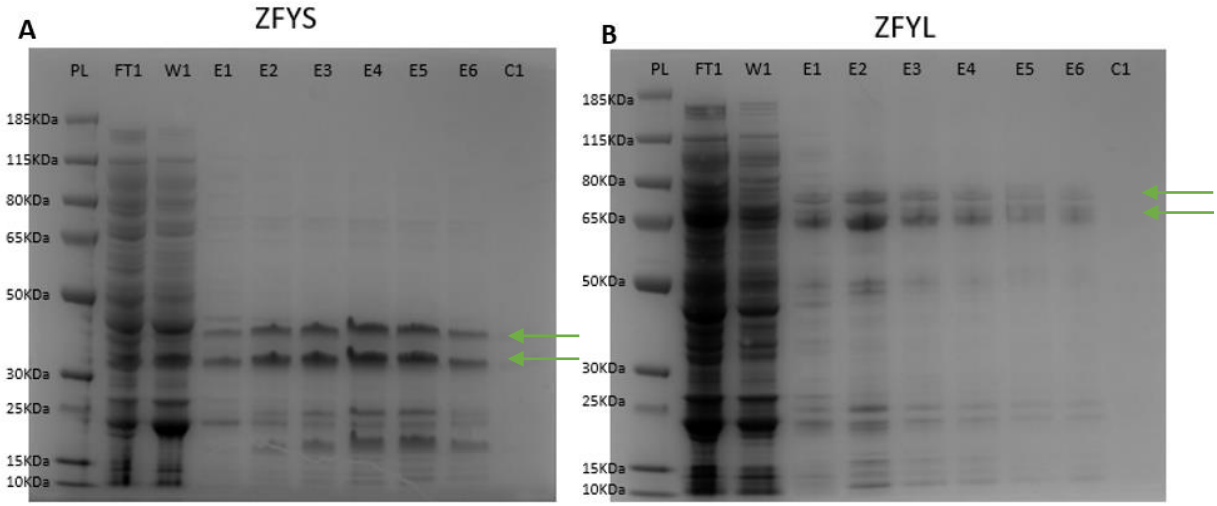


Figure 5.20: 4-12% Bis-Tris SDS-Page gel nickel column purification. A: *ZFYS* purification fractions & **B:** *ZFYL* purification fractions. This nickel column used the optimised wash buffer containing a low concentration of imidazole. **PL:** Protein Ladder, **FT1:** 1st flow through, **W1:** 1st wash through fraction, **E1-E6:** elution fractions, **C1:** 1st clean fraction.

Adding 50mM imidazole to the wash buffer seems to be successful, as the elution fractions are much cleaner, and the *ZFYS* and *ZFYL* bands are much more visible (green arrows) as seen in **Figure 5.20**. The imidazole in the wash buffer has majorly helped remove loosely bound contaminants. Moreover, the double banding pattern is very evident for both constructs. However, it should be noted that unfortunately, both *ZFY* variants do seem to be coming out in the flow through and wash, suggesting the protein might be binding weakly to the column.

As shown in **Figure 5.20**, adding imidazole to the wash buffer has dramatically improved elution purity by stripping background contaminants as intended. Target *ZFYS* and *ZFYL* bands exhibit much higher visibility against largely diminished co-purifying proteins. The double banding pattern is very evident for both constructs in **Figure 5.20**, with both lower bands dominating in abundance. However, the continued presence of some *ZFY* in the initial flow-through and wash fractions implies a portion still fails to adequately bind resin.

Following nickel column purification, buffers were reconfigured for subsequent *ZFY* dialysis and anion exchange chromatography. This was in an attempt to improve *ZFY*'s tolerance to salt changes as problems had previously been noted with precipitation. Both the buffer composition and pH were altered to address this issue.

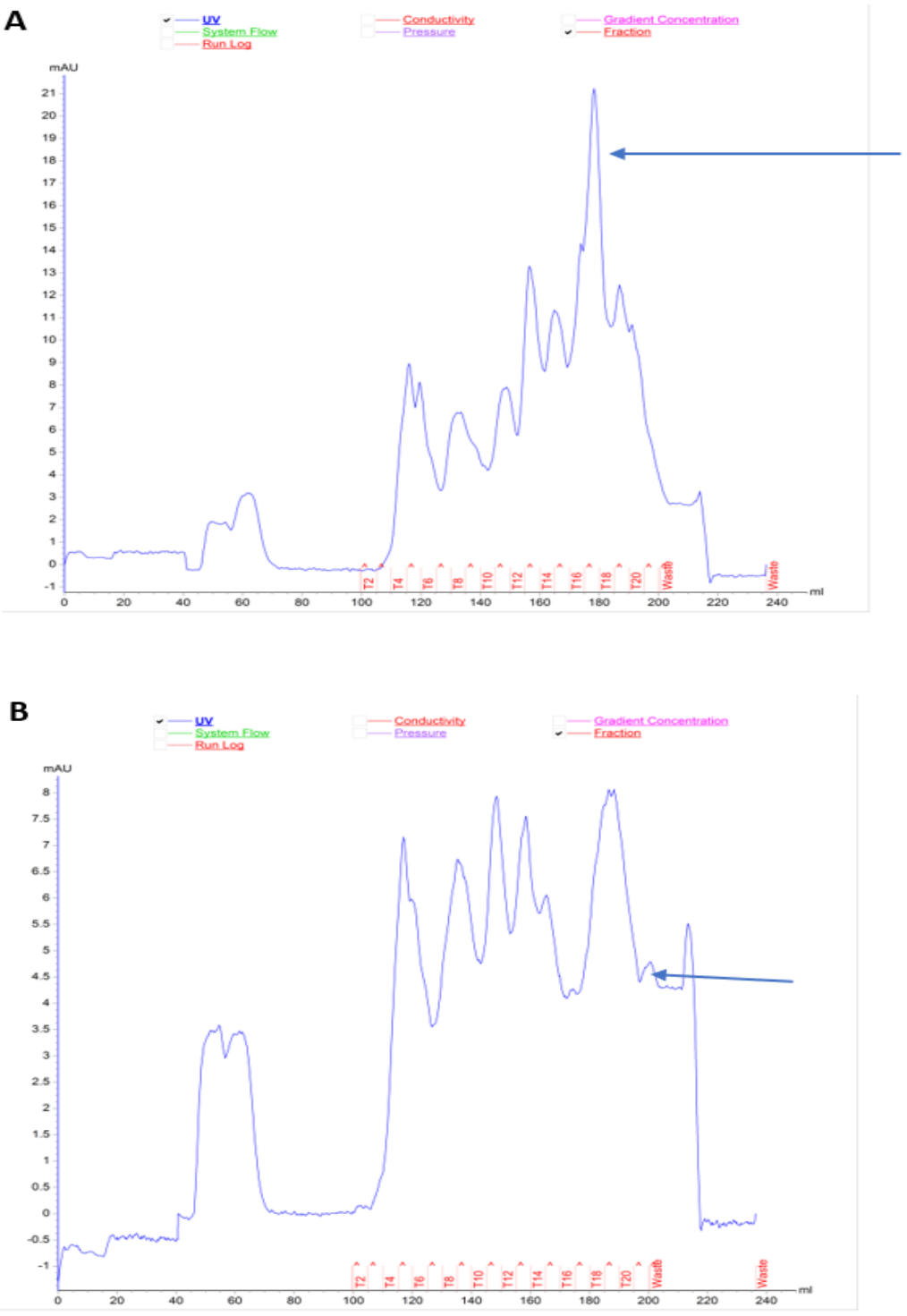


Figure 5.21: AKTA anion exchange graphs using the new Bis-Tris buffer. A: *ZFYS* construct & **B:** *ZFYL* construct. mAU is plotted on the Y-axis and measures the absorbance, corresponding to the amount of protein present. Arrows indicate the protein peak associated with *ZFYS* and *ZFYL*, respectively.

After modifying the buffers, a distinct peak pattern emerged, differing from the previous graphs in **Figure 5.17**. The increased number of peaks prompted a detailed

analysis of the fractions to determine which peaks corresponded to *ZFYS* and *ZFYL* expression. Subsequent examination on an SDS-page gel revealed that *ZFYS* and *ZFYL* were eluting in later fractions. *ZFYS* was detected in fractions T17 and T18 (~400mM NaCl), while *ZFYL* appeared in fractions T19 and T20 (~500mM NaCl). The mAU analysis indicated a substantial expression of *ZFYS* at approximately ~21mAU, in contrast to a lower expression of *ZFYL* at around ~8mAU. Consistently, *ZFYL* exhibited significantly lower expression levels compared to *ZFYS*. Despite this, the buffer alteration proved effective in reducing protein loss through precipitation, contributing to overall improved yields with the implemented changes.

These changes in protocol meant enough *ZFYS* had been purified to perform end-to-end sequencing of the protein to determine where the protein is being targeted for cleavage. This truncated version of *ZFYS* is expressed at a much higher level than the full-length and is therefore, the most abundant protein at the end of the protocol which is unwanted.

MGSSHHHHHH SSGLVPRGSH MLEMDEDEFE LQPQEPNSFF DGIVDDAGKI EHDGSTGVTI
DAESEMDPCK VDSTCPEVIK VYIFKADPGE DDLGGTVDIV ESEPENDHGV ELLDQNSSIR
VPREKMVYMT VNDSQQEDED LNVAEIADEV YMEVIVGEED AAVAAAAAAV HEQQIDEDEM
KTFVPIAWAA AYGNNSDGIE NRNGTASALL HIDESAGLGR LAKQKPKKKR RPDSRQYQT

Figure 5.22: End-to-end sequencing by top-down mass spectrometry of purified *ZFYS* protein. Highlighted red text are the amino acids missing in the purified protein submitted for sequencing.

The end-to-end sequencing results presented in **Figure 5.22,** reveal the presence of a 186 amino acid protein instead of the anticipated 239 amino acid protein. The final 53 amino acids are cleaved along with the starting amino acid methionine. Methionine is often cleaved in an *E. coli* system and this occurrence is expected. However, the cleavage of the last 53 amino acids at a Tryptophan codon, has resulted in the continuous production of a truncated protein. Tryptophan is always encoded by a UGG codon, which closely resembles two stop codons, UGA and UAG. While the construct was sequence verified by GenScript before shipping, it is possible that it subsequently acquired a premature stop mutation at some point during the experiments. However, the presence of doublet bands in the *ZFYL* experiments would imply that both constructs had independently acquired the same premature stop mutation. An alternative possibility is that some feature of *ZFY* sequence triggers misreading of this tryptophan codon as a stop codon in bacterial cells.

Due to the persistence in *ZFY* termination, the decision to order new synthetically engineered constructs which are codon optimised was made. *E. coli* codon optimisation would hopefully prevent protein termination and protein degradation.

5.3.7 Codon Optimisation

Following the arrival of the synthesised bacterial codon-optimised constructs, an expression test was performed to determine any major differences between the constructs.

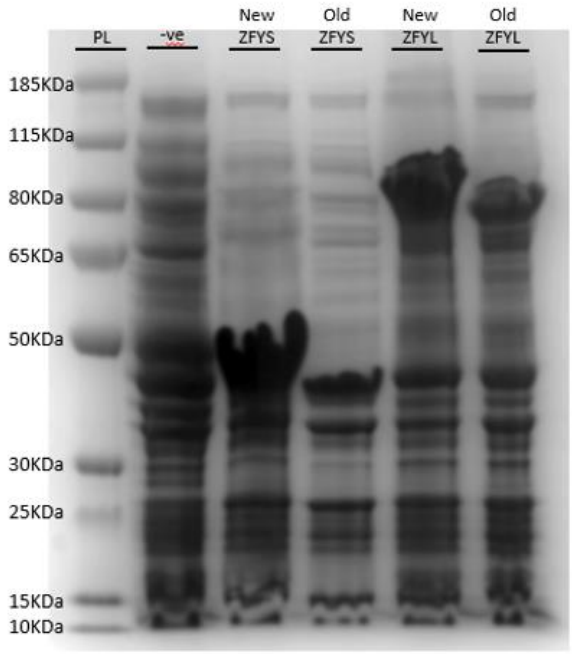


Figure 5.23: Expression test to compare the protein expression using the non-optimised and optimised constructs. PL: Protein Ladder, **-ve:** negative control (*E. coli)*, **New *ZFYS*:** Optimised *ZFYS* construct, **Old *ZFYS*:** Non-optimised *ZFYS* construct, **New *ZFYL*:** Optimised *ZFYL* construct, **Old *ZFYL*:** Non-optimised construct. 4-12% Bis-tris SDS-Page gel. 10mM IPTG induction.

In **Figure 5.23**, a noticeable contrast in expression levels is evident between the non-optimised and optimised constructs. Both *ZFYS* and *ZFYL* optimised constructs exhibit a higher quantity of protein at the anticipated size compared to the non-optimised counterparts. Regarding the production of full-length proteins versus shorter truncated ones, it is challenging to discern due to the substantial amount of protein present. Nevertheless, based on **Figure 5.23**, there is a possibility that the optimised constructs yield a protein of slightly larger size than the non-optimised ones, hinting at the potential expression of full-length proteins. These points therefore hint towards improved protein production in the Rosetta cell system when the constructs are codon optimised.

Following this, protein expression and purification were performed as previously done using the improved protocol.

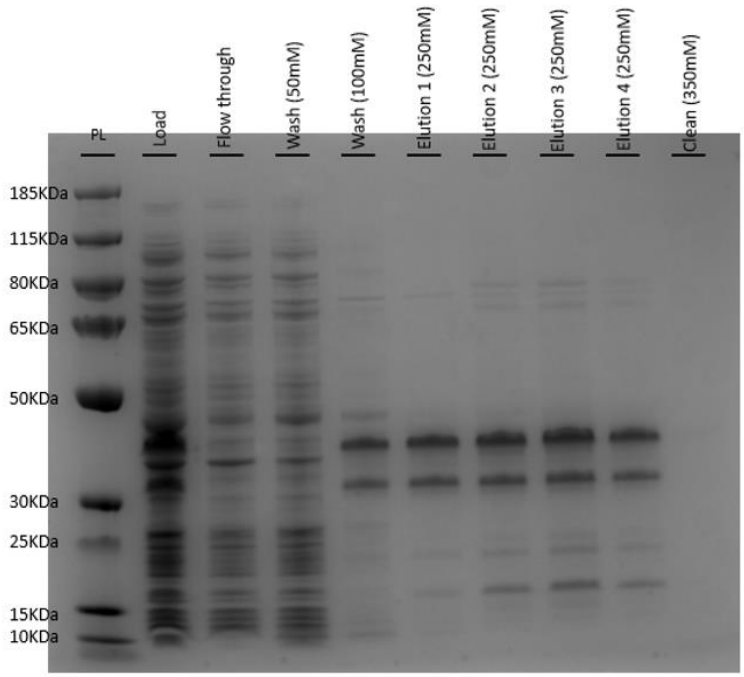


Figure 5.24: Codon optimised *ZFYS* nickel column purification using previously optimised conditions. PL: Protein Ladder, **Load:** Supernatant loaded onto the column, **Flow though:** supernatant flow through, **Wash (50mM):** wash buffer containing 50mM imidazole flow through, **Wash (100mM):** wash buffer containing 100mM imidazole flow through, **Elution 1-4:** Elution fractions, **Clean (350mM):** column stripped with 350mM imidazole. 4-12% Bis-Tris SDS-Page gel.

The new optimised *ZFYS* construct still results in the expression of the full and truncated protein as previously seen in the non-optimised constructs shown in **Figure 5.24**, however, the full-length protein now seems to be expressed slightly more than the truncated. This suggests that the non-optimised construct favoured the truncated protein form, whilst the optimised construct seems to favour the full-length protein slightly more. This observation was replicated across both *ZFYS* and *ZFYL* constructs, for both the native human sequence and the codon-optimised sequences. This indicates that for unknown reasons, this specific tryptophan codon is consistently misread as a stop codon in *E. coli* cells. In contrast, the western blotting from the mammalian cell work (**Figure 4.8** in Chapter 3) did not exhibit this issue. In this bacterial work, even though we are still seeing two bands, the protein yield seems much greater compared to the non-optimised constructs. With this nickel column purification, an additional wash step was performed with 100mM imidazole in the wash buffer to help with the purity of the *ZFYS*. Upon examining the elution fractions,

it is evident that there are fewer contaminating bands. However, the addition of 100mM imidazole leads to the elution of *ZFYS* from the column, resulting in a loss of *ZFYS*. Despite this, the elution fractions exhibit improved cleanliness. Although a portion of *ZFYS* appears in the flow-through and the first wash, the impact is not significant.

The upper *ZFYS* band in **Figure 5.24** underwent mass spectrometry, confirming it as the full-length protein without any cleavage events.

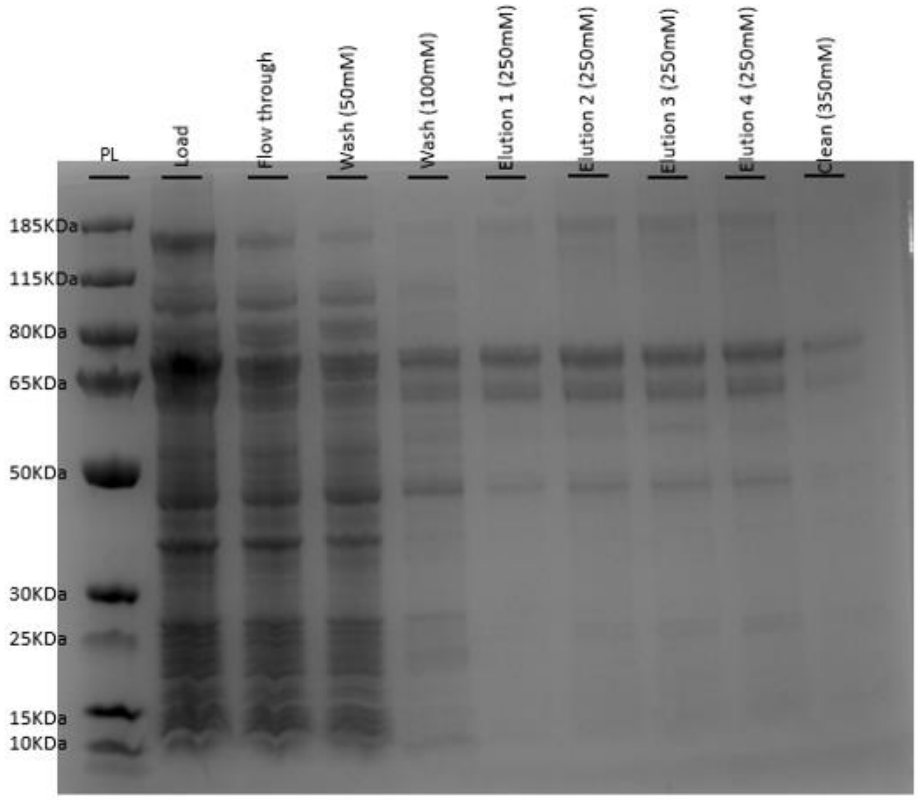


Figure 5.25: Codon optimised *ZFYL* nickel column purification using previously optimised conditions. PL: Protein Ladder, **Load:** Supernatant loaded onto the column, **Flow though:** supernatant flow through, **Wash (50mM):** wash buffer containing 50mM imidazole flow through, **Wash (100mM):** wash buffer containing 100mM imidazole flow through, **Elution 1-4:** Elution fractions, **Clean (350mM):** column stripped with 350mM imidazole. 4-12% Bis-Tris SDS-Page gel.

As seen in **Figure 5.25**, like with the new *ZFYS* optimised constructs, the *ZFYL* yield has significantly improved, yet the doubling banding pattern for *ZFYL* is very clear. A large proportion of *ZFYL* is present in the elution fractions, both the full-length and truncated versions, however, *ZFYL* is being lost at all stages on the nickel column. A large amount of *ZFYL* (both forms) is coming out in the washes as well as the flow through when imidazole is absent. This suggests that the His-tag is not very accessible, possibly due to the protein folding on itself due to the electrostatic interactions. *ZFYL* has a larger negative charge compared to *ZFYS* which might

explain why the amount of *ZFYL* loss in the flow-through is greater. This loss in the flow through was not noticed before due to the low yield of protein previously.

To attempt to address the issue of *ZFYL* protein loss in the flow-through, modifications were made to the salt concentration and pH of the lysis buffer. The salt concentration of the lysis buffer was changed from 100mM NaCl to 300mM NaCl to see if this alteration helped with the binding of *ZFYL* to the nickel column.

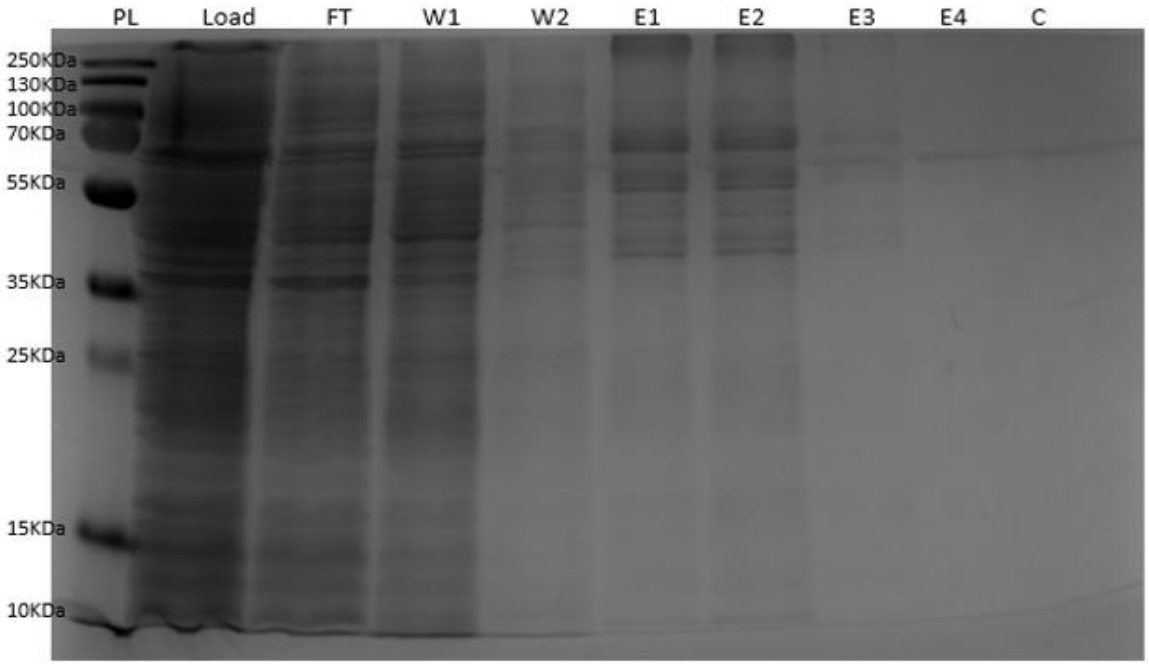


Figure 5.26: 12% Tris-glycine gel showing *ZFYL* nickel purification fractions after the alterations to the buffers had been made. PL: Protein Ladder, **Load:** Supernatant loaded onto the column, **W1-W2:** wash fractions (50mM imidazole), **E1-E4:** Elution fractions, **C:** Clean fraction.

With the salt alterations made, **Figure 5.26** shows a potential reduction in the amount of protein lost in the flow-through and wash steps, but this is hard to confirm due to there being potentially a lower protein yield using these new buffers. This unsuccessful change continues to suggest a possible His-tag accessibility problem with the *ZFYL* construct.

Following this trial, the lower salt concentration buffers were used for *ZFYL* rather than the high salt buffer in an effort to improve protein binding. Protein purification and expression were continued as previously done.

5.3.8 Anti-*ZFY* Antibody Testing

The western blots shown in Figures 5.6 to 5.11 above utilised an anti-His antibody for verification. To further confirm the identity of the targeted *ZFY* proteins, an anti-*ZFY* antibody from Abcam was tested. The precise epitope for this antibody is commercially sensitive and not released by the company, who state only that it binds within the initial 1-60 amino acids of *ZFY*. Notably, most of this stated range falls within the second coding exon that is not part of the *ZFYS* acidic domain. Consequently, western blotting was conducted to determine whether the antibody recognises both variants or exclusively identifies *ZFYL*.

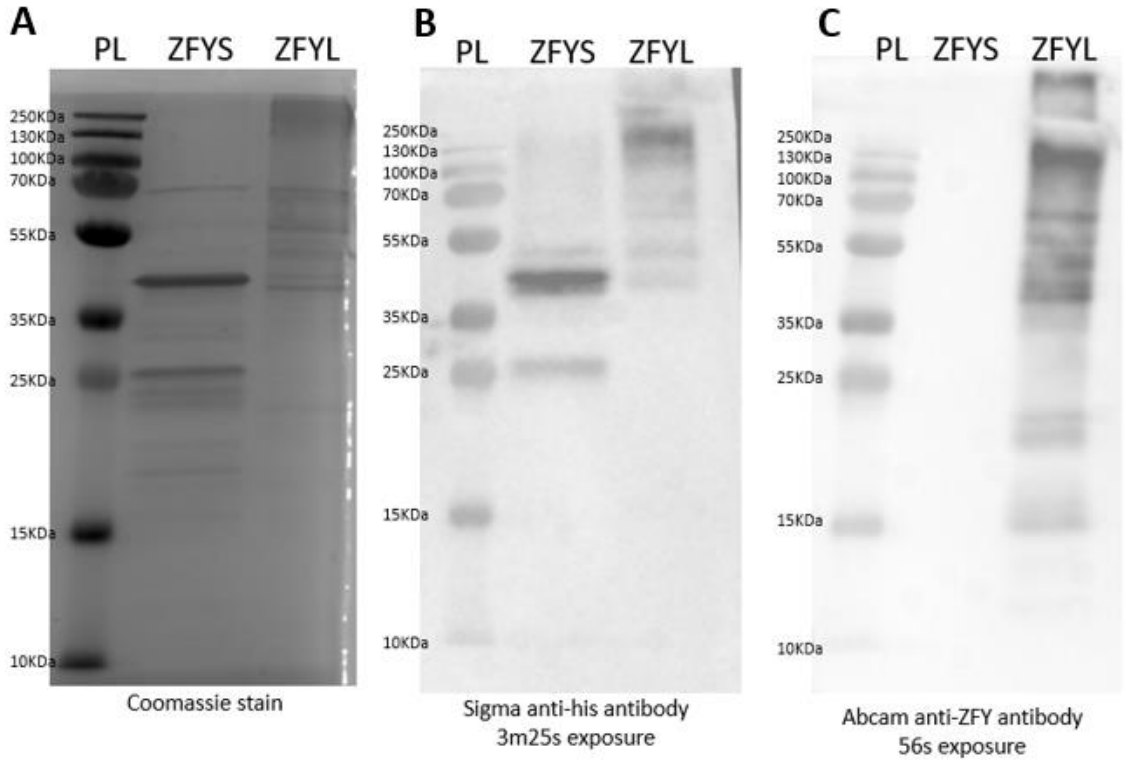


Figure 5.27: 12% Tris-Glycine gels of *ZFYS* and *ZFYL* proteins to test for antibody specificity. A: Coomassie-stained *ZFYS* and *ZFYL* samples following full purification including nickel column purification and ion exchange. **B:** Western blot of the *ZFYS* and *ZFYL* samples using an anti-his antibody. **C:** Western blot of the *ZFYS* and *ZFYL* samples using the new anti-*ZFY* antibody.

Figure 5.27A shows the high expression of the full-length *ZFYS* protein, with a lower abundance of the truncated *ZFYS* protein. **Figure 5.27B** shows a strong western signal using the His-antibody. Even following codon optimisation, the purification process still leads to low *ZFYL* expression, making Coomassie staining results unclear, however, the His-antibody signal in **Figure 5.27B** confirms the presence of *ZFYL*. The anti-*ZFY* antibody western shown in **Figure 5.27C** indicates no signal for *ZFYS* but reveals a strong signal for *ZFYL*, greater than the signal seen with the His-

antibody. **Figure 5.27C** thus demonstrates that the commercial anti-*ZFY* antibody binds to a region within the second coding exon and is specific for *ZFYL*.

5.3.9 BL21 PlysS Freeze-Thaw Lysis Method

Following continuous alterations to the lysis method including switching the order of purification by first completing ion exchange and then following this with nickel column purification, yields of *ZFYL* remained low and degradation products were still very abundant. These problems were identified to occur during the lysis process and further after.

It was then thought that the lysis method needed to be altered. This led to the expression system changing again to BL21 PlysS *E. coli* cells. BL21 PlysS contains a PlysS plasmid encoding T7 lysozyme making lysis easier. This means that a freeze-thaw method should be sufficient to break open the cells and allow for the lysis process to be carried out at cold conditions reducing degradation as a result of higher temperatures.

Two lysis buffers were trialled; (1) the original sodium phosphate buffer and (2) a new 10mM Tris-HCl lysis buffer identified in papers for use in this kind of lysis protocol.

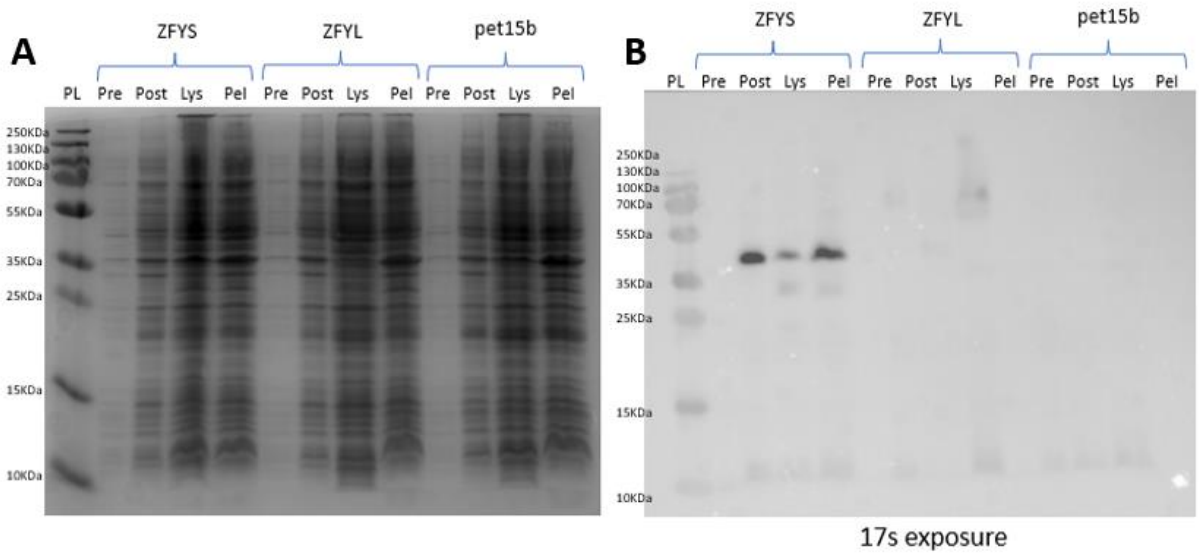


Figure 5.28: 12% Tris-Glycine SDS-Page gel of freeze-thaw lysis protocol using the sodium phosphate lysis buffer. A: Coomassie-stained gel showing the freeze-thaw lysis method for both the *ZFYS*, *ZFYL* and pet15b constructs. **B:** Western blot showing the freeze-thaw lysis method for both the *ZFYS*, *ZFYL* and pet15b constructs. **PL:** Protein Ladder, **Pre:** Pre-induction sample, **Post:** Post-induction sample, **Lys:** protein Lysate, **Pel:** Pellet.

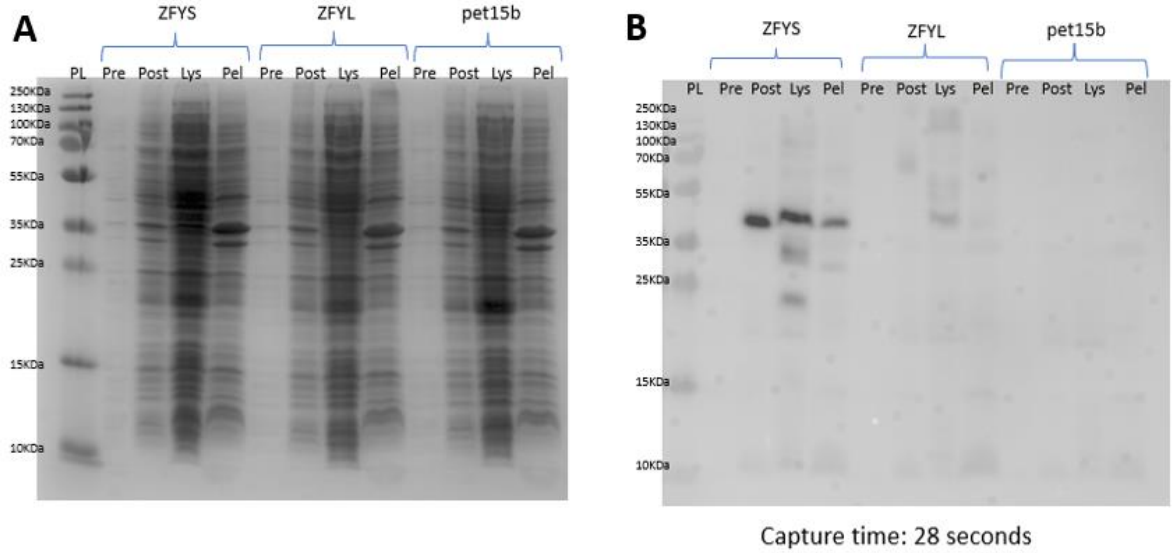


Figure 5.29: 12% Tris-Glycine SDS-Page gel of freeze-thaw lysis protocol using the 10mM Tris-HCl lysis buffer. A: Coomassie stained gel showing the freeze-thaw lysis method for both the *ZFYS*, *ZFYL* and pet15b constructs. **B:** Western blot showing the freeze-thaw lysis method for both the *ZFYS*, *ZFYL* and pet15b constructs. **PL:** Protein Ladder, **Pre:** Pre-induction sample, **Post:** Post-induction sample, **Lys:** protein Lysate, **Pel:** Pellet.

The PlysS results shown in **Figures 5.28** and **5.29** show that both versions of the proteins are expressed (short more than long) but undergo rapid degradation during lysis, resulting in the spectrum of smaller bands consistently seen throughout this series of experiments. However, the PlysS cells seem to hardly express *ZFYL* at all

as very little signal is seen in the post-induction sample by Coomassie staining or western blotting. Whilst *ZFYS* is expressed well, it is clear to see that degradation is occurring during lysis. This is made evident by the lower his-signal bands even though the sample was kept at low temperatures. Furthermore, the lysis doesn't seem to be very efficient since there is a strong signal in the pellet lane, suggesting that the lysis is not complete.

When comparing the lysis buffers in **Figure 5.28** and **5.29,** it seems that the Tris-HCl buffer is better at breaking the cells apart, as for *ZFYS* there is more protein present in the lysate compared to the pellet. However, the Tris-HCl buffer also seems to possibly amplify the amount of degraded *ZFYS* protein compared to the sodium phosphate buffer. Taking all the above experiments to date, we reach the following conclusions:

1) Both *ZFYS* and *ZFYL* can be expressed in *E coli*, with expression confirmed by Coomassie staining, His-tag Western blot, anti-*ZFY* Western blot (for *ZFYL* only), mass spectrometric detection of *ZFY* peptides and full-length sequencing (for *ZFYS* only)
2) Both isoforms appear susceptible not only to rapid degradation during purification but also to premature translation termination, occurring at a specific tryptophan codon
3) *ZFYL* is consistently expressed at lower levels than *ZFYS* in every *E. coli* system tried, and as the purification protocol progresses the yield continues to fall to undetectable levels
4) *ZFYS* yields are better, with *ZFYS* protein levels remaining detectable throughout various stages of purification, but with major problems with protein degradation as mentioned above

Overall, degradation and yield problems were consistent throughout, with the *ZFYL* yield being too low for most further experimentation, and thus this line of work was abandoned. We note *ZFYL* was also expressed at lower levels in the mammalian cells, suggesting a possible general problem with *ZFYL* expression may be due to the size of the construct. Following the programme of work described above, we returned to the (more successful) mammalian cell culture system in order to investigate the binding partners of *ZFY*.

5.3.10 GFP-Pull Down

The *ZFY*-GFP constructs used for the successful mammalian transfections in chapter 4 were used to perform a GFP-pull down. Transfections were done as previously described and lysates were prepared. This change in method also means that the entire *ZFYS* and *ZFYL* proteins are being studied, not just the acidic domain.

Initial pull-down trials showed by Coomassie staining a low yield of both *ZFY* construct in the elution fractions, consequently, wells of cells were pooled to increase the protein yield. This was probably a result of protein being lost during washing steps and a generally lower protein yield compared to the control. Additional care was taken during the washing steps to ensure the cell pellet was undisturbed and remained intact.

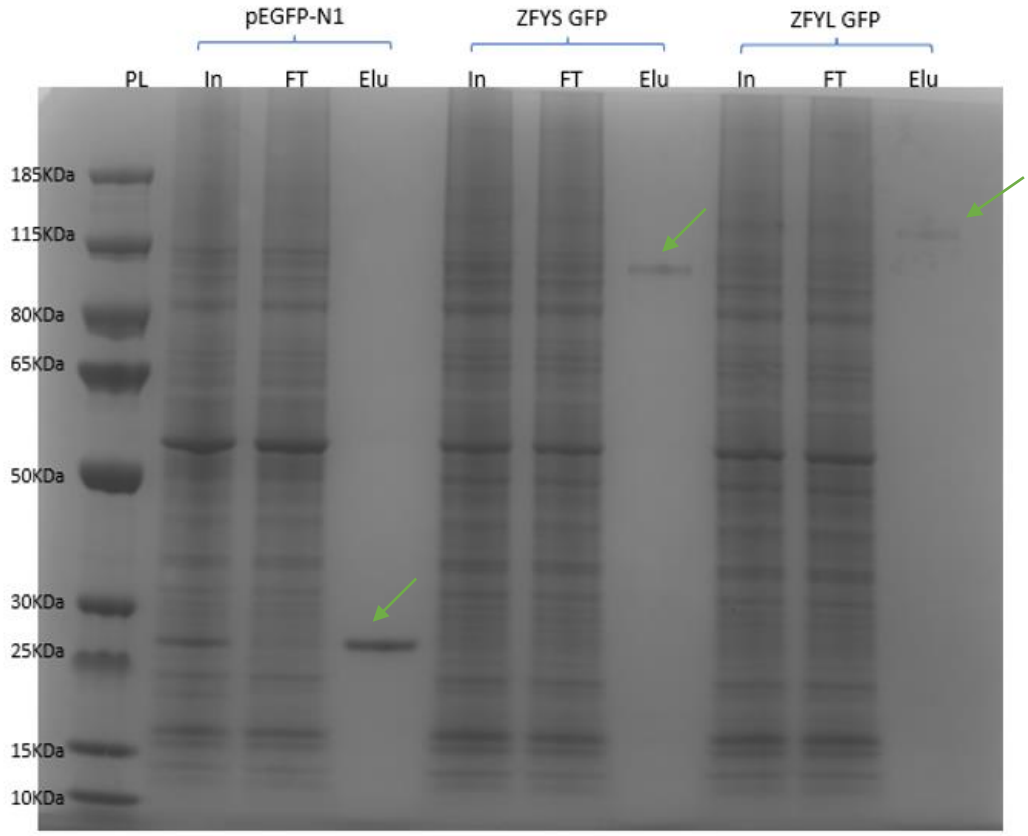


Figure 5.30: Coomassie stained 4-12% Bis-Tris SDS-Page gel showing the GFP-Pull down after combining multiple lysates for each construct transfected into HEK293 cells. PL: Protein Ladder, **In:** Input Sample, **FT:** flow through, **Elu:** Elution.

In **Figure 5.30**, bands are identified in the elution fractions for all the constructs at the expected sizes for each construct (green arrows), indicating that the GFP-Trap beads have successfully pulled down both the *ZFYS*-GFP and *ZFYL*-GFP constructs, as well as the free GFP from the control pEGFP-N1 transfections. These bands are present in the input, indicating a highly efficient pulldown and complete capture of GFP-tagged protein from the lysates.

No other bands were detectable by Coomassie staining of the pulldowns, indicating that any potential binding proteins are present at low stoichiometry. This however does not preclude the possibility of identifying low abundance / transient interacting partners. Therefore, two further replicate transfections/pulldowns were performed for each of *ZFYS* and *ZFYL*. The pulldown fractions were concentrated by gel electrophoresis and submitted for mass spectrometry-based proteomics, with the pEGFP-N1 construct used as a negative control.

5.3.11 Proteomics

Proteomics is the investigation of a protein's interactions, functions, composition and structure (Al-Amrani *et al*., 2021). Mass spectrometry is widely used to characterise proteins and has evolved into a global tool for proteomics research. Using mass spectrometry unique peptide sequences are identified and matched to proteins. Applying this approach to *ZFYS* and *ZFYL* samples generated lists of potential interacting proteins.

Initial data filtering removed the most common contaminants such as Keratin and albumin (Table 1:(Hodge *et al*., 2013)). Further filtering removed the hits of *ZFY* and GFP themselves, and any proteins identified to have a higher abundance in the pEGFP-N1 negative control. This, therefore, removed any entities adhering non-specifically and would refine the data to look at the functional relationship between *ZFYL* and *ZFYS*.

Our initial analysis was based on the first of two replicate pulldowns, following exclusion of common contaminants as described above. This preliminary analysis revealed 55 potential protein interactions (*Supplementary* **Table 8**). However, proteins identified purely by one sole unique peptide exhibit high false discovery rates. Therefore, a further stringency threshold was set to only include protein hits with at least 2 unique peptide hits. This threshold limit reduced the number of potential interacting proteins to 39. Of these 39 proteins, all candidate interacting proteins were pulled down to some extent by both *ZFYS* and *ZFYL* **(Table 5.8)**. This both increases our confidence that these are genuine interacting partners and backs up the RNA-Seq data suggesting that *ZFYS* and *ZFYL* have qualitatively similar but quantitatively different functions.

Table 5.8: List of the 39 potential interactors collected from n=1. These genes meet the criterial of having at least 2 unique peptide hits.

Description	Gene Name
Poly [ADP-ribose] polymerase 1	PARP1
Nucleolin	NCL
Heat shock cognate 71 kDa protein	HSPA8
Heat shock 70 kDa protein 1A	HSPA1A
60S acidic ribosomal protein P0	RPLP0
40S ribosomal protein S3	RPS3
Heterogeneous nuclear ribonucleoproteins C1/C2	HNRNPC
60S ribosomal protein L18	RPL18
ATP synthase subunit alpha_ mitochondrial	ATP5F1A
Actin_ cytoplasmic 1	ACTB
L-lactate dehydrogenase B chain	LDHB
Probable 28S rRNA (cytosine(4447)-C(5))-methyltransferase	NOP2
Elongation factor 1-alpha 1	EEF1A1
Tubulin beta chain	TUBB
Polyubiquitin-B	UBB
Heterogeneous nuclear ribonucleoprotein U	HNRNPU
Heterogeneous nuclear ribonucleoprotein R	HNRNPR
Heterogeneous nuclear ribonucleoproteins A2/B1	HNRNPA2B1
60S ribosomal protein L7	RPL7
60S ribosomal protein L4	RPL4
60S ribosomal protein L12	RPL12
Clathrin heavy chain	CLTC
40S ribosomal protein S8	RPS8
Heterogeneous nuclear ribonucleoprotein H	HNRNPH1
Histone H4	H4C1
60S ribosomal protein L14	RPL14
60S ribosomal protein L23a	RPL23A
ATP-dependent RNA helicase DDX3X	DDX3X
ATP-dependent RNA helicase DDX50	DDX50
Tubulin alpha-3C chain	TUBA3C
Histone H2A type 1-B/E	H2AC4
Nucleophosmin	NPM1
Heterogeneous nuclear ribonucleoprotein K	HNRNPK
60S ribosomal protein L29	RPL29
116 kDa U5 small nuclear ribonucleoprotein component	EFTUD2
60S ribosomal protein L11	RPL11
Interleukin enhancer-binding factor 3	ILF3
28S ribosomal protein S29_ mitochondrial	DAP3
Insulin-like growth factor 2 mRNA-binding protein 1	IGF2BP1

Notably, while relative abundance comparisons are only semi-quantitative given the potential variation in bait protein amounts introduced to mass spectrometry, 24 of the putative interacting proteins displayed higher abundance in the *ZFYL* pulldown compared to 15 that were more abundant in the *ZFYS* pulldown. However, within this list of interacting partners, we were particularly interested in any interactions with partners involved in RNA metabolism or RNA transcription. Interestingly of the 39 protein hits, 12 form part of the 60S and 40S ribosomal subunits and 6 are heterogeneous nuclear ribonucleoproteins (hnRNPs). The ribosome is the site of translation and hnRNPs represent a large RNA-binding protein family with roles in splicing, mRNA stabilisation and transcriptional and translational regulation suggesting roles pointing towards mRNA and protein synthesis.

Additional proteins identified in this first replicate represented categories fitting closely with nuclear activities include histones and heat shock factors influencing chromatin architecture and stress responses. A cluster of proteins with potential interest was observed, namely Y-box binding proteins. Y-box binding proteins are a family of DNA- and RNA-binding proteins with roles in DNA transcription, mRNA splicing and translation. They may thus represent a family of transcriptional coactivators working in tandem with *ZFY* for transcriptional activation. This protein family has also been heavily investigated concerning tumorigenesis and their use as potential biomarkers. Y-box binding protein (YB-1) and Y-box binding protein (YB-3) were both identified as potential interactions with both *ZFYS* and *ZFYL*. Although their unique peptide hits fall below the threshold, a reasonable number of peptide hits was noted allowing for further confirmatory analysis by western blotting.

To confirm YB-1 as a potential interactor of *ZFY*, follow-up western blots on independent lysates were carried out using a YB-1 antibody. Pre- and post-pull-down western verification was conducted on both *ZFY* and pEGFP-N1 control samples to gauge YB-1 signals across the samples.

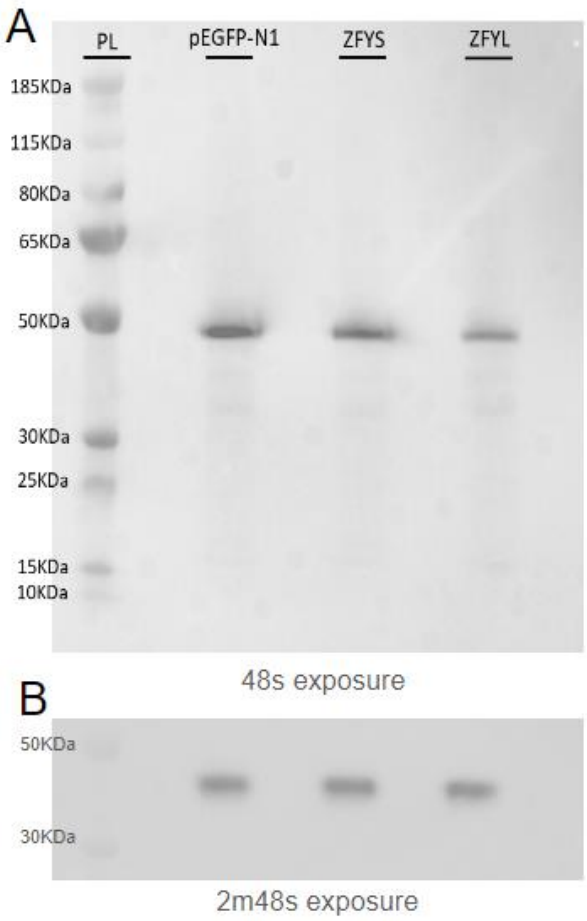


Figure 5.31: YB-1 western blot of the input HEK293 cell lysates following overexpression of pEGFP-N1, *ZFYS* and *ZFYL* constructs, before GFP-pull down. A: Membrane probed with YB-1 antibody, expected molecular weight = 50KDa. **B:** Membrane probed with Beta-actin for loading control, expected molecular weight = 43KDa. Pl = Protein ladder.

From **Figure 5.31**, YB-1 expression is evident across all lysates before performing the GFP-pull-down protocol, confirming its expression in HEK293 cells and that this is consistent across all samples. To then check the mass spectrometry-based proteomics, a western blot was performed on the independent set of mammalian transfected cells post GFP pull-down to determine if it was co-purified with the *ZFY* constructs.

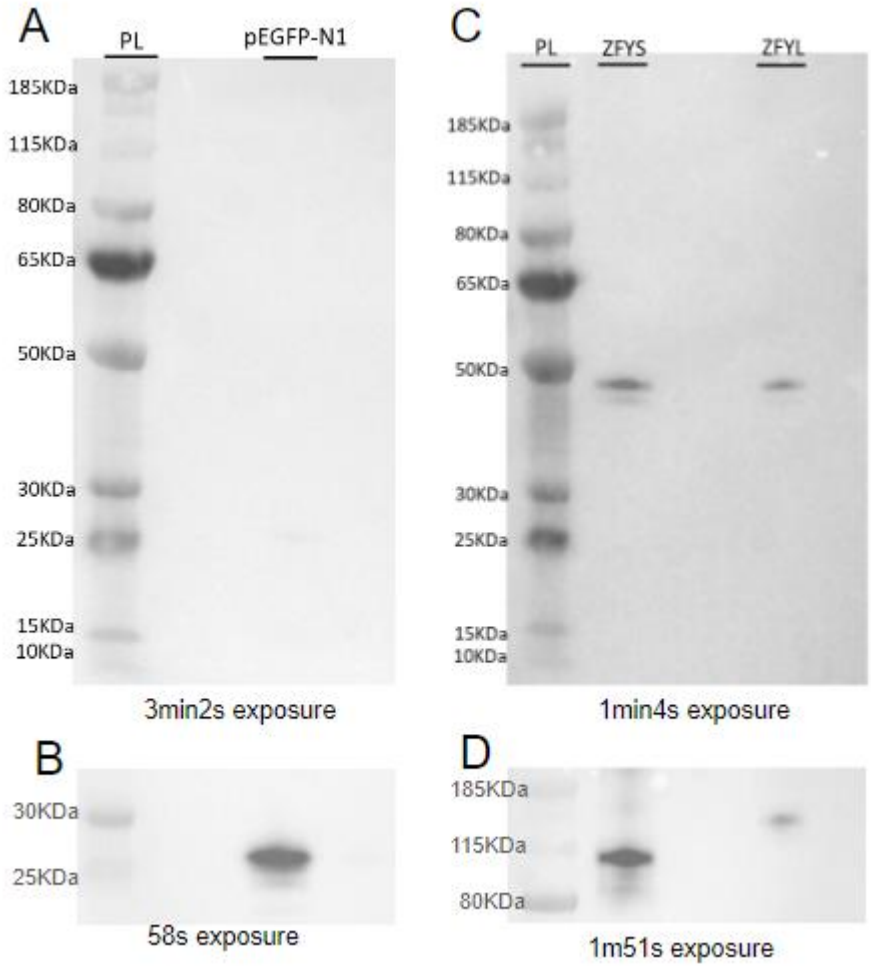


Figure 5.32: YB-1 western blot of HEK293 lysates post GFP-pulldown. A: pEGFP-N1 probed with YB-1 antibody, **B:** pEGFP-N1 probed with GFP-antibody, **C:** *ZFY* constructs probed with YB-1 antibody, **D:** *ZFY* constructs probed with GFP-antibody. Molecular weights; YB-1 = 50KDa, pEGFP-N1 = 27KDa, *ZFYS* = 96KDa, *ZFYL* = 118KDa. PL: Protein ladder.

Following the GFP-pulldown of the samples, **Figure 5.32C** validates the co-precipitation of YB-1 with *ZFYS* and *ZFYL*. Notably, YB-1 is not pulled down in the pEGFP-N1 control, indicating that there is specific binding of YB-1 to *ZFYS* and *ZFYL*. Beta-actin was used as a loading control pre-pulldown as seen in **Figure 5.31B**, however, beta-actin is not pulled down, so instead GFP was used as a control post-

pulldown seen in **Figure 5.32B** and **Figure D**. Whilst the GFP signal is not even across the three samples, it is clear that YB-1 does not bind pEGFP-N1 due to the large abundance of GFP signal, but no YB-1 signal.

Following this preliminary analysis of the first replicate pulldown, a second biological replicate pulldown was performed, and mass spectrometry proteomics performed as above. This second set of mass spectrometry results returned 61 potential hits following the initial filtering and returned 42 proteins once the two-peptide threshold was set (see *supplementary* **Table 9** *and* **Table 10***)*. Although the YB1 pulldown was not replicated in the second spectrometry replicate due to the low sensitivity of this technique, we note that this had previously been confirmed by Western blotting in an independent set of lysates.

Finally, we combined the two proteomics pulldown datasets, to identify duplicate results that were consistently observed. 19 proteins were identified to meet the filtering requirements across both datasets (**Table 5.9**, see *supplementary* ***Table 11*** for peptide and confidence scores**)**. Overall, similar protein families seen in the first set were replicated in the second including histone proteins, heat shock proteins, 40S/60S ribosomal subunit proteins and heterogeneous nuclear ribonucleoproteins. While the interacting genes listed in **Table 5.9** do not show differential expression in the RNA-Seq data, other family members, including HSPA12B, TUBB2A, TUBB4A, and TUBB4B, exhibit differential expression in the RNA-Seq results. Despite the absence of crossover in the list, the molecular functions indicated suggest that ZFY plays a significant role, potentially accounting for all the downstream pathways identified in the RNA-Seq data.

Table 5.9: Duplicated protein hits identified by mass spectrometry proteomics. This table shows the proteins that meet filtering requirements (≥2 unique peptide hits) and were found in both pulldown replicates. Molecular function and tissue specificity were found using The Human Protein Atlas (Uniprot). Interactions include interacting proteins found in either one of the datasets.

Gene Name	Molecular Function	Reactome Pathway	Tissue Specificity	Interactions
Heat shock cognate 71 kDa protein (HSPA8)	Autophagy, Host-virus interaction, mRNA processing, mRNA splicing, Stress response, Transcription, Transcription regulation	Metabolism of Proteins	Low tissue specificity	HSP90AA1
Nucleolin (NCL)	DNA-binding, RNA-	Metabolism of	Low tissue	HNRNPU,

	binding	RNA, rRNA processing	specificity	HNRNPK, HNRNPD
Poly[ADP-ribose] polymerase 1 (PARP1)	Allosteric enzyme, DNA-binding, Glycosyltransferase, Nucleotidyltransferase, Transferase	Metabolism of Proteins	Low tissue specificity	
Tubulin beta chain (TUBB)	GTP-binding, Magnesium, Metal-binding, Nucleotide-binding		Low tissue specificity	TUBA1B
Tubulin alpha-3C chain (TUBA3C)	GTP-binding, Magnesium, Metal-binding, Nucleotide-binding	Metabolism of Proteins	Testis enhanced (spermatid development)	
Heterogeneous nuclear ribonucleoproteins C1/C2 (HNRNPC)	RNA-binding, mRNA processing, mRNA splicing	Metabolism of RNA, Metabolism of Proteins	Low tissue specificity	HNRNPA2B1
ATP synthase subunit alpha_mitochondrial (ATP5F1A)	ATP-binding, Nucleotide-binding, ATP synthesis, Hydrogen ion transport, Ion transport, Transport	Metabolism of Proteins	Tongue enhanced	
Polyubiquitin B (UBB)	Protein degradation, chromatin maintenance, gene expression regulation, stress response	Metabolism of RNA, rRNA processing, Eukaryotic Translation Elongation, Metabolism of Proteins	Low tissue specificity	
L-lactate dehydrogenase B chain (LDHB)	Oxidoreductase		Heart muscle enhanced	
Probable 28S rRNA (NOP2)	Methyltransferase, RNA-binding, Transferase, Ribosome biogenesis, rRNA processing	Metabolism of RNA, rRNA processing	Low tissue specificity	
60S ribosomal protein L7 (RPL7)	Ribonucleoprotein, Ribosomal protein, RNA-binding	Metabolism of RNA, rRNA processing, Eukaryotic Translation Elongation,	Low tissue specificity	RPL29, RPS3, RPL19, RPL35, RPL8

		Eukaryotic Translation Termination, Metabolism of Proteins		
Histone H4 (H4C1)	DNA-binding	Metabolism of Proteins	Bone marrow, lymphoid tissue enhanced	H2AC4
ATP-dependent RNA helicase (DDX3X)	Apoptosis, Chromosome partition, Host-virus interaction, Immunity, Innate immunity, Ribosome biogenesis, Transcription, Transcription regulation, Translation regulation, DNA-binding, RNA-binding		Low tissue specificity (testis-late spermatids enriched)	
60S ribosomal protein L11 (RPL11)	Ribonucleoprotein, Ribosomal protein, RNA-binding, rRNA-binding	Metabolism of RNA, rRNA processing, Eukaryotic Translation Elongation, Eukaryotic Translation Termination, Metabolism of Proteins	Low tissue specificity	RPL19, RPL35, RPL8
28S ribosomal protein S29_mitochondrial (DAP3)	GTP-binding, Nucleotide-binding, Apoptosis	Metabolism of Proteins	Low tissue specificity	
Heterogeneous nuclear ribonucleoproteins A2/B1 (HNRNPA2B1)	RNA-binding, mRNA processing, mRNA splicing, mRNA transport, Transport	Metabolism of RNA	Low tissue specificity	
Heterogeneous nuclear ribonucleoprotein U (HNRNPU)	Chromatin regulator, DNA-binding, RNA-binding, Repressor, Cell cycle, Cell division, Differentiation, Mitosis, mRNA processing, mRNA splicing, Transcription, Transcription regulation	Metabolism of RNA	Low tissue specificity	

Ribosomal protein L18 (RPL18)	Ribonucleoprotein, Ribosomal protein	Metabolism of RNA, rRNA processing, Eukaryotic Translation Elongation, Eukaryotic Translation Termination, Metabolism of Proteins	Low tissue specificity	
Ribosomal protein L14 (RPL14)	Ribonucleoprotein, Ribosomal protein	Metabolism of RNA, rRNA processing, Eukaryotic Translation Elongation, Eukaryotic Translation Termination, Metabolism of Proteins	Low tissue specificity	RPL4

Among these 19 duplicated findings, Reactome pathway analysis revealed that several were associated with RNA metabolism, encompassing RNA processing, transcription, translation, and protein quality control. Further molecular functions include splicing, cell cycle regulation, differentiation and immunity. This indicates that *ZFY* encompasses a range of molecular functions potentially linked to male development. Multiple hits formed part of the ribosomal protein family and hnRNP family, with other proteins including histone and heat shock proteins. Although some of the protein interactions seem unrelated to others, there were a few instances where proteins were discovered to interact with others identified within one of the datasets creating a potential network of interactions. Comprehending the linkage within the *ZFY* network and its consequential functions is intricate and requires additional interpretation and confirmation for thorough understanding.

5.4 Discussion

Transcription factors serve as the principal controllers of gene expression by selectively binding to particular DNA sequences and enlisting transcription regulatory proteins (Mitsis *et al*., 2020);(G. Wang *et al*., 2015). Their activity is governed by diverse factors such as epigenetic mechanisms, gene regulatory elements, and molecular cofactors (Mitsis *et al*., 2020);(G. Wang *et al*., 2015). Most eukaryotic transcription factors are thought to recruit cofactors such as “coactivators” and “corepressors” (Lambert *et al*., 2018). Such cofactors are frequently large multi-subunit protein complexes or multi-domain proteins commonly containing domains involved in chromatin binding, nucleosome remodelling and histone modifications such as other transcription factors and RNA polymerases (Lambert *et al*., 2018). It has been proven that transcription factors rarely operate independently (Spitz & Furlong, 2012). Consequently, their roles should be viewed within a more integrated, combinatorial framework (Spitz & Furlong, 2012). This chapter aimed to identify the potential regulatory partners of both *ZFY* variants binding to the AAD.

Following many trials to express both variants of *ZFY*, continuous yield, purity and degradation issues persisted. Alterations to the expression system, buffers, lysis, column purification and codon optimisation led to improvements but eventually, it was determined that the expression and purification of *ZFY* were beyond our capacity.

Originally, the human *ZFY* acidic domain sequence was inserted into a pET-15b vector backbone and transformed into BL21 *E. coli* cells. The initial induction of both *ZFY* variants showed low effectiveness. However, noticeable enhancements were observed upon increasing the IPTG concentration to 0.8mM and extending the induction time overnight. This pattern resembled the cell lysis process, where initial attempts were not entirely successful. Nonetheless, adjustments in incubation and sonication times led to improvements. Nickel column purification was used to take advantage of the His-tagged construct. At first, both protein variants' presence was detectable solely via western blotting, owing to its higher sensitivity compared to Coomassie staining, which highlighted the challenges of low yield. Nevertheless, this approach revealed that both *ZFY* variants exhibited a higher molecular weight than anticipated, likely due to the proteins' high negative charge. In the initial protein expression and purification attempts, both variants exhibited a double banding pattern, with *ZFYS* showing a more pronounced effect, due to its consistently higher protein yield. It was presumed that this could be the result of a cleavage event resulting in shorter byproducts. This was confirmed by mass spectrometry for *ZFYS*, but not *ZFYL* due to the continuous yield problem. Mass spectrometry of *ZFYS* revealed the presence of two higher abundant weights; 20414.5Da and 26684.5Da.

A minor size difference exists between the observed higher weight and the anticipated size of *ZFYS*, and the exact cause remains unclear, a potential suggestion is protein modification by *E. coli*. Additionally, the reason for the presence of a band at 20414.5 Da is puzzling, as sequence analysis does not identify any potential codon that *E. coli* might have difficulty expressing. Due to this, it was decided to move into a different *E. coli* expression system. The Rosetta *E. coli* strain is a BL21 derivative designed to enhance the expression of eukaryotic proteins that contain proteins that *E. coli* rarely use. This BL21 derivative would hopefully help boost the protein yield and stop protein truncation.

Immediately after moving into the Rosetta cells an increase in protein yield was seen for both variants. Purity was still a limiting factor so further alterations to buffers and purification methodology were made. By adding a low level of imidazole to the wash buffer, loosely bound contaminants were removed, and cleaner elution fractions were seen. Following nickel column purification, a second column purification was added; anion exchange. Anion exchange was chosen due to *ZFY*'s high negative charge. Anion exchange contributed to enhancing the purity of *ZFY*. Nevertheless, the yield consistently diminished with each step of the protocol, resulting in a final yield lower than anticipated. Furthermore, the short byproducts were still evident suggesting that the *E. coli* struggles with this non-codon optimised sequence. End-to-end sequencing showed that *ZFY* was terminating at a tryptophan codon resulting in the absence of the last 53 amino acids. Tryptophan is the rarest amino acid in wild-type *E. coli* (Pezo *et al*., 2013), however, Tryptophan is generally not considered a problematic codon in *E. coli*. It is encoded by the sequence UGG, with two of the stop codons, UAG and UGA, bearing a striking resemblance. A single base change could potentially trigger the formation of a stop codon, prematurely halting protein synthesis. It is unlikely that this represents a mutation in the plasmids encoding the construct, as this issue was seen in both *ZFYS* and *ZFYL* constructs and persisted following codon optimisation (see below). Thus, any putative stop mutation would have to have arisen four times independently. Thus, the most likely explanation is that some feature either of the mRNA or the nascent polypeptide triggers misreading of the UGG codon as a stop codon.

With the issues persisting in the Rosetta cells, it was decided to get the sequences codon optimised for *E. coli.* Codon optimisation substitutes codons with synonymous codons that are more frequently used in the host organism leading to higher levels of recombination protein expression (Jenkins *et al*., 2023). After getting the sequences optimised, they were transformed into the Rosetta cells. Again, marked improvements in yield were noted, especially for the long variant. However, codon optimisation did

not stop the early termination of *ZFY*, but the abundance of the full-length versions was greater than the truncated versions. Subsequently, it was suggested that *ZFY* might be particularly prone to degradation within *E. coli*. BL21 pLysS cells were employed to investigate this theory, leveraging their elevated lysozyme production, which could enable the exploration of alternative lysis techniques.

A freeze-thaw lysis procedure was implemented, ensuring that the protein remained consistently chilled to mitigate any heat-related degradation. Nonetheless, this approach only exacerbated the ongoing degradation problem observed earlier. It appears that upon cell lysis, *ZFY* is promptly targeted for degradation, yielding truncated byproducts. This validation prompted the conclusion of recombinant protein expression in *E. coli*, opting instead for mammalian cell transfection protocols to extract *ZFY* for subsequent pull-down methodology.

To determine *ZFY*'s interacting factors and target proteins, a pull-down methodology was used through a GFP-tag. Subsequently, mass spectrometry-based proteomics was used to identify the proteins pulled down alongside *ZFY* and thus potential *ZFY* interactors. Two sets of duplicate samples were sent for proteomic analysis, revealing a spectrum of identified proteins including heterogeneous nuclear ribonucleoproteins, heat shock proteins, histone proteins, and 60S and 40S ribosomal subunit proteins. These protein groups play various roles in transcription and translation processes, encompassing functions such as transcription regulation, mRNA synthesis and stabilisation, DNA repair, protein folding and transport, as well as overseeing protein quality control regulation.

Multiple hnRNPs were identified across both repeats and were found to interact with both *ZFYS* and *ZFYL*. hnRNPs represent a large family of RNA-binding proteins regulating alternative splicing, mRNA stabilisation and transcriptional and translation regulation (Geuens *et al*., 2016). While the members of the family share general features, they differ in domain composition and functional properties. hnRNP interest has increased in disease research due to their association with many types of cancer and their potential role in tumorigenesis (Geuens *et al*., 2016).

After filtering, six hnRNPs were identified in the first dataset and three in the second. Notably, three of these hnRNPs were found duplicated in both datasets, they were HNRNPC1/C2, HNRNPA2/B1 and HNRPU. HNRNPC1/C2 has roles in splicing, translational regulation and transcript sorting with links to diseases like Alzheimer's Disease, Fragile X Syndrome and Cancer (Geuens *et al*., 2016). HNRNPA2/B1 functions in splicing and mRNA stability with links to Alzheimer's Disease, Cancer and Amyotrophic Lateral Sclerosis. Prior research has demonstrated that HNRNPA2/B1 functions as a catalyst for cancer progression via the PI3K/Akt, *WNT*/β-catenin,

MAPK/ERK, and additional signalling pathways (Lu *et al*., 2022). Given the wide array of targets associated with HNRNPA2/B1 and their diverse functions, it may also play a role in initiating an RNA switch to regulate the activity of miRNAs or lncRNAs in cancer cells (Yin et al., 2021);(Lu et al., 2022). HNRNPA2/B1 has also been shown to potentially serve as an oncogenic trigger due to its involvement in alternative splicing (Lu *et al*., 2022). Finally, HNRNPU has been shown to play a role in splicing and transcription regulation (Geuens *et al*., 2016), with links to several neurodevelopment disorders (Mastropasqua *et al*., 2022). The hnRNPs bind together with several other transcription factors to promoter and enhancer sequences to direct transcription (Geuens *et al*., 2016). *ZFY* may bind other transcriptional factors such as hnRNPs to direct transcription during male development and spermatogenesis.

Multiple ribosomal proteins were identified across both repeats and were found to interact with both *ZFYS* and *ZFYL*. Ribosomal proteins constitute the structural components of the ribosome, playing a vital role in both ribosome assembly and function (Kang *et al*., 2021). Anomalies in ribosome biogenesis, translation, and specific ribosomal proteins have been associated with various human diseases collectively referred to as ribosomopathies, which have also been implicated in the progression of cancer later in life (Kang *et al*., 2021). Ribosome biogenesis is thus highly regulated and controlled by a plethora of transcription factors, small nucleolar (snoRNAs) and RNA polymerases all collaborating to promote transcription, modification and processing of rRNAs, ribosomal protein synthesis, ribosome assembly and subsequently protein synthesis (Kang *et al*., 2021).

In the proteomics dataset, the initial set contained a combined total of 13 ribosomal proteins encompassing both the 40S and 60S subunits, whereas the subsequent dataset identified 18 ribosomal proteins. After combining the dataset, four large ribosomal subunit proteins were found to be duplicated; RPL7, RPL11, RPL14 and RPL18. RPL11 is highly investigated due to its role as a tumour suppressor (Kayama *et al*., 2016);(Fumagalli *et al*., 2009);(J. Chen *et al*., 2023). This highly conserved 60S ribosomal protein is not only involved in protein synthesis but also in cell cycle progression and cell fate determination. It has been shown that RPL11 binds and inhibits MDM2 ubiquitin ligase, promoting the stability of p53 thereby acting as a tumour suppressor (Kayama *et al*., 2016);(Fumagalli *et al*., 2009);(J. Chen *et al*., 2023). RPL14 has been linked to several types of cancer, including oesophageal squamous cell carcinomas, as well as lung and oral cancers, indicating a potential involvement in tumorigenesis. Nonetheless, the specific role of RPL14 remains largely unexplored. (Z. Zhang *et al*., 2021). Less research has been completed on

RPL7 and RPL18, so their specific function and clinical association is poorly understood.

Ribosomal proteins are crucial for both ribosome assembly and protein translation. Their individual functions are increasingly recognised and investigated (X. Zhou *et al*., 2015). Due to their association with ribosomal biogenesis, they are vital for cell growth, proliferation, differentiation and development. However, many of these proteins have been associated with the activation of the tumour suppressor p53 pathway, particularly in response to ribosomal stress (X. Zhou *et al*., 2015). The biogenesis of ribosomes begins in the nucleus and requires four rRNAs, 80 ribosomal proteins and ~70 snRNAs, with the final assembly and maturation finishing when the molecules are exported to the cytoplasm (Jiao *et al*., 2023). This process is regulated by several signalling pathways including mTOR, Myc and noncoding RNA (ncRNA) many of which are enhanced in cancers (Jiao *et al*., 2023). *ZFY* interacting with ribosomal proteins could indicate a much wider functioning network and potentially explains why it continues to persist on the Y chromosome, due to potential male development roles including the regulation of male-specific protein synthesis.

The interaction with ribosomes is unlikely to occur directly during translation as *ZFY* is a nuclear protein, and translation takes place in the cytoplasm. However, it is possible that *ZFY* interacts somehow with nascent ribosomes in the nucleolus. Is it also possible that this interaction may be an experimental artefact given that ribosomal proteins necessarily interact with negatively charged DNA and RNA and thus may bind non-specifically to the negatively-charged *ZFY* acidic domain during extraction. However, the consistent identification of specific ribosomal proteins across four experiments (two *ZFYS* and two *ZFYL* pulldowns) without consistently pulling down other highly abundant basic proteins such as the core histones argues against this kind of non-specific charge interaction.

The interaction between *ZFY* and ribosomal proteins might provide an explanation for the poor expression in bacterial systems – if the nascent *ZFY* peptide binds to ribosomes during translation, this could lead both to overall low levels of translation and to translational stalling / premature termination as observed in our *E coli* work.

Another protein identified is HSPA8, which belongs to the heat shock protein family. This molecular chaperone protein has been shown to have an integral role in cellular stress responses and is overexpressed in many cancers (J. Li & Ge, 2021);(Kobzeva *et al*., 2023). This may be attributed to its numerous interacting proteins involved in regulating protein quality control, including processes such as ubiquitination, response to unfolded proteins, and protein folding (J. Li & Ge, 2021);(Kobzeva *et al*., 2023). H4C1 serves as a fundamental element within the nucleosome, which fulfils

three primary roles: (1) compacting and organising genomic material, (2) serving as a signalling hub for chromatin-template processes, and (3) contributing to the formation of higher-order chromatin structures (Dhar *et al*., 2017);(McGinty & Tan, 2015). This suggests that *ZFY* may have diverse roles, spanning from transcription to protein synthesis and quality control.

However, confusion emerged when mitochondrial proteins such as DAP3 were found to potentially interact with *ZFY*. The mitochondrial genome is under strict maternal inheritance, meaning that deleterious mutations in mitochondrial DNA (mtDNA) can be harmful to males but not females (Ågren *et al*., 2020);(Wade & Fogarty, 2021). *ZFY* is under strict paternal inheritance. However, studies indicate that the motility of sperm, and consequently successful reproduction, relies on genes situated on the Y chromosome and within the mitochondrial genome (Wade & Fogarty, 2021). Although the precise nature of mitochondrial-Y interactions and their significance to male fitness remains unclear, evidence from *Drosophila melanogaster* suggests that loci within the mitochondrial genome can influence the expression of numerous autosomal loci in males, a phenomenon not observed in females (Dean *et al*., 2015);(Ågren *et al*., 2020). This could elucidate the interactions between *ZFY* and mitochondrial genes, as they may share common functions crucial for sperm motility and male fertility. DAP3 is a mitochondrial ribosomal protein, that oversees mitochondrial-encoded protein synthesis and mitochondrial dynamics (Xiao *et al*., 2015). By playing a key role in regulating mitochondrial function (Xiao *et al*., 2015), a potential link to sperm development could be made. Though there is no documented connection between DAP3 and spermatogenesis, DAP3 is expressed within the testis. Mitochondria in sperm form the "mitochondrial sheath" and play a crucial role in sperm structure and function (Hirata *et al*., 2002). They are vital for fertility as they provide the energy necessary for sperm motility with abnormal mitochondrial DNA resulting in infertility (Hirata *et al*., 2002).

Although not listed in **Table 5.9**, YB-1 was initially identified in the first proteomic dataset and was considered a potential protein of interest due to its associations with cancer. However, this finding was not replicated in the subsequent proteomics dataset. Nevertheless, confirmation through western blot analysis on a third sample verified the presence of YB-1 post-pulldown. YB-1 belongs to the highly conserved Y-box family, known for regulating gene transcription by binding to double- or single-stranded Y boxes within the promoters of various eukaryotic organisms (Homer *et al*., 2005). It is expressed throughout spermatogenesis, maintaining consistent expression levels without observable changes (Kretov, 2022). YB-1's transcriptional targets include genes involved in cell death, such as FAS and TP53, as well as those

related to cell proliferation, such as EGFR, MMP-2, and DNA topoisomerase IIα (Homer *et al*., 2005). Moreover, YB-1 has been shown to bind RNA and stimulate RNA splicing, suggesting pleiotropic functions (Homer *et al*., 2005). Additionally, Homer *et al* demonstrated that YB-1 inhibits p53's ability to induce cell death and transactivate cell death genes but does not affect p53's ability to transactivate CDKN1A or MDM2, necessary for cell cycle arrest (Homer *et al*., 2005). This partially explains the association between YB-1 and drug resistance, as well as poor tumour prognosis. Furthermore, bioinformatics analysis by Zhan *et al* revealed increased YB-1 expression in HNSC, correlated with a poorer prognosis (Zhan *et al*., 2022). Another study further showed that YB-1 facilitates tumorigenesis in hepatocellular carcinoma via the *WNT*/β-catenin pathway (Chao *et al*., 2017), suggesting a potential link to the elevated expression of *WNT* family members observed in the RNA-Seq dataset if *ZFY* interacts with YB-1.

DDX3X, a DEAD-box RNA helicase, evades X inactivation and has been recognised as a potential binding partner for *ZFY* in this experiment. Its homologue on the Y chromosome, DDX3Y, resides in the AZFa region(Dicke et al., 2023). With 92% sequence similarity between the two homologues (Kotov *et al*., 2017);(Dicke et al., 2023), there's a possibility that in experiments using male cell lines, DDX3Y could emerge as a *ZFY* target especially since evidence suggests they might be interchangeable in certain circumstances(Dicke et al., 2023). DDX3X exhibits ubiquitous expression and is associated with various cellular functions, including RNA metabolism, DNA damage response, apoptosis, *WNT*/β-catenin signalling, and tumorigenesis(Dicke et al., 2023). DDX3X has been found to have a major role in RNA metabolism, regulating almost all the stages including transcription, pre-mRNA splicing, RNA export and translation(Dicke et al., 2023). This means that DDX3X is a major molecule of interest for disease and cancer. Whereas, DDX3Y is only expressed in spermatocytes and is vital for successful spermatogenesis(Dicke et al., 2023);(Ditton et al., 2004);(Gueler et al., 2012);. Deletions in the DDX3Y gene have been linked to azoospermia and Sertoli Cell-Only Syndrome emphasising the importance of DDX3Y in the maintenance and development of early male germ cells (Dicke et al., 2023);(Ditton et al., 2004);(Gueler et al., 2012). *ZFY* expression has been observed in early spermatocytes, ceasing at the initiation of MSCI. This could imply a possible feedback mechanism wherein *ZFY* activates DDX3Y expression before MSCI, facilitating the continuation of spermatogenesis. Additionally, *ZFY* may participate in interactions with DDX3X/Y, facilitating male-specific RNA metabolism functions essential for spermatogenesis and male development via the *WNT*/β-

catenin signalling pathway. This further elucidates the prevalence of *WNT* family members observed in the RNA-Seq analysis.

In general, numerous proteins listed in **Table 5.9** are associated with RNA metabolism, RNA processing, transcription, and translation highlighted by Reactome pathway analysis. As a transcription factor, *ZFY* is anticipated to bind to DNA sequences to regulate gene transcription. It could be proposed that *ZFY* activates the expression of genes involved in RNA metabolism to facilitate the advancement of spermatogenesis and male development. Numerous identified proteins in **Table 5.9** also serve as transcription factors, suggesting that *ZFY* might operate in a more complex manner and potentially be implicated in activating numerous downstream pathways.

In summary, *ZFY* appears to have numerous potential interactors and downstream targets with a plethora of functions. Moreover, the extensive number of differentially expressed genes detected in the RNA-Seq analysis might be accounted for by the multitude of signalling pathways mentioned here that are downstream of the identified targets of *ZFY*. *ZFY* seems to have roles ranging from DNA repair, transcription, translation, protein quality control and sperm mitochondrial roles. The multitude of roles discussed here could elucidate *ZFY*'s continued presence on the Y chromosome over generations, due to its crucial importance in male development and fertility.

6. Discussion & Future Works

This thesis is composed of four results chapters all forming part of an overarching investigation into *ZFY*'s function as a transcriptional activator, and how this relates to its structure and evolution.

6.1 Phylogenetic analysis of *ZFY* reveals strong negative selection, specific conserved motifs in the acidic domain, and accelerated evolution in rodents

The human Y chromosome only represents 2-3% of the haploid gene (Quintana-Murci & Fellous, 2001), however, *ZFY* continues to be encoded on the slowly degenerating chromosome. This strongly suggests that *ZFY* plays an essential role in some aspect of male biology. Consistent with this, phylogenetic analysis in this thesis – which included *ZFY* and *ZFX* sequences from multiple mammalian species across the eutherian radiation together with marsupial and non-mammalian outgroups - found that *ZFY* is under strong negative selection pressure. Sequence conservation was strongest within the DNA-binding domain, most likely to ensure that *ZFY* maintains a consistent set of downstream target genes shared with *ZFX* (see also section on gene conversion below). The sequence of the acidic transactivation domain was less conserved, however there was gross conservation of the charge and hydrophobicity structure within this domain, despite underlying changes in the specific amino acids present. This is consistent with current models of how activation domains function through multivalent transient interactions with the transcription machinery.

Specifically, some patches of high conservation within the AAD matched the consensus 9aaTAD motifs known to play a role in transactivation (S. Piskacek *et al*., 2007) These short 9aaTAD regions are common domain regions in the transactivation domains of transcription factors ranging from yeast to mammalian cells (S. Piskacek *et al*., 2007). The hydrophobic nature of these clusters are crucial for interaction with multiple transcription mediators but are able to do so with different binding affinity (M. Piskacek *et al*., 2016). Whilst there is a large variability in 9aaTAD character they are all universally recognised by transcription machinery mediators. This conservation has been linked to the 9aaTAD domain occurrence and functionality is down to transcriptional mediators such as TAF9 and XIX domain in MED15 (M. Piskacek *et al*., 2016). Not all of the predicted 9aaTAD sequences were highly conserved, with some being specifically lost in rodents. This may indicate a change in function or a change in selection pressure within this clade (discussed further below).

Intriguingly, in addition to the known 9aaTAD transactivation motif, we identified additional short hydrophobic motifs that were highly conserved but were a poor match to the 9aaTAD consensus. There were highly conserved (i.e. present in all land

vertebrates including Xenopus) hydrophobic stretches at residues -270-276 and ~376-386 (*Supplementary* **Figure 1**), with global consensus sequences VIKVYIF and F(M/V)PIAWAAAY respectively. Neither matches a 9aaTAD sequence suggesting there may be other functional classes of activation motif present in the *ZFY* AAD. The functionality of each of the various potential activating motifs identified will require experimental validation to determine their activity in reporter assays.

Investigations into possible gene conversions were also performed due to previous reports of interchromosomal gene conversion in this gene family (Hayashida *et al*., 1992);(Pamilo & Bianchi, 1993);(Drouin *et al*., 1999);(Slattery *et al*., 2000);(Bidon *et al*., 2015). Gene conversions were identified by two methods: firstly, phylogenetic tree analysis, in which gene conversions are indicated by the pairing of a *ZFY* and *ZFX* sequences within a given species; and secondly by the use of GENECONV software, which performs a similar analysis on a more granular basis using local alignment. A full-length alignment of *ZFY* **(Figure 2.5)** indicated only one potential gene conversion, in elephant. However, when the nucleotide alignment was subsequently split into coding exons 1-6 **(Figure 2.11A)** and coding exon 7 **(Figure 2.11B)** a different story was seen. While analysis of coding exons 1-6 showed no detectable gene conversion even in elephant, analysis of coding exon 7 showed multiple additional potential gene conversion in horse, pig, marmot, and stoat. Moreover, the rodents (mice and rats), artiodactyls (cow, goat, deer) and primates (humans, chimpanzees, gorillas, macaques, baboons, snub-nosed monkeys and marmosets) also showed pairing of *ZFY* and *ZFX* sub-trees, suggesting gene conversions at the root of each of these clades. Geneconv was employed to corroborate the identified potential gene conversions, resulting in the replication of gene conversions in rats, elephants, and stoats. Details of other potential conversions could not be replicated by Geneconv, likely because it has low statistical power in the face of very high rates of gene conversion (Mansai & Innan, 2010).

Overall, the findings indicate that there is strong negative selection to maintain *ZFY* gene function. This is particularly intense in the DBD and in specific motifs within the AAD (potentially the site of protein/protein interactions with binding partners). Together with this, there is frequent and recurrent gene conversion between *ZFX* and *ZFY*, but that the extent of these conversions is limited to coding exon 7 encoding the DBD. This further indicates an imperative to preserve the functionality of the DNA binding domain and ensure that the X and Y copies consistently target the same recognition sequences.

Set against this story of overall conservation, however, as previously observed by Tucker *et al* (Tucker *et al*., 2003) it was found that rodent *ZFY* sequences, in particular

mice and rat *ZFY* seem to be undergoing much more rapid evolution than *ZFY* sequences in other mammalian species. This was made evident by the long branch lengths of these animals and their large substitution rate per million years shown in **Table 2.5** in Section 2.3.6.2. This could indicate that rodent *ZFY* is under a different selection pressure and as a result could eventually have differing roles to other *ZFY* sequences. In particular, in this thesis we show for the first time that this more rapid evolution is not restricted to the final exon of the sequence as studied by Tucker *et al* but is also seen in the initial exons comprising the transactivation domain. The reason for this acceleration is not clear: Tucker *et al* suggested that the shift to testis-specific expression for *ZFY* in rodent species has led to relaxed selective constraint. Future work could test this directly by defining more precisely which species show accelerated *ZFY* evolution and which species have testis-specific *ZFY* expression.

6.2 The testis-specific splicing *ZFY* is conserved in non-eutherian species and is likely to be regulated by *RBMY*

A unique feature of *ZFY* is that it undergoes testis-specific alternative splicing to exist as two variants; a testis-specific short variant and a ubiquitous long variant. This is seen in humans, mice and other mammalian species. During the project, chapter 3 aimed to see if this testis-specific splicing event was evident in non-eutherian species, where *ZFY* is located on an autosome. Using publicly available RNA-Seq data it was found that a testis-specific splicing event was occurring in both chicken and opossum. Opossum seems to exhibit similar splicing characteristics to humans, with the second coding exon being selectively skipped in a proportion of transcripts in the testis. Chickens in contrast seem to be experiencing a different testis-specific splicing event with a potential novel transcript observed that joins several novel exons to the terminal exon encoding the DBD. Although strongly suggested by the public dataset, the existence of this novel *ZFY*-related transcript will require experimental validation.

Focusing specifically on human *ZFY*, this thesis shows for the first time that in a cell culture reporter system, expression of *RBMY* increases the level of exon skipping for the second coding exon of *ZFY*. Expression data has indicated that *ZFYS* is only expressed in cell types which also express *RBMY*, the earlier premeiotic cells (spermatogonia and primary spermatocytes) (Skrisovska *et al*., 2007). Therefore, if *RBMY* really does regulate *ZFY* splicing, it would explain the testis-specific nature of *ZFYS* as that is where *RBMY* is being expressed. Furthermore, as described by Decarpentrie *et al*, a *RBMY* deficient (AZFb deletion) human patient showed a phenotype similar to that of mouse *ZFY* overexpression (Decarpentrie *et al*., 2012). This can be explained by *RBMY* regulating *ZFY* splicing, since a deficiency in *RBMY*

would result in more *ZFYL*, but less *ZFYS*. This would lead to increased *ZFYL* expression, resulting in a *ZFY* overexpression phenotype characterised by meiotic arrest.

Further circumstantial evidence of a functional link between *ZFY* and *RBMY* is shown by deletion model work in mice. Despite many attempts it has not yet been possible to obtain a stably transmitting mouse line with a specific deficiency of *RBMY*. While several lines were generated (termed Y^{d1} through to Y^{d6}) with deletions of part or all of the *RBMY* gene cluster, all of these also unexpectedly showed epigenetic silencing of Sry due to spreading of centric heterochromatin, with consequent sex reversal (Capel *et al*., 1993);(Laval *et al*., 1995). Complementing the Sry deficiency with an autosomal transgene resulted in fertile XY^{d1},Sry male mice with abnormal sperm morphology. Subsequently, Vernet *et al* showed that the Y^{d1} deletion also epigenetically silences *ZFY*2 in these males and in sex-reversed XY^{d1} females (Vernet, Szot, *et al*., 2014).

Collectively the data tentatively suggest that while in humans *RBMY* deletion leads to sterility, in mice *RBMY* deletion is compatible with male germ cell development if *Zfy2* is also silenced. Under our hypothesis, in XY^{d1},Sry male mice, expression of *ZFYS* is reduced due to the absence of *RBMY*. This would be expected to lead to an overdose of *ZFYL* in premeiotic cells – as we believe is the case in men with AZFb deletions – triggering apoptosis and infertility. However, epigenetic silencing of *Zfy2* in XYd1,Sry males reduces the net *ZFYL* expression prior to meiosis, rescuing the *ZFYL* overexpression phenotype, allowing germ cells to survive and complete the meiotic divisions. Since *Zfy1* remains active in these males, spermatid development can then complete since *Zfy1* alone is sufficient for assisted fertility, albeit with some sperm morphological abnormalities (Yamauchi et al., 2022).

Overall, this thesis shows that the testis-specific splicing shift between *ZFYS* and *ZFYL* is conserved in marsupials and must therefore have preceded acquisition of *ZFY* by the eutherian Y chromosome. Moreover, the splicing appears to be regulated by *RBMY* (which is Y-borne in both marsupials and eutherians). Finally, while chicken does not show the same splicing event as mammals, there is an exciting possibility that an alternative testis-specific isoform exists in chicken also. If so, this means that mammals and birds have independent strategies that nevertheless result in the production of a low-function “DNA-binding-only” form of *ZFY* specifically in testis.

A final intriguing data point is that in the RNA-Seq work (see following section), both *ZFYL* and *ZFYS* downregulated the expression of RBMXL2, which in turn is a close relative of both RBMX and *RBMY*. As with all the transcriptional changes in response to *ZFY* expression, this downregulation was more pronounced for *ZFYL* than *ZFYS*.

Since HEK293 cells are female, we cannot tell whether *ZFY* expression directly regulates *RBMY* transcription. However, this is something to investigate in the future. If *ZFYL* does downregulate *RBMY* it would form a self-reinforcing feedback loop in which production of *ZFYL* would decrease *RBMY* expression and led to further increases in *ZFYL* expression at the expense of *ZFYS*, leading to a pronounced "switch" between isoforms dependent on cell differentiation.

6.3 *ZFYS* is a "weaker version" of *ZFYL*, but both have important spermatogenic roles.

The roles of both *ZFY* variants have been previously investigated with much of the current literature focusing on mouse studies limiting the amount of available research in humans. Whilst there is considerable sequence similarity between mouse and human *ZFY*, rodent *ZFY* as previously described is rapidly evolving and at a much greater speed compared to human *ZFY*. This creates enough reason to believe that mouse *ZFY* and human *ZFY* could potentially have differing biological functions, making investigations into human *ZFY* variant functions more crucial.

To begin to address the functions of *ZFY* in a human system, an overexpression model in human cell culture was developed. By extracting RNA from HEK293 cells overexpressing the *ZFY* variants, the aim was to identify changes in the cell's transcriptomes that would indicate *ZFY*'s potential biological functions. RNA sequencing was performed, and a downstream differential analysis pipeline was performed. Pathway enrichment analysis was subsequently completed to identify enriched pathways and infer the role of *ZFY* when expressed. It was also to scrutinise the distinction between *ZFYS* and *ZFYL* and ascertain their distinct respective functions.

One of the first observations made was that *ZFYS* is a potentially "weaker" transcription factor compared to *ZFYL*. While both variants targeted many of the same targets, *ZFYL* seemed to do so to a greater extent, with greater L2FC changes identified. What is not known is if *ZFYS* is a competitive inhibitor of *ZFYL* as these two compete for the same binding sites, further *in vivo* methodology needs to be performed to interpret the physiological consequences of *ZFYS* not only in terms of how it directly regulates genes but how it interacts with and possibly reduces the activity of *ZFYL* in the same cell.

Enrichment analysis using Reactome was performed, and potentially upregulated pathways were identified. Both *ZFYS* and *ZFYL* seem to target collagen-associated biological pathways ($p<0.05$), extracellular matrix pathways ($p<0.05$) and potassium channels pathways (not significant). Collagen is a major scaffolding protein within the

extracellular matrix, the support structure for cells, tissues and organs (Pompili *et al*., 2021). Within the processes of spermatogenesis and fertilisation two key extracellular matrices, the basement membrane, a modified form of extracellular matrix and the zona pellucida, the egg extracellular matrix protecting the plasma membrane are crucial (Siu & Yan Cheng, 2008);(Litscher & Wassarman, 2020).

6.3.1 The Extracellular Matrix is Crucial during Spermatogenesis.

Sertoli cells and spermatogonia rest on the basement membrane in the testis at varying stages of the seminiferous epithelial cycle, relying on both structural and hormonal supports (Siu & Yan Cheng, 2008). It is therefore very likely that the extracellular matrix has a major role in spermatogenesis regulation, particularly spermatogonia and germ cell regulation. Further to this, the basement membrane is in contact with the underlying collagen network and alongside the lymphatic network this forms the tunica propria of the seminiferous tubules which is crucial to the production of spermatozoa with help from the Leydig cells. This suggests the critical function of the extracellular matrix in spermatogenesis (Siu & Yan Cheng, 2008). Abnormal basement membrane structures have been detected in infertile men and have been seen in the testes of men with cryptorchidism, vasectomy and varicoceles (Siu & Cheng, 2004).

In vitro studies have shown the importance of the extracellular matrix in Sertoli function specifically through the regulation of Sertoli cell morphology and behaviour (Siu & Cheng, 2004). The extracellular matrix seems to control the differentiation, cell growth and migration of the Sertoli cells during the progression of spermatogenesis. Further to this Leydig cell proliferation, testosterone production and gene expression have also been shown to be affected by the extracellular matrix. All of these processes are essential to the proper progression of spermatogenesis (Siu & Cheng, 2004).

Spermatogonia represent the germ cell stem cell population that lies on the basement membrane surrounded by somatic Sertoli cells (W. H. Walker, 2010). Sertoli cells support spermatogonia development through complex endocrine and paracrine inputs (W. H. Walker, 2010). Three classes of spermatogonia lie on the basement membrane; stem cell spermatogonia, proliferative spermatogonia, and differentiating spermatogonia (Creasy & Chapin, 2013). Both stem cell spermatogonia and proliferative spermatogonia are responsible for renewing their cell number and ensuring there is a committed pool of spermatogonia for differentiation (Creasy & Chapin, 2013). Differentiating spermatogonia upon signalling, migrate away from the basement membrane and begin to mature into preleptotene spermatocytes

continuing on to spermatozoa (Gruber *et al*., 2010);(Mruk & Cheng, 2015). This process has to be highly regulated by several factors such as hormones, temperature and oxygen availability ensuring tight regulation of spermatogenesis and the basement membrane (Gruber *et al*., 2010).

Not only does the extracellular matrix have an integral role in the spermatogenesis process but it also forms the blood-testis barrier (BTB). This barrier is unlike other mammalian barriers as it is composed only of specialised junctions between adjacent Sertoli cells near the basement membrane. However, the BTB is one of the tightest blood-tissue barriers in the mammalian body, dividing the seminiferous epithelium into the basal and apical compartments **(Figure 6.1)** (Cheng & Mruk, 2012). The apical compartment is where meiosis I and meiosis II of spermatogenesis occur, but within the basal compartment of the epithelium spermatogonial renewal, differentiation and cell cycle progression up to meiosis occurs. During spermatogenesis, the BTB is not static and continues to undergo reconstruction to allow the transit of preleptotene spermatocytes across the barrier (Cheng & Mruk, 2012). However, the reconstruction of the BTB must not compromise the immunological barrier produced by the BTB to prevent the production of antibodies against meiotic and postmeiotic germ cells. This means a timely degeneration of the BTB above transiting cells and production of the new BTB behind the spermatocytes in transit is crucial (Cheng & Mruk, 2012). The testes are immune-privileged regions and the BTB sequesters any detrimental immune responses in autoantigenic germ cells. Therefore, the BTB plays a pivotal role in preserving the optimal microenvironment necessary for germ cell development and maturation (Mital *et al*., 2011). Additionally, it serves to sequester autoantigens within germ cells, thereby preventing harmful autoimmune reactions and safeguarding germ cells against cytotoxic substances (Mital *et al*., 2011).

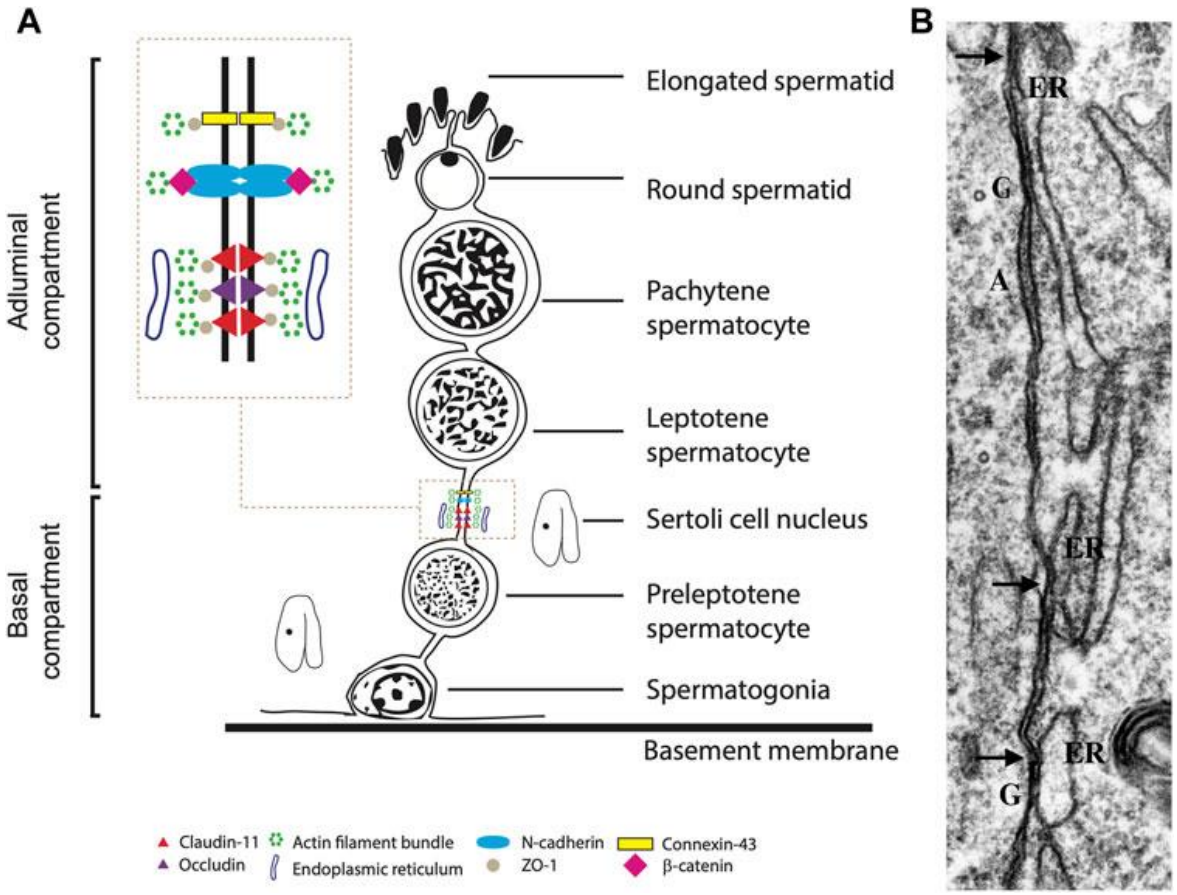


Figure 6.1: A: Illustration of the blood testis barrier. B: Transmission electron microscopy of the Sertoli cell junctions. (Luaces *et al*., 2023)

The extracellular matrices are regulated by matrix metalloproteinases (MMPs), endopeptidase enzymes that are capable of degrading numerous pericellular substances and virtually all structural matrix proteins (Sternlicht & Werb, 2001)(Cabral-Pacheco *et al*., 2020). The degradation of the ECM is greatly important due to its links to spermatogenesis, as well as embryonic development and angiogenesis in other tissues. The alteration of MMP expression can result in abnormal extracellular matrix degradation and subsequent disease progression including cancer progression (Cabral-Pacheco *et al*., 2020). *ZFYL* and *ZFYS* expression seems to result in the differential expression of MMPs with 8 and 5 identified respectively. MMPs are inhibited by tissue inhibitors of MMPs (TIMPs), endogenous protein regulators (Yao *et al*., 2011)(Cabral-Pacheco *et al*., 2020). TIMPs are present in the extracellular matrix and function by blocking MMPs through the formation of a 1:1 stoichiometric complex. The balance between MMPs and TIMPs within the testis is crucial for controlling germ cell development and Sertoli cells (Yao *et al*., 2011). *ZFYS* and *ZFYL* both target and upregulated *TIMP-2*. Testicular TIMP-2 has been identified to be secreted by Sertoli and Leydig cells, with functions related

to germ cell apoptosis regulation, germ cell migration and tissue restructuring during testicular development (Yao *et al*., 2011). This indicates that both *ZFY* variants potentially target crucial extracellular matrix regulators and thus function to aid in the maintenance and homeostasis of the extracellular matrix in the testis during spermatogenesis.

However, this is not the only extracellular matrix of importance during the processes of fertilisation. The zona pellucida is the extracellular matrix responsible for protecting the egg's plasma membrane.

6.3.1.1 Fertilisation and the Egg Extracellular Matrix

For fertilisation of the egg to be successful the egg and spermatozoa plasma membrane must fuse. Therefore the spermatozoa must undergo capacitation, a pre-requirement for fertilisation as only capacitated spermatozoa are able to undergo the acrosome reaction to fertilise the egg (Jin & Yang, 2017). However, only around 20-40% of the sperm subpopulation become capacitated, selection for capacitation is unknown (Aldana *et al*., 2021). Capacitation is a series of biochemical and physiological changes that the spermatozoa must undergo. These include changes in the membrane properties, intracellular ion concentration, enzyme activity and protein modifications. The changes induce the stimulation of the acrosome reaction in preparation for spermatozoa penetration of the zona pellucida, the egg's extracellular matrix (Jin & Yang, 2017).

The zona pellucida is a thick extracellular matrix that encloses the mammalian oocytes, eggs, and early embryos and is vital for oogenesis, fertilisation and preimplantation development (Litscher & Wassarman, 2020);(Wassarman & Litscher, 2022). The zona pellucida is capable of inducing the acrosome reaction of spermatozoa facilitating the completion of fertilisation (Gupta, 2021). The acrosome reaction is a key step during the interaction of gametes and ultimately allows the spermatozoa to penetrate the zona pellucida and fuse with the oocyte membrane. If spermatozoa are unable to perform this key reaction, fertilisation will not occur (Brucker & Lipford, 1995).

The acrosome reaction is a Ca^{2+}-dependent exocytotic process (Yanagimachi, 2011). While the mechanism of modifications that trigger the acrosome reaction to begin is generally unknown, it is known that there is an upregulation of intracellular concentration of Ca^{2+} and an increase in intracellular pH (Aldana *et al*., 2021). The control of the pH is of vital importance to the functioning of the sperm Ca^{2+} and K^{+} channels, and if this is not tightly controlled spontaneous acrosome reactions can occur (Aldana *et al*., 2021). Under normal conditions, the acrosome is an acidic

vesicle of lysosomal/Golgi origin containing hydrolytic enzymes. The elevation of pH and Ca^{2+} and acrosomal release of Ca^{2+} are thought to trigger the acrosomal reaction. The acrosome swells resulting in the deformation of the outer acrosomal membrane, allowing the subsequent interaction and docking with the plasma membrane. The fusion of the acrosomal membrane and the plasma membrane triggers exocytosis (Aldana *et al*., 2021).

Not only is Ca^{2+} vital to acrosome exocytosis, but K^{+} is also vital to membrane hyperpolarisation during sperm capacitation, motility hyperactivation and acrosome exocytosis (Delgado-Bermúdez *et al*., 2024). K^{+} channels are regulated by calmodulin, the cytosolic Ca^{2+}-binding protein via the activation of phosphorylation cascades essential to motility hyperactivation and acrosomal exocytosis (Delgado-Bermúdez *et al*., 2024). Therefore, for successful fertilisation extracellular matrix degradation and potassium channels are both crucial. While the potassium channels were not significantly enriched in the analysis there could be a suggestive link due to the other related pathways being upregulated.

The potential links to *ZFY*'s role in spermatogenesis and fertilisation through the regulation of the extracellular matrix have been discussed, however, the upregulation of pathways governing extracellular matrix regulation may also suggest involvement in tumour microenvironment shaping and cancer progression

6.3.1.2 The Tumour Microenvironment Shaping and Cancer Progression

Solid tumours are highly heterogenous environments and are a combination of tumour cells, vasculature, extracellular matrix, stromal and immune cells (Henke *et al*., 2020). The extracellular matrix of solid tumours greatly differs from normal organs and influences not only malignancy and growth of the tumour but also controls the response towards cancer therapy (Henke *et al*., 2020). Throughout tumour progression, carcinogenic cells work to recruit host stromal cells which create a unique microenvironment resulting in the remodelling of the extracellular matrix consequently promoting tumour invasion (Popova & Jücker, 2022). MMPs catalyse the remodelling of the extracellular matrix through the modification and cross-linking of extracellular matrix proteins increasing stiffness and composition alteration (Popova & Jücker, 2022).

The deposition of the extracellular matrix is a hallmark of cancer (Popova & Jücker, 2022). The most common alteration associated with this is the increase in the deposition of collagen, resulting in the collagen fibres of tumours often being straightened and reorganised. The most important component of the extracellular matrix is collagen (C. Walker *et al*., 2018). Collagen I and IV have both been shown

to promote invadopodia extension and tumour cell migration with other collagens able to bind and activate receptor tyrosine kinases discoidin domain receptors that have been suggested to facilitate metastases and cancer aggressiveness. Overall, collagen subtypes are often upregulated in cancer as they are crucial to the majority of the steps in tumour progression, such as proliferation, invasion, angiogenesis and metastasis (Popova & Jücker, 2022).

Therefore, the extracellular matrix has a key role in cancer progression, with collagen being a vital component of the matrix, changes in the deposition or degradation of collagen can lead to an imbalance of extracellular homeostasis. From RNA-Seq data, it is suggested that *ZFY* potentially has roles in the extracellular matrix and collagen control which could suggest that changes in *ZFY* expression could result in cancer progression.

6.3.2 *WNT* Signalling: a Key Cascade Regulating Development

The *WNT* signalling pathway, widely preserved throughout evolution, plays a pivotal role in governing various cellular functions (Koni *et al*., 2020). These encompass determining cell destiny, orchestrating organ formation in embryonic stages, maintaining equilibrium in adult tissues, facilitating cell movement, establishing cell polarity, and perpetuating the renewal of stem cells (Koni *et al*., 2020). *WNT* proteins make up the important signalling molecules regulating the diverse cellular processes (Kestler & Kühl, 2008).

Due to the complexity of the *WNT* cascade, it was further subdivided into two different branches; the canonical *WNT*/β-catenin and the non-canonical β-catenin-independent pathways, with the latter being further subdivided into two additional branches; the planar cell polarity and the *WNT*/calcium pathways (Koni *et al*., 2020). Across all the branches of *WNT* signalling, a plethora of functions are performed with the help of 19 *WNT* glycoproteins alongside other proteins (Patel *et al*., 2019). These *WNT* proteins have been classified into various types. Highly transforming members include *WNT*1, *WNT*3, *WNT*3a, and *WNT7*a. Intermediately transforming or non-transforming members encompass *WNT*2, *WNT*4, *WNT*5a, *WNT*5b, *WNT*6, *WNT7*b, and *WNT*11 (Patel *et al*., 2019).

The canonical pathway is defined by an accumulation of intracellular β-catenin which leads to its translocation to the nucleus where it is capable of controlling gene expression (Ackers & Malgor, 2018). Whereas the non-canonical is β-catenin-independent and controls gene expression from alternative intracellular mechanisms. The non-canonical pathway is capable of inhibiting the canonical pathway. Both the canonical and non-canonical pathways control key metabolic pathways such as

mTOR and insulin signalling explaining their link to the pathogenesis of diabetes. However, their clinical relationship expands beyond metabolic processes and diseases (Ackers & Malgor, 2018). Both the canonical and non-canonical *WNT*/β-catenin pathways have been shown to contribute to cancer development (Koni *et al*., 2020).

While *WNT* signalling is crucial to many biological processes and is associated with multiple clinical diseases and cancer development, studies have shown links to roles in the testis and spermatogenesis. *WNT* signalling within the testes potentially has an inhibitory role in testis formation while it seems also to be vital to late stages of spermatogenesis (Singh *et al*., 2019).

6.3.2.1 *WNT* Signalling and Spermatogenesis

Mouse models have contributed to the identification of the potential roles of *WNT* signalling in spermatogenesis and male fertility (Kerr *et al*., 2014). Within the testicular epithelium, the presence of many *WNT* ligand isoforms has been detected including; *WNT1, WNT3A, WNT4 and WNT7* (Covarrubias *et al*., 2015). They were explicitly expressed either in adult tissues or during testicular development. Other members of the *WNT*/β-catenin signalling pathway have been discovered to be expressed in the testes such as the Frizzled 3,4 and DVL 1, 2 and 3. (Covarrubias *et al*., 2015).

WNT signalling is vital in fetal life for adult fertility, with *WNT*4 shown as being essential for the normal development of the male fetal reproductive tract (Jeays-Ward *et al*., 2024);(Kerr *et al*., 2014). This was confirmed as mutant *WNT*4 mice testes had abnormal Sertoli cell differentiation (Jeays-Ward *et al*., 2024);(Kerr *et al*., 2014). Not only is *WNT*4 crucial to Sertoli cell function, but *WNT*3 has also been shown to be critical for Sertoli cell function. *WNT*3 mice knockdowns showed significantly reduced litter sizes compared to the age-matched controls, as well as sperm count, and testicular size (Basu *et al*., 2017). Within the *WNT*3 knockdown mice, testicular sections showed abnormal tubular structure, reduced presence of sperm and sloughed-off germ cells. This suggests that *WNT*3 is important for Sertoli-cell mediated regulation of spermatogenesis and thus is vital for male fertility (Basu *et al*., 2017). Another example was *WNT7*a deletions in mice which showed the inhibition of the Müllerian duct regression which results in the retention of the female reproductive tract tissues in adult males, consequences of this include the impediment of the sperm passage at ejaculation (Parr & McMahon, 1998);(Kerr *et al*., 2014). *WNT7*a expression is highest in spermatids and is one of the strongest targets of *ZFYL* and aligns with *ZFYL*s' expression, suggesting an important link between these two

genes. In boar, *WNT*1 was shown to play a role in the achievement of *in vitro* sperm capacitation as well as in progesterone-induced *in vitro* acrosome exocytosis (Covarrubias *et al*., 2015). *WNT* signalling has also been shown to be vital to spermatogonial stem cell self-renewal, and this is potentially regulated by *WNT*5a (Reh *et al*., 2011). In mice, *WNT*5a and *WNT*5a receptors were identified in Sertoli cells and spermatogonial stem cells respectively indicating *WNT*5a could be an extrinsic factor supporting stem cell renewal via the non-canonical pathway (Reh *et al*., 2011). Whilst not all of these were identified in this data set it does suggest that several *WNT* proteins have major functioning roles within spermatogenesis and explains why a high number of *WNT* proteins were upregulated by both *ZFYS* and *ZFYL*.

Furthermore, continuous expression of active β-catenin isoform in Sertoli cells was found to result in germ cell and Sertoli cell depletion (Kerr *et al*., 2014). Further to this incorrect activation of *WNT* signalling was also shown to impair germ cell development on top of Sertoli cell apical extension loss and BTB integrity loss. However, the deletion of β-catenin resulted in no phenotypic changes. This shows that maintaining *WNT* signalling is crucial as continuous activation of *WNT* signalling in Sertoli cells results in spermatogenic defects while deletion of β-catenin does not affect the function of Sertoli cells. *WNT* proteins seem to have multiple sites of action that influence testis development and maintain male fertility (Kerr *et al*., 2014).

6.4 *ZFYL* Shows Enrichment of Presynaptic Function Pathways

During the analysis of *ZFYL*-specific functions, enrichment analysis highlighted intriguing pathways, particularly those associated with neurons, specifically focusing on presynaptic functions. While none of these pathways passed FDR significance, they are still worth highlighting. While a human neuron and sperm are very distinct cells both morphologically and functionally, evidence has suggested that there are more similarities between the human brain and testis than thought (Matos *et al*., 2021). When looking at the functional level of human neurons and sperm, shared characteristics are present, including the processes of exocytosis, the presence of similar receptors and similar signalling pathways (Matos *et al*., 2021). Sperm have been shown to express a repertoire of signalling receptors including; GABAergic, dopaminergic, noradrenergic, cholinergic and others (Ramırez-Reveco *et al*., 2017). While the mechanisms involved are different, both neurons and sperms are capable of activating other cells (Matos *et al*., 2021). Sperms can activate oocytes to produce a diploid embryo following the fusion of their plasma membranes while neurons are capable of activating other neurons or somatic effector cells. Another similarity they

share is exocytic processes. Neurons use exocytosis to release neurotransmitters from synaptic vesicles essential for neuron communication. The synaptic vesicles are comparable to the acrosome of the sperm packaged with hydrolytic enzymes. The release of these hydrolytic enzymes is vital for the breakdown of the zona pellucida and subsequent fusion of the sperm and oocyte. However, the major difference is that neurons continuously undergo exocytosis when in contrast sperm only do this once (Matos *et al*., 2021). These pathways are therefore all essential to the successful fertilisation of the oocyte and the production of a diploid embryo.

The *ZFYL* enriched pathways linked to presynaptic function suggest roles later on at the point of fertilisation which corroborates with *ZFYL*'s predominant expression post-meiosis in spermatogenesis. Subsequent analysis, then looked into *ZFYS* pathway enrichment to try and ascertain the different roles of these two variants.

6.5 *ZFYS* Activates a Key Cancer Pathway Driver

ZFYS is a testis-specific variant with expression identified in a head and neck cancer cell line. This is what led to the hypothesis that *ZFYS* is a potential cancer-testis gene. One of the key aims of this thesis was to investigate the role of *ZFYS* and through RNA-Seq identify potentially enriched pathways that could explain *ZFYS*'s expression in a head and neck squamous cell carcinoma cell line.

ErbB signalling showed increased enrichment upon the overexpression of *ZFYS*, whereas this effect was not observed with the overexpression of *ZFYL*. ErbB signalling encompasses four transmembrane tyrosine kinase receptors, ErbB receptors (*ErbB1/EGFR/HER1, ErbB2/HER2, ErbB3/HER3 and ErbB4/HER4*), which are responsible for cell differentiation, migration, mitogenesis and survival (Arteaga, 2011);(Appert-Collin *et al*., 2015). These receptors form functional dimers following their activation by epidermal growth factor (EGF)-family growth factors (Citri & Yarden, 2006). Downstream effectors of ErbB signalling include the MAPK/ERK, PI3K-AKT and phospholipase C gamma pathways (Jacobi *et al*., 2017). Together these signalling pathways have major roles including cell proliferation, apoptosis, angiogenesis, cell adhesion and motility, embryonic development, and organogenesis (Jacobi *et al*., 2017).

However, the dysregulation of these tyrosine kinase receptors has been linked to cell transformation and cancer with studies showing that alongside downstream pathways they can regulate epithelial-mesenchymal transition, migration, and tumour invasion by extracellular matrix modulation (Arteaga, 2011);(Appert-Collin *et al*., 2015);(Kumagai *et al*., 2021). As previously mentioned the extracellular matrix has a major role in tumour progression and tumour microenvironment shaping (Arteaga,

2011);(Appert-Collin *et al*., 2015). Mutant profiles of tumours have unique ErbB gene expression profiles and they have been identified as key for cancer cell proliferation and survival (Jacobi *et al*., 2017).

This enrichment analysis could hint at *ZFYS* having a cancerous function when wrongly expressed outside the testes potentially through the over-activation of ErbB signalling. However, to confirm this further *in vitro* and *in vivo* methodology would need to be performed. Proteomics was subsequently used to identify the direct binding partners of both *ZFY* variants, with key protein hints being the hnRNP and ribosomal protein families. Both these families are associated with tumorigenesis (Pecoraro *et al*., 2021);(Sudhakaran & Doseff, 2023).

6.6 Proteomics Analysis shows links to DNA and RNA metabolism

Proteomics analysis following pull-down methodology suggested that both variants of *ZFY* are capable of binding a variety of hnRNPs and ribosomal proteins. This was a surprising result as it was anticipated that this would identify interactions with other transcription factors and/or components of the Mediator complex as seen for other known acidic domains of other transcription factors (M. Piskacek *et al*., 2016);(S. Piskacek *et al*., 2007). hnRNPs are a family of RNA-binding proteins with functions including nucleic acid metabolism (transcription, 5’ capping and polyadenylation), nascent transcript packaging, alternative splicing and translation regulation (Ping Han *et al*., 2010). 80 ribosomal proteins form the eukaryotic ribosome, the translation machinery required for protein synthesis from mRNA (X. Zhou *et al*., 2015). Ribosome biogenesis and protein translation are both key processes that are finely tuned with and vital for cell growth, proliferation, differentiation, and development (X. Zhou *et al*., 2015).

Both these protein families are vital for development and encompass processes from nucleic acid metabolism to protein synthesis. This could indicate that both variants are crucial for overall protein expression and potentially link transcription and translation. While there is a lack of papers showing these two protein families interacting with each other, *ZFY* could be the linker between hnRNP RNA metabolism and subsequent ribosomal biogenesis and protein synthesis on the ribosome. This indicates that while *ZFY* is vital in spermatogenesis and male reproduction, *ZFY* might also be crucial to male-specific protein expression.

As mentioned in 3.1.1, hnRNPs inhibit alternative splicing, which could indicate a role in the regulation of *ZFY* splicing. Furthermore, both hnRNPs and ribosome proteins have been linked to the cancer-immune landscape (Pecoraro *et al*., 2021);(Sudhakaran & Doseff, 2023). hnRNPs have been shown to be involved in the

diversity of cancer proteomes through their functions in alternative splicing and translation. Meaning they are capable of promoting the expression of cancer-associated genes through controlling transcription factors, chromatin remodelling or directly binding to DNA (Sudhakaran & Doseff, 2023). Ribosomal biogenesis is a crucial step in cellular regulation, and greater ribosome production has been identified in tumour cells (Pecoraro *et al*., 2021). This increase in ribosome biogenesis leads to alterations in homeostasis which can lead to nucleolar and ribosomal stress (Pecoraro *et al*., 2021). This indicates that the abnormal expression of ZFY outside the testis might contribute to tumorigenesis by altering hnRNP and ribosomal protein functions.

6.7 Suggestions for Future Work

There are many directions that this project could continue in and alterations that could be made to add to the knowledge obtained in this project.

To further delve into the function of the 9aaTAD regions of *ZFY*, cell culture work could be carried out to confirm the function of the potential interaction domains identified in the 9aaTAD screen completed in chapter 2. Does mutagenesis of these domains alter *ZFY* activity? To add to the results from chapter 2/3, further research into understanding the expression of *ZFYS* in marsupials would be required. During germ cell differentiation, is there a shift from short to long isoforms, as seen in placental mammals? Finally, from chapter 3, a potential novel chicken transcript was identified, so further investigations into the confirmation of this transcript would be necessary – is it really an alternative “low-function” *ZFY* lacking an acidic domain?

Chapter 3 showed preliminary evidence for *RBMY* being the splicing regulator of *ZFY*, however, to consolidate this, a more physiological system is needed to examine the effects of *RBMY* on endogenously expressed *ZFY*. For example, using NIKS or another male cell line which natively expresses *ZFY*, transfection with *RBMY* could be used to monitor changes in *ZFY* splicing. Another experiment worth pursuing would be to confirm and understand the potential regulatory loop, specifically confirming whether *ZFY* regulates expression RBMX/*RBMY*/*RBMXL2* from their endogenous promoters.

In this work, HEK293 cells were selected for their ease of growth and maintenance. Additionally, being female cells, HEK293 lack endogenous *ZFY* expression, ensuring observed changes resulted from exogenous *ZFY*. However, later cancer correlation analysis revealed numerous Y-linked genes correlating with *ZFY* expression. As HEK293 cells do not express these Y-linked targets, this was not further explored. Males and females differ substantially, and as a Y-linked gene, *ZFY* likely plays male-

specific roles unobservable in HEK293 cells. While convenient, using a female cell line provided an incomplete picture of *ZFY* functionality. Additionally, HEK293 cells are embryonic, which distinguishes them from other somatic cell lines. Being embryonic, they are more proliferative and exhibit greater plasticity, allowing them to adapt to various experimental conditions and potentially express a broader range of genes. While this makes HEK293 cells valuable for general research, their embryonic nature may limit their ability to accurately represent the behaviour of somatic cells in more specialised or tissue-specific studies. Future studies should explore male cell lines to elucidate additional male-specific *ZFY* activities. Specifically for testis function, ideally a germ cell could be utilised for further investigation. Spermatogonia are transfectable cells, whilst spermatids are not but spermatid expression data from knockdown mice does exist. For more cancer related functions, the two main areas of focus would be to focus on squamous epithelium/skin cells by using NIKS and then to also further look into what is happening in lymphocytes with/without the loss of the Y chromosome. In future experiments, the overexpression of ZFY via transfection into HEK293 cells will need adjustment, as its current level is approximately 10,000 times higher than other genes and thus diminishes its clinical relevance and complicates comparisons with other cell lines.

Due to *ZFY* and *ZFX*'s high sequence similarity (>90%), it results in frequent cross-reactivity and hinders conclusive findings. Gelfand *et al* described the universal complications of homologous X and Y cross-reactivity, citing *ZFY* immunopositivity in female breast cancer samples on the Human Protein Atlas (Gelfand & Ambati, 2023). They explained that many protein-based resources cannot differentiate X and Y chromosomes, incorrectly reporting positive immunoreactivity in female cells and tissues. Overcoming this limitation will require either rigorously validated reagents capable of unambiguously distinguishing between these highly related gene pairs (Gelfand & Ambati, 2023), or selective epitope tagging of the endogenous copies of *ZFX* and *ZFY*. Incorporating a tag into the constitutive vs alternatively spliced exons would also help resolve differential functionality of *ZFYL* vs *ZFYS*.

It should be noted that the discussions in this thesis are mainly based on GO enrichment analysis, a statistical association, which can lead to the overinterpretation of results (Gaude & Dessimoz, 2016);(Reimand et al., 2019). One key issue is its reliance on existing annotations, which can be incomplete or biased toward well-studied genes, leading to an overrepresentation of certain pathways. Additionally, GO terms can be broad or overlapping, making it difficult to pinpoint specific biological processes. The statistical methods used often assume independence between genes, which is not always accurate in biological systems. Furthermore, enrichment

analyses are typically based on predefined gene sets and do not account for the dynamic nature of gene regulation across different conditions. These limitations can lead to both false positives and an incomplete understanding of the biological context (Gaude & Dessimoz, 2016);(Reimand et al., 2019).

To explore the identified genes from GO terms and determine their potential role, techniques such as gene knockouts and CRISPR-Cas9 gene editing could be used to assess the gene's effect on specific biological processes in the presence and absence of *ZFY*, i.e. using a cell line that endogenously expresses *ZFY* and a cell line that does not endogenously express *ZFY.*

Originally, Chapter 5 set out to express and purify *ZFY* in an *E. coli* system to subsequently examine its cofactors in a testis lysate. However, due to challenges related to yield and purity, this objective could not be fulfilled. An alternative method using HEK293 cells was used but this setback highlights the limitations associated with utilising a female cell line lysate, which fails to capture the male-specific functions of *ZFY* effectively. If time was not a limiting factor further optimisations to recombinant *ZFY* protein expression could have been made. For example, while *E. coli* is a fast, inexpensive and robust system there are some disadvantages in their use for recombinant protein expression (Francis & Page, 2010). *E. coli* can often result in unfolded/misfolded proteins and cannot perform some post-translational modifications which can lead to the insoluble expression of some proteins (Francis & Page, 2010). One potential alternative expression system could be yeast, which might improve *ZFY* expression, however, a similar toxic effect could be seen. Common suggestions for increasing unstable protein expression in *E. coli* include low-temperature induction, chaperone co-expression and vector choice. Lower induction temperatures have been shown to improve the production of folded, soluble proteins (Francis & Page, 2010). Proteins that pose potential toxicity to *E. coli* growth will be marked for degradation. In such instances, coexpression with a partner protein might alleviate this issue. *ZFY* could indeed be toxic to the cells; however, confirming *ZFY*'s partner proteins would be necessary to address this concern. Further changes to IPTG concentration and media composition could be made to aid in recombinant protein expression. Furthermore, these further changes would hopefully allow for further NMR analysis of *ZFY* and its structure. As in this thesis, we only managed to explore 1D NMR results at extremely low protein concentrations.

Ni *et al* performed ChIP-Seq analysis in the 22Rv1 cell line, a prostate cancer cell line. The publicly available ChIP-Seq data (GSE145160) included the analysis of *ZFX*, *ZNF711*, and *ZFY* in 22Rv1, with control analysis performed in HEK293T cells but only for *ZFX* and *ZNF711*. With more time, the addition of a *ZFY* control would be

optimal for a more in-depth ChIP-Seq analysis of *ZFY*, to identify potential changes in targets in a cancerous vs non-cancerous cell line. By combining ChIP-Seq, RNA-Seq and proteomics datasets direct targets of *ZFY* could be distinguished. ChIP-Seq data could be used in tandem with RNA-Seq to confirm YBX1/2/3 interactions with *ZFY* to validate the proteomics. Furthermore, utilising their single and double knockout methodology of *ZFX* and *ZNF711* paired with the over-expression of *ZFY* in HEK293 cells, would allow for a more focused look at *ZFY* by completely removing any of the *ZFY*/*ZFX*/*ZNF711* shared binding targets. This would also make the over-expression experiments more clinically relevant by altering the amount of exogenous *ZFY* incorporated into the cells, as we previously mentioned problems with extremely high *ZFY* TPM values making the results less clinically relevant. By making these alterations we would hopefully see a reduction in transcriptomic changes but also refine these changes to make downstream analysis more cohesive.

To delve deeper into exploring the potential oncogenic effects of *ZFY*, it would be crucial to employ a more extensive panel of cancer cell lines to address additional questions. However, due to time constraints and limited availability of cell lines, only a limited number of cell lines that met the criteria were accessible for this thesis' timeframe.

Finally, another key experiment would focus on the point that *ZFYS* and *ZFYL* will both compete for the same binding sites in the genome. Using an *in vivo* methodology, it would be interesting to interpret the physiological consequences of *ZFYS* not only in terms of how it directly regulates genes but also how it interacts with and potentially reduces the activity of *ZFYL* in the same cells.

6.8 Final Conclusions

This thesis aimed to gain an understanding of the significance of *ZFY*, potentially explaining why it has continued to persist on the slowly degrading Y chromosome. A key aim was to gain insight into the evolution of *ZFY* as it evolved from an autosomal gene to a sex chromosome gene at the placental mammal divergence. Using a wide range of animal species, this thesis showed that *ZFY* is under negative selection, with novel findings showing the conservation of the 9aaTAD regions within the AAD. However, the rodent species seem to be rapidly diverging away from other placental mammals, with greater genetic changes potentially influencing the role of *ZFY*. Moreover, investigations into the splicing variation in autosomal Zf* species, showed novel similarities between the autosomal and sex chromosome *ZFY* splicing expression within the testis. A potential novel transcript was also observed in chicken testis, linking novel exons to the terminal DBD exon. While this thesis set out to

confirm *RBMY* as the testis-specific splice factor responsible for generating the testis-specific short form, we cannot definitively confirm *RBMY* as the cause of exon skipping in this thesis, however, the preliminary results are promising and require further investigation. However, by combining this data with the RNA-Seq data collected, a potential novel negative feedback loop regulating *ZFY* splicing has been suggested.

To gain insight into the downstream targets of *ZFYS* and *ZFYL*, as well as their specific protein partners, transcriptomics and proteomics were performed. However, due to the plethora of interactors and transcriptome changes, pinpointing the direct role of both *ZFY* variants was difficult, however, here some possible novel downstream pathways regulated by *ZFYL* and *ZFYS* have been identified including *WNT* signalling, ErbB signalling and extracellular matrix remodelling. Proteomics looking for protein interactors found unexpected novel findings. Interactions with other transcription factors were expected, but interactions with nucleolar and ribosomal components were identified instead. Finally, while in this thesis *ZFYS* cannot be confidently confirmed as a cancer-testis gene, there is some preliminary evidence to suggest a potential link to tumorigenesis through ECM remodelling and ErbB signalling, but these processes would need to be further broken down and investigated.

Overall, while *ZFYS*' research interest diminished, this gene should attract renewed attention due to its biological potential, particularly with links to spermatogenesis, meiosis quality control, fertility and as a possible cancer-testis gene.

7. References

Abbott, J. K., Nordén, A. K., & Hansson, B. (2017). Sex chromosome evolution: historical insights and future perspectives. *Proceedings: Biological Sciences*, *284*(1854), 20162806. https://doi.org/10.1098/rspb.2016.2806

Ackers, I., & Malgor, R. (2018). Interrelationship of canonical and non-canonical *WNT* signalling pathways in chronic metabolic diseases. *Diabetes and Vascular Disease Research*, *15*(1), 3–13. https://doi.org/10.1177/1479164117738442

Adeola, F. (2018). Normalization of Gene Expression by Quantitative RT-PCR in Human Cell Line: comparison of 12 Endogenous Reference Genes. *Ethiopian Journal of Health Sciences*, *28*(6), 741. https://doi.org/10.4314/EJHS.V28I6.9

Agrawal, A. A., Conner, J. K., & Rasmann, S. (2010). Tradeoffs and negative correlations in evolutionary ecology. In *Evolution since Darwin: the first 150 years* (pp. 243–268).

Ågren, J. A., Munasinghe, M., & Clark, A. G. (2020). Mitochondrial-Y chromosome epistasis in Drosophila melanogaster. *Proceedings of The Royal Society B*, *287*(1937), 20200469. https://doi.org/10.1098/rspb.2020.0469

Ai, H., Yang, H., Li, L., Ma, J., Liu, K., & Li, Z. (2023). Cancer/testis antigens: promising immunotherapy targets for digestive tract cancers. *Frontiers in Immunology*, *14*. https://doi.org/10.3389/fimmu.2023.1190883

Akhade, V. S., Dighe, S. N., Kataruka, S., & Rao, M. R. S. (2016). Mechanism of *WNT* signaling induced down regulation of mrhl long non-coding RNA in mouse spermatogonial cells. *Nucleic Acids Research*, *44*(1), 387–401. https://doi.org/10.1093/nar/gkv1023

Akhtar Ali, M., Younis, S., Wallerman, O., & Sjöblom, T. (2015). Transcriptional modulator *ZBED6* affects cell cycle and growth of human colorectal cancer cells. *PNAS*, *112*(25), 7743–7748. https://doi.org/10.1073/pnas.1509193112

Al-Amrani, S., Al-Jabri, Z., Al-Zaabi, A., Alshekaili, J., & Al-Khabori, M. (2021). Proteomics: Concepts and applications in human medicine. *World Journal of Biological Chemistry*, *12*(5), 57–69. https://doi.org/10.4331/wjbc.v12.i5.57

Alavattam, K. G., Maezawa, S., Paul, ·, Andreassen, R., Satoshi, ·, & Namekawa, H. (2022). Meiotic sex chromosome inactivation and the XY body: a phase separation hypothesis. *Cellular and Molecular Life Sciences*, *79*, 18. https://doi.org/10.1007/s00018-021-04075-3

Aldana, A., Carneiro, J., Martínez-Mekler, G., & Darszon, A. (2021). Discrete Dynamic Model of the Mammalian Sperm Acrosome Reaction: The Influence of Acrosomal pH and Physiological Heterogeneity. *Frontiers in Physiology*, *12*, 682790. https://doi.org/10.3389/fphys.2021.682790

Amaral, A., Castillo, J., Ramalho-Santos, J., & Oliva, R. (2014). The combined human

sperm proteome: cellular pathways and implications for basic and clinical science. *HUman Reproduction Update*, *20*(1), 40–62. https://doi.org/10.1093/humupd/dmt046

Anisimova, M., & Liberles, D. A. (2012). Detecting and understanding natural selection. In G. M. Cannarozzi & A. Schneider (Eds.), *Codon Evolution: Mechanisms and Models* (pp. 73–96). OUP Oxford.

Antonio Pé rez-Mancera, P., Bermejo-Rodríguez, C., Gonzá lez-Herrero, I., Herranz, M., Flores, T., Jimé nez, R., & Sá nchez-García, I. (2007). Adipose tissue mass is modulated by SLUG (SNAI2). *Human Molecular Genetics*, *16*(23), 2972–2986. https://doi.org/10.1093/hmg/ddm278

Appert-Collin, A., Hubert, P., Crémel, G., & Bennasroune, A. (2015). Role of ErbB Receptors in Cancer Cell Migration and Invasion. *Frontiers in Pharmacology*, *6*, 283. https://doi.org/10.3389/fphar.2015.00283

Arteaga, C. L. (2011). ERBB receptors in cancer: signaling from the inside. *Breast Cancer Research*, *13*, 304. https://doi.org/10.1186/bcr2829

Asgari, R., Mansouri, K., Abdolmaleki, A., & Bakhiari, M. (2021). Association of matrix metalloproteinases with male reproductive functions; with focus on MMP2, 7, and 9. *Meta Gene*, *29*, 100906. https://doi.org/10.1016/j.mgene.2021.100906

Atton, G., Gordon, K., Brice, G., Keeley, V., Riches, K., Ostergaard, P., Mortimer, P., & Mansour, S. (2015). The lymphatic phenotype in Turner syndrome: an evaluation of nineteen patients and literature review. *European Journal of Human Genetics*, *23*(12), 1634–1639. https://doi.org/10.1038/ejhg.2015.41

Bachtrog, D. (2013). Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature Reviews Genetics*, *14*, 113–124. https://doi.org/10.1038/nrg3366

Bachtrog, D., Kirkpatrick, M., Mank, J. E., Mcdaniel, S. F., Pires, J. C., Rice, W. R., & Valenzuela, N. (2011). Are all sex chromosomes created equal? *Trends in Genetics*, *27*(9), 350–357. https://doi.org/10.1016/j.tig.2011.05.005

Bachtrog, D., Mank, J. E., Peichel, C. L., Kirkpatrick, M., Otto, S. P., Ashman, T.-L., Hahn, M. W., Kitano, J., Mayrose, I., Ming, R., Perrin, N., Ross, L., Valenzuela, N., & Vamosi, J. C. (2014). Sex Determination: Why So Many Ways of Doing It? *PLoS Biology*, *12*(7), e1001899. https://doi.org/10.1371/journal.pbio.1001899

Bailey, S. F., Alonso Morales, L. A., & Kassen, R. (2021). Effects of Synonymous Mutations beyond Codon Bias: The Evidence for Adaptive Synonymous Substitutions from Microbial Evolution Experiments. *Genome Biology and Evolution*, *13*(9), evab141. https://doi.org/10.1093/gbe/evab141

Baralle, D., & Baralle, D. (2005). Splicing in action: assessing disease causing sequence changes. *Journal of Medical Genetics*, *42*, 737–748.

https://doi.org/10.1136/jmg.2004.029538

Barasc, H., Mary, N., Letron, R., Calgaro, A., Dudez, A., Bonnet, N., Lihbib-Mansais, Y., Yerle, M., Ducos, A., & Pinton, A. (2012). Y-Autosome Translocation Interferes with Meiotic Sex Inactivation and Expression of Autosomal Genes: A Case Study in the Pig. *Sexual Development*, *6*(1–3), 143–150. https://doi.org/10.1159/000331477

Barfield, J., Yeung, C., & Cooper, T. (2005). Characterization of potassium channels involved in volume regulation of human spermatozoa. *Molecular Human Reproduction*, *11*(12), 891–897. https://doi.org/10.1093/molehr/gah208

Basu, S., Arya, S. P., Usmani, A., Pradhan, B. S., Sarkar, R. K., Ganguli, N., Shukla, M., Mandal, K., Singh, S., Sarda, K., & Majumdar, S. S. (2017). Defective *WNT*3 expression by testicular Sertoli cells compromise male fertility. *Cell and Tissue Research*, *371*, 351–363. https://doi.org/10.1007/s00441-017-2698-5

Beggs, J. D. (1978). Transformation of yeast by a replicating hybrid plasmid. *Nature*, *275*, 104–109.

Bellefroid, E. J., Poncelet, D. A., Lecocq, P. J., Revelant, O., & Martial, J. A. (1991). The evolutionarily conserved Kruppel-associated box domain defines a subfamily of eukaryotic multifingered proteins (DNA-binding proteins/sequence conservation/ceil differentiation). *Proceedings of the National Academy of Sciences of the United States of America*, *88*, 3608–3612. https://doi.org/10.1073/pnas.88.9.3608

Berruti, G., & Paiardi, C. (2011). Acrosome biogenesis. *Spermatogenesis*, *1*(2), 95–98. https://doi.org/10.4161/SPMG.1.2.16820

Bi, M., Wassler, M. J., & Hardy, D. M. (2002). Sperm Adhesion to the Extracellular Matrix of the Egg. In *Fertilization* (pp. 153–180). https://doi.org/10.1016/B978-012311629-1/50007-3

Bidon, T., Schreck, N., Hailer, F., Nilsson, M. A., & Janke, A. (2015). Genome-Wide Search Identifies 1.9 Mb from the Polar Bear Y Chromosome for Evolutionary Analyses. *Genome Biology and Evolution*, *7*(7), 2010–2022. https://doi.org/10.1093/gbe/evv103

Birtle, Z., Ponting, C. P., & Bateman, A. (2006). BIOINFORMATICS DISCOVERY NOTE Meisetz and the birth of the KRAB motif. *Bioinformatics*, *22*(23), 2841–2845. https://doi.org/10.1093/bioinformatics/btl498

Blackmon, H., & Demuth, J. P. (2015). The fragile Y hypothesis: Y chromosome aneuploidy as a selective pressure in sex chromosome and meiotic mechanism evolution. *BioEssays*, *37*(9), 942–950. https://doi.org/10.1002/bies.201500040

Boija, A., Klein, I. A., Sabari, B. R., Dall'Agnese, A., Coffey, E. L., Zamudio, A. V., Li, C. H., Shrinivas, K., Manteiga, J. C., Hannett, N. M., Abraham, B. J., Afeyan, L. K., Guo, Y. E., Rimel, J. K., Fant, C. B., Schuijers, J., Lee, T. I., Taatjes, D. J., & Young, R. A.

(2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell*, *175*(7), 1842–1855. https://doi.org/10.1016/j.cell.2018.10.042

Bornhorst, J. A., & Falke, J. J. (2000). Purification of Proteins Using Polyhistidine Affinity Tags. *Methods in Enzymology*, *326*, 245–254. https://doi.org/10.1016/s0076-6879(00)26058-8

Borsani, G., Tonlorenzi, R., Simmler, M. C., Dandolo, L., Arnaud, D., Capra, V., Grompe, M., Pizzuti, A., Muzny, D., Lawrence, C., Willard, H. F., Avner, P., & Ballabio, A. (1991). Characterization of a murine gene expressed from the inactive X chromosome. *Nature*, *351*, 325–329. https://doi.org/10.1038/351325a0

Briand, L., Marcion, G., Kriznik, A., Heydel, J., Artur, Y., Garrido, C., Seigneuric, R., & Neiers, F. (2016). A self-inducible heterologous protein expression system in Escherichia coli. *Scientific Reports*, *6*, 33037. https://doi.org/10.1038/srep33037

Brockdorff, N., Ashworth, A., Kay, G. F., Cooper, P., Smith, S., McCabe, V. M., Norris, D. P., Penny, G. D., Patel, D., & Rastan, S. (1991). Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. *Nature*, *351*, 329–331. https://doi.org/10.1038/351329a0

Bromham, L. (2009). Why do species vary in their rate of molecular evolution? *Biology Letters*, *5*(3), 401–404. https://doi.org/10.1098/rsbl.2009.0136

Brriton-Jones, C., & Haines, C. (2000). Microdeletions on the long arm of the Y chromosome and their association with male-factor infertility. *Hong Kong Medical Journal*, *6*(2), 184–189.

Brucker, C., & Lipford, G. (1995). The human sperm acrosome reaction: physiology and regulatory mechanisms. An update. *Human Reproduction Update*, *1*(1), 51–62. https://doi.org/10.1093/humupd/1.1.51

Brymora, A., Valova, V. A., & Robinson, P. J. (2004). Protein-Protein Interactions Identified by Pull-Down Experiments and Mass Spectrometry. *Current Protocols in Cell Biology*, *22*(1), 17.5.1-17.5.51. https://doi.org/10.1002/0471143030.cb1705s22

Cabral-Pacheco, G. A., Garza-Veloz, I., Rosa, C. C.-D. la, Ramirez-Acuña, J. M., Perez-Romero, B. A., Guerrero-Rodriguez, J. F., Martinez-Avila, N., & Martinez-Fierro, M. L. (2020). The Roles of Matrix Metalloproteinases and Their Inhibitors in Human Diseases. *International Journal of Molecular Sciences*, *21*(24), 9739. https://doi.org/10.3390/ijms21249739

Cancer Research. (2024). *Leukaemia (all subtypes combined) incidence statistics*. Cancer Research UK. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia/incidence#heading-Zero

Cannarella, R., Condorelli, R. A., Mongioì, L. M., Vignera, S. La, & Calogero, A. E. (2020).

Molecular Sciences Molecular Biology of Spermatogenesis: Novel Targets of Apparently Idiopathic Male Infertility. *International Journal of Molecular Sciences*, *21*(5), 1728. https://doi.org/10.3390/ijms21051728

Capel, B., Rasberry, C., Dyson, J., Bishop, C. E., Simpson, E., Vivian, N., Lovell-Badge, R., Rastan, S., & Cattanach, B. M. (1993). Deletion of Y chromosome sequences located outside the testis determining region can cause XY female sex reversal. *Nature Genetics*, *5*, 301–307. https://doi.org/10.1038/ng1193-301

Cardoso-Moreira, M., Halbert, J., Valloton, D., Velten, B., Chen, C., Shao, Y., Liechti, A., Ascenção, K., Rummel, C., Ovchinnikova, S., Mazin, P. V., Xenarios, I., Harshman, K., Mort, M., Cooper, D. N., Sandi, C., Soares, M. J., Ferreira, P. G., Afonso, S., … Kaessmann, H. (2019). Gene expression across mammalian organ development. *Nature*, *571*(7766), 505. https://doi.org/10.1038/S41586-019-1338-5

Carey, S. B., Aközbek, L., & Harkess, A. (2022). The contributions of Nettie Stevens to the field of sex chromosome biology. *Philosophical Transactions of the Royal Society B*, *377*(1850), 20210215. https://doi.org/10.1098/rstb.2021.0215

Carter, T. D., & Outten, F. W. (2021). Ni-NTA Affinity Chromatography to Characterize Protein–Protein Interactions During Fe-S Cluster Biogenesis. *Methods in Molecular Biology*, *2353*, 125–136. https://doi.org/10.1007/978-1-0716-1605-5_7

Casola, C., Zekonyte, U., Philips, A. D., Cooper, D. N., & Han, M. W. (2012). Interlocus gene conversion events introduce deleterious mutations into at least 1% of human genes associated with inherited disease. *Genome Research*, *22*(3), 429–435. https://doi.org/10.1101/gr.127738.111

Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G., & Raschellà, G. (2017). Zinc-finger proteins in health and disease. *Cell Death Discovery*, *3*, 17071. https://doi.org/10.1038/cddiscovery.2017.71

Chae, W.-J., & Bothwell, A. L. (2018). Canonical and Non-Canonical *WNT* Signaling in Immune Cells. *Trends in Immunology*, *39*(10), 830–847. https://doi.org/10.1016/j.it.2018.08.006

Chao, H.-M., Huang, H.-X., Chang, P.-H., Tseng, K.-C., Miyajima, A., & Chern, E. (2017). Y-box binding protein-1 promotes hepatocellular carcinoma-initiating cell progression and tumorigenesis via *WNT*/β-catenin pathway. *Oncotarget*, *8*(2), 2604–2616. https://doi.org/10.18632/oncotarget.13733

Charlesworth, B. (2003). The organization and evolution of the human Y chromosome. *Genome Biology*, *4*, 226. https://doi.org/10.1186/gb-2003-4-9-226

Chen, J., Lei, C., Zhang, H., Huang, X., Yang, Y., Liu, J., Jia, Y., Shi, H., Zhang, Y., Zhang, J., & Du, J. (2023). RPL11 promotes non-small cell lung cancer cell proliferation by regulating endoplasmic reticulum stress and cell autophagy. *BMC*

Molecular and Cell Biology, *24*(7). https://doi.org/s12860-023-00469-2

Chen, Y., Chen, Z., Tang, Y., & Xiao, Q. (2021). The involvement of noncanonical *WNT* signaling in cancers. *Biomedicine & Pharmacotherapy*, *133*, 110946. https://doi.org/10.1016/j.biopha.2020.110946

Cheng, C. Y., & Mruk, D. D. (2012). The Blood-Testis Barrier and Its Implications for Male Contraception. *Pharmacological Reviews*, *64*(1), 16–64. https://doi.org/10.1124/pr.110.002790

Chuang, J. H., & Li, H. (2004). Functional Bias and Spatial Organization of Genes in Mutational Hot and Cold Regions in the Human Genome. *PLoS Biology*, *2*(2), e29. https://doi.org/10.1371/journal.pbio.0020029

Citri, A., & Yarden, Y. (2006). EGF–ERBB signalling: towards the systems level. *Nature Reviews Molecular Cell Biology*, *7*, 505–516. https://doi.org/10.1038/nrm1962

Cloutier, J. M., & Turner, J. M. A. (2010). Meiotic sex chromosome inactivation. *Current Biology*, *20*(22), R962–R963. https://doi.org/10.1016/j.cub.2010.09.041

Colaco, S., & Modi, D. (2018). Genetics of the human Y chromosome and its association with male infertility. *Reproductive Biology and Endocrinology*, *16*(14). https://doi.org/10.1186/s12958-018-0330-5

Cook, M. B., McGlynn, K. A., Devesa, S. S., Freedman, N. D., & Anderson, W. F. (2012). Sex Disparities in Cancer Mortality and Survival. *Cancer Epidemiology, Biomarkers and Prevention*, *20*(8), 1629–1637. https://doi.org/10.1158/1055-9965.EPI-11-0246

Corda, G., & Sala, A. (2017). Non-canonical *WNT*/PCP signalling in cancer: Fzd6 takes centre stage. *Oncogenesis*, *6*(e364). https://doi.org/10.1038/oncsis.2017.69

Coskun, O. (2016). Separation techniques: Chromatography. *Northern Clinics of Istanbul*, *3*(2), 156–160. https://doi.org/10.14744/nci.2016.32757

Covarrubias, A., Yeste, M., Salazar, E., Ramírez-Reveco, A., Rodriguez Gil, J., & Concha, I. (2015). The *WNT*1 ligand/Frizzled 3 receptor system plays a regulatory role in the achievement of the ‘in vitro’ capacitation and subsequent ‘in vitro’ acrosome exocytosis of porcine spermatozoa. *Andrology*, *3*(2), 357–367. https://doi.org/10.1111/andr.12011

Creasy, D. M., & Chapin, R. E. (2013). Chapter 59 - Male Reproductive System. In *Haschek and Rousseaux’s Handbook of Toxicologic Pathology (Third Edition)* (pp. 2493–2598). https://doi.org/10.1016/B978-0-12-415759-0.00059-5

D’Souza, G., Westra, W. H., Wang, S. J., Van Zante, A., Wentz, A., Kluz, N., Rettig, E., Ryan, W. R., Ha, P. K., Kang, H., Bishop, J., Quon, H., Kiess, A. P., Richmon, J. D., Eisele, D. W., & Fakhry, C. (2017). Differences in the Prevalence of Human Papillomavirus (HPV) in Head and Neck Squamous Cell Cancers by Sex, Race, Anatomic Tumor Site, and HPV Detection Method. *JAMA Oncology*, *3*(3), 169–177.

https://doi.org/10.1001/jamaoncol.2016.3067

Dean, R., Lemos, B., & Dowling, D. (2015). Context-dependent effects of Y chromosome and mitochondrial haplotype on male locomotive activity in Drosophila melanogaster. *Journal of Evolutionary Biology*, *28*(10), 1861–1871. https://doi.org/10.1111/jeb.12702

Decarpentrie, F., Vernet, N., Mahadevaiah, S. K., Longepied, G., Streichemberger, E., Aknin-Seifer, I., Ojarikre, O. A., Burgoyne, P. S., Metzler-Guillemain, C., & Mitchell, M. J. (2012). Human and mouse ZFY genes produce a conserved testis-specific transcript encoding a zinc finger protein with a short acidic domain and modified transactivation potential. *Human Molecular Genetics*, *21*(12), 2631–2645. https://doi.org/10.1093/HMG/DDS088

Delgado-Bermúdez, A., Yeste, M., Bonet, S., & Pinart, E. (2024). Physiological role of potassium channels in mammalian germ cell differentiation, maturation, and capacitation. *Andrology*, 1–18. https://doi.org/10.1111/andr.13606

Desai, M. M., & Fisher, D. S. (2007). Beneficial Mutation–Selection Balance and the Effect of Linkage on Positive Selection. *Genetics*, *176*(3), 1759–1798. https://doi.org/10.1534/genetics.106.067678

Dhar, S., Gursoy-Yuzugllu, O., Parasuram, R., & Price, B. D. (2017). The tale of a tail: histone H4 acetylation and the repair of DNA breaks. *Philosophical Transactions of the Royal Society B*, *372*(20160284). https://doi.org/10.1098/rstb.2016.0284

Dicke, A.-K., Pilatz, A., Wyrwoll, M. J., Punab, M., Ruckert, C., Nagirnaja, L., Aston, K. I., Conrad, D. F., Persio, S. Di, Neuhaus, N., Fietz, D., Laan, M., Stallmeyer, B., & Tüttelmann, F. (2023). DDX3Y is likely the key spermatogenic factor in the AZFa region that contributes to human non-obstructive azoospermia. *Communications Biology*, *6*, 350. https://doi.org/10.1038/s42003-023-04714-4

Ditton, H. J., Zimmer, J., Kamp, C., Meyts, E. R.-D., & Vogt, P. H. (2004). The AZFa gene DBY (DDX3Y) is widely transcribed but the protein is limited to the male germ cells by translation control. *Human Molecular Genetics*, *13*(19), 2333–2341. https://doi.org/10.1093/hmg/ddh240

Dobbs, R. W., Sofer, L., & Ohlander, S. (2018). Y Chromosome Microdeletions. In *Encyclopedia of Reproduction (Second Edition)* (pp. 238–241). https://doi.org/10.1016/B978-0-12-801238-3.64538-5

Dong, Z., & Chen, Y. (2013). Transcriptomics: Advances and approaches. *Science China Life Sciences*, *56*(10), 960–967. https://doi.org/10.1007/s11427-013-4557-2.

Drouin, G., Prat, F., Ell, M., & Clarke, G. D. (1999). Detecting and characterizing gene conversions between multigene family members. *Molecular Biology and Evolution*, *16*(10), 1369–1390. https://doi.org/10.1093/oxfordjournals.molbev.a026047

Ehrmann, I., Crichton, J. H., Gazzara, M. R., James, K., Liu, Y., Grellscheid, S. N., Curk, T., de Rooij, D., Steyn, J. S., Cockell, S., Adams, I. R., Barash, Y., & Elliott, D. J. (2019). An ancient germ cell-specific RNA-binding protein protects the germline from cryptic splice site poisoning. *ELife*, *8*. https://doi.org/10.7554/ELIFE.39304

El-Gamal, M. I., Mewafi, N. H., Abdelmotteleb, N. E., Emara, M. A., Tarazi, H., Sbenati, R. M., Madkour, M. M., Zaraei, S.-O., Shahin, A. I., & Anbar, H. S. (2021). A Review of HER4 (ErbB4) Kinase, Its Impact on Cancer, and Its Inhibitors. *Molecules*, *26*(23), 7376. https://doi.org/10.3390/molecules26237376

Ellis, P. J. I., & Affara, N. A. (2009). Spermatogenesis and sex chromosome gene content: An evolutionary perspective. *Human Fertility*, *9*(1), 1–7. https://doi.org/10.1080/14647270500230114

Emerson, R., & Thomas, J. (2009). Adaptive Evolution in Zinc Finger Transcription. *PLOS Genetics*, *5*(1). https://doi.org/10.1371/journal.pgen.1000325

Engelstädter, J. (2008). Muller's Ratchet and the Degeneration of Y Chromosomes: A Simulation Study. *Genetics*, *180*(2), 957–967. https://doi.org/10.1534/genetics.108.092379

Ercan, S. (2015). Mechanisms of X Chromosome Dosage Compensation. *Journal of Genomics*, *3*, 1–19. https://doi.org/10.7150/jgen.10404

Esteves, S. C. (2015). Clinical management of infertile men with nonobstructive azoospermia. *Asian Journal of Andrology*, *17*(3), 459–470. https://doi.org/10.4103/1008-682X.148719

Ezaz, T., Stiglec, R., Veyrunes, F., & Marshall Graves, J. A. (2006). Relationships between Vertebrate ZW and XY Sex Chromosome Systems. *Current Biology*, *16*, r736–r743. https://doi.org/10.1016/j.cub.2006.08.021

Fekete, S., Beck, A., Veuthey, J.-L., & Guillarme, D. (2015). Ion-exchange chromatography for the characterization of biopharmaceuticals. *Journal of Pharmaceutical and Biomedical Analysis*, *113*, 43–55. https://doi.org/10.1016/j.jpba.2015.02.037

Felinski, E., Kim, J., Lu, J., & Quinn, P. (2001). Recruitment of an RNA Polymerase II Complex Is Mediated by the Constitutive Activation Domain in CREB, Independently of CREB Phosphorylation. *Molecular and Cellular Biology*, *21*(4), 1001–1010. https://doi.org/10.1128/MCB.21.4.1001-1010.2001

Ferreira, M. E., Hermann, S., Prochasson, P., Workman, J. L., Berndt, K. D., & Wright, A. P. (2005). Mechanism of Transcription Factor Recruitment by Acidic Activators. *Journal of Biological Chemistry*, *280*(23), 21779–21784. https://doi.org/10.1074/jbc.M502627200

Flesch, F. M., & Gadella, B. M. (2000). Dynamics of the mammalian sperm plasma

membrane in the process of fertilization. *Biochemica et Biophysica Acta*, *1469*(3), 197–235. https://doi.org/10.1016/S0304-4157(00)00018-6

Foley, G., Mora, A., Ross, C. M., Bottoms, S., Sützl, L., Lamprecht, M. L., Zaugg, J., Essebier, A., Balderson, B., Newell, R., Thomson, R. E. S., Kobe, B., Barnard, R. T., Guddat, L., Schenk, G., Carsten, J., Gumulya, Y., Rost, B., Haltrich, D., … Bodén, M. (2022). Engineering indel and substitution variants of diverse and ancient enzymes using Graphical Representation of Ancestral Sequence Predictions (GRASP). *PLOS Computational Biology*, *18*(10), e1010633. https://doi.org/10.1371/journal.pcbi.1010633

Francis, D. M., & Page, R. (2010). Strategies to Optimize Protein Expression in E. coli. *Current Protocols in Protein Science*, *61*(1), 5241–52429. https://doi.org/10.1002/0471140864.ps0524s61

Frietze, S., & Farnham, P. J. (2011). Transcription Factor Effector Domains. In *A Hanbook of Transcription Factors* (pp. 261–277). https://doi.org/10.1007/978-90-481-9069-0_12

Fu, Y., Wei, Y., Zhou, Y., Wu, H., Hong, Y., Long, C., Wang, J., Wu, Y., Wu, S., Shen, L., & Wei, G. (2021). *WNT*5a Regulates Junctional Function of Sertoli cells Through PCP-mediated Effects on mTORC1 and mTORC2. *Endocrinology*, *162*(10), bqab149. https://doi.org/10.1210/endocr/bqab149

Fumagalli, S., Cara, A. Di, Neb-Gulati, A., Natt, F., Schwemberger, S., Hall, J., Babcock, G. F., Bernardi, R., Pandolfi, P. P., & Thomas, G. (2009). Absence of nucleolar disruption after impairment of 40S ribosome biogenesis reveals an rpL11-translation-dependent mechanism of p53 induction. *Nature Cell Biology*, *11*, 501–508. https://doi.org/10.1038/ncb1858

Funami, K., Matsumoto, M., Enokizono, Y., Ishii, N., Tatematsu, M., Oshiumi, H., Inagak, F., & Seya, T. (2015). Identification of a Regulatory Acidic Motif as the Determinant of Membrane Localization of TICAM-2. *The Jourunal of Immunology*, *195*(9), 4456–4465. https://doi.org/10.4049/jimmunol.1402628

Furlan, G., & Galupa, R. (2022). Mechanisms of Choice in X-Chromosome Inactivation. *Cells*, *11*(3), 535. https://doi.org/10.3390/cells11030535

Furman, B. L. S., & Evans, B. J. (2018). Divergent Evolutionary Trajectories of Two Young, Homomorphic, and Closely Related Sex Chromosome Systems. *Genome Biology and Evolution*, *10*(3), 742–755. https://doi.org/10.1093/gbe/evy045

Furman, B. L. S., Metzger, D. C. H., Darolti, I., Wright, A. E., Sandkam, B. A., Almeida, P., Shu, J. J., & Mank, J. E. (2020). Sex Chromosome Evolution: So Many Exceptions to the Rules. *Genome Biology and Evolution*, *12*(6), 750–763. https://doi.org/10.1093/gbe/evaa081.

Gashti, N. G., Gilani, M. A. S., & Abbasi, M. (2021). Sertoli cell-only syndrome: etiology and clinical management. *Journal of Assisted Reproduction and Genetics*, *38*(3), 559–572. https://doi.org/10.1007/s10815-021-02063-x

Gaude, P., & Dessimoz, C. (2016). Gene Ontology: Pitfalls, Biases, and Remedies. *The Gene Ontology Handbook*, *1446*, 189–205. https://doi.org/10.1007/978-1-4939-3743-1_14

Gelfand, B. D., & Ambati, J. (2023). Y chromosome proteins in female tissues. *Science*, *382*(6666), 39–40. https://doi.org/10.1126/science.ade7187

Geuens, T., Bouhy, D., & Timmerman, V. (2016). The hnRNP family: insights into their role in health and disease. *Human Genetics*, *135*, 851–867. https://doi.org/10.1007/s00439-016-1683-5

Gilbert, S. F. (2000). *Developmental Biology. 6th edition*. Sunderland (MA): Sinauer Associates. https://www.ncbi.nlm.nih.gov/books/NBK10095/

Gonzalez, D. H. (2016). Introduction to Transcription Factor Structure and Function. In *Plant Transcription Factors* (pp. 3–11). Academic Press. https://doi.org/10.1016/B978-0-12-800854-6.00001-4

Gonzalez, H., Hagerling, C., & Werb, Z. (2018). Roles of the immune system in cancer: from tumor initiation to metastatic progression. *Genes and Development*, *32*(19–20), 1267–1284. https://doi.org/10.1101/gad.314617.118

Gormley, M., Creaney, G., Schache, A., Ingarfield, K., & Conway, D. I. (2022). Reviewing the epidemiology of head and neck cancer: definitions, trends and risk factors. *British Dental Journal*, *233*, 780–786. https://doi.org/10.1038/s41415-022-5166-x

Griswold, M. D. (2016). Spermatogenesis: The Commitment to Meiosis. *Physiol Rev*, *96*, 1–17. https://doi.org/10.1152/physrev.00013.2015.-Mam

Grover, A., Pande, A., Choudhary, K., Gupta, K., & Sundar, D. (2010). Re-programming DNA-binding specificity in zinc finger proteins for targeting unique address in a genome. *Systems and Synthetic Biology*, *4*(4), 323–329. https://doi.org/10.1007/s11693-011-9077-4

Gruber, M., Mathew, L. K., Runge, A. C., Garcia, J. A., & Simon, M. C. (2010). EPAS1 Is Required for Spermatogenesis in the Postnatal Mouse Testis. *Biology of Reproduction*, *82*(6), 1227–1236. https://doi.org/10.1095/biolreprod.109.079202

Guan, Y., Zhu, Q., Huang, D., Zhao, S., Lo, L. J., & Peng, J. (2015). An equation to estimate the difference between theoretically predicted and SDS PAGE-displayed molecular weights for an acidic peptide. *Scientific Reports*, *5*, 13370. https://doi.org/10.1038/srep13370

Gueler, B., Sonne, S. B., Zimmer, J., Hilscher, B., Hilscher, W., Græm, N., Meyts, E. R.-D., & Vogt, P. H. (2012). AZFa protein DDX3Y is differentially expressed in human

male germ cells during development and in testicular tumours: new evidence for phenotypic plasticity of germ cells. *Human Reproduction*, *27*(6), 1547–1555. https://doi.org/10.1093/humrep/des047

Gupta, S. K. (2021). Human Zona Pellucida Glycoproteins: Binding Characteristics With Human Spermatozoa and Induction of Acrosome Reaction. *Frontiers in Cell and Development Biology*, *9*, 619868. https://doi.org/10.3389/fcell.2021.619868

Hake, L., & O'Connor, C. (2008). Genetic Mechanisms of Sex Determination. *Nature Education*, *1*(1), 25.

Hayashida, H., Kuma, K., & Miyata, T. (1992). Interchromosomal Gene Conversion as a Possible Mechanism for Explaining Divergence Patterns of ZFY-Related Genes. *Journal of Molecular Evolution*, *35*, 181–183. https://doi.org/10.1007/BF00183228

Henke, E., Nandigama, R., & Ergün, S. (2020). Extracellular Matrix in the Tumor Microenvironment and Its Impact on Cancer Therapy. *Frontiers in Molecular Biosciences*, *6*, 160. https://doi.org/10.3389/fmolb.2019.00160

Hermann, B. P., Cheng, K., Singh, A., Cruz, L. R. La, Mutoji, K. N., Chen, I., Gildersleeve, H., Lehle, J. D., Mayo, M., Law, N. C., Oatley, M. J., Velte, E. K., Bryan, A., Fritze, D., Silber, S., Geyer, C. B., Oatley, J. M., & Mccarrey, J. R. (2018). The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids. *Cell Rep*, *25*(6), 1650–1667. https://doi.org/10.1016/j.celrep.2018.10.026.

Hirata, S., Hoshi, K., Shoda, T., & Mabuchi, T. (2002). Spermatozoon and mitochondrial DNA. *Reproductive Medicine and Biology*, *1*(2), 41–47. https://doi.org/10.1046/j.1445-5781.2002.00007.x

Hoang, D. T., Chernomor, O., Haeseler, A. von, Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hodge, K., Have, S. Ten, Hutton, L., & Lamond, A. I. (2013). Cleaning up the masses: Exclusion lists to reduce contamination with HPLC-MS/MS. *Journal of Proteomics*, *88*, 92–103. https://doi.org/10.1016/j.jprot.2013.02.023

Holmes, R., Keating, M., Cork, A., Trujillo, J., McCredie, K., & Freireich, E. (1985). Loss of the Y chromosome in acute myelogenous leukemia: a report of 13 patients. *Cancer Genetics and Cytogenetics*, *17*(3), 269–278. https://doi.org/10.1016/0165-4608(85)90018-4

Holmlund, H., Yamauchi, Y., Ruthig, V. A., Cocquet, J., & Ward, M. A. (2023). Return of the forgotten hero: the role of Y chromosome-encoded Zfy in male reproduction. *Molecular Human Reproduction*, *29*(8), gaad025. https://doi.org/10.1093/molehr/gaad025.

Homer, C., Knight, D. A., Hananeia, L., Sheard, P., Risk, J., Lasham, A., Royds, J. A., & Braithwaite, A. W. (2005). Y-box factor YB1 controls p53 apoptotic function. *Oncogene*, *24*, 8314–8325. https://doi.org/10.1038/sj.onc.1208998

Hughes, A. L., Friedman, R., Rivailler, P., & French, J. O. (2008). Synonymous and Nonsynonymous Polymorphisms versus Divergences in Bacterial Genomes. *Molecular Biology and Evolution*, *25*(10), 2199–2209. https://doi.org/10.1093/molbev/msn166

Hughes, J. F., & Page, D. C. (2015). The Biology and Evolution of Mammalian Y Chromosomes. *Annual Review of Genetics*, *49*, 507–527. https://doi.org/10.1146/annurev-genet-112414-055311

Huson, D. H., & Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution*, *23*(2), 254–267. https://doi.org/10.1093/molbev/msj030

Hynes, N. E., & MacDonald, G. (2009). ErbB receptors and signaling pathways in cancer. *Current Opinion in Cell Biology*, *21*(2), 177–184. https://doi.org/10.1016/j.ceb.2008.12.010

Isernia, C., Malgieri, G., Russo, L., D'Abrosca, G., Baglivo, I., Pedone, P. V, & Fattorusso, R. (2020). Zinc Fingers. *Metal Ions in Life Science*, *23*(20). https://doi.org/10.1515/9783110589757-018

Iuchi, S. (2001). Three classes of C 2 H 2 zinc finger proteins. *Cellular and Molecular Life Sciences CMLS*, *58*, 625–635. https://doi.org/10.1007/PL00000885

Jackson, S. S., Marks, M. A., Katki, H., Cook, M. B., Hyun, N., Freedman, N. D., Kahle, L. L., Castle, P. E., Graubard, B. I., & Chaturvedi, A. K. (2022). Sex disparities in the incidence of 21 cancer types: Quantification of the contribution of risk factors. *Cancer*, *128*(19), 3531–3540. https://doi.org/10.1002/cncr.34390

Jacobi, N., Seeboeck, R., Hofmann, E., & Eger, A. (2017). ErbB Family Signalling: A Paradigm for Oncogene Addiction and Personalized Oncology. *Cancers*, *9*(4), 33. https://doi.org/10.3390/cancers9040033

Janečka, J. E., Davis, B. W., Ghosh, S., Paria, N., Das, P. J., Orlando, L., Schubert, M., Nielsen, M. K., Stout, T. A. E., Brashear, W., Li, G., Johnson, C. D., Metz, R. P., Zadjali, A. M. Al, Love, C. C., Varner, D. D., Bellott, D. W., Murphy, W. J., Chowdhary, B. P., & Raudsepp, T. (2018). Horse Y chromosome assembly displays unique evolutionary features and putative stallion fertility genes. *Nature Communications*, *9*, 2945. https://doi.org/10.1038/s41467-018-05290-6

Jay, A., Reitz, D., Namekawa, S. H., & Heyer, W.-D. (2021). Cancer testis antigens and genomic instability: More than immunology. *DNA Repair*, *108*, 103214. https://doi.org/10.1016/j.dnarep.2021.103214

Jaya, F. R., Brito, B. P., & Darling, A. E. (2023). Evaluation of recombination detection methods for viral sequencing. *Virus Evolution*, *9*(2), vead066. https://doi.org/10.1093/ve/vead066

Jeays-Ward, K., Dandonneau, M., & Swain, A. (2024). *WNT*4 is required for proper male as well as female sexual development. *Developmental Biology*, *276*, 431–440. https://doi.org/10.1016/j.ydbio.2004.08.049

Jenkins, M. C., Parker, C., Brien, C., Campos, P., Tucker, M., & Miska, K. (2023). Effects of codon optimization on expression in Escherichia coli of protein-coding DNA sequences from the protozoan Eimeria. *Journal of Microbiological Methods*, *211*, 106750. https://doi.org/10.1016/j.mimet.2023.106750

Jiang, W., & Chen, L. (2021). Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing. *Computational and Structural Biotechnology Journal*, *19*, 183–195. https://doi.org/10.1016/j.csbj.2020.12.009

Jiao, L., Liu, Y., Yu, X.-Y., Pan, X., Zhang, Y., Tu, J., Song, Y.-H., & Li, Y. (2023). Ribosome biogenesis in disease: new players and therapeutic targets. *Signal Transfuction and Targeted Therapy*, *8*(15). https://doi.org/10.1038/s41392-022-01285-4

Jin, S.-K., & Yang, W.-X. (2017). Factors and pathways involved in capacitation: how are they regulated? *Oncotarget*, *8*(2), 3600–3627. https://doi.org/10.18632/oncotarget.12274

Johnson, D., Burtness, B., Leemans, C. R., Wai Yan Lui, V., Bauman, J., & Grandis, J. (2020). Head and neck squamous cell carcinoma. *Nat Rev Dis Primers*, *6*, 92. https://doi.org/https://doi.org/10.1038/s41572-020-00224-3

Kahrl, A. F., Snook, R. R., & Fitzpatrick, J. L. (2021). Fertilization mode drives sperm length evolution across the animal tree of life. *Nature Ecology and Evolution*, *5*, 1153–1164. https://doi.org/10.1038/s41559-021-01488-y

Kang, J., Brajanovski, N., Chan, K. T., Xuan, J., Pearson, R. B., & Sanij, E. (2021). Ribosomal proteins and human diseases: molecular mechanisms and targeted therapy. *Signal Transfuction and Targeted Therapy*, *6*(323). https://doi.org/10.1038/s41392-021-00728-8

Karlin, S., & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(6), 2264–2268. https://doi.org/10.1073/pnas.87.6.2264

Kayama, K., Watanabe, S., Takafuji, T., Tsuji, T., Hironaka, K., Matsumoto, M., Nakayama, K. I., Enari, M., Kohno, T., Shiraishi, K., Kiyono, T., Yoshida, K., Sugimoto, N., & Fujita, M. (2016). GRWD1 negatively regulates p53 via the RPL11–

MDM2 pathway and promotes tumorigenesis. *EMBO Reports*, *18*, 123–137. https://doi.org/10.15252/embr.201642444

Kent-First, M., Muallem, A., Shultz, J., Pryor, J., Roberts, K., Nolten, W., Meisner, L., Chandley, A., Gouchy, G., Jorgensen, L., Havighurst, T., & Grosch, J. (1999). Defining regions of the Y-chromosome responsible for male infertility and identification of a fourth AZF region (AZFd) by Y-chromosome microdeletion detection. *Molecular Reproduction and Development*, *53*(1), 27–41. https://doi.org/10.1002/(SICI)1098-2795(199905)53:1<27::AID-MRD4>3.0.CO;2-W

Kerr, G. E., Young, J. C., Horvay, K., Abud, H. E., & Loveland, K. L. (2014). Regulated *WNT*/Beta-Catenin Signaling Sustains Adult Spermatogenesis in Mice. *Biology of Reproduction*, *90*(1), 1–12. https://doi.org/10.1095/biolreprod.112.105809

Kestler, H. A., & Kühl, M. (2008). From individual *WNT* pathways towards a *WNT* signalling network. *Philosophical Transactions of the Royal Society B*, *363*, 1333–1347. https://doi.org/10.1098/rstb.2007.2251

Kido, T., & Lau, Y.-F. C. (2015). Roles of the Y chromosome genes in human cancers. *Asian Journal of Andrology*, *17*(3), 373–380. https://doi.org/10.4103/1008-682X.150842

Kim, E., Magen, A., & Ast, G. (2006). Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, *35*(1), 125–131. https://doi.org/10.1093/nar/gkl924

Kim, H.-S., & Takenaka. (2000). Evolution of the X-linked Zinc Finger Gene and the Y-linked Zinc Finger Gene in Primates. *Molecules and Cells*, *10*(5), 512–518. https://doi.org/10.1007/S1016-8478(23)17511-X

Kinene, T., Wainaina, J., Maina, S., & Boykin, L. M. (2016). Rooting Trees, Methods for,. In *Encyclopedia of Evolutionary Biology* (pp. 489–493). https://doi.org/10.1016/B978-0-12-800049-6.00215-8

Kobzeva, K. A., Soldatova, M. O., Stetskaya, T. A., Soldatov, V. O., Deykin, A. V., Freidin, M. B., Bykanova, M. A., Churnosov, M. I., Polonikov, A. V., & Bushueva, O. Y. (2023). Association between HSPA8 Gene Variants and Ischemic Stroke: A Pilot Study Providing Additional Evidence for the Role of Heat Shock Proteins in Disease Pathogenesis. *Genes*, *14*(6), 1171. https://doi.org/10.3390/genes14061171

Koch, S., Acebron, S. P., Herbst, J., Hatiboglu, G., & Niehrs, C. (2015). Post-transcriptional *WNT* Signaling Governs Epididymal Sperm Maturation. *Cell*, *163*, 1225–1236. https://doi.org/10.1016/j.cell.2015.10.029

Komiya, Y., & Habas, R. (2008). *WNT* signal transduction pathways. *Organogenesis*, *4*(2), 68–75. https://doi.org/10.4161/org.4.2.5851

Koni, M., Pinnaro, V., & Felice Brizzi, M. (2020). The *WNT* Signalling Pathway: A Tailored

Target in Cancer. *International Journal of Molecular Sciences*, *21*(20), 7697. https://doi.org/10.3390/ijms21207697

Koopman, P., Ashworth, A., & Lovell-Badge, R. (1991). The ZFY gene family in humans and mice. *Trends in Genetics*, *7*(4), 132–136. https://doi.org/10.1016/0168-9525(91)90458-3.

Kotha, S. R., & Staller, M. V. (2023). Clusters of acidic and hydrophobic residues can predict acidic transcriptional activation domains from protein sequence. *Genetics*, *225*(2). https://doi.org/10.1093/genetics/iyad131

Kotov, A. A., Olenkina, O. M., Godneeva, B. K., Adashev, V. E., & Olenina, L. V. (2017). Progress in understanding the molecular functions of DDX3Y (DBY) in male germ cell development and maintenance. *BioScience Trends*, *11*(1), 46–53. https://doi.org/10.5582/bst.2016.01216

Kretov, D. (2022). Role of Y-Box Binding Proteins in Ontogenesis. *Biochemistry Moscow*, *87*, 71–85. https://doi.org/10.1134/S0006297922140061

Kristofich, J. C., Morgenthaler, A. B., Kinney, W. R., Ebmeier, C. C., Snyder, D. J., Old, W. M., Cooper, V. S., & Copley, S. D. (2018). Synonymous mutations make dramatic contributions to fitness when growth is limited by a weak-link enzyme. *PLOS Genetics*, *14*(8), e1007615. https://doi.org/10.1371/journal.pgen.1007615

Kryazhimskiy, S., & Plotkin, J. B. (2008). The Population Genetics of dN/dS. *PLOS Genetics*, *4*(12), e1000304. https://doi.org/10.1371/journal.pgen.1000304

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbour Protocols*, *2015*(11), 951–969. https://doi.org/10.1101/pdb.top084970

Kumagai, S., Koyama, S., & Nishikawa, H. (2021). Antitumour immunity regulated by aberrant ERBB family signalling. *Nature Reviews Cancer*, *21*, 181–197. https://doi.org/10.1038/s41568-020-00322-0

Kumar, S., & Subramanian, S. (2002). Mutation rates in mammalian genomes. *PNAS*, *99*(2), 803–808. https://doi.org/10.1073/pnas.022629899

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., & Hedges, S. B. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, *39*(8), msac174. https://doi.org/10.1093/molbev/msac174

Lahn, B., & Page, D. (1999). *Four Evolutionary Strata on the Human X Chromosome*. Science. https://doi.org/10.1126/science.286.5441.964.

Lahn, B. T., Pearson, N. M., & Jegalian, K. (2001). The human Y chromosome, in the light of evolution. *Nature Reviews Genetics*, *2*, 207–216. https://doi.org/10.1038/35056058

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X.,

Taipale, J., Hughes, T. R., & Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, *172*, 650–665. https://doi.org/10.1016/j.cell.2018.01.029

Lanfranco, F., Kamischke, A., Zitzmann, M., & Nieschlag, E. (2004). Klinefelter's Syndrome. *The Lancet*, *364*(9430), 273–283. https://doi.org/10.1016/S0140-6736(04)16678-6

Laval, S. H., GLENISTER, P. H., RASBERRY, C., THORNTON, C. E., MAHADEVALAH, S. K., COOKE, H. J., BURGOYNE, P. S., & CATTANACH, B. M. (1995). Y chromosome short arm-Sxr recombination in XSxr/Y males causes deletion of Rbm and XY female sex reversal. *Proceedings of the National Academy of Sciences of the United States of America*, *92*, 10403–10407. https://doi.org/10.1073/pnas.92.22.10403

Layman, C. (2012). Disorders of the Hypothalamic-Pituitary-Gonadal Axis. In *Handbook of Neuroendocrinology* (pp. 659–683). https://doi.org/10.1016/B978-0-12-375097-6.10030-7

Lebeuf-Taylor, E., McCloskey, N., Bailey, S. F., Hinz, A., & Kassen, R. (2019). The distribution of fitness effects among synonymous mutations in a gene under directional selection. *ELife*, *8*, e45952. https://doi.org/10.7554/eLife.45952

Li, D., Wang, T.-W., Aratani, S., Omori, S., Tamatani, M., Johmura, Y., & Nakanishi, M. (2023). Transcriptomic characterization of Lonrf1 at the single-cell level under pathophysiological conditions. *Journal of Biochemistry*, *173*(6), 459–469. https://doi.org/10.1093/jb/mvad021

Li, J., & Ge, Z. (2021). High HSPA8 expression predicts adverse outcomes of acute myeloid leukemia. *BMC Cancer*, *21*(475). https://doi.org/10.1186/s12885-021-08193-w

Li, L., Gao, Y., Chen, H., Jesus, T., Tang, E., Li, N., Lian, Q., Ge, R., & Chenga, C. Y. (2017). Cell polarity, cell adhesion, and spermatogenesis: role of cytoskeletons. *F1000 Research*, *6*, 1565. https://doi.org/10.12688/f1000research.11421.1

Li, X.-F., Ren, P., Shen, W.-Z., Jin, X., & Zhang, J. (2020). The expression, modulation and use of cancer-testis antigens as potential biomarkers for cancer immunotherapy. *American Journal of Translational Research*, *12*(11), 7002–7019.

Li, X., Han, M., Zhang, H., Liu, F., Pan, Y., Zhu, J., Liao, Z., Chen, X., & Zhang, B. (2022). Structures and biological functions of zinc finger proteins and their roles in hepatocellular carcinoma. *Biomarker Research*, *10*(2). https://doi.org/10.1186/s40364-021-00345-1

Lin, R., Zhang, Y., Pradhan, K., & Li, L. (2020). *TICAM2*-related pathway mediates neutrophil exhaustion. *Scientific Reports*, *10*(14397). https://doi.org/10.1038/s41598-020-71379-y

Lin, Y.-C., Boone, M., Meuris, L., Lemmens, I., Roy, N. Van, Soete, A., Reumers, J., Moisse, M., Plaisance, S., Drmanac, R., Chen, J., Speleman, F., Lambrechts, D., Van De Peer, Y., Tavernier, J., & Callewaert, N. (2014). Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nature Communications*, *5*, 4767. https://doi.org/10.1038/ncomms5767

Litscher, E. S., & Wassarman, P. M. (2020). Zona Pellucida Proteins, Fibrils, and Matrix. *Annual Review of Biochemistry*, *20*(89), 695–715. https://doi.org/10.1146/annurev-biochem-011520-105310

Liu, J., Xiao, Q., Xiao, J., Niu, C., Li, Y., Zhang, X., Zhou, Z., Shu, G., & Yin, G. (2022). *WNT*/β-catenin signalling: function, biological mechanisms, and therapeutic opportunities. *Signal Transfuction and Targeted Therapy*, *7*(3). https://doi.org/10.1038/s41392-021-00762-6

Liu, S., Sima, Z., Liu, X., & Chen, H. (2022). Zinc Finger Proteins: Functions and Mechanisms in Colon Cancer. *Cancers*, *14*(21), 5242. https://doi.org/10.3390/cancers14215242

Ljungström, V., Mattisson, J., Halvardson, J., Pandzic, T., Davies, H., Rychlicka-Buniowska, E., Danielsson, M., Lacaze, P., Cavelier, L., Dumanski, J. P., Baliakas, P., & Forsberg, L. A. (2022). Loss of Y and clonal hematopoiesis in blood—two sides of the same coin? *Leukemia*, *36*, 889–891. https://doi.org/10.1038/s41375-021-01456-2

Loda, A., & Heard, E. (2019). Xist RNA in action: Past, present, and future. *PLOS Genetics*, *15*(9), e1008333. https://doi.org/10.1371/journal.pgen.1008333

Louche, A., Salcedo, S. P., & Bigot, S. (2017). Protein–Protein Interactions: Pull-Down Assays. In *Bacterial Protein Secretion Systems* (pp. 247–255). Humana Press. https://doi.org/10.1007/978-1-4939-7033-9_20

Love, M. I., Anders, S., Kim, V., & Huber, W. (2015). RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000 Research*, *4*, 1070. https://doi.org/10.12688/f1000research.7035.1

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*, 550. https://doi.org/10.1186/s13059-014-0550-8

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLOS Computational Biology*, *13*(5), e1005457. https://doi.org/10.1371/journal.pcbi.1005457

Lu, Y., Wang, X., Gu, Q., Wang, J., Sui, Y., Wu, J., & Feng, J. (2022). Heterogeneous nuclear ribonucleoprotein A/B: an emerging group of cancer biomarkers and therapeutic targets. *Cell Death Discovery*, *8*(337). https://doi.org/10.1038/s41420-

022-01129-8

Luaces, J., Toro-Urrego, N., Otero-Losada, M., & Capani, F. (2023). What do we know about blood-testis barrier? current understanding of its structure and physiology. *Frontiers in Cell and Development Biology*, *11*, 1114769. https://doi.org/10.3389/fcell.2023.1114769

Mangs, A. H., & Morris, B. J. (2007). The Human Pseudoautosomal Region (PAR): Origin, Function and Future. *Current Genomics*, *8*, 129–136. https://doi.org/10.2174/138920207780368141

Mansai, S. P., & Innan, H. (2010). The Power of the Methods for Detecting Interlocus Gene Conversion. *Genetics*, *184*(2), 517–527. https://doi.org/10.1534/genetics.109.111161

Manterola, M., Page, J., Vasco, C., Berríos, S., Parra, M. T., Viera, A., Rufas, J. S., Zuxxotti, M., Garagna, S., & Fernández-Donoso, R. (2009). A High Incidence of Meiotic Silencing of Unsynapsed Chromatin Is Not Associated with Substantial Pachytene Loss in Heterozygous Male Mice Carrying Multiple Simple Robertsonian Translocations. *PLOS Genetics*, *5*(8), e1000625. https://doi.org/10.1371/journal.pgen.1000625

Mardon, G., Luoh, S.-W., Simpson, E. M., Gill, G., Brown, L. G., & Page', D. C. (1990). Mouse Zfx protein is similar to Zfy-2: each contains an acidic activating domain and 13 zinc fingers. *Molecular and Cellular Biology*, *10*(2), 681–688. https://doi.org/10.1128/MCB.10.2.681-688.1990

Mastropasqua, F., Oksanen, M., Soldini, C., Alatar, S., Arora, A., Ballarino, R., Molinari, M., Agostini, F., Poulet, A., Watts, M., Rabkina, I., Becker, M., Li, D., Anderlid, B.-M., Isaksson, J., Remnelius, K. L., Moslem, M., Jacob, Y., Falk, A., … Tammimie, K. (2022). Deficiency of Heterogeneous Nuclear Ribonucleoprotein U leads to delayed neurogenesis. *BioRxiv*, 507275. https://doi.org/10.1101/2022.09.14.507275

Matlin, A. J., Clark, F., & Smith, C. W. J. (2005). Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, *6*, 386–398. https://doi.org/10.1038/nrm1645

Matos, B., Publicover, S. J., Castro, L. F. C., Esteves, P. J., & Fardilha, M. (2021). Brain and testis: more alike than previously thought? *Open Biology*, *11*(6), 200322. https://doi.org/10.1098/rsob.200322

McFarlane, R. J., Feichtinger, J., & Larcombe, L. (2014). Cancer germline gene activation. *Cell Cycle*, *13*(14), 2151–2152. https://doi.org/10.4161/cc.29661

McGinty, R. K., & Tan, S. (2015). Nucleosome Structure and Function. *Chemical Reviews*, *115*(6), 2255–2273. https://doi.org/10.1021/cr500373h

Mei, A. H.-C., Tung, K., Han, J., Perumal, D., Laganà, A., Keats, J., Auclair, D., Chari, A.,

Jagannath, S., Parekh, S., & Cho, H. J. (2020). MAGE-A inhibit apoptosis and promote proliferation in multiple myeloma through regulation of BIM and p21Cip1. *Oncotarget*, *11*(7), 727–739. https://doi.org/10.18632/oncotarget.27488

Meizel, S. (2004). The sperm, a neuron with a tail: 'neuronal' receptors in mammalian sperm. *Biological Reviews*, *79*(4), 713–732. https://doi.org/10.1017/S1464793103006407

Melcher, K. (2000). The strength of acidic activation domains correlates with their affinity for both transcriptional and non-transcriptional protiens. *Journal of Molecular Biology*, *301*(5), 1097–1112. https://doi.org/10.1006/JMBI.2000.4034

Mital, P., Hinton, B. T., & Dufour, J. M. (2011). The Blood-Testis and Blood-Epididymis Barriers Are More than Just Their Tight Junctions. *Biology of Reproduction*, *84*(5), 851–858. https://doi.org/10.1095/biolreprod.110.087452

Mitsis, T., Efthimiadou, A., Bacopoulou, F., Vlachakis, D., Chrousos, G., & Eliopoulos, E. (2020). Transcription factors and evolution: An integral part of gene expression. *World Academy of Sciences Journal*, *2*(1), 3–8. https://doi.org/10.3892/wasj.2020.32

Mruk, D. D., & Cheng, C. Y. (2015). The Mammalian Blood-Testis Barrier: Its Biology and Regulation. *Endocrine Reviews*, *36*(5), 564–591. https://doi.org/10.1210/er.2014-1101

Mueller, R. F., & Young, I. D. (1995). *Emery's Elements of Medical Genetics* (Ninth).

Muller, H. J. (1918). Genetic Variability, Twin Hybrids and Constant Hybrids, in a Case of Balanced Lethal Factors. *Genetics*, *3*(5), 422–499. https://doi.org/10.1093/genetics/3.5.422

Munjal, G., Hanmandlu, M., & Srivastava, S. (2018). Phylogenetics Algorithms and Applications. *Springer Nature*, *10*(904), 187–194. https://doi.org/10.1007/978-981-13-5934-7_17

Murat, F., Mbengue, N., Boeg Winge, S., Trefzer, T., Leushkin, E., Sepp, M., Cardoso-Moreira, M., Schmidt, J., Schneider, C., Mößinger, K., Brüning, T., Lamanna, F., Riera Belles, M., Conrad, C., Kondova, I., Bontrop, R., Behr, R., Khaitovich, P., Pääbo, S., … Kaessmann, H. (2023). The molecular evolution of spermatogenesis across mammals. *Nature*, *613*, 308–316. https://doi.org/10.1038/s41586-022-05547-7

Muro, R., Nitta, T., Kitajima, M., Okada, T., & Suzuki, H. (2018). *RASAL3*-mediated T cell survival is essential for inflammatory responses. *Biochemical and Biophysical Research Communications*, *496*(1), 25–30. https://doi.org/10.1016/j.bbrc.2017.12.159

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., & Scheffler, K. (2013). FUBAR: A Fast, Unconstrained Bayesian AppRoximation for

Inferring Selection. *Molecular Biology and Evolution*, *30*(5), 1196–1205. https://doi.org/10.1093/molbev/mst030

Nabholz, B., Glemin, S., & Galtier, N. (2008). Strong Variations of Mitochondrial Mutation Rate across Mammals–the Longevity Hypothesis. *Molecular Biology and Evolution*, *25*(1), 120–130. https://doi.org/ff10.1093/molbev/msm248ff. ffhal-03327979f

Nakasuji, T., Ogonuki, N., Chiba, T., Kato, T., Shiozawa, K., Yamatoya, K., Tanaka, H., Kondo, T., Miyado, K., Miyasaka, N., Kubota, T., Ogura, A., & Asahara, H. (2017). Complementary Critical Functions of Zfy1 and Zfy2 in Mouse Spermatogenesis and Reproduction. *PLOS Genetics*, *13*(1), e1006578. https://doi.org/10.1371/journal.pgen.1006578

Navarro-Costa, P., Gonçalves, J., & Plancha, C. E. (2010). The AZFc region of the Y chromosome: at the crossroads between genetic diversity and male infertility. *Human Reproduction Update*, *16*(5), 525–542. https://doi.org/10.1093/humupd/dmq005

Navarro-Costa, P., Plancha, C. E., & Gonçalves, J. (2010). Genetic Dissection of the AZF Regions of the Human Y Chromosome: Thriller or Filler for Male (In)fertility? *Journal of Biomedicine and Biotechnology*, *936569*. https://doi.org/10.1155/2010/936569

Neumann, A., Meinke, S., Goldammer, G., Strauch, M., Schubert, D., Timmermann, B., Heyd, F., & Preußner, M. (2020). Alternative splicing coupled mRNA decay shapes the temperature-dependent transcriptome. *EMBO Reports*, *21*(12). https://doi.org/10.15252/EMBR.202051369

Ni, W., Perez, A. A., Schreiner, S., Nicolet, C. M., & Farnham, P. J. (2020). Characterization of the ZFX family of transcription factors that bind downstream of the start site of CpG island promoters. *Nucleic Acids Research*, *48*(11), 5986–6000. https://doi.org/10.1093/NAR/GKAA384

Nin, D. S., & Deng, L.-W. (2023). Biology of Cancer-Testis Antigens and Their Therapeutic Implications in Cancer. *Cells*, *12*(6), 926. https://doi.org/10.3390/cells12060926

Nowicka-Bauer, K., & Szymczak-Cendlak, M. (2021). Structure and Function of Ion Channels Regulating Sperm Motility—An Overview. *International Journal of Molecular Sciences*, *22*(6), 3259. https://doi.org/10.3390/ijms22063259

O'Donnell, L., Stanton, P., & de Krester, D. M. (2017). Endocrinology of the Male Reproductive System and Spermatogenesis. In F. KR, A. B, & B. MR (Eds.), *Endotext*.

Oehninger, S., & Kruger, T. F. (2021). Sperm Morphology and its disorders in the context of infertility. In *F&S Reviews* (pp. 75–92). https://doi.org/10.1016/j.xfnr.2020.09.002

Oncogene. (2024). National Human Genome Research Institute. https://www.genome.gov/genetics-glossary/Oncogene

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, *12*, 87–89. https://doi.org/10.1038/nrg2934

Page, D. C., C., E. M., Fisher, B. M., & Brown, L. G. (1990). Additional deletion in sex-determining region of human Y chromosome resolves paradox of X,t(Y;22) female. *Nature*, *346*, 279–281. https://doi.org/10.1038/346279a0

Pai, S. G., Carneiro, B. A., Mota, J. M., Costa, R., Leite, C. A., Barroso-Sousa, R., Kaplan, J. B., Chae, Y. K., & Giles, F. J. (2017). *WNT*/beta-catenin pathway: modulating anticancer immune response. *Journal of Hematology and Oncology*, *10*(101). https://doi.org/10.1186/s13045-017-0471-6

Palmer, D. H., Rogers, T. F., Dean, R., & Wright, A. E. (2019). How to identify sex chromosomes and their turnover. *Molecular Ecology*, *28*(21), 4709–4724. https://doi.org/10.1111/mec.15245

Pamilo, P., & Bianchi, N. (1993). Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Molecular Biology and Evolution*, *10*(2), 271–281. https://doi.org/10.1093/oxfordjournals.molbev.a040003

Panning, B. (2008). X-chromosome inactivation: the molecular basis of silencing. *Journal of Biology*, *7*(30). https://doi.org/10.1186/jbiol95

Parr, B., & McMahon, A. (1998). Sexually dimorphic development of the mammalian reproductive tract requires *WNT*-7a. *Nature*, *395*, 707–710. https://doi.org/10.1038/27221

Pastén, K., Bastian, Y., Roa-Espitia, A. L., Maldonado-García, D., Mendoza-Hernández, G., Ortiz-García, C. I., Mújica, A., & Hernández-González, E. O. (2014). ADAM15 participates in fertilization through a physical interaction with acrogranin. *Reproduction*, *148*(6), 623–634. https://doi.org/10.1530/REP-14-0179

Patel, S., Alam, A., Pant, R., & Chattopadhyay, S. (2019). *WNT* Signaling and Its Significance Within the Tumor Microenvironment: Novel Therapeutic Insights. *Frontiers in Immunology*, *10*, 2872. https://doi.org/10.3389/fimmu.2019.02872

Pecoraro, A., Pagano, M., Russo, G., & Russo, A. (2021). Ribosome Biogenesis and Cancer: Overview on Ribosomal Proteins. *International Journal of Molecular Sciences*, *22*(11), 5496. https://doi.org/10.3390/ijms22115496

Persikov, A., Osada, R., & Singh, M. (2009). Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics*, *25*(1), 22–29. https://doi.org/10.1093/bioinformatics/btn580

Persikov, A., & Singh, M. (2014). De Novo Prediction of DNA-binding Specificities for Cys2His2 Zinc Finger Proteins. *Nucleic Acids Research*, *42*(1), 97–108. https://doi.org/10.1093/nar/gkt890

Pezo, V., Louis, D., Guérineau, V., Le Caer, J., Gaillon, L., Mutzel, R., & Marlière, P.

(2013). A Metabolic Prototype for Eliminating Tryptophan From The Genetic Code. *Scientific Reports*, *3*, 1539. https://doi.org/10.1038/srep01359

Pieraccioli, M., Nicolai, S., Antonov, A., Somers, J., Malewicz, M., Melino, G., & Raschellà, G. (2016). ZNF281 contributes to the DNA damage response by controlling the expression of XRCC2 and XRCC4. *Oncogene*, *35*, 2592–2601. https://doi.org/10.1038/onc.2015.320

Pierce, A., Miller, G., Arden, R., & Gottfredson, L. S. (2009). Why is intelligence correlated with semen quality? Biochemical pathways common to sperm and neuron function and their vulnerability to pleiotropic mutations. *Communicative and Integrative Biology*, *2*(5), 385–387. https://doi.org/10.4161/cib.2.5.8716

Pinart, E. (2022). Ion Channels of Spermatozoa: Structure, Function, and Regulation Mechanisms. *International Journal of Molecular Sciences*, *23*(11), 5880. https://doi.org/10.3390/ijms23115880

Ping Han, S., Hang Tang, Y., & Smith, R. (2010). Functional diversity of the hnRNPs: past, present and perspectives. *Biochemical Journal*, *430*(3), 379–392. https://doi.org/10.1042/BJ20100396

Pinto, F. M., Odriozola, A., Candenas, L., & Subirán, N. (2023). The Role of Sperm Membrane Potential and Ion Channels in Regulating Sperm Function. *Internationl Journal of Molecular Sciences*, *24*(8), 6995. https://doi.org/10.3390/ijms24086995

Piskacek, M. (2009). Common Transactivation Motif 9aaTAD recruits multiple general co-activatorsTAF9, MED15, CBP and p300. *Nature Precedings*. https://doi.org/10.1038/npre.2009.3488.1

Piskacek, M., Havelka, M., Rezacova, M., & Knight, A. (2016). The 9aaTAD Transactivation Domains: From Gal4 to p53. *PLoS One*, *12*(11), 9. https://doi.org/10.1371/journal.pone.0162842

Piskacek, S., Gregor, M., Nemethova, M., Grabner, M., Kovarik, P., & Piskacek, M. (2007). Nine-amino-acid transactivation domain: Establishment and prediction utilities. *Genomics*, *89*(6), 756–768. https://doi.org/10.1016/j.ygeno.2007.02.003

Pompili, S., Latella, G., Gaudio, E., Sferra, R., & Vetuschi, A. (2021). The Charming World of the Extracellular Matrix: A Dynamic and Protective Network of the Intestinal Wall. *Frontiers in Medicine*, *8*, 610189. https://doi.org/10.3389/fmed.2021.610189

Popova, N. V, & Jücker, M. (2022). The Functional Role of Extracellular Matrix Proteins in Cancer. *Cancers*, *14*(1), 238. https://doi.org/10.3390/cancers14010238

Posada, D., & Crandall, K. (2001). Selecting the best-fit model of nucleotide substitution. *Systematic Biology*, *50*(4), 580–601.

Pouyet, F., Aeschacher, S., Thiery, A., & Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias

demographic inferences. *ELife*, *7*(e36317). https://doi.org/10.7554/eLife.36317

Powell, S. F., Vu, L., Spanos, W. C., & Pyeon, D. (2021). The Key Differences between Human Papillomavirus-Positive and -Negative Head and Neck Cancers: Biological and Clinical Implications. *Cancers*, *13*(20), 5206. https://doi.org/10.3390/cancers13205206

Prelich, G. (2012). Gene Overexpression: Uses, Mechanisms, and Interpretation. *Genetics*, *190*(3), 841–854. https://doi.org/10.1534/genetics.111.136911

Puga Molina, L. C., Luque, G. M., Balestrini, P. A., Marin-Briggiler, C. I., Romarowki, A., & Buffone, M. G. (2018). Molecular Basis of Human Sperm Capacitation. *Frontiers in Cell and Development Biology*, *6*. https://doi.org/10.3389/fcell.2018.00072

Pulix, M., Lukashchuk, V., Smith, D. C., & Dickson, A. J. (2021). Molecular characterization of HEK293 cells as emerging versatile cell factories. *Current Opinion in Biotechnology*, *71*, 18–24. https://doi.org/10.1016/j.copbio.2021.05.001

Quintana-Murci, L., & Fellous, M. (2001). The human Y chromosome: the biological role of a “functional wasteland.” *Journal of Biomedicine and Biotechnology*, *1*, 1. http://jbb.hindawi.com

Ramırez-Reveco, A., Villarroel-Espındola, F., Rodriguez-Gil, J., & Concha, I. I. (2017). Neuronal signaling repertoire in the mammalian sperm functionality. *Biology of Reproduction*, *96*(3), 505–524. https://doi.org/10.1095/biolreprod.116.144154

Ramm, S. A., Schärer, L., Ehmcke, J., & Wistuba, J. (2014). Sperm competition and the evolution of spermatogenesis. *Molecular Human Reproduction*, *20*(12), 1169–1179. https://doi.org/10.1093/MOLEHR/GAU070

Ranke, M. B., & Saenger, P. (2001). Turner’s syndrome. *The Lancet*, *358*(9278), 309–314. https://doi.org/10.1016/S0140-6736(01)05487-3

Rapaport, F., Boisson, B., Gregor, A., & Patin, E. (2021). Negative selection on human genes underlying inborn errors depends on disease outcome and both the mode and mechanism of inheritance. *PNAS*, *118*(3), e2001248118. https://doi.org/10.1073/pnas.2001248118

Rascio, F., Spadaccino, F., Rocchetti, M. T., Castellano, G., Stallone, G., Netti, G. S., & Ranieri, E. (2021). The Pathogenic Role of PI3K/AKT Pathway in Cancer Onset and Drug Resistance: An Updated Review. *Cancers*, *13*(16), 3949. https://doi.org/10.3390/cancers13163949

RBMY1A1. (2024). The Human Protein Atlas. https://www.proteinatlas.org/ENSG00000234414-*RBMY1A1*

Reh, J. R., Zhang, X., & Nagano, M. C. (2011). *WNT*5a is a cell-extrinsic factor that supports self-renewal of mouse spermatogonial stem cells. *Journal of Cell Science*, *124*(14), 2357–2366. https://doi.org/10.1242/jcs.080903

Reimand, J., Isser, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., Merico, D., & Bader, G. D. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols*, *14*(2), 482–517. https://doi.org/10.1038/s41596-018-0103-9

Rey, R. A. (2021). Noncanonical *WNT* Signaling in the Integrity of the Blood-Testis Barrier and Sperm Release. *Endocrinology*, *162*(10), bqab159. https://doi.org/10.1210/endocr/bqab159

Rhie, A., Nurk, S., Cechova, M., Hoyt, S. J., & Taylor, D. J. (2023). The complete sequence of a human Y chromosome. *Nature*, *621*, 344–354. https://doi.org/10.1038/s41586-023-06457-y

Romano, R., Ellis, L. S., Yu, N., Bellizzi, J., Brown, T. C., Korah, R., Carling, T., Costa-Guda, J., & Arnold, A. (2017). Mutational Analysis of ZFY in Sporadic Parathyroid Adenomas. *Journal of the Endrocine Society*, *1*(4), 313–316. https://doi.org/10.1210/js.2017-00031

Rosano, G. L., & Ceccarelli, E. A. (2014). Recombinant protein expression in Escherichia coli: advances and challenges. *Frontiers in Microbiology*, *5*, 172. https://doi.org/10.3389/fmicb.2014.00172

Royo, H., Polikiewicz, G., Mahadevaiah, S. K., Prosser, H., Mitchell, M., Bradley, A., De Rooij, D. G., Burgoyne, P. S., & Turner, J. M. A. (2010). Report Evidence that Meiotic Sex Chromosome Inactivation Is Essential for Male Fertility. *Current Biology*, *20*, 2117–2123. https://doi.org/10.1016/j.cub.2010.11.010

Rubin, J. B. (2022). The spectrum of sex differences in cancer. *Trends in Cancer*, *8*(4), 303–315. https://doi.org/10.1016/j.trecan.2022.01.013

Ruiz-Herrera, A., Waters, P. D., & Mable, B. (2022). Fragile, unfaithful and persistent Ys-on how meiosis can shape sex chromosome evolution. *Heredity*, *129*, 22–30. https://doi.org/10.1038/s41437-022-00532-2

Ruiz, A. J., Castaneda, C., Raudsepp, T., & Tibary, A. (2019). Azoospermia and Y Chromosome-Autosome Translocation in a Friesian Stallion. *Journal of Equine Veterinary Science*, *82*(102781). https://doi.org/10.1016/j.jevs.2019.07.002

Russell, L., Ettlin, R., Sinha Hikim, A., & Clegg, E. (1990). *Histological and Histopathological Evaluation of The Testis*.

Sabatini, M. E., & Chiocca, S. (2020). Human papillomavirus as a driver of head and neck cancers. *British Journal of Cancer*, *122*, 306–314. https://doi.org/10.1038/s41416-019-0602-7

Saito, S., Cao, D.-Y., Victor, A. R., Peng, Z., Wu, H.-Y., & Okwan-Duodu, D. (2021). *RASAL3* Is a Putative RasGAP Modulating Inflammatory Response by Neutrophils.

Frontiers in Immunology, *12*. https://doi.org/10.3389/fimmu.2021.744300

Sakamoto, T., & Innan, H. (2022). Muller's ratchet of the Y chromosome with gene conversion. *Genetics*, *220*(1), iyab204. https://doi.org/10.1093/genetics/iyab204

Salmaninejad, A., Zamani, M. R., Pourvahedi, M., Golchehre, Z., Bereshneh, A. H., & Rezaei, N. (2016). Cancer/Testis Antigens: Expression, Regulation, Tumor Invasion, and Use in Immunotherapy of Cancers. *Immunological Investigations*, *45*(7), 619–640. https://doi.org/10.1080/08820139.2016.1197241

Sammut, S. J., Feichtinger, J., Stuart, N., Wakeman, J. A., Larcombe, L., & McFarlane, R. J. (2014). A novel cohort of cancer-testis biomarker genes revealed through meta-analysis of clinical data sets. *Oncoscience*, *1*(5), 349–359. https://doi.org/10.18632/oncoscience.37

Sanborn, A. L., Yeh, B. T., Feigerle, J. T., Hao, C. V, Townshend, R. J., Aiden, E. L., Dror, R. O., & Kornberg, R. D. (2021). Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. *ELife*, *10*, e68068. https://doi.org/10.7554/eLife.68068

Sawyer, S. A. (1989). Statistical tests for detecting gene conversion. *Molecular Biology and Evolution*, *6*, 526–538. https://doi.org/10.1093/oxfordjournals.molbev.a040567

Sawyer, S. A. (1999). *GENECONV: A computer package for the statistical detection of gene conversion*. http://www.math.wustl.edu/~sawyer

Scanlan, M. J., Gure, A. O., Jungbluth, A. A., Old, L. J., & Chen, Y.-T. (2002). Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. *Immunological Reviews*, *188*(1), 22–32. https://doi.org/10.1034/j.1600-065X.2002.18803.x

Schneider-Gadicke, A., Beer-Romero, P., Brown, L. G., Mardon, G., Luoh, S., & Page, D. C. (1989). Putative transcription activator with alternative isoforms encoded by human ZFX gene. *Nature*, *342*, 708–711. https://doi.org/https://doi.org/10.1038/342708a0

Scott, A. D., & Baum, D. A. (2016). Phylogenetic Tree. In R. M. Kliman (Ed.), *Encyclopedia of Evolutionary Biology* (pp. 270–276). https://doi.org/10.1016/B978-0-12-800049-6.00203-1

Segre, J. A., Bauer, C., & Fuchs, E. (1999). Klf4 is a transcription factor required for establishing the barrier function of the skin. *Nature Genetics*, *4*(3), 356–360. https://doi.org/10.1038/11926.

Selkirk, C. (2004). Ion-exchange Chromatography. In *Protein Purification Protocols* (pp. 125–131). Humana Press. https://doi.org/10.1385/1-59259-655-X:125

Shukla, K. K., Mahdi, A. A., & Rajender, S. (2013). Ion Channels in Sperm Physiology and Male Fertility and Infertility. *Journal of Andrology*, *33*(5), 777–788.

https://doi.org/10.2164/jandrol.111.015552

Shvetsova, E., Sofronova, A., Monajemi, R., Gagalova, K., Draisma, H. H. M., White, S. J., Santen, G. W. E., Lopes, S. M. C. de S., Heijmans, B. T., Meurs, J. van, Jansen, R., Franke, L., Szymon M. Kiełbasa, J. T. den D., Hoen, P. A. C. 't, Consortium, B., & Consortium, G. (2018). Skewed X-inactivation is common in the general female population. *European Journal of Human Genetics*, *27*, 455–465. https://doi.org/10.1038/s41431-018-0291-3

Sidorenko, J., Kassam, I., Kemper, K. E., Zeng, J., Lloyd-Jones, L. R., Montgomery, G. W., Gibson, G., Metspalu, A., Esko, T., Yang, J., McRae, A. F., & Visscher, P. M. (2019). The effect of X-linked dosage compensation on complex trait variation. *Nature Communications*, *10*(3009). https://doi.org/10.1038/s41467-019-10598-y

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higginsa, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*, 539. https://doi.org/10.1038/msb.2011.75

Sinclair, A. (1988). Sequences homologous to ZFY, a candidate human sex-determining gene, are autosomal in marsupials. *Nature*, *336*(6201), 780–783. https://doi.org/10.1038/336780a0

Singh, V., Joshi, M., Singh, K., & Singh, R. (2019). *WNT* signaling in spermatogenesis and male infertility. In *Molecular Signalling in Spermatogenesis and Male Infertility* (p. Chapter 9).

Siu, M. K., & Cheng, C. Y. (2004). Extracellular matrix: Recent advances on its role in junction dynamics in the seminiferous epithelium during spermatogenesis. *Biology of Reproduction*, *71*(2), 375–391. https://doi.org/10.1095/biolreprod.104.028225

Siu, M. K., & Yan Cheng, C. (2008). Extracellular Matrix and Its Role in Spermatogenesis. *Advances in Experimental Medicine and Biology*, *636*, 74–91. https://doi.org/10.1007/978-0-387-09597-4_5

Skrisovska, L., Bourgeois, C. F., Stefl, R., Grellscheid, S.-N., Kister, L., Wenter, P., Elliott, D. J., Stevenin, J., & Allain, F. H.-T. (2007). The testis-specific human protein RBMY recognizes RNA through a novel mode of interaction. *EMBO Reports*, *8*(4), 372–379. https://doi.org/10.1038/sj.embor.7400910

Slattery, J. P., Sanner-Wachter, L., & O'Brien, S. J. (2000). Novel gene conversion between X-Y homologues located in the nonrecombining region of the Y chromosome in Felidae (Mammalia). *Proceedings of the National Academy of Sciences of the United States of America*, *97*(10), 5307–5312. https://doi.org/10.1073/pnas.97.10.5307.

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthès, P., Kokkinaki, M., Nef, S., Gnirke, A., Dym, M., deMassy, B., Mikkelsen, T. S., & Kaessmann, H. (2013). Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Reports*, *3*(6), 2179–2190. https://doi.org/10.1016/J.CELREP.2013.05.031

Spitz, F., & Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, *13*, 613–626. https://doi.org/10.1038/nrg3207

Staller, M. V, Ramirez, E., Holehouse, A. S., Pappu, R. V, & Cohen, B. A. (2021). Design principles of acidic transcriptional activation domains. *BioRxiv*. https://doi.org/10.1101/2020.10.28.359026

Staller, M. V, Ramirez, E., Kotha, S. R., Holehouse, A. S., Pappu, R. V, & Cohen, B. A. (2022). Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains. *Cell Systems*, *13*(4), 334–345. https://doi.org/10.1016/j.cels.2022.01.002

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. A., & Soreq, H. (2005). Function of alternative splicing. *Gene*, *344*, 1–20. https://doi.org/10.1016/j.gene.2004.10.022

Sternlicht, M. D., & Werb, Z. (2001). HOW MATRIX METALLOPROTEINASES REGULATE CELL BEHAVIOR. *Annual Review of Cell and Developmental Biology*, *17*, 463–516. https://doi.org/10.1146/annurev.cellbio.17.1.463

Stevens, N. (1905). Studies in Spermatogenesis. *Washington, D.C. : Carnegie Institution of Washington*, *36*.

Stocco, D. M. (2001). StAR PROTEIN AND THE REGULATION OF STEROID HORMONE BIOSYNTHESIS. *Annual Review of Physiology*, *63*, 193–213. https://doi.org/10.1146/annurev.physiol.63.1.193.

Subrini, J., & Turner, J. (2021). Y chromosome functions in mammalian spermatogenesis. *ELife*, *10*, e67345. https://doi.org/10.7554/eLife.67345

Sudhakaran, M., & Doseff, A. I. (2023). Role of Heterogeneous Nuclear Ribonucleoproteins in the Cancer-Immune Landscape. *International Journal of Molecular Sciences*, *24*(6), 5086. https://doi.org/10.3390/ijms24065086

Supplitt, S., Karpinski, P., Sasiadek, M., & Laczmanska, I. (2021). Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine. *International Journal of Molecular Sciences*, *22*(3), 1422. https://doi.org/10.3390/ijms22031422

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*, W609–W612. https://doi.org/10.1093/nar/gkl315

Takase, H. M., & Nusse, R. (2016). Paracrine *WNT*/β-catenin signaling mediates proliferation of undifferentiated spermatogonia in the adult mouse testis. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(11), E1489–E1497. https://doi.org/10.1073/pnas.1601461113

Talbot, P., Shur, B. D., & Myles, D. G. (2003). Cell Adhesion and Fertilization: Steps in Oocyte Transport, Sperm-Zona Pellucida Interactions, and Sperm-Egg Fusion. *Biology of Reproduction*, *68*(1), 1–9. https://doi.org/10.1095/biolreprod.102.007856

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis version 11. *Molecular Biology and Evolution*, *38*, 3022–3027. https://doi.org/10.1093/molbev/msab120

Tapia, J., Macias-Garcia, B., Miro-Moran, A., Ortega-Ferrusola, C., Salido, G., Peña, F., & Aparicio, I. (2012). The Membrane of the Mammalian Spermatozoa: Much More Than an Inert Envelope. *Reproduction in Domestic Animals*, *47*(s3), 65–75. https://doi.org/10.1111/j.1439-0531.2012.02046.x

Tartaglia, N. R., Howell, S., Sutherland, A., Wilson, R., & Wilson, L. (2010). A review of trisomy X (47,XXX). *Orphanet Journal of Rare Diseases*, *5*(8). https://doi.org/10.1186/1750-1172-5-8

Tazi, J., Bakkour, N., & Stamm, S. (2009). Alternative Splicing and Disease. *Biochimica et Biophysica Acta*, *1792*(1), 14–26. https://doi.org/10.1016/j.bbadis.2008.09.017

Tegel, H., Tourle, S., Ottosson, J., & Persson, A. (2010). Increased levels of recombinant human proteins with the Escherichia coli strain Rosetta(DE3). *Protein Expression and Purification*, *69*(2), 159–167. https://doi.org/10.1016/j.pep.2009.08.017

Tenorio, F., Neto, L., Bach, P. V., Najari, B. B., Li, P. S., & Goldstein, M. (2016). Spermatogenesis in humans and its affecting factors. *Seminars in Cell & Developmental Biology*, *59*, 10–26. https://doi.org/10.1016/j.semcdb.2016.04.009

The Human Protein Atlas. (2024a). *WNT7A*. The Human Protein Atlas. https://www.proteinatlas.org/ENSG00000154764-*WNT7*A

The Human Protein Atlas. (2024b). *ZFY*. The Human Protein Atlas. https://www.proteinatlas.org/ENSG00000067646-ZFY/tissue

Thomas, P., & Smart, T. G. (2005). HEK293 cell line: A vehicle for the expression of recombinant proteins. *Journal of Pharmacological and Toxicological Methods*, *51*(3), 187–200. https://doi.org/10.1016/j.vascn.2004.08.014

Thompson, J., Higgins, D., & Gibson, T. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*, 4673–4680. https://doi.org/10.1093/nar/22.22.4673

Tian, C., Zhu, R., Zhu, L., Qiu, T., Cao, Z., & Kang, T. (2014). Potassium channels:

structures, diseases, and modulators. *Chemical Biology & Drug Design*, *83*(1), 1–26. https://doi.org/10.1111/cbdd.12237.

Tokmakov, A. A., Kurotani, A., & Sato, K.-I. (2021). Protein pI and Intracellular Localization. *Frontiers in Molecular Biosciences*, *8*, 775736. https://doi.org/10.3389/fmolb.2021.775736

Tricoli, J. V, & Bruce Bracken, R. (1993). ZFY Gene Expression and Retention in Human Prostate Adenocarcinoma. *GENES. CHROMOSOMES & CANCER*, *6*(2), 65–72. https://doi.org/10.1002/gcc.2870060202

Triezenberg, S. J. (1995). Structure and function of transcriptional activation domains. *Current Opinion in Genetics & Development*, *5*(2), 190–196. https://doi.org/10.1016/0959-437X(95)80007-7

Trifinopoulos, J., Nguyen, L.-T., Haeseler, A. von, & Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research*, *44*(W1), W232–W235. https://doi.org/10.1093/nar/gkw256

Trujillo, B. C. (2019). *RBMY and ZFY: Identification of two testis-specific genes in HPV-oropharyngeal squamous cell carcinomas*. University of Kent.

Tsuei, D.-J., Hsu, H.-C., Lee, P.-H., Jeng, Y.-M., Pu, Y.-S., Chen, C.-N., Lee, Y.-C., Chou, W.-C., Chang, C.-J., Ni, Y.-H., & Chang, M.-H. (2004). RBMY, a male germ cell-specific RNA-binding protein, activated in human liver cancers and transforms rodent fibroblasts. *Oncogene*, *23*, 5815–5822. https://doi.org/10.1038/sj.onc.1207773

Tucker, P. K., Adkins, R. M., & Rest, J. S. (2003). Differential Rates of Evolution for the ZFY-Related Zinc Finger Genes, Zfy, Zfx, and Zfa in the Mouse Genus Mus. *Molecular Biology and Evolution*, *20*(6), 999–1005. https://doi.org/10.1093/MOLBEV/MSG112

Turner, J. M. A. (2007). Meiotic Sex Chromosome Inactivation. *Development (Cambridge, England)*, *134*(10), 1823–1831.

Turner, J. M. A. (2015). Meiotic Silencing in Mammals. *Annual Review of Genetics*, *14*, 20. https://doi.org/10.1146/annurev-genet-112414-055145

Udupa, A., Kotha, S. R., & Staller, M. V. (2024). Commonly asked questions about transcriptional activation domains. *Current Opinion in Structural Biology*, *84*, 102732. https://doi.org/10.1016/j.sbi.2023.102732

Ule, J., & Blencowe, B. J. (2019). Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Molecular Cell Review*, *76*(2), 329–345. https://doi.org/10.1016/j.molcel.2019.09.017

van der Bruggen, P., Traversari, C., Chomez, P., Lurquin, C., De Plaen, E., Van den Eynde, B., Knuth, A., & Boon, T. (1991). A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science*, *254*(5038), 1643–1647.

https://doi.org/10.1126/science.1840703

Veikkolainen, V., Vaparanta, K., Halkilahti, K., Iljin, K., Sundvall, M., & Elenius, K. (2011). Function of ERBB4 is determined by alternative splicing. *Cell Cycle*, *10*(16), 2647–2657. https://doi.org/10.4161/cc.10.16.17194

Vernet, N., Mahadevaiah, S. K., de Rooij, D. G., Burgoyne, P. S., & Ellis, P. J. I. (2016). Zfy genes are required for efficient meiotic sex chromosome inactivation (MSCI) in spermatocytes. *Human Molecular Genetics*, *25*(24), 5300. https://doi.org/10.1093/HMG/DDW344

Vernet, N., Mahadevaiah, S. K., Decarpentrie, F., Longepied, G., De Rooij, D. G., Burgoyne, P. S., & Mitchell, M. J. (2016). Mouse Y-Encoded Transcription Factor Zfy2 Is Essential for Sperm Head Remodelling and Sperm Tail Development. *PLOS One*, *11*(1), e0145398. https://doi.org/10.1371/journal.pone.0145398

Vernet, N., Mahadevaiah, S. K., Ellis, P. J. I., Rooij, D. G. de, & Burgoyne, P. S. (2012). Spermatid development in XO male mice with varying Y chromosome short-arm gene content: evidence for a Y gene controlling the initiation of sperm morphogenesis. *Reproduction*, *144*(4), 433–445. https://doi.org/10.1530/REP-12-0158

Vernet, N., Mahadevaiah, S. K., Ojarikre, O. A., Longepied, G., Prosser, H. M., Bradley, A., Mitchell, M. J., & Burgoyne, P. S. (2011). The Y-Encoded Gene Zfy2 Acts to Remove Cells with Unpaired Chromosomes at the First Meiotic Metaphase in Male Mice. *Current Biology*, *21*(9), 787–793. https://doi.org/10.1016/j.cub.2011.03.057

Vernet, N., Mahadevaiah, S. K., Yamauchi, Y., Decarpentrie, F., & Mitchell, M. J. (2014). Mouse Y-Linked Zfy1 and Zfy2 Are Expressed during the Male-Specific Interphase between Meiosis I and Meiosis II and Promote the 2 nd Meiotic Division. *PLOS Genetics*, *10*(6), e1004444. https://doi.org/10.1371/journal.pgen.1004444

Vernet, N., Szot, M., Mahadevaiah, S. K., Ellis, P. J. I., Decarpentrie, F., Ojarikre, O. A., Rattigan, Á., Taketo, T., & Burgoyne, P. S. (2014). The expression of Y-linked Zfy2 in XY mouse oocytes leads to frequent meiosis 2 defects, a high incidence of subsequent early cleavage stage arrest and infertility. *Development (Cambridge, England)*, *141*(4), 855–866. https://doi.org/10.1242/dev.091165

Vickram, S., Rohini, K., Srinivasan, S., Veenakumari, D. N., Archana, K., Anbarasu, K., Jeyanthi, P., Thanigaivel, S., Gulothungan, G., Rajendiran, N., & Srikumar, P. S. (2021). Role of Zinc (Zn) in Human Reproduction: A Journey from Initial Spermatogenesis to Childbirth. *International Journal of Molecular Science*, *22*(4), 2188. https://doi.org/10.3390/ijms22042188

Vicoso, B. (2019). Molecular and evolutionary dynamics of animal sex-chromosome turnover. *Nature Ecology and Evolution*, *3*, 1632–1641.

https://doi.org/10.1038/s41559-019-1050-8

Vog, P. H., Edelmann, A., Kirsch, S., Henegariu, O., Hirschmann, P., Kiesewetter, F., Köhn, F. M., Schill, W. B., Farah, S., Ramos, C., Hartmann, M., Hartschuh, W., Meschede, D., Behre, H. M., Castel, A., Nieschlag, E., Weidner, W., Gröne, H.-J., Jung, A., … Haidl, G. (1996). Human Y Chromosome Azoospermia Factors (AZF) Mapped to Different Subregions in Yq11. *Human Molecular Genetics*, *5*(7), 933–943. https://doi.org/10.1093/hmg/5.7.933

Vogt, P. ., Falcao, C. ., Hanstein, R., & Zimmer, J. (2008). The AZF proteins. *International Journal of Andrology*, *31*, 383–394. https://doi.org/10.1111/j.1365-2605.2008.00890.x

Vogt, P. H., Bender, U., Deibel, B., Kiesewetter, F., Zimmer, J., & Strowitzki, T. (2021). Human AZFb deletions cause distinct testicular pathologies depending on their extensions in Yq11 and the Y haplogroup: new cases and review of literature. *Cell and Bioscience*, *11*, 60. https://doi.org/10.1186/s13578-021-00551-2

Wade, M. J., & Fogarty, L. (2021). Adaptive co-evolution of mitochondria and the Y-chromosome: A resolution to conflict between evolutionary opponents. *Ecology and Evolution*, *11*(23), 17307–17313. https://doi.org/10.1002/ece3.8366

Walker, C., Mojares, E., & del Rio Hernández, A. (2018). Role of Extracellular Matrix in Development and Cancer Progression. *International Journal of Molecular Sciences*, *19*(10), 3028. https://doi.org/10.3390/ijms19103028

Walker, W. H. (2010). Non-classical actions of testosterone and spermatogenesis. *Philosophical Transactions of the Royal Society B*, *365*(1546), 1557–1569. https://doi.org/10.1098/rstb.2009.0258

Wang, G., Wang, F., Huang, Q., Li, Y., Liu, Y., & Wang, Y. (2015). Understanding Transcription Factor Regulation by Integrating Gene Expression and DNase I Hypersensitive Sites. *BioMed Research International*, 757530. https://doi.org/10.1155/2015/757530

Wang, X., Jiang, L., Wallerman, O., & Wlesh, N. (2013). Transcription factor *ZBED6* affects gene expression, proliferation, and cell death in pancreatic beta cells. *PNAS*, *110*(40), 15997–16002. https://doi.org/10.1073/pnas.1303625110

Ward, J. P., Gubin, M. M., & Schreiber, R. D. (2016). The Role of Neoantigens in Naturally Occurring and Therapeutically Induced Immune Responses to Cancer. *Advances in Immunology*, *130*, 25–74. https://doi.org/10.1016/bs.ai.2016.01.001

Wassarman, P. M., & Litscher, E. S. (2022). Female fertility and the zona pellucida. *ELife*, *11*, e76106. https://doi.org/10.7554/eLife.76106

Waters, P. D., & Ruiz-Herrera, A. (2020). Meiotic Executioner Genes Protect the Y from Extinction. *Trends in Genetics*, *36*(10), 728–738.

https://doi.org/10.1016/j.tig.2020.06.008

Wilson, E. B. (1906). Studies on chromosomes. III. The sexual differences of the chromosomegroups in Hemiptera, with some considerations on the determination and inheritance of sex. *Journal of Experimental Zoology*, *3*(1), 1–40. https://doi.org/10.1002/jez.1400030102

Wilson Sayres, M. A., Lohmueller, K. E., & Nielsen, R. (2014). Natural Selection Reduced Diversity on Human Y Chromosomes. *PLOS Genetics*, *10*(1), e1004064. https://doi.org/10.1371/journal.pgen.1004064

Wingfield, P. T. (2015). Overview of the Purification of Recombinant Proteins. *Current Protocols in Protein Science*, *80*, 6.1.1-6.1.35. https://doi.org/10.1002/0471140864.ps0601s80

Xiao, L., Xian, H., Yee Lee, K., Xiao, B., Wang, H., Yu, F., Shen, H.-M., & Liou, Y.-C. (2015). Death-associated Protein 3 Regulates Mitochondrial-encoded Protein Synthesis and Mitochondrial Dynamics. *Journal of Biological Chemistry*, *41*, 24961–24974. https://doi.org/10.1074/jbc.M115.673343

Yamauchi, Y., Matsumura, T., Bakse, J., Holmlund, H., Blanchet, G., Carrot, E., Ikawa, M., & Ward, M. A. (2022). Loss of mouse Y chromosome gene Zfy1 and Zfy2 leads to spermatogenesis impairment, sperm defects, and infertility. *Biology of Reproduction*, *106*(6), 1312–1326. https://doi.org/10.1093/biolre/ioac057

Yamauchi, Y., Riel, J. M., Ruthig, V., & Ward, M. A. (2015). Mouse Y-Encoded Transcription Factor Zfy2 Is Essential for Sperm Formation and Function in Assisted Fertilization. *PLoS Genet*, *11*(12), 1005476. https://doi.org/10.1371/journal.pgen.1005476

Yanagimachi, R. (2011). Mammalian Sperm Acrosome Reaction: Where Does It Begin Before Fertilization? *Biology of Reproduction*, *85*(1), 4–5. https://doi.org/10.1095/biolreprod.111.092601

Yao, P.-L., Lin, Y.-C., & Richburg, J. H. (2011). Transcriptional Suppression of Sertoli Cell Timp2 in Rodents Following Mono-(2-ethylhexyl) Phthalate Exposure Is Regulated by CEBPA and MYC. *Biology of Reproduction*, *85*(6), 1203–1215. https://doi.org/10.1095/biolreprod.111.093484

Yarden, Y., & Sliwkowski, M. X. (2001). Untangling the ErbB signalling network. *Nature Reviews Molecular Cell Biology*, *2*, 127–137. https://doi.org/10.1038/35052073

Yin, M., Cheng, M., Liu, C., Wu, K., Xiong, W., Fang, J., Li, Y., & Zhang, B. (2021). HNRNPA2B1 as a trigger of RNA switch modulates the miRNA-mediated regulation of CDK6. *IScience*, *24*(11), 103345. https://doi.org/10.1016/j.isci.2021.103345

Young, J. C., Kerr, G., Micati, D., Nielsen, J. E., Meyts, E. R.-D., Abud, H. E., & Loveland, K. L. (2020). *WNT* signalling in the normal human adult testis and in male germ cell

neoplasms. *Human Reproduction*, *35*(9), 1991–2003. https://doi.org/10.1093/humrep/deaa150

Younis, I., Berg, M., Kaida, D., Dittmar, K., Wang, C., & Dreyfuss, G. (2010). Rapid-Response Splicing Reporter Screens Identify Differential Regulators of Constitutive and Alternative Splicing. *Molecular and Cellular Biology*, *30*(7), 1718–1728. https://doi.org/10.1128/MCB.01301-09

Yu, X.-W., Wei, Z.-T., Jiang, Y.-T., & Zhang, S.-L. (2015). Y chromosome azoospermia factor region microdeletions and transmission characteristics in azoospermic and severe oligozoospermic patients. *International Journal of Clinical and Experimental Medicine*, *8*(9), 14634–14646.

Zeng, S., Chen, L., Liu, X., Tang, H., Wu, H., & Liu, C. (2023). Single-cell multi-omics analysis reveals dysfunctional *WNT* signaling of spermatogonia in non-obstructive azoospermia. *Frontiers in Endocrinology*, *14*. https://doi.org/10.3389/fendo.2023.1138386

ZFY: The Human Protein Atlas. (2024). https://www.proteinatlas.org/ENSG00000067646-ZFY

Zhan, Y., Chen, X., Zheng, H., Luo, J., Yang, Y., Ning, Y., Wang, H., Zhang, Y., Zhou, M., Wang, W., & Fan, S. (2022). YB1 associates with oncogenetic roles and poor prognosis in nasopharyngeal carcinoma. *Scientific Reports*, *12*(3699). https://doi.org/10.1038/s41598-022-07636-z

Zhang, D., Wang, Y., Lin, H., Sun, Y., Wang, M., Jia, Y., Yu, X., Jiang, H., Xu, W., Sun, J., & Xu, Z. (2020). Function and therapeutic potential of G protein-coupled receptors in epididymis. *British Journal of Pharmacology*, *177*(24), 5489–5508. https://doi.org/10.1111/bph.15252

Zhang, Q., Zhao, L., Yang, Y., Li, S., Liu, Y., & Chen, C. (2022). Mosaic loss of chromosome Y promotes leukemogenesis and clonal hematopoiesis. *JCI Insight*, *7*(3), e153768. https://doi.org/10.1172/jci.insight.153768

Zhang, Y., & Wang, X. (2020). Targeting the *WNT*/β-catenin signaling pathway in cancer. *Journal of Hematology and Oncology*, *13*(165). https://doi.org/10.1186/s13045-020-00990-3

Zhang, Z., Zhang, Y., Qiu, Y., Mo, W., & Yang, Z. (2021). Human/eukaryotic ribosomal protein L14 (RPL14/eL14) overexpression represses proliferation, migration, invasion and EMT process in nasopharyngeal carcinoma. *Bioengineered*, *12*(1), 217502186. https://doi.org/10.1080/21655979.2021.1932225

Zhou, C., & Parsons, J. L. (2020). The radiobiology of HPV-positive and HPV-negative head and neck squamous cell carcinoma. *Expert Reviews in Molecular Medicine*, *22*(e3), 1–11. https://doi.org/10.1017/erm.2020.4

Zhou, X., Liao, W.-J., Liao, J.-M., Liao, P., & Lu, H. (2015). Ribosomal proteins: functions beyond the ribosome. *Journal of Molecular Cell Biology*, *7*(2), 92–104. https://doi.org/10.1093/jmcb/mjv014

Zickler, D., & Kleckner, N. (1998). THE LEPTOTENE-ZYGOTENE TRANSITION OF MEIOSIS. *Annual Review of Genetics*, *32*, 619–697. https://doi.org/10.1146/annurev.genet.32.1.619

Zimta, A.-A., Bogdan Tigu, A., Braicu, C., Stefan, C., Ionescu, C., & Berindan-Neagoe, I. (2020). An emerging class of long non-coding RNA wit oncogenic role arises from the snoRNA host genes. *Frontiers in Oncology*, *10*(389). https://doi.org/10.3389/fonc.2020.00389

8. Supplementary Data

Table 1: Placental mammal *ZFY* nucleotide and protein sequences collected from the NCBI search engine. Included in the table are the corresponding accession numbers for the species nucleotide and protein CDS used in the analysis.

Binomial Nomenclature	Common Species Nomenclature	Taxonomic Classification	Protein name	Database for CDS sequence & Database for Protein	Database DNA Accession	Database Protein Accession
Homo sapiens	Human	Chordata/ Mammalia/ Primate/ Hominidae/ Homo	*ZFY*	NCBI	NM_003411.4	NP_003402.2
Pan troglodytes	Chimpanzee	Chordata/ Mammalia/ Primate/ Hominidae/ Pan	*ZFY*	NCBI	XM_009445712.3	XP_009443987.1
Gorilla gorilla	Gorilla	Chordata/ Mammalia/ Primate/ Hominidae/ Gorilla	*ZFY*	NCBI	AH014841.2	AAX94761.1
Macaca mulatta	Rhesus monkey	Chordata/ Mammalia/ Primate/ Hominidae/ Macaca	*ZFY*	NCBI	XM_015128596.2	XP_014984082.1
Papio anubis	Olive baboon	Chordata/ Mammalia/ Primate/ Cercopithecoidea / Papio	*ZFY*	NCBI	XM_031661108.1	XP_031516968.1
Rhinopithecus roxellana	Golden snub-noised monkey	Chordata/ Mammalia/ Primate/ Cercopithecoidea / Rhinopithecus	*ZFY*	NCBI	XM_030926312.1	XP_030782172.1
Callithrix jacchus	White-tufted-ear marmoset	Chordata/ Mammalia/ Primate/ Callitrichidae/ Callithrix	*ZFY*	NCBI	XM_035289933.1	XP_035145824.1
Marmota marmota marmota	Alphine marmot	Chordata/ Mammalia/ Rodentia/ Sciuridae/ Marmota	*ZFY*	NCBI	XM_015488020.1	XP_015343506.1
Mus musculus	Mouse	Chordata/ Mammalia/ Rodentia/ Muridae/ Mus	*ZFY*1	NCBI	NM_009570.4	NP_033596.3
			*ZFY*2	NCBI	NM_009571.2	NP_033597.2

Rattus norvegicus	Brown rat	Chordata/ Mammalia/ Rodentia/ Muridae/ Rattus	*ZFY*2	NCBI	XM_039100724.1	XP_038956652.1
Bos taurus	Cattle	Chordata/ Mammalia/ Artiodactyla/ Bovidae/ Bos	*ZFY*	NCBI	NM_177491.1	NP_803457.1
Capra hircus	Goat	Chordata/ Mammalia/ Artiodactyla/ Bovidae/ Capra	*ZFY*	NCBI	XM_018044894.1	XP_017900383.1
Odocoileus virginianus texanus	White-tailed deer	Chordata/ Mammalia/ Artiodactyla/ Cervidae/ Odocoileus	*ZFY*	NCBI	XM_020903648.1	XP_020759307.1
Sus scrofa	Pig	Chordata/ Mammalia/ Artiodactyla/ Suidae/ Sus	*ZFY*	NCBI	XM_021080936.1	XP_0209365.1
Canis lupus familiaris	Dog	Chordata/ Mammalia/ Carnivoral/ Canidae/ Canis	*ZFY*	NCBI	XM_038451383.1	XP_038307311.1
Mustela erminea	Short-tailed weasel	Chordata/ Mammalia/ Carnivoral/ Mustelidae/ Mustela	*ZFY*	NCBI	XM_032331909.1	XP_032187800.1
Loxodonta africana	African savanna elephant	Chordata/ Mammalia/ Proboscidea/ Elephantidae/ Loxodonta	*ZFY*	NCBI	GATM01000012.1	JAC06687.1
Equus caballus	Horse	Chordata/ Mammalia/ Perissaodactyla/ Equidae/ Equus	*ZFY*	NCBI	No accession	No accession

Table 2: Placental mammal *ZFX* nucleotide and protein sequences collected from the NCBI search engine. Included in the table are the corresponding accession numbers for the species nucleotide and protein CDS used in the analysis. Equus caballus was collected from the ENSEMBL database.

Binomial Nomenclature	Common Species Nomenclature	Taxonomic Classification	Protein name	Database for CDS sequence & Database for Protein	Database DNA Accession	Database Protein Accession
Homo sapiens	Human	Chordata/ Mammalia/ Primate/	*ZFX*	NCBI	NM_003410.4	NP_003401.2

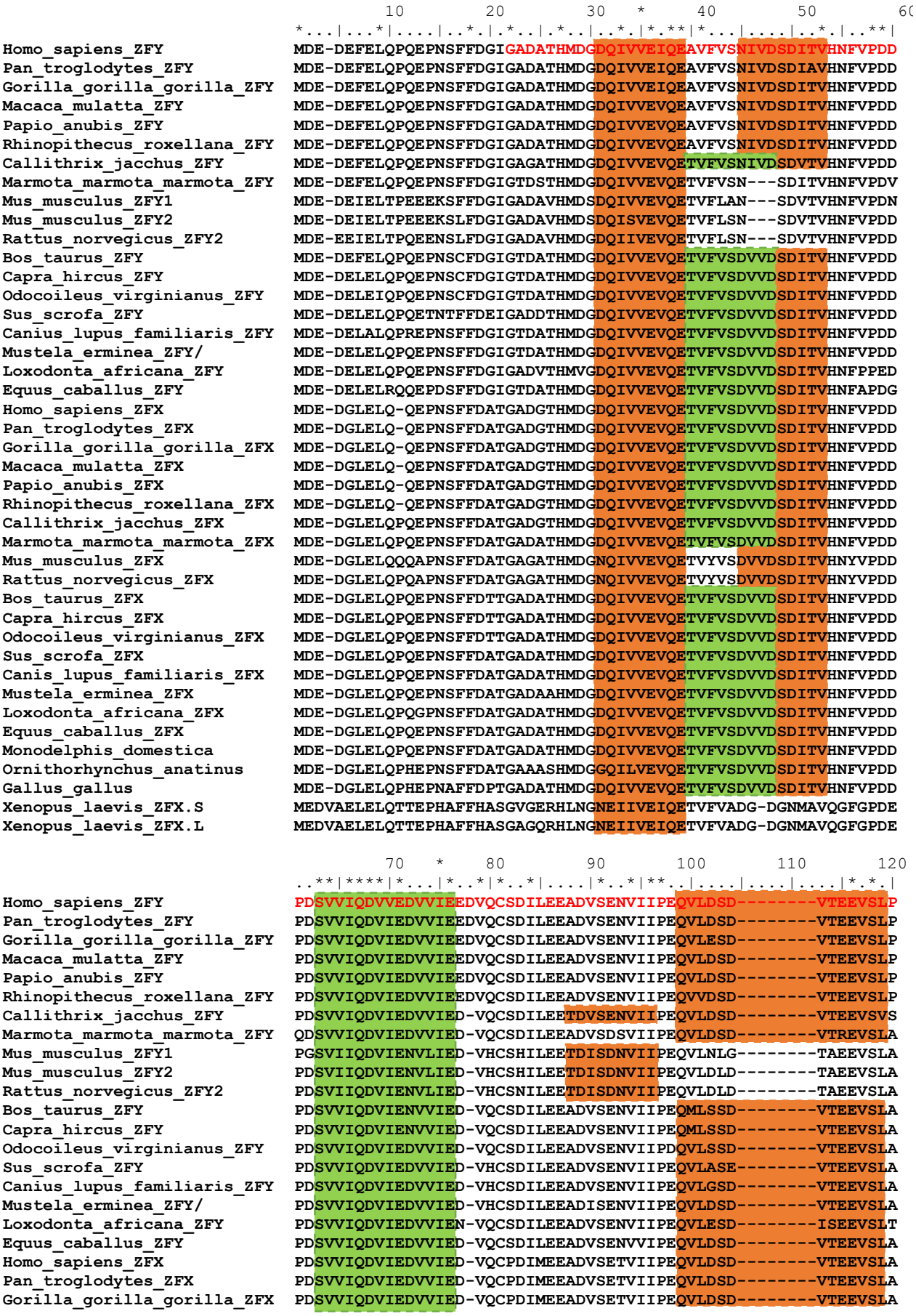
		Hominidae/ Homo				
Pan troglodytes	Chimpanzee	Chordata/ Mammalia/ Primate/ Hominidae/ Pan	*ZFX*	NCBI	XM_016943494.1	XP_016798983.1
Gorilla gorilla	Gorilla	Chordata/ Mammalia/ Primate/ Hominidae/ Gorilla	*ZFX*	NCBI	XM_031005909.1	XP_030861769.1
Macaca mulatta	Rhesus monkey	Chordata/ Mammalia/ Primate/ Hominidae/ Macaca	*ZFX*	NCBI	XM_015127106.2	XP_014982592.1
Papio anubis	Olive baboon	Chordata/ Mammalia/ Primate/ Cercopithecoidea / Papio	*ZFX*	NCBI	XM_031660467.1	XP_031516327.1
Rhinopithecus roxellana	Golden snub-noised monkey	Chordata/ Mammalia/ Primate/ Cercopithecoidea / Rhinopithecus	*ZFX*	NCBI	XM_030933878.1	XP_030789738.1
Callithrix jacchus	White-tufted-ear marmoset	Chordata/ Mammalia/ Primate/ Callitrichidae/ Callithrix	*ZFX*	NCBI	XM_035289564.1	XP_035145455.1
Marmota marmota marmota	Alphine marmot	Chordata/ Mammalia/ Rodentia/ Sciuridae/ Marmota	*ZFX*	NCBI	XM_015488093.1	XP_015343579.1
Mus musculus	Mouse	Chordata/ Mammalia/ Rodentia/ Muridae/ Mus	*ZFX*	NCBI	NM_011768.2	NP_035898.2
Rattus norvegicus	Brown rat	Chordata/ Mammalia/ Rodentia/ Muridae/ Rattus	*ZFX*	NCBI	XM_006257030.4	XP_006257092.1
Bos taurus	Cattle	Chordata/ Mammalia/ Artiodactyla/ Bovidae/ Bos	*ZFX*	NCBI	NM_177490.1	NP_803456.1
Capra hircus	Goat	Chordata/ Mammalia/ Artiodactyla/ Bovidae/ Capra	*ZFX*	NCBI	XM_018043810.1	XP_017899299.1

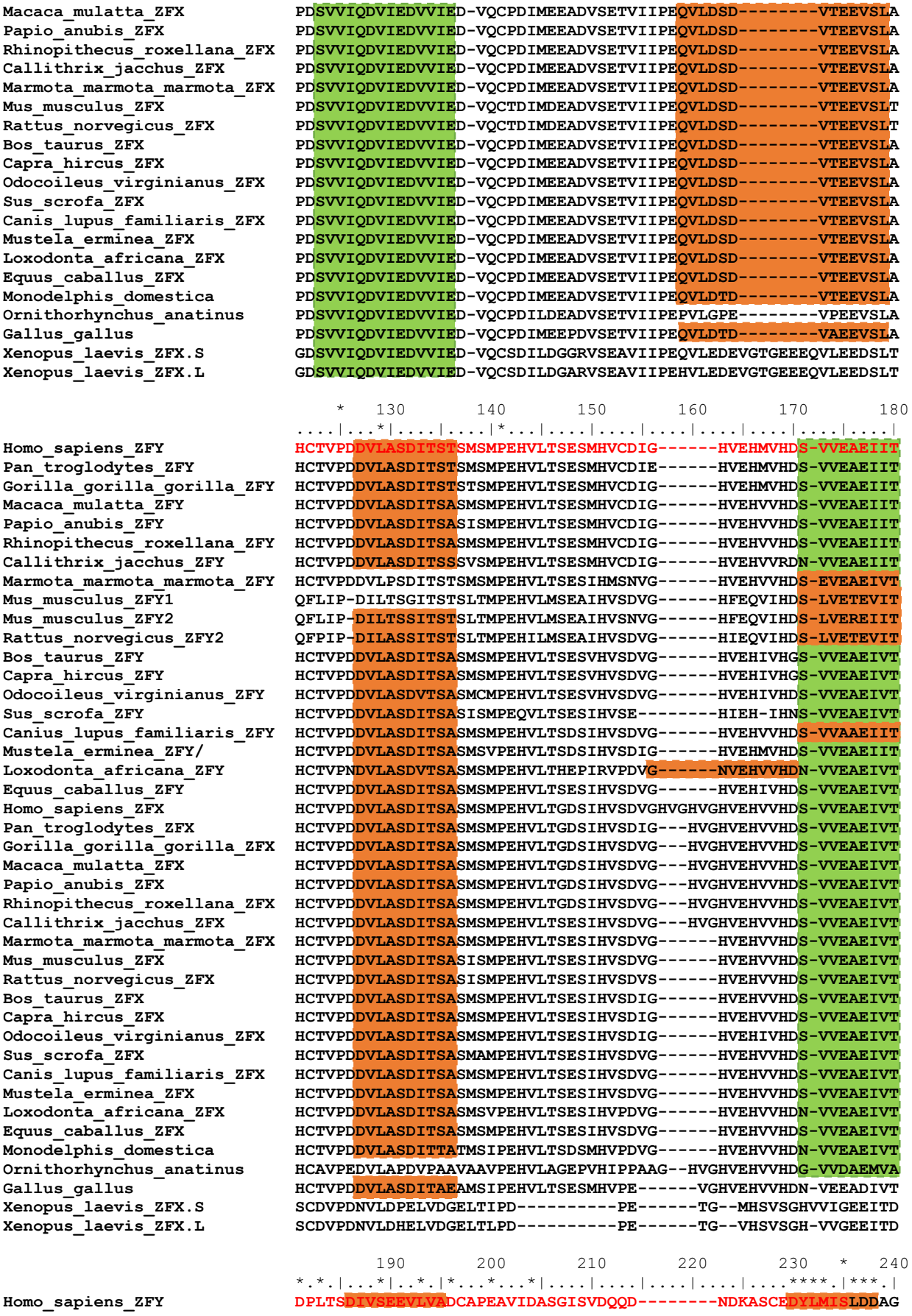
Odocoileus virginianus texanus	White-tailed deer	Chordata/ Mammalia/ Artiodactyla/ Cervidae/ Odocoileus	*ZFX*	NCBI	XM_020903072.1	XP_020758731.1
Sus scrofa	Pig	Chordata/ Mammalia/ Artiodactyla/ Suidae/ Sus	*ZFX*	NCBI	XM_021080656.1	XP_020936315.1
Canis lupus familiaris	Dog	Chordata/ Mammalia/ Carnivoral/ Canidae/ Canis	*ZFX*	NCBI	XM_038449670.1	XP_038305598.1
Mustela erminea	Short-tailed weasel	Chordata/ Mammalia/ Carnivoral/ Mustelidae/ Mustela	*ZFX*	NCBI	XM_032329779.1	XP_032185670.1
Loxodonta africana	African savanna elephant	Chordata/ Mammalia/ Proboscidea/ Elephantidae/ Loxodonta	*ZFX*	NCBI	XM_010595734.2	XP_010594036.1
Equus caballus	Horse	Chordata/ Mammalia/ Perissaodactyla/ Equidae/ Equus	*ZFX*	Ensembl	*ZFX*-201 ENSECAT00000034055.2	ENSECAP00000040431.1

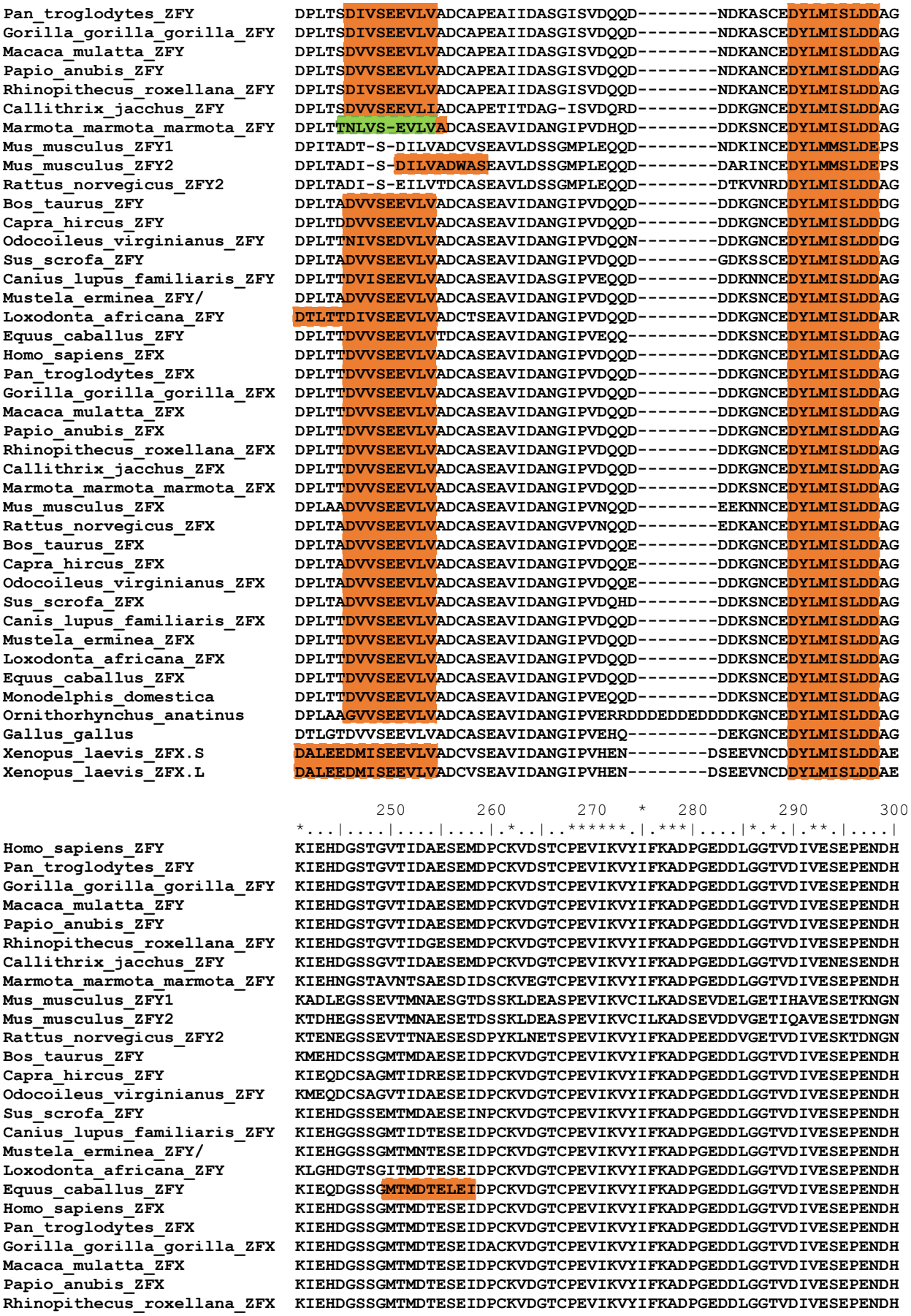
Table 3: Autosomal Zf* nucleotide and protein sequences collected from the NCBI search engine. Included in the table are the corresponding accession numbers for the species nucleotide and protein CDS used in the analysis. Gallus gallus and Xenopus laevis were only used for protein sequence analysis, so no DNA accession numbers are included for these species.

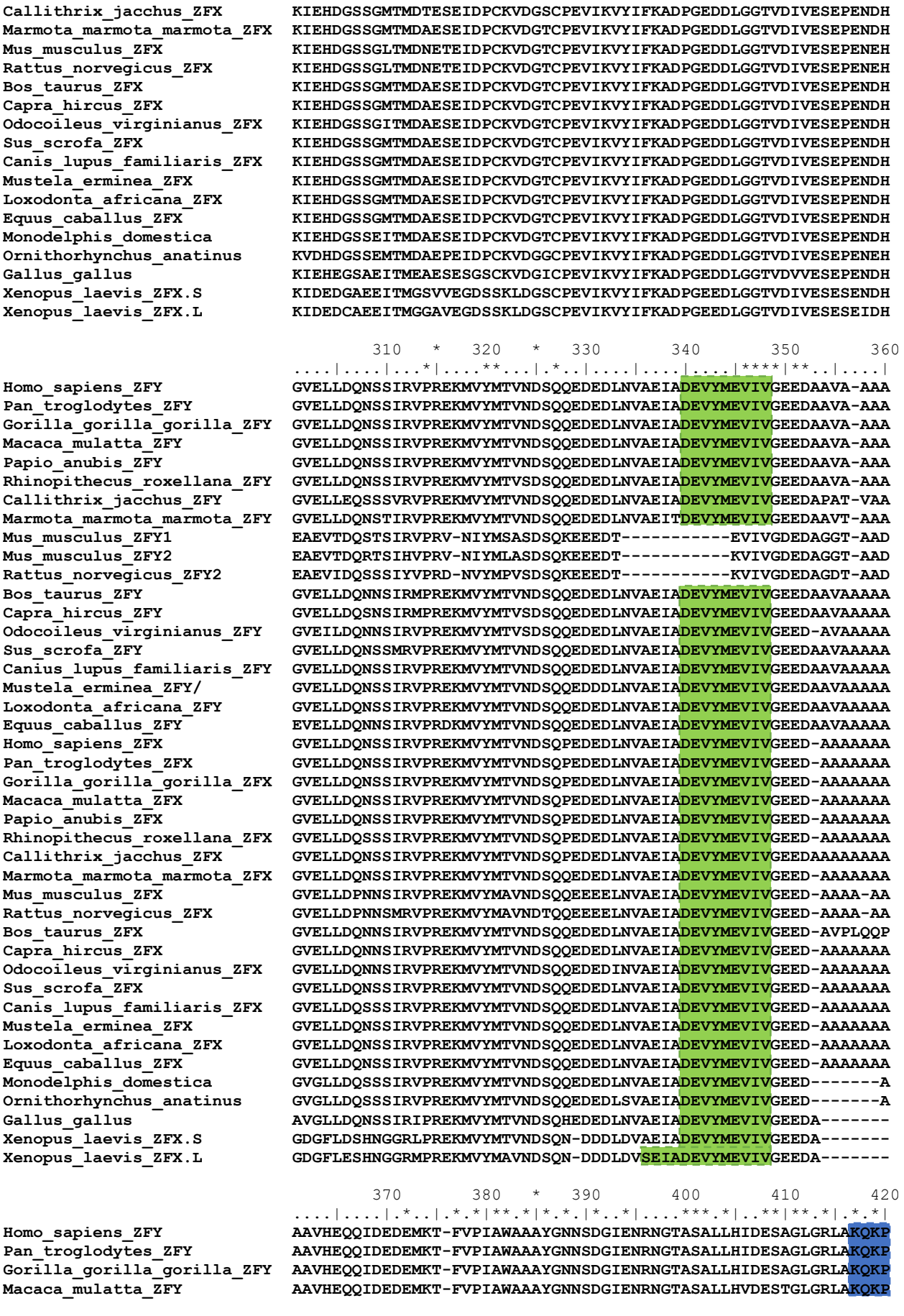
Binomial Nomenclature	Common Species Nomenclature	Taxonomic Classification	Protein name	Database for CDS sequence & Database for Protein	Database DNA Accession	Database Protein Accession
Ornithorhynchus anatinus	Platypus	Chordata/ Mammalia/ Monotremata/ Ornithorhynchidae/ Ornithorhynchus	*ZFX*	NCBI	XM_029079877.2	XP_028953710.1
Monodelphis domestica	Gray short-tailed opossum	Chordata/ Mammalia/ Didelphimorphia/ Didelphidae/ Monodelphis	*ZFX*	NCBI	XM_016433377.1	XP_016288863.1
Gallus gallus	Chicken	Chordata/ Aves/ Galliformes/ Phasiandiae Gallus	*ZFX*	NCBI	-	XP_015127980.1

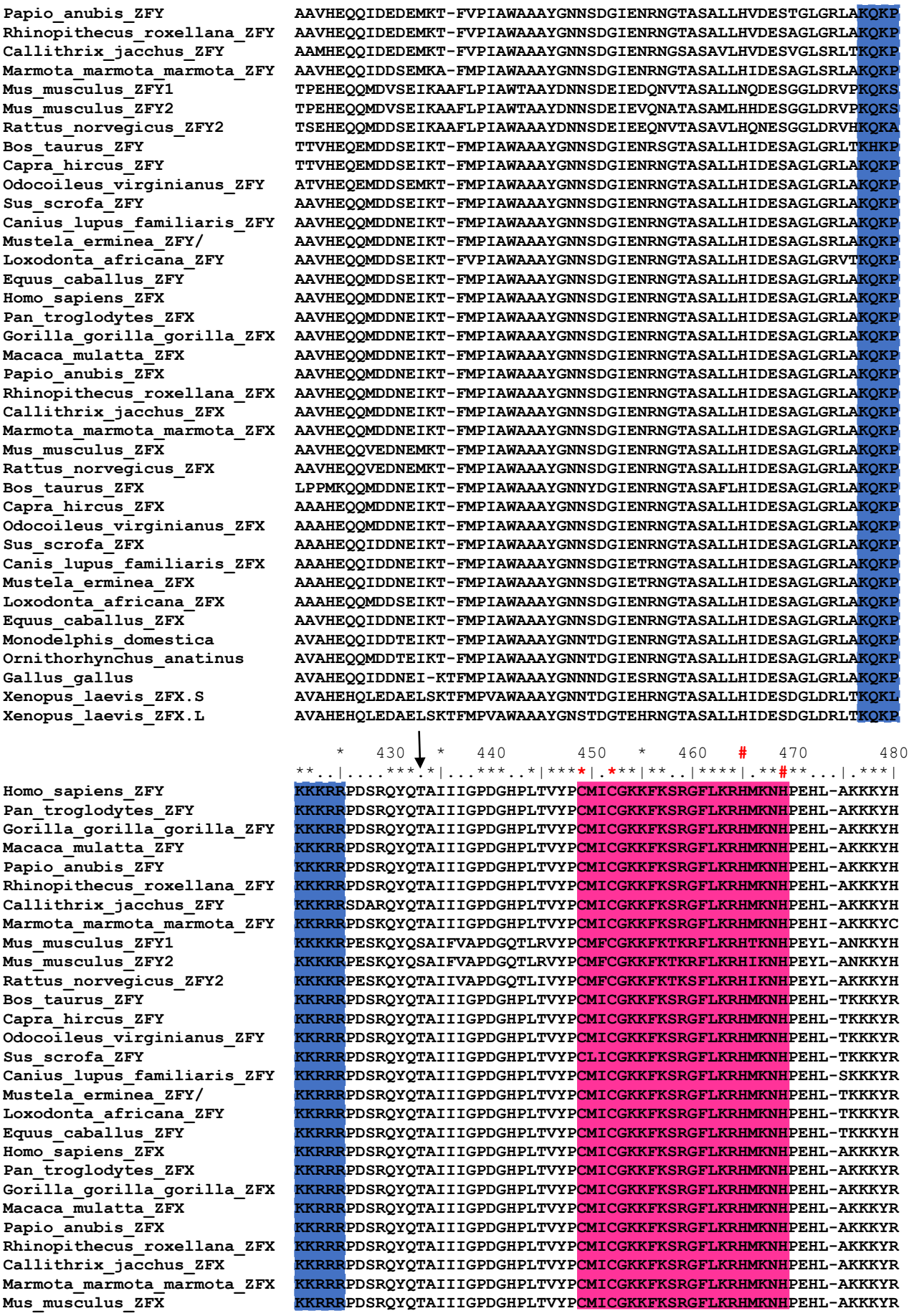
Xenopus laevis	African clawed frog	Chordata/ Amphibia/ Anura/ Pipidae/ Xenopus	*ZFX*.S	NCBI	-	NP_001081639.1
			ZFX.L	NCBI	-	XP_018101727.1

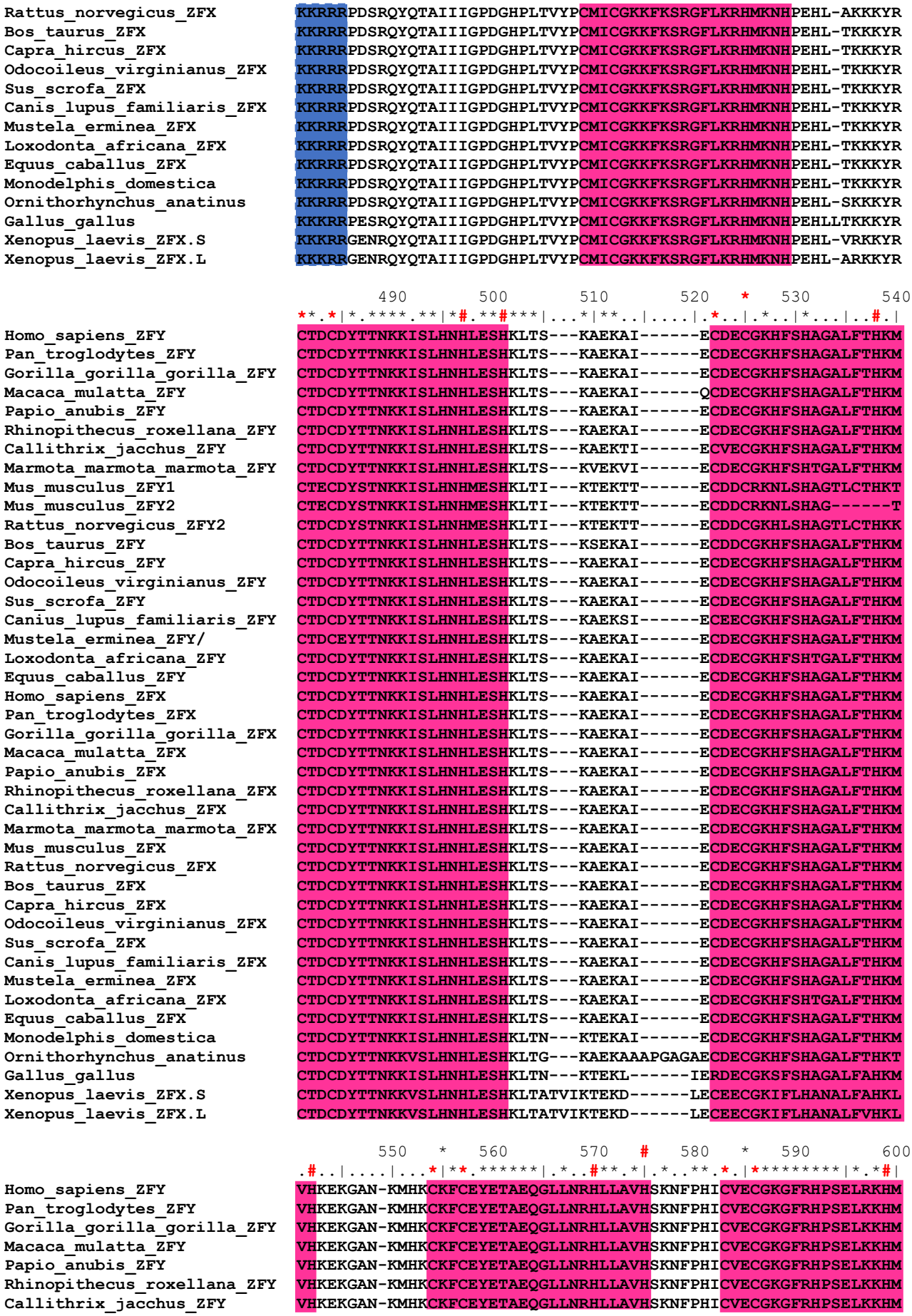

10 20 30 * 40 50 60
Homo_sapiens_ZFY MDE-DEFELQPQEPNSFFDGIGADATHMDGDQIVVEIQEAVFVSNIVDSDITVHNFVPDD
Pan_troglodytes_ZFY MDE-DEFELQPQEPNSFFDGIGADATHMDGDQIVVEIQEAVFVSNIVDSDIAVHNFVPDD
Gorilla_gorilla_gorilla_ZFY MDE-DEFELQPQEPNSFFDGIGADATHMDGDQIVVEIQEAVFVSNIVDSDITVHNFVPDD
Macaca_mulatta_ZFY MDE-DEFELQPQEPNSFFDGIGADATHMDGDQIVVEVQEAVFVSNIVDSDITVHNFVPDD
Papio_anubis_ZFY MDE-DEFELQPQEPNSFFDGIGADATHMDGDQIVVEVQEAVFVSNIVDSDITVHNFVPDD
Rhinopithecus_roxellana_ZFY MDE-DEFELQPQEPNSFFDGIGADATHMDGDQIVVEVQEAVFVSNIVDSDITVHNFVPDD
Callithrix_jacchus_ZFY MDE-DEFELQPQEPNSFFDGIGAGATHMDGDQIVVEVQETVFVSNIVDSDVTVHNFVPDD
Marmota_marmota_marmota_ZFY MDE-DEFELQPQEPNSFFDGIGTDSTHMDGDQIVVEVQETVFVSN---SDITVHNFVPDV
Mus_musculus_ZFY1 MDE-DEIELTPEEEKSFFDGIGADAVHMDSDQIVVEVQETVFLAN---SDVTVHNFVPDN
Mus_musculus_ZFY2 MDE-DEIELTPEEEKSLFDGIGADAVHMDSDQISVEVQETVFLSN---SDVTVHNFVPDD
Rattus_norvegicus_ZFY2 MDE-EEIELTPQEENSLFDGIGADAVHMDGDQIIVEVQETVFLSN---SDVTVHNFVPDD
Bos_taurus_ZFY MDE-DEFELQPQEPNSCFDGIGTDATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Capra_hircus_ZFY MDE-DELELQPQEPNSCFDGIGTDATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Odocoileus_virginianus_ZFY MDE-DELEIQPQEPNSCFDGIGTDATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Sus_scrofa_ZFY MDE-DELELQPQETNTFFDEIGADDTHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Canius_lupus_familiaris_ZFY MDE-DELALQPREPNSFFDGIGTDATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Mustela_erminea_ZFY/ MDE-DELELQPQEPNSFFDGIGTDATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Loxodonta_africana_ZFY MDE-DELELQPQEPNSFFDGIGADVTHMVGDQIVVEVQETVFVSDVVDSDITVHNFPPED
Equus_caballus_ZFY MDE-DELELRQQEPDSFFDGIGTDATHMDGDQIVVEVQETVFVSDVVDSDITVHNFAPDG
Homo_sapiens_ZFX MDE-DGLELQ-QEPNSFFDATGADGTHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Pan_troglodytes_ZFX MDE-DGLELQ-QEPNSFFDATGADGTHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Gorilla_gorilla_gorilla_ZFX MDE-DGLELQ-QEPNSFFDATGADGTHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Macaca_mulatta_ZFX MDE-DGLELQ-QEPNSFFDATGADGTHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Papio_anubis_ZFX MDE-DGLELQ-QEPNSFFDATGADGTHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Rhinopithecus_roxellana_ZFX MDE-DGLELQ-QEPNSFFDATGADGTHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Callithrix_jacchus_ZFX MDE-DGLELQPQEPNSFFDATGADGTHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Marmota_marmota_marmota_ZFX MDE-DGLELQPQEPNSFFDATGADATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Mus_musculus_ZFX MDE-DGLELQQQAPNSFFDATGAGATHMDGNQIVVEVQETVYVSDVVDSDITVHNYVPDD
Rattus_norvegicus_ZFX MDE-DGLELQPQAPNSFFDATGAGATHMDGNQIVVEVQETVYVSDVVDSDITVHNYVPDD
Bos_taurus_ZFX MDE-DGLELQPQEPNSFFDTTGADATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Capra_hircus_ZFX MDE-DGLELQPQEPNSFFDTTGADATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Odocoileus_virginianus_ZFX MDE-DGLELQPQEPNSFFDTTGADATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Sus_scrofa_ZFX MDE-DGLELQPQEPNSFFDATGADATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Canis_lupus_familiaris_ZFX MDE-DGLELQPQEPNSFFDATGADATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Mustela_erminea_ZFX MDE-DGLELQPQEPNSFFDATGADAAHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Loxodonta_africana_ZFX MDE-DGLELQPQGPNSFFDATGADATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Equus_caballus_ZFX MDE-DGLELQPQEPNSFFDATGADATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Monodelphis_domestica MDE-DGLELQPQEPNSFFDATGADATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Ornithorhynchus_anatinus MDE-DGLELQPHEPNSFFDATGAAASHMDGGQILVEVQETVFVSDVVDSDITVHNFVPDD
Gallus_gallus MDE-DGLELQPHEPNAFFDPTGADATHMDGDQIVVEVQETVFVSDVVDSDITVHNFVPDD
Xenopus_laevis_ZFX.S MEDVAELELQTTEPHAFFHASGVGERHLNGNEIIVEIQETVFVADG-DGNMAVQGFGPDE
Xenopus_laevis_ZFX.L MEDVAELELQTTEPHAFFHASGAGQRHLNGNEIIVEIQETVFVADG-DGNMAVQGFGPDE
70 * 80 90 100 110 120
Homo_sapiens_ZFY PDSVVIQDVVEDVVIEEDVQCSDILEEADVSENVIIPEQVLDSD-------VTEEVSLP
Pan_troglodytes_ZFY PDSVVIQDVIEDVVIEEDVQCSDILEEADVSENVIIPEQVLDSD-------VTEEVSLP
Gorilla_gorilla_gorilla_ZFY PDSVVIQDVIEDVVIEEDVQCSDILEEADVSENVIIPEQVLESD-------VTEEVSLP
Macaca_mulatta_ZFY PDSVVIQDVIEDVVIEEDVQCSDILEEADVSENVIIPEQVLDSD-------VTEEVSLP
Papio_anubis_ZFY PDSVVIQDVIEDVVIEEDVQCSDILEEADVSENVIIPEQVLDSD-------VTEEVSLP
Rhinopithecus_roxellana_ZFY PDSVVIQDVIEDVVIEEDVQCSDILEEADVSENVIIPEQVVDSD-------VTEEVSLP
Callithrix_jacchus_ZFY PDSVVIQDVIEDVVIED-VQCSDILEETDVSENVIIPEQVLDSD-------VTEEVSVS
Marmota_marmota_marmota_ZFY QDSVVIQDVIEDVVIED-VQCSDILEEADVSDSVIIPEQVLDSD-------VTREVSLA
Mus_musculus_ZFY1 PGSVIIQDVIENVLIED-VHCSHILEETDISDNVIIPEQVLNLG-------TAEEVSLA
Mus_musculus_ZFY2 PDSVIIQDVIENVLIED-VHCSHILEETDISDNVIIPEQVLDLD-------TAEEVSLA
Rattus_norvegicus_ZFY2 PDSVIIQDVIENVLIED-VHCSNILEETDISDNVIIPEQVLDLD-------TAEEVSLA
Bos_taurus_ZFY PDSVVIQDVIENVVIED-VQCSDILEEADVSENVIIPEQMLSSD-------VTEEVSLA
Capra_hircus_ZFY PDSVVIQDVIENVVIED-VQCSDILEEADVSENVIIPEQMLSSD-------VTEEVSLA
Odocoileus_virginianus_ZFY PDSVVIQDVIEDVVIED-VQCSDILEEADVSENVIIPDQVLSSD-------VTEEVSLA
Sus_scrofa_ZFY PDSVVIQDVIEDVVIED-VHCSDILEEADVSENVIIPEQVLASE-------VTEEVSLA
Canius_lupus_familiaris_ZFY PDSVVIQDVIEDVVIED-VHCSDILEEADVSENVIIPEQVLGSD-------VTEEVSLA
Mustela_erminea_ZFY/ PDSVVIQDVIEDVVIED-VHCSDILEEADISENVIIPEQVLDSD-------VTEEVSLA
Loxodonta_africana_ZFY PDSVVIQDVIEDVVIEN-VQCSDILEEADVSENVIIPEQVLESD-------ISEEVSLT
Equus_caballus_ZFY PDSVVIQDVIEDVVIED-VQCSDILEEADVSENVVIPEQVLDSD-------VTEEVSLA
Homo_sapiens_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD-------VTEEVSLA
Pan_troglodytes_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD-------VTEEVSLA
Gorilla_gorilla_gorilla_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD-------VTEEVSLA

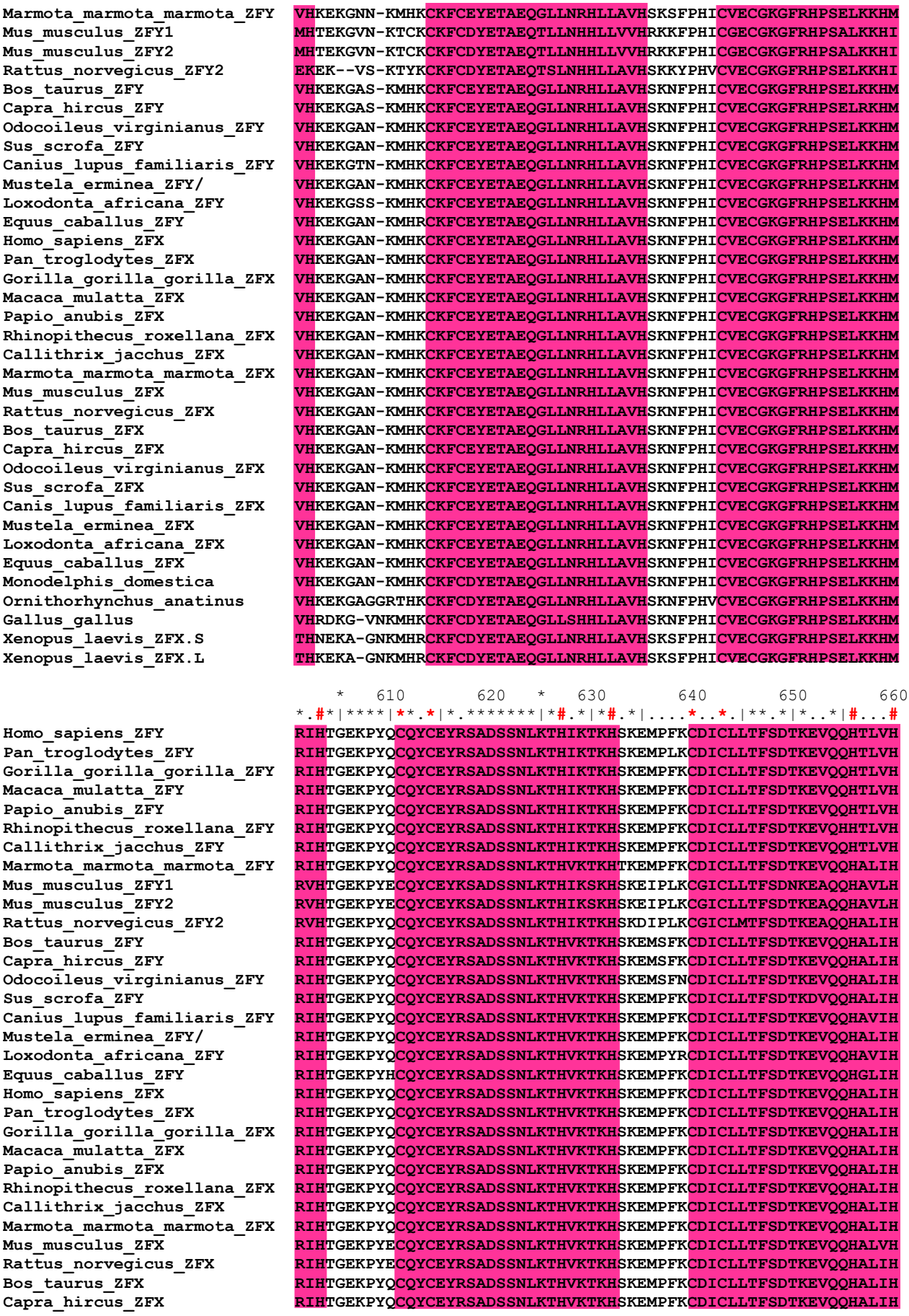

Macaca_mulatta_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD--------VTEEVSLA
Papio_anubis_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD--------VTEEVSLA
Rhinopithecus_roxellana_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD--------VTEEVSLA
Callithrix_jacchus_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD--------VTEEVSLA
Marmota_marmota_marmota_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD--------VTEEVSLA
Mus_musculus_ZFX PDSVVIQDVIEDVVIED-VQCTDIMDEADVSETVIIPEQVLDSD--------VTEEVSLT
Rattus_norvegicus_ZFX PDSVVIQDVIEDVVIED-VQCTDIMDEADVSETVIIPEQVLDSD--------VTEEVSLT
Bos_taurus_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD--------VTEEVSLA
Capra_hircus_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD--------VTEEVSLA
Odocoileus_virginianus_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD--------VTEEVSLA
Sus_scrofa_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD--------VTEEVSLA
Canis_lupus_familiaris_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD--------VTEEVSLA
Mustela_erminea_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD--------VTEEVSLA
Loxodonta_africana_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD--------VTEEVSLA
Equus_caballus_ZFX PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDSD--------VTEEVSLA
Monodelphis_domestica PDSVVIQDVIEDVVIED-VQCPDIMEEADVSETVIIPEQVLDTD--------VTEEVSLA
Ornithorhynchus_anatinus PDSVVIQDVIEDVVIED-VQCPDILDEADVSETVIIPEPVLGPE--------VPEEVSLA
Gallus_gallus PDSVVIQDVIEDVVIED-VQCPDIMEEPDVSETVIIPEQVLDTD--------VAEEVSLA
Xenopus_laevis_ZFX.S GDSVVIQDVIEDVVIED-VQCSDILDGGRVSEAVIIPEQVLEDEVGTGEEEQVLEEDSLT
Xenopus_laevis_ZFX.L GDSVVIQDVIEDVVIED-VQCSDILDGARVSEAVIIPEHVLEDEVGTGEEEQVLEEDSLT
130
140
150
160
170
180
Homo_sapiens_ZFY HCTVPDDVLASDITSTSMSMPEHVLTSESMHVCDIG------HVEHMVHDS-VVEAEIIT
Pan_troglodytes_ZFY HCTVPDDVLASDITSTSMSMPEHVLTSESMHVCDIE------HVEHMVHDS-VVEAEIIT
Gorilla_gorilla_gorilla_ZFY HCTVPDDVLASDITSTSTSMPEHVLTSESMHVCDIG------HVEHMVHDS-VVEAEIIT
Macaca_mulatta_ZFY HCTVPDDVLASDITSASMSMPEHVLTSESMHVCDIG------HVEHVVHDS-VVEAEIIT
Papio_anubis_ZFY HCTVPDDVLASDITSASISMPEHVLTSESMHVCDIG------HVEHVVHDS-VVEAEIIT
Rhinopithecus_roxellana_ZFY HCTVPDDVLASDITSASMSMPEHVLTSESMHVCDIG------HVEHVVHDS-VVEAEIIT
Callithrix_jacchus_ZFY HCTVPDDVLASDITSSSVSMPEHVLTSESMHVCDIG------HVEHVVRDN-VVEAEIIT
Marmota_marmota_marmota_ZFY HCTVPDDVLPSDITSTSMSMPEHVLTSESIHMSNVG------HVEHVVHDS-EVEAEIVT
Mus_musculus_ZFY1 QFLIP-DILTSGITSTSLTMPEHVLMSEAIHVSDVG------HFEQVIHDS-LVETEVIT
Mus_musculus_ZFY2 QFLIP-DILTSSITSTSLTMPEHVLMSEAIHVSNVG------HFEQVIHDS-LVEREIIT
Rattus_norvegicus_ZFY2 QFPIP-DILASSITSTSLTMPEHILMSEAIHVSDVG------HIEQVIHDS-LVETEVIT
Bos_taurus_ZFY HCTVPDDVLASDITSASMSMPEHVLTSESVHVSDVG------HVEHIVHGS-VVEAEIVT
Capra_hircus_ZFY HCTVPDDVLASDITSASMSMPEHVLTSESVHVSDVG------HVEHIVHGS-VVEAEIVT
Odocoileus_virginianus_ZFY HCTVPDDVLASDVTSASMCMPEHVLTSESVHVSDVG------HVEHIVHDS-VVEAEIVT
Sus_scrofa_ZFY HCTVPDDVLASDITSASISMPEQVLTSESIHVSE--------HIEH-IHNS-VVEAEIVT
Canius_lupus_familiaris_ZFY HCTVPDDVLASDITSASMSMPEHVLTSDSIHVSDVG------HVEHVVHDS-VVAAEIIT
Mustela_erminea_ZFY/ HCTVPDDVLASDITSASMSVPEHVLTSDSIHVSDIG------HVEHMVHDS-VVEAEIIT
Loxodonta_africana_ZFY HCTVPNDVLASDVTSASMSMPEHVLTHEPIRVPDVG------NVEHVVHDN-VVEAEIVT
Equus_caballus_ZFY HCTVPDDVLASDITSASMSMPEHVLTSESIHVSDVG------HVEHIVHDS-VVEAEIVT
Homo_sapiens_ZFX HCTVPDDVLASDITSASMSMPEHVLTGDSIHVSDVGHVGHVGHVEHVVHDS-VVEAEIVT
Pan_troglodytes_ZFX HCTVPDDVLASDITSASMSMPEHVLTGDSIHVSDIG---HVGHVEHVVHDS-VVEAEIVT
Gorilla_gorilla_gorilla_ZFX HCTVPDDVLASDITSASMSMPEHVLTGDSIHVSDVG---HVGHVEHVVHDS-VVEAEIVT
Macaca_mulatta_ZFX HCTVPDDVLASDITSASMSMPEHVLTGDSIHVSDVG---HVGHVEHVVHDS-VVEAEIVT
Papio_anubis_ZFX HCTVPDDVLASDITSASMSMPEHVLTGDSIHVSDVG---HVGHVEHVVHDS-VVEAEIVT
Rhinopithecus_roxellana_ZFX HCTVPDDVLASDITSASMSMPEHVLTGDSIHVSDVG---HVGHVEHVVHDS-VVEAEIVT
Callithrix_jacchus_ZFX HCTVPDDVLASDITSASMSMPEHVLTSESIHVSDVG---HVGHVEHVVHDS-VVEAEIVT
Marmota_marmota_marmota_ZFX HCTVPDDVLASDITSASMSMPEHVLTSESIHVSDVG------HVEHVVHDS-VVEAEIVT
Mus_musculus_ZFX HCTVPDDVLASDITSASISMPEHVLTSESIHVSDVG------HVEHVVHDS-VVEAEIVT
Rattus_norvegicus_ZFX HCTVPDDVLASDITSASISMPEHVLTSESIHVSDVS------HVEHVVHDS-VVEAEIVT
Bos_taurus_ZFX HCTVPDDVLASDITSASMSMPEHVLTSESIHVSDIG------HVEHVVHDS-VVEAEIVT
Capra_hircus_ZFX HCTVPDDVLASDITSASMSMPEHVLTSESIHVSDIG------HVEHVVHDS-VVEAEIVT
Odocoileus_virginianus_ZFX HCTVPDDVLASDITSASMSMPEHVLTSESIHVSDIG------HVEHIVHDS-VVEAEIVT
Sus_scrofa_ZFX HCTVPDDVLASDITSASMAMPEHVLTSESIHVSDVG------HVEHVVHDS-VVEAEIVT
Canis_lupus_familiaris_ZFX HCTVPDDVLASDITSASMSMPEHVLTSESIHVSDVG------HVEHVVHDS-VVEAEIVT
Mustela_erminea_ZFX HCTVPDDVLASDITSASMSMPEHVLTSESIHVSDVG------HVEHVVHDS-VVEAEIVT
Loxodonta_africana_ZFX HCTVPDDVLASDITSASMSVPEHVLTSESIHVPDVG------HVEHVVHDN-VVEAEIVT
Equus_caballus_ZFX HCTVPDDVLASDITSASMSMPEHVLTSESIHVSDVG------HVEHVVHDS-VVEAEIVT
Monodelphis_domestica HCTVPDDVLASDITTATMSIPEHVLTSDSMHVPDVG------HVEHVVHDN-VVEAEIVT
Ornithorhynchus_anatinus HCAVPEDVLAPDVPAAVAAVPEHVLAGEPVHIPPAAG--HVGHVEHVVHDG-VVDAEMVA
Gallus_gallus HCTVPDDVLASDITAEAMSIPEHVLTSESMHVPE------VGHVEHVVHDN-VEEADIVT
Xenopus_laevis_ZFX.S SCDVPDNVLDPELVDGELTIPD----------PE------TG--MHSVSGHVVIGEEITD
Xenopus_laevis_ZFX.L SCDVPDNVLDHELVDGELTLPD----------PE------TG--VHSVSGH-VVGEEITD
190
200
210
220
230
240
Homo_sapiens_ZFY DPLTSDIVSEEVLVADCAPEAVIDASGISVDQQD--------NDKASCEDYLMISLDDAG

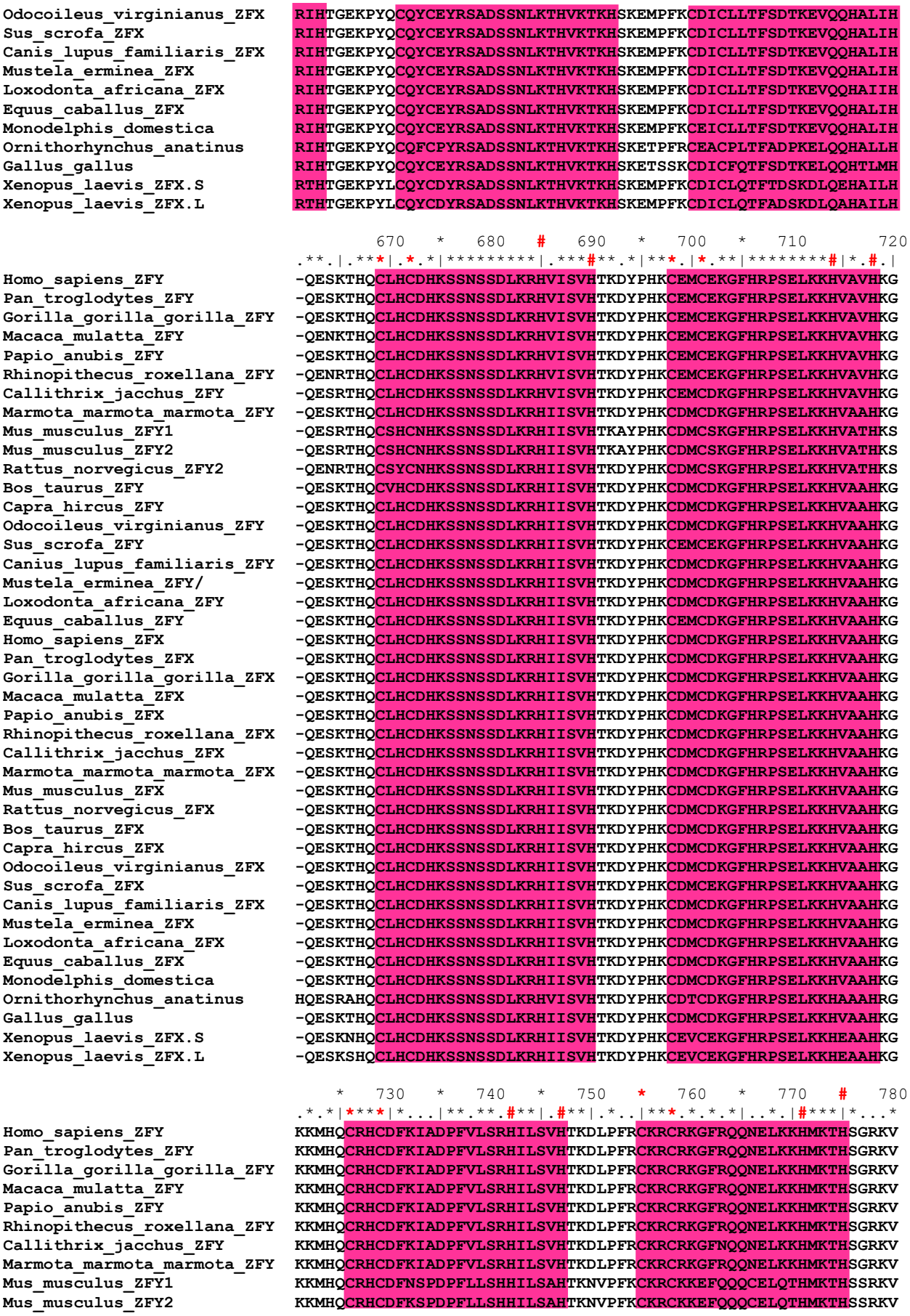

Pan_troglodytes_ZFY DPLTSDIVSEEVLVADCAPEAIIDASGISVDQQD--------NDKASCEDYLMISLDDAG
Gorilla_gorilla_gorilla_ZFY DPLTSDIVSEEVLVADCAPEAIIDASGISVDQQD--------NDKASCEDYLMISLDDAG
Macaca_mulatta_ZFY DPLTSDVVSEEVLVADCAPEAIIDASGISVDQQD--------NDKANCEDYLMISLDDAG
Papio_anubis_ZFY DPLTSDVVSEEVLVADCAPEAIIDASGISVDQQD--------NDKANCEDYLMISLDDAG
Rhinopithecus_roxellana_ZFY DPLTSDIVSEEVLVADCAPEAIIDASGISVDQQD--------NDKANCEDYLMISLDDAG
Callithrix_jacchus_ZFY DPLTSDVVSEEVLIADCAPETITDAG-ISVDQRD--------DDKGNCEDYLMISLDDAG
Marmota_marmota_marmota_ZFY DPLTTNLVS-EVLVADCASEAVIDANGIPVDHQD--------DDKSNCEDYLMISLDDAG
Mus_musculus_ZFY1 DPITADT-S-DILVADCVSEAVLDSSGMPLEQQD--------NDKINCEDYLMMSLDEPS
Mus_musculus_ZFY2 DPLTADI-S-DILVADWASEAVLDSSGMPLEQQD--------DARINCEDYLMMSLDEPS
Rattus_norvegicus_ZFY2 DPLTADI-S-EILVTDCASEAVLDSSGMPLEQQD--------DTKVNRDDYLMISLDDAG
Bos_taurus_ZFY DPLTADVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKGNCEDYLMISLDDDG
Capra_hircus_ZFY DPLTDDVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKGNCEDYLMISLDDDG
Odocoileus_virginianus_ZFY DPLTTNIVSEDVLVADCASEAVIDANGIPVDQQN--------DDKGNCEDYLMISLDDDG
Sus_scrofa_ZFY DPLTADVVSEEVLVADCASEAVIDANGIPVDQQD--------GDKSSCEDYLMISLDDAG
Canius_lupus_familiaris_ZFY DPLTTDVISEEVLVADCASEAVIDASGIPVEQQD--------DDKNNCEDYLMISLDDAG
Mustela_erminea_ZFY/ DPLTADVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKSNCEDYLMISLDDAG
Loxodonta_africana_ZFY DTLTTDIVSEEVLVADCTSEAVIDANGIPVDQQD--------DDKGNCEDYLMISLDDAR
Equus_caballus_ZFY DPLTTDVVSEEVLVTDCASEAVIDANGIPVEQQ---------DDKSNCEDYLMISLDDAG
Homo_sapiens_ZFX DPLTTDVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKGNCEDYLMISLDDAG
Pan_troglodytes_ZFX DPLTTDVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKGNCEDYLMISLDDAG
Gorilla_gorilla_gorilla_ZFX DPLTTDVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKGNCEDYLMISLDDAG
Macaca_mulatta_ZFX DPLTTDVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKGNCEDYLMISLDDAG
Papio_anubis_ZFX DPLTTDVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKGNCEDYLMISLDDAG
Rhinopithecus_roxellana_ZFX DPLTTDVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKGNCEDYLMISLDDAG
Callithrix_jacchus_ZFX DPLTTDVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKGNCEDYLMISLDDAG
Marmota_marmota_marmota_ZFX DPLTTDVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKSNCEDYLMISLDDAG
Mus_musculus_ZFX DPLAADVVSEEVLVADCASEAVIDANGIPVNQQD--------EEKNNCEDYLMISLDDAG
Rattus_norvegicus_ZFX DPLTADVVSEEVLVADCASEAVIDANGVPVNQQD--------EDKANCEDYLMISLDDAG
Bos_taurus_ZFX DPLTADVVSEEVLVADCASEAVIDANGIPVDQQE--------DDKGNCEDYLMISLDDAG
Capra_hircus_ZFX DPLTADVVSEEVLVADCASEAVIDANGIPVDQQE--------DDKGNCEDYLMISLDDAG
Odocoileus_virginianus_ZFX DPLTADVVSEEVLVADCASEAVIDANGIPVDQQE--------DDKGNCEDYLMISLDDAG
Sus_scrofa_ZFX DPLTADVVSEEVLVADCASEAVIDANGIPVDQHD--------DDKSNCEDYLMISLDDAG
Canis_lupus_familiaris_ZFX DPLTTDVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKSNCEDYLMISLDDAG
Mustela_erminea_ZFX DPLTTDVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKSNCEDYLMISLDDAG
Loxodonta_africana_ZFX DPLTTDVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKSNCEDYLMISLDDAG
Equus_caballus_ZFX DPLTTDVVSEEVLVADCASEAVIDANGIPVDQQD--------DDKSNCEDYLMISLDDAG
Monodelphis_domestica DPLTTDVVSEEVLVADCASEAVIDANGIPVEQQD--------DDKSNCEDYLMISLDDAG
Ornithorhynchus_anatinus DPLAAGVVSEEVLVADCASEAVIDANGIPVERRDDDEDDEDDDDKGNCEDYLMISLDDAG
Gallus_gallus DTLGTDVVSEEVLVADCASEAVIDANGIPVEHQ---------DEKGNCEDYLMISLDDAG
Xenopus_laevis_ZFX.S DALEEDMISEEVLVADCVSEAVIDANGIPVHEN--------DSEEVNCDDYLMISLDDAE
Xenopus_laevis_ZFX.L DALEEDMISEEVLVADCVSEAVIDANGIPVHEN--------DSEEVNCDDYLMISLDDAE
250 260 270 * 280 290 300
...|....|....|....|...|..******.|.***|....|*.*.|.**.|....|
Homo_sapiens_ZFY KIEHDGSTGVTIDAESEMDPCKVDSTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Pan_troglodytes_ZFY KIEHDGSTGVTIDAESEMDPCKVDSTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Gorilla_gorilla_gorilla_ZFY KIEHDGSTGVTIDAESEMDPCKVDSTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Macaca_mulatta_ZFY KIEHDGSTGVTIDAESEMDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Papio_anubis_ZFY KIEHDGSTGVTIDAESEMDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Rhinopithecus_roxellana_ZFY KIEHDGSTGVTIDGESEMDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Callithrix_jacchus_ZFY KIEHDGSSGVTIDAESEMDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVENESENDH
Marmota_marmota_marmota_ZFY KIEHNGSTAVNTSAESDIDSCKVEGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Mus_musculus_ZFY1 KADLEGSSEVTMNAESGTDSSKLDEASPEVIKVCILKADSEVDELGETIHAVESETKNGN
Mus_musculus_ZFY2 KTDHEGSSEVTMNAESETDSSKLDEASPEVIKVCILKADSEVDDVGETIQAVESETDNGN
Rattus_norvegicus_ZFY2 KTENEGSSEVTTNAESESDPYKLNETSPEVIKVYIFKADPEEDDVGETVDIVESKTDNGN
Bos_taurus_ZFY KMEHDCSSGMTMDAESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Capra_hircus_ZFY KIEQDCSAGMTIDRESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Odocoileus_virginianus_ZFY KMEQDCSAGVTIDAESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Sus_scrofa_ZFY KIEHDGSSEMTMDAESEINPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Canius_lupus_familiaris_ZFY KIEHGGSSGMTIDTESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Mustela_erminea_ZFY/ KIEHGGSSGMTMNTESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Loxodonta_africana_ZFY KLGHDGTSGITMDTESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Equus_caballus_ZFY KIEQDGSSGMTMDTELEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Homo_sapiens_ZFX KIEHDGSSGMTMDTESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Pan_troglodytes_ZFX KIEHDGSSGMTMDTESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Gorilla_gorilla_gorilla_ZFX KIEHDGSSGMTMDTESEIDACKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Macaca_mulatta_ZFX KIEHDGSSGMTMDTESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Papio_anubis_ZFX KIEHDGSSGMTMDTESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Rhinopithecus_roxellana_ZFX KIEHDGSSGMTMDTESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH

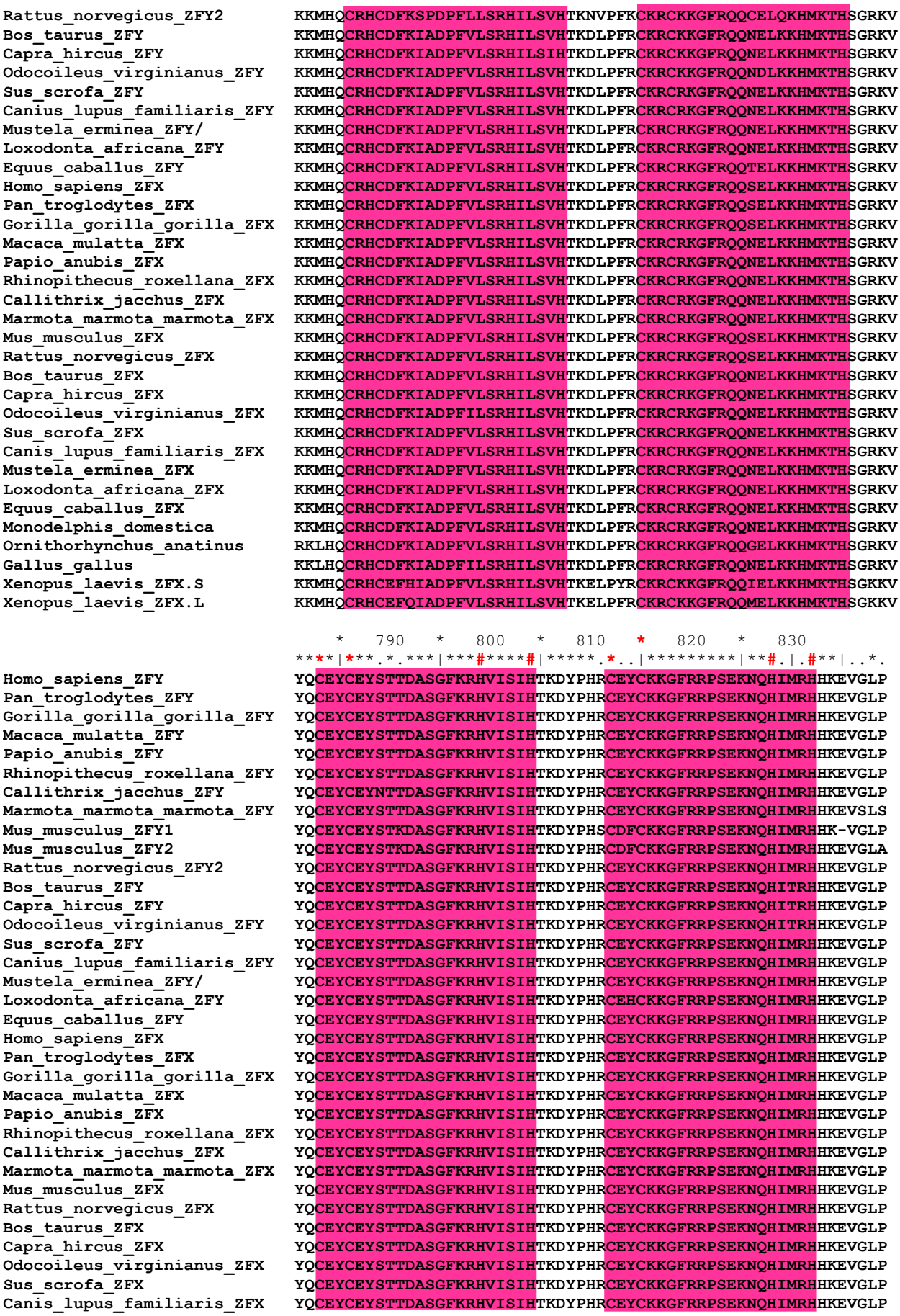

Callithrix_jacchus_ZFX KIEHDGSSGMTMDTESEIDPCKVDGSCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Marmota_marmota_marmota_ZFX KIEHDGSSGMTMDAESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Mus_musculus_ZFX KIEHDGSSGLTMDNETEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENEH
Rattus_norvegicus_ZFX KIEHDGSSGLTMDNETEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENEH
Bos_taurus_ZFX KIEHDGSSGMTMDAESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Capra_hircus_ZFX KIEHDGSSGMTMDAESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Odocoileus_virginianus_ZFX KIEHDGSSGITMDAESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Sus_scrofa_ZFX KIEHDGSSGMTMDAESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Canis_lupus_familiaris_ZFX KIEHDGSSGMTMDAESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Mustela_erminea_ZFX KIEHDGSSGMTMDAESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Loxodonta_africana_ZFX KIEHDGSSGMTMDAESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Equus_caballus_ZFX KIEHDGSSGMTMDAESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Monodelphis_domestica KIEHDGSSEITMDAESEIDPCKVDGTCPEVIKVYIFKADPGEDDLGGTVDIVESEPENDH
Ornithorhynchus_anatinus KVDHDGSSEMTMDAEPEIDPCKVDGGCPEVIKVYIFKADPGEDDLGGTVDIVESEPENEH
Gallus_gallus KIEHEGSAEITMEAESESGSCKVDGICPEVIKVYIFKADPGEDDLGGTVDVVESEPENDH
Xenopus_laevis_ZFX.S KIDEDGAEEITMGSVVEGDSSKLDGSCPEVIKVYIFKADPGEEDLGGTVDIVESESENDH
Xenopus_laevis_ZFX.L KIDEDCAEEITMGGAVEGDSSKLDGSCPEVIKVYIFKADPGEEDLGGTVDIVESESEIDH
310 * 320 * 330 340 350 360
Homo_sapiens_ZFY GVELLDQNSSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEEDAAVA-AAA
Pan_troglodytes_ZFY GVELLDQNSSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEEDAAVA-AAA
Gorilla_gorilla_gorilla_ZFY GVELLDQNSSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEEDAAVA-AAA
Macaca_mulatta_ZFY GVELLDQNSSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEEDAAVA-AAA
Papio_anubis_ZFY GVELLDQNSSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEEDAAVA-AAA
Rhinopithecus_roxellana_ZFY GVELLDQNSSIRVPREKMVYMTVSDSQQEDEDLNVAEIADEVYMEVIVGEEDAAVA-AAA
Callithrix_jacchus_ZFY GVELLEQSSSVRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEEDAPAT-VAA
Marmota_marmota_marmota_ZFY GVELLDQNSTIRVPREKMVYMTVNDSQQEDEDLNVAEITDEVYMEVIVGEEDAAVT-AAA
Mus_musculus_ZFY1 EAEVTDQSTSIRVPRV-NIYMSASDSQKEEEDT----------EVIVGDEDAGGT-AAD
Mus_musculus_ZFY2 EAEVTDQRTSIHVPRV-NIYMLASDSQKEEEDT----------KVIVGDEDAGGT-AAD
Rattus_norvegicus_ZFY2 EAEVIDQSSSIYVPRD-NVYMPVSDSQKEEEDT----------KVIVGDEDAGDT-AAD
Bos_taurus_ZFY GVELLDQNNSIRMPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEEDAAVAAAAA
Capra_hircus_ZFY GVELLDQSNSIRMPREKMVYMTVSDSQQEDEDLNVAEIADEVYMEVIVGEEDAAVAAAAA
Odocoileus_virginianus_ZFY GVEILDQNNSIRVPREKMVYMTVSDSQQEDEDLNVAEIADEVYMEVIVGEED-AVAAAAA
Sus_scrofa_ZFY GVELLDQNSSMRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEEDAAVAAAAA
Canius_lupus_familiaris_ZFY GVELLDQNSSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEEDAAVAAAAA
Mustela_erminea_ZFY/ GVELLDQNSSIRVPREKMVYMTVNDSQQEDDDLNVAEIADEVYMEVIVGEEDAAVAAAAA
Loxodonta_africana_ZFY GVELLDQNSSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEEDAAVAAAAA
Equus_caballus_ZFY EVELLDQNNSIRVPRDKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEEDAAVAAAAA
Homo_sapiens_ZFX GVELLDQNSSIRVPREKMVYMTVNDSQPEDEDLNVAEIADEVYMEVIVGEED-AAAAAAA
Pan_troglodytes_ZFX GVELLDQNSSIRVPREKMVYMTVNDSQPEDEDLNVAEIADEVYMEVIVGEED-AAAAAAA
Gorilla_gorilla_gorilla_ZFX GVELLDQNSSIRVPREKMVYMTVNDSQPEDEDLNVAEIADEVYMEVIVGEED-AAAAAAA
Macaca_mulatta_ZFX GVELLDQNSSIRVPREKMVYMTVNDSQPEDEDLNVAEIADEVYMEVIVGEED-AAAAAAA
Papio_anubis_ZFX GVELLDQNSSIRVPREKMVYMTVNDSQPEDEDLNVAEIADEVYMEVIVGEED-AAAAAAA
Rhinopithecus_roxellana_ZFX GVELLDQSSSIRVPREKMVYMTVNDSQPEDEDLNVAEIADEVYMEVIVGEED-AAAAAAA
Callithrix_jacchus_ZFX GVELLDQNSSIRVPREKMVYMTVNDSQPEDEDLNVAEIADEVYMEVIVGEEDAAAAAAAA
Marmota_marmota_marmota_ZFX GVELLDQNSSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEED-AAAAAAA
Mus_musculus_ZFX GVELLDPNNSIRVPREKMVYMAVNDSQQEEEELNVAEIADEVYMEVIVGEED-AAAA-AA
Rattus_norvegicus_ZFX GVELLDPNNSMRVPREKMVYMAVNDTQQEEEELNVAEIADEVYMEVIVGEED-AAAA-AA
Bos_taurus_ZFX GVELLDQNNSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEED-AVPLQQP
Capra_hircus_ZFX GVELLDQNNSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEED-AAAAAAA
Odocoileus_virginianus_ZFX GVELLDQNNSIRVPREKMVYMTVNDSQQEDEDINVAEIADEVYMEVIVGEED-AAAAAAA
Sus_scrofa_ZFX GVELLDQNSSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEED-AAAAAAA
Canis_lupus_familiaris_ZFX GVELLDQSSSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEED-AAAAAAA
Mustela_erminea_ZFX GVELLDQNSSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEED-AAAAAAA
Loxodonta_africana_ZFX GVELLDQNSSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEED-AAAAAAA
Equus_caballus_ZFX GVELLDQNSSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEED-AAAAAAA
Monodelphis_domestica GVGLLDQSSSIRVPREKMVYMTVNDSQQEDEDLNVAEIADEVYMEVIVGEED-------A
Ornithorhynchus_anatinus GVGLLDQSSSIRVPREKMVYMTVNDSQQEDEDLSVAEIADEVYMEVIVGEED-------A
Gallus_gallus AVGLLDQNSSIRIPREKMVYMTVNDSQHEDEDLNVAEIADEVYMEVIVGEEDA-------
Xenopus_laevis_ZFX.S GDGFLDSHNGGRLPREKMVYMTVNDSQN-DDDLDVAEIADEVYMEVIVGEEDA-------
Xenopus_laevis_ZFX.L GDGFLESHNGGRMPREKMVYMAVNDSQN-DDDLDVSEIADEVYMEVIVGEEDA-------
370 380 * 390 400 410 420
Homo_sapiens_ZFY AAVHEQQIDEDEMKT-FVPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Pan_troglodytes_ZFY AAVHEQQIDEDEMKT-FVPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Gorilla_gorilla_gorilla_ZFY AAVHEQQIDEDEMKT-FVPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Macaca_mulatta_ZFY AAVHEQQIDEDEMKT-FVPIAWAAAYGNNSDGIENRNGTASALLHVDESTGLGRLAKQKP


Papio_anubis_ZFY AAVHEQQIDEDEMKT-FVPIAWAAAYGNNSDGIENRNGTASALLHVDESTGLGRLAKQKP
Rhinopithecus_roxellana_ZFY AAVHEQQIDEDEMKT-FVPIAWAAAYGNNSDGIENRNGTASALLHVDESAGLGRLAKQKP
Callithrix_jacchus_ZFY AAMHEQQIDEDEMKT-FVPIAWAAAYGNNSDGIENRNGSASAVLHVDESVGLSRLTKQKP
Marmota_marmota_marmota_ZFY AAVHEQQIDDSEMKA-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLSRLAKQKP
Mus_musculus_ZFY1 TPEHEQQMDVSEIKAAFLPIAWTAAYDNNSDEIEDQNVTASALLNQDESGGLDRVPKQKS
Mus_musculus_ZFY2 TPEHEQQMDVSEIKAAFLPIAWTAAYDNNSDEIEVQNATASAMLHHDESGGLDRVPKQKS
Rattus_norvegicus_ZFY2 TSEHEQQMDDSEIKAAFLPIAWAAAYDNNSDEIEEQNVTASAVLHQNESGGLDRVHKQKA
Bos_taurus_ZFY TTVHEQEMDDSEIKT-FMPIAWAAAYGNNSDGIENRSGTASALLHIDESAGLGRLTKHKP
Capra_hircus_ZFY TTVHEQEMDDSEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Odocoileus_virginianus_ZFY ATVHEQEMDDSEMKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Sus_scrofa_ZFY AAVHEQQMDDSEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Canius_lupus_familiaris_ZFY AAVHEQQMDDNEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Mustela_erminea_ZFY/ AAVHEQQMDDNEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLSRLAKQKP
Loxodonta_africana_ZFY AAVHEQQMDDSEIKT-FVPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRVTKQKP
Equus_caballus_ZFY AAVHEQQMDDSEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Homo_sapiens_ZFX AAVHEQQMDDNEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Pan_troglodytes_ZFX AAVHEQQMDDNEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Gorilla_gorilla_gorilla_ZFX AAVHEQQMDDNEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Macaca_mulatta_ZFX AAVHEQQMDDNEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Papio_anubis_ZFX AAVHEQQMDDNEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Rhinopithecus_roxellana_ZFX AAVHEQQMDDNEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Callithrix_jacchus_ZFX AAVHEQQMDDNEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Marmota_marmota_marmota_ZFX AAVHEQQMDDNEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Mus_musculus_ZFX AAVHEQQVEDNEMKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Rattus_norvegicus_ZFX AAVHEQQVEDNEMKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Bos_taurus_ZFX LPPMKQQMDDNEIKT-FMPIAWAAAYGNNYDGIENRNGTASAFLHIDESAGLGRLAKQKP
Capra_hircus_ZFX AAAHEQQMDDNEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Odocoileus_virginianus_ZFX AAAHEQQMDDNEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Sus_scrofa_ZFX AAAHEQQIDDNEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Canis_lupus_familiaris_ZFX AAAHEQQIDDNEIKT-FMPIAWAAAYGNNSDGIETRNGTASALLHIDESAGLGRLAKQKP
Mustela_erminea_ZFX AAAHEQQIDDNEIKT-FMPIAWAAAYGNNSDGIETRNGTASALLHIDESAGLGRLAKQKP
Loxodonta_africana_ZFX AAAHEQQMDDSEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Equus_caballus_ZFX AAVHEQQIDDNEIKT-FMPIAWAAAYGNNSDGIENRNGTASALLHIDESAGLGRLAKQKP
Monodelphis_domestica AVAHEQQIDDTEIKT-FMPIAWAAAYGNNTDGIENRNGTASALLHIDESAGLGRLAKQKP
Ornithorhynchus_anatinus AVAHEQQMDDTEIKT-FMPIAWAAAYGNNTDGIENRNGTASALLHIDESAGLGRLAKQKP
Gallus_gallus AVAHEQQIDDNEI-KTFMPIAWAAAYGNNNDGIESRNGTASALLHIDESAGLGRLAKQKP
Xenopus_laevis_ZFX.S AVAHEHQLEDAELSKTFMPVAWAAAYGNNTDGIEHRNGTASALLHIDESDGLDRLTKQKL
Xenopus_laevis_ZFX.L AVAHEHQLEDAELSKTFMPVAWAAAYGNSTDGTEHRNGTASALLHIDESDGLDRLTKQKP
* 430 * 440 450 * 460 # 470 480
Homo_sapiens_ZFY KKKRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYH
Pan_troglodytes_ZFY KKKRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYH
Gorilla_gorilla_gorilla_ZFY KKKRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYH
Macaca_mulatta_ZFY KKKRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYH
Papio_anubis_ZFY KKKRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYH
Rhinopithecus_roxellana_ZFY KKKRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYH
Callithrix_jacchus_ZFY KKKRRSDARQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYH
Marmota_marmota_marmota_ZFY KKRRRPDSKQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHI-AKKKYC
Mus_musculus_ZFY1 KKKKRPESKQYQSAIFVAPDGQTLRVYPCMFCGKKFKTKRFLKRHTKNHPEYL-ANKKYH
Mus_musculus_ZFY2 KKKKRPESKQYQSAIFVAPDGQTLRVYPCMFCGKKFKTKRFLKRHIKNHPEYL-ANKKYH
Rattus_norvegicus_ZFY2 KKKKRPESKQYQTAIIVAPDGQTLIVYPCMFCGKKFKTKSFLKRHIKNHPEYL-AKKKYH
Bos_taurus_ZFY KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Capra_hircus_ZFY KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Odocoileus_virginianus_ZFY KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Sus_scrofa_ZFY KKRRRPDSRQYQTAIIIGPDGHPLTVYPCLICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Canius_lupus_familiaris_ZFY KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-SKKKYR
Mustela_erminea_ZFY/ KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Loxodonta_africana_ZFY KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Equus_caballus_ZFY KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYH
Homo_sapiens_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYR
Pan_troglodytes_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYR
Gorilla_gorilla_gorilla_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYR
Macaca_mulatta_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYR
Papio_anubis_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYR
Rhinopithecus_roxellana_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYR
Callithrix_jacchus_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYR
Marmota_marmota_marmota_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYR
Mus_musculus_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYR


Rattus_norvegicus_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-AKKKYR
Bos_taurus_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Capra_hircus_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Odocoileus_virginianus_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Sus_scrofa_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Canis_lupus_familiaris_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Mustela_erminea_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Loxodonta_africana_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Equus_caballus_ZFX KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Monodelphis_domestica KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-TKKKYR
Ornithorhynchus_anatinus KKRRRPDSRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-SKKKYR
Gallus_gallus KKKRRPESRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHLLTKKKYR
Xenopus_laevis_ZFX.S KKKRRGENRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-VRKKYR
Xenopus_laevis_ZFX.L KKKRRGENRQYQTAIIIGPDGHPLTVYPCMICGKKFKSRGFLKRHMKNHPEHL-ARKKYR
490 500 510 520 530 540
Homo_sapiens_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Pan_troglodytes_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Gorilla_gorilla_gorilla_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Macaca_mulatta_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------QCDECGKHFSHAGALFTHKM
Papio_anubis_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Rhinopithecus_roxellana_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Callithrix_jacchus_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KAEKTI------ECVECGKHFSHAGALFTHKM
Marmota_marmota_marmota_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KVEKVI------ECDECGKHFSHTGALFTHKM
Mus_musculus_ZFY1 CTECDYSTNKKISLHNHMESHKLTI---KTEKTT------ECDDCRKNLSHAGTLCTHKT
Mus_musculus_ZFY2 CTECDYSTNKKISLHNHMESHKLTI---KTEKTT------ECDDCRKNLSHAG------T
Rattus_norvegicus_ZFY2 CTDCDYSTNKKISLHNHMESHKLTI---KTEKTT------ECDDCGKHLSHAGTLCTHKK
Bos_taurus_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KSEKAI------ECDDCGKHFSHAGALFTHKM
Capra_hircus_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Odocoileus_virginianus_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Sus_scrofa_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Canius_lupus_familiaris_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KAEKSI------ECEECGKHFSHAGALFTHKM
Mustela_erminea_ZFY/ CTDCEYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Loxodonta_africana_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHTGALFTHKM
Equus_caballus_ZFY CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Homo_sapiens_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Pan_troglodytes_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Gorilla_gorilla_gorilla_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Macaca_mulatta_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Papio_anubis_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Rhinopithecus_roxellana_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Callithrix_jacchus_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Marmota_marmota_marmota_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Mus_musculus_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Rattus_norvegicus_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Bos_taurus_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Capra_hircus_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Odocoileus_virginianus_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Sus_scrofa_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Canis_lupus_familiaris_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Mustela_erminea_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Loxodonta_africana_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHTGALFTHKM
Equus_caballus_ZFX CTDCDYTTNKKISLHNHLESHKLTS---KAEKAI------ECDECGKHFSHAGALFTHKM
Monodelphis_domestica CTDCDYTTNKKISLHNHLESHKLTN---KTEKAI------ECDECGKHFSHAGALFTHKM
Ornithorhynchus_anatinus CTDCDYTTNKKVSLHNHLESHKLTG---KAEKAAAPGAGAECDECGKHFSHAGALFTHKT
Gallus_gallus CTDCDYTTNKKISLHNHLESHKLTN---KTEKL------IERDECGKSFSHAGALFAHKM
Xenopus_laevis_ZFX.S CTDCDYTTNKKVSLHNHLESHKLTATVIKTEKD------LECEECGKIFLHANALFAHKL
Xenopus_laevis_ZFX.L CTDCDYTTNKKVSLHNHLESHKLTATVIKTEKD------LECEECGKIFLHANALFVHKL
550 560 570 580 590 600
Homo_sapiens_ZFY VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELRKHM
Pan_troglodytes_ZFY VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Gorilla_gorilla_gorilla_ZFY VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Macaca_mulatta_ZFY VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Papio_anubis_ZFY VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Rhinopithecus_roxellana_ZFY VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Callithrix_jacchus_ZFY VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM


Marmota_marmota_marmota_ZFY VHKEKGNN-KMHKCKFCEYETAEQGLLNRHLLAVHSKSFPHICVECGKGFRHPSELKKHM
Mus_musculus_ZFY1 MHTEKGVN-KTCKCKFCDYETAEQTLLNHHLLVVHRKKFPHICGECGKGFRHPSALKKHI
Mus_musculus_ZFY2 MHTEKGVN-KTCKCKFCDYETAEQTLLNHHLLVVHRKKFPHICGECGKGFRHPSALKKHI
Rattus_norvegicus_ZFY2 EKEK--VS-KTYKCKFCDYETAEQTSLNHHLLAVHSKKYPHVCVECGKGFRHPSELKKHI
Bos_taurus_ZFY VHKEKGAS-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Capra_hircus_ZFY VHKEKGAS-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELRKHM
Odocoileus_virginianus_ZFY VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Sus_scrofa_ZFY VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Canius_lupus_familiaris_ZFY VHKEKGTN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Mustela_erminea_ZFY/ VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Loxodonta_africana_ZFY VHKEKGSS-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Equus_caballus_ZFY VHKEKGAN-KMHRCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Homo_sapiens_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Pan_troglodytes_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Gorilla_gorilla_gorilla_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Macaca_mulatta_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Papio_anubis_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Rhinopithecus_roxellana_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Callithrix_jacchus_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Marmota_marmota_marmota_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Mus_musculus_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Rattus_norvegicus_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Bos_taurus_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Capra_hircus_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Odocoileus_virginianus_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Sus_scrofa_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Canis_lupus_familiaris_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Mustela_erminea_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Loxodonta_africana_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Equus_caballus_ZFX VHKEKGAN-KMHKCKFCEYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Monodelphis_domestica VHKEKGAN-KMHKCKFCDYETAEQGLLNRHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Ornithorhynchus_anatinus VHKEKGAGGRTHKCKFCDYETAEQGLLNRHLLAVHSKNFPHVCVECGKGFRHPSELKKHM
Gallus_gallus VHRDKG-VNKMHKCKFCDYETAEQGLLSHHLLAVHSKNFPHICVECGKGFRHPSELKKHM
Xenopus_laevis_ZFX.S THNEKA-GNKMHRCKFCDYETAEQGLLNRHLLAVHSKSFPHICVECGKGFRHPSELKKHM
Xenopus_laevis_ZFX.L THKEKA-GNKMHRCKFCDYETAEQGLLNRHLLAVHSKSFPHICVECGKGFRHPSELKKHM
* 610 620 * 630 640 650 660
.#|****|**.*|*.*******|*#.*|*#.*|....*..*.|**.*|*..*|#...#
Homo_sapiens_ZFY RIHTGEKPYQCQYCEYRSADSSNLKTHIKTKHSKEMPFKCDICLLTFSDTKEVQQHTLVH
Pan_troglodytes_ZFY RIHTGEKPYQCQYCEYRSADSSNLKTHIKTKHSKEMPLKCDICLLTFSDTKEVQQHTLVH
Gorilla_gorilla_gorilla_ZFY RIHTGEKPYQCQYCEYRSADSSNLKTHIKTKHSKEMPFKCDICLLTFSDTKEVQQHTLVH
Macaca_mulatta_ZFY RIHTGEKPYQCQYCEYRSADSSNLKTHIKTKHSKEMPFKCDICLLTFSDTKEVQQHTLVH
Papio_anubis_ZFY RIHTGEKPYQCQYCEYRSADSSNLKTHIKTKHSKEMPFKCDICLLTFSDTKEVQQHTLVH
Rhinopithecus_roxellana_ZFY RIHTGEKPYQCQYCEYRSADSSNLKTHIKTKHSKEMPFKCDICLLTFSDTKEVQHHTLVH
Callithrix_jacchus_ZFY RIHTGEKPYQCQYCEYRSADSSNLKTHIKTKHSKEMPFKCDICLLTFSDTKEVQQHTLVH
Marmota_marmota_marmota_ZFY RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHTKEMPFKCDICLLTFSDTKEVQQHALIH
Mus_musculus_ZFY1 RVHTGEKPYECQYCEYKSADSSNLKTHIKSKHSKEIPLKCGICLLTFSDNKEAQQHAVLH
Mus_musculus_ZFY2 RVHTGEKPYECQYCEYKSADSSNLKTHIKSKHSKEIPLKCGICLLTFSDTKEAQQHAVLH
Rattus_norvegicus_ZFY2 RVHTGEKPYQCQYCEYKSADSSNLKTHIKTKHSKDIPLKCGICLMTFSDTKEAQQHALIH
Bos_taurus_ZFY RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMSFKCDICLLTFSDTKEVQQHALIH
Capra_hircus_ZFY RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMSFKCDICLLTFSDTKEVQQHALIH
Odocoileus_virginianus_ZFY RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMSFNCDICLLTFSDTKEVQQHALIH
Sus_scrofa_ZFY RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKDVQQHALIH
Canius_lupus_familiaris_ZFY RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHAVIH
Mustela_erminea_ZFY/ RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Loxodonta_africana_ZFY RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPYRCDICLLTFSDTKEVQQHAVIH
Equus_caballus_ZFY RIHTGEKPYHCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHGLIH
Homo_sapiens_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Pan_troglodytes_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Gorilla_gorilla_gorilla_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Macaca_mulatta_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Papio_anubis_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Rhinopithecus_roxellana_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Callithrix_jacchus_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Marmota_marmota_marmota_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Mus_musculus_ZFX RIHTGEKPYECQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALVH
Rattus_norvegicus_ZFX RIHTGEKPYECQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Bos_taurus_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Capra_hircus_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH


Odocoileus_virginianus_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Sus_scrofa_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Canis_lupus_familiaris_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Mustela_erminea_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Loxodonta_africana_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHAIIH
Equus_caballus_ZFX RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCDICLLTFSDTKEVQQHALIH
Monodelphis_domestica RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKEMPFKCEICLLTFSDTKEVQQHALIH
Ornithorhynchus_anatinus RIHTGEKPYQCQFCPYRSADSSNLKTHVKTKHSKETPFRCEACPLTFADPKELQQHALLH
Gallus_gallus RIHTGEKPYQCQYCEYRSADSSNLKTHVKTKHSKETSSKCDICFQTFSDTKELQQHTLMH
Xenopus_laevis_ZFX.S RTHTGEKPYLCQYCDYRSADSSNLKTHVKTKHSKEMPFKCDICLQTFTDSKDLQEHAILH
Xenopus_laevis_ZFX.L RTHTGEKPYLCQYCDYRSADSSNLKTHVKTKHSKEMPFKCDICLQTFADSKDLQAHAILH
670 * 680 # 690 * 700 * 710 720
Homo_sapiens_ZFY -QESKTHQCLHCDHKSSNSSDLKRHVISVHTKDYPHKCEMCEKGFHRPSELKKHVAVHKG
Pan_troglodytes_ZFY -QESKTHQCLHCDHKSSNSSDLKRHVISVHTKDYPHKCEMCEKGFHRPSELKKHVAVHKG
Gorilla_gorilla_gorilla_ZFY -QESKTHQCLHCDHKSSNSSDLKRHVISVHTKDYPHKCEMCEKGFHRPSELKKHVAVHKG
Macaca_mulatta_ZFY -QENKTHQCLHCDHKSSNSSDLKRHVISVHTKDYPHKCEMCEKGFHRPSELKKHVAVHKG
Papio_anubis_ZFY -QESKTHQCLHCDHKSSNSSDLKRHVISVHTKDYPHKCEMCEKGFHRPSELKKHVAVHKG
Rhinopithecus_roxellana_ZFY -QENRTHQCLHCDHKSSNSSDLKRHVISVHTKDYPHKCEMCEKGFHRPSELKKHVAVHKG
Callithrix_jacchus_ZFY -QESRTHQCLHCDHKSSNSSDLKRHVISVHTKDYPHKCEMCDKGFHRPSELKKHVAVHKG
Marmota_marmota_marmota_ZFY -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Mus_musculus_ZFY1 -QESRTHQCSHCNHKSSNSSDLKRHIISVHTKAYPHKCDMCSKGFHRPSELKKHVATHKS
Mus_musculus_ZFY2 -QESRTHQCSHCNHKSSNSSDLKRHIISVHTKAYPHKCDMCSKGFHRPSELKKHVATHKS
Rattus_norvegicus_ZFY2 -QENRTHQCSYCNHKSSNSSDLKRHIISVHTKDYPHKCDMCSKGFHRPSELKKHVATHKS
Bos_taurus_ZFY -QESKTHQCVHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Capra_hircus_ZFY -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Odocoileus_virginianus_ZFY -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Sus_scrofa_ZFY -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCEMCEKGFHRPSELKKHVAAHKG
Canius_lupus_familiaris_ZFY -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Mustela_erminea_ZFY/ -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Loxodonta_africana_ZFY -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Equus_caballus_ZFY -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCEMCDKGFHRPSELKKHVAAHKG
Homo_sapiens_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Pan_troglodytes_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Gorilla_gorilla_gorilla_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Macaca_mulatta_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Papio_anubis_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Rhinopithecus_roxellana_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Callithrix_jacchus_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Marmota_marmota_marmota_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Mus_musculus_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Rattus_norvegicus_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Bos_taurus_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Capra_hircus_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Odocoileus_virginianus_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Sus_scrofa_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCEKGFHRPSELKKHVAAHKG
Canis_lupus_familiaris_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Mustela_erminea_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Loxodonta_africana_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Equus_caballus_ZFX -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Monodelphis_domestica -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Ornithorhynchus_anatinus HQESRAHQCLHCDHKSSNSSDLKRHVISVHTKDYPHKCDTCDKGFHRPSELKKHAAAHRG
Gallus_gallus -QESKTHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCDMCDKGFHRPSELKKHVAAHKG
Xenopus_laevis_ZFX.S -QESKNHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCEVCEKGFHRPSELKKHEAAHKG
Xenopus_laevis_ZFX.L -QESKSHQCLHCDHKSSNSSDLKRHIISVHTKDYPHKCEVCEKGFHRPSELKKHEAAHKG
* 730 * 740 * 750 * 760 * 770 # 780
Homo_sapiens_ZFY KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Pan_troglodytes_ZFY KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Gorilla_gorilla_gorilla_ZFY KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Macaca_mulatta_ZFY KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Papio_anubis_ZFY KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Rhinopithecus_roxellana_ZFY KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Callithrix_jacchus_ZFY KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFNQQNELKKHMKTHSGRKV
Marmota_marmota_marmota_ZFY KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Mus_musculus_ZFY1 KKMHQCRHCDFNSPDPFLLSHHILSAHTKNVPFKCKRCKKEFQQQCELQTHMKTHSSRKV
Mus_musculus_ZFY2 KKMHQCRHCDFKSPDPFLLSHHILSAHTKNVPFKCKRCKKEFQQQCELQTHMKTHSSRKV


Rattus_norvegicus_ZFY2 KKMHQCRHCDFKSPDPFLLSRHILSVHTKNVPFKCKRCKKGFRQQCELQKHMKTHSGRKV
Bos_taurus_ZFY KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCKKGFRQQNELKKHMKTHSGRKV
Capra_hircus_ZFY KKMHQCRHCDFKIADPFVLSRHILSIHTKDLPFRCKRCKKGFRQQNELKKHMKTHSGRKV
Odocoileus_virginianus_ZFY KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCKKGFRQQNDLKKHMKTHSGRKV
Sus_scrofa_ZFY KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Canius_lupus_familiaris_ZFY KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Mustela_erminea_ZFY/ KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Loxodonta_africana_ZFY KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Equus_caballus_ZFY KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQTELKKHMKTHSGRKV
Homo_sapiens_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQSELKKHMKTHSGRKV
Pan_troglodytes_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQSELKKHMKTHSGRKV
Gorilla_gorilla_gorilla_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQSELKKHMKTHSGRKV
Macaca_mulatta_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Papio_anubis_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Rhinopithecus_roxellana_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Callithrix_jacchus_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Marmota_marmota_marmota_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Mus_musculus_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQSELKKHMKTHSGRKV
Rattus_norvegicus_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQSELKKHMKTHSGRKV
Bos_taurus_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Capra_hircus_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Odocoileus_virginianus_ZFX KKMHQCRHCDFKIADPFILSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Sus_scrofa_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Canis_lupus_familiaris_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Mustela_erminea_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Loxodonta_africana_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Equus_caballus_ZFX KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Monodelphis_domestica KKMHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Ornithorhynchus_anatinus RKLHQCRHCDFKIADPFVLSRHILSVHTKDLPFRCKRCRKGFRQQGELKKHMKTHSGRKV
Gallus_gallus KKLHQCRHCDFKIADPFILSRHILSVHTKDLPFRCKRCRKGFRQQNELKKHMKTHSGRKV
Xenopus_laevis_ZFX.S KKMHQCRHCEFHIADPFVLSRHILSVHTKELPYRCKRCKKGFRQQIELKKHMKTHSGKKV
Xenopus_laevis_ZFX.L KKMHQCRHCEFQIADPFVLSRHILSVHTKELPFRCKRCKKGFRQQMELKKHMKTHSGKKV
* 790 * 800 * 810 * 820 * 830
|.*.***|***#****#|*****.*..|*********|**#.|.#**|..*.
Homo_sapiens_ZFY YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Pan_troglodytes_ZFY YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Gorilla_gorilla_gorilla_ZFY YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Macaca_mulatta_ZFY YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Papio_anubis_ZFY YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Rhinopithecus_roxellana_ZFY YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Callithrix_jacchus_ZFY YQCEYCEYNTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Marmota_marmota_marmota_ZFY YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVSLS
Mus_musculus_ZFY1 YQCEYCEYSTKDASGFKRHVISIHTKDYPHSCDFCKKGFRRPSEKNQHIMRHHK-VGLP
Mus_musculus_ZFY2 YQCEYCEYSTKDASGFKRHVISIHTKDYPHRCDFCKKGFRRPSEKNQHIMRHHKEVGLA
Rattus_norvegicus_ZFY2 YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Bos_taurus_ZFY YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHITRHHKEVGLP
Capra_hircus_ZFY YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHITRHHKEVGLP
Odocoileus_virginianus_ZFY YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHITRHHKEVGLP
Sus_scrofa_ZFY YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Canius_lupus_familiaris_ZFY YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Mustela_erminea_ZFY/ YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Loxodonta_africana_ZFY YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEHCKKGFRRPSEKNQHIMRHHKEVGLP
Equus_caballus_ZFY YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Homo_sapiens_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Pan_troglodytes_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Gorilla_gorilla_gorilla_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Macaca_mulatta_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Papio_anubis_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Rhinopithecus_roxellana_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Callithrix_jacchus_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Marmota_marmota_marmota_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Mus_musculus_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Rattus_norvegicus_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Bos_taurus_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Capra_hircus_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Odocoileus_virginianus_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Sus_scrofa_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP
Canis_lupus_familiaris_ZFX YQCEYCEYSTTDASGFKRHVISIHTKDYPHRCEYCKKGFRRPSEKNQHIMRHHKEVGLP

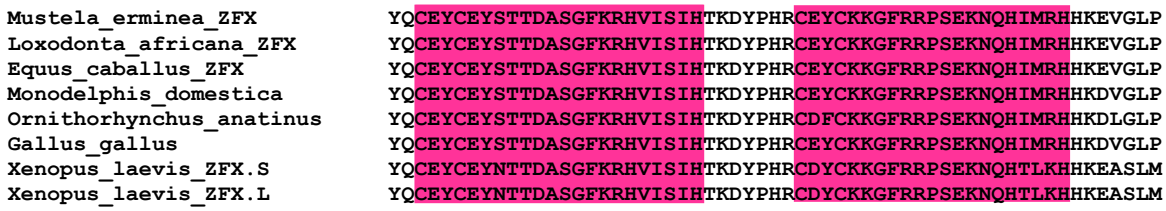


Figure 1: Complete protein alignment of ZFY/X and Zf* sequences collected.

Key:

9aaTAD – not a perfect 100% match

9aaTAD – perfect match

Nuclear localization motif (motif containing highly positively charged proteins (lysines/arginines)

Zinc Finger (* Conserved C residue # Conserved H residue)

- Homo Sapiens red highlighted sequence is the spliced region of ZFYS

↓ End of Acidic domain and start of DNA binding domain

Black Asterisk (*) = conserved amino acid

Table 4: Pairwise Inner Fragments detected by Geneconv. The table displays potential conversions identified using Geneconv. Permutation (sim) p-value and KA p-values, not multiple-comparison corrected are listed for each identified conversion. * Indicates significant p-values <0.05.

Sequence Names	Sim Pvalue	KA Pvalue	Aligned Offsets			Num. Poly	Num. Diffs	Total Mats	Mismatch Penalty
			Begin	End	Length				
Rattus norvegicus *Zfy2*;Rattus norvegicus *Zfx*	0.0037*	0.00282*	2353	2426	74	21	0	437	None
Mustela erminea *ZFY*;Mustela erminea *ZFX*	0.0030*	0.00416*	2032	2279	248	79	0	123	None
Mustela erminea *ZFY*;Mustela erminea *ZFX*	0.0159*	0.01972*	1813	1997	185	67	0	123	None
Loxodonta Africana *ZFY*;Loxodonta Africana *ZFX*	0.0016*	0.00282*	1931	2189	259	82	0	123	None

Loxodonta Africana *ZFY*;Loxodonta Africana *ZFX*	0.0126*	0.01523*	2191	2397	207	69	0	123	None
Loxodonta Africana *ZFY*;Loxodonta Africana *ZFX*	0.0218*	0.02551*	1602	1766	165	65	0	123	None

Table 5: Pairwise Outer Fragments detected by Geneconv. The table displays potential conversions identified using Geneconv. Permutation (sim) p-value and KA p-values, not multiple-comparison corrected are listed for each identified conversion. * Indicates significant p-values <0.05.

Sequence Names	Sim Pvalue	KA Pvalue	Aligned Offsets			Num. Poly	Num. Mat	Total Mats	Mismatch Penalty
			Begin	End	Length				
Marmota marmota marmota *ZFY*	0.0345*	0.03840*	1032	1044	13	6	6	0	None
Mus musculus *Zfy2*	0.0105*	0.01346*	913	945	33	15	15	0	None
Bos taurus *ZFY*	0.0000*	0.00000*	1050	1067	18	14	14	0	None
Bos taurus *ZFY*	0.0003*	0.00044*	1035	1048	14	8	8	0	None

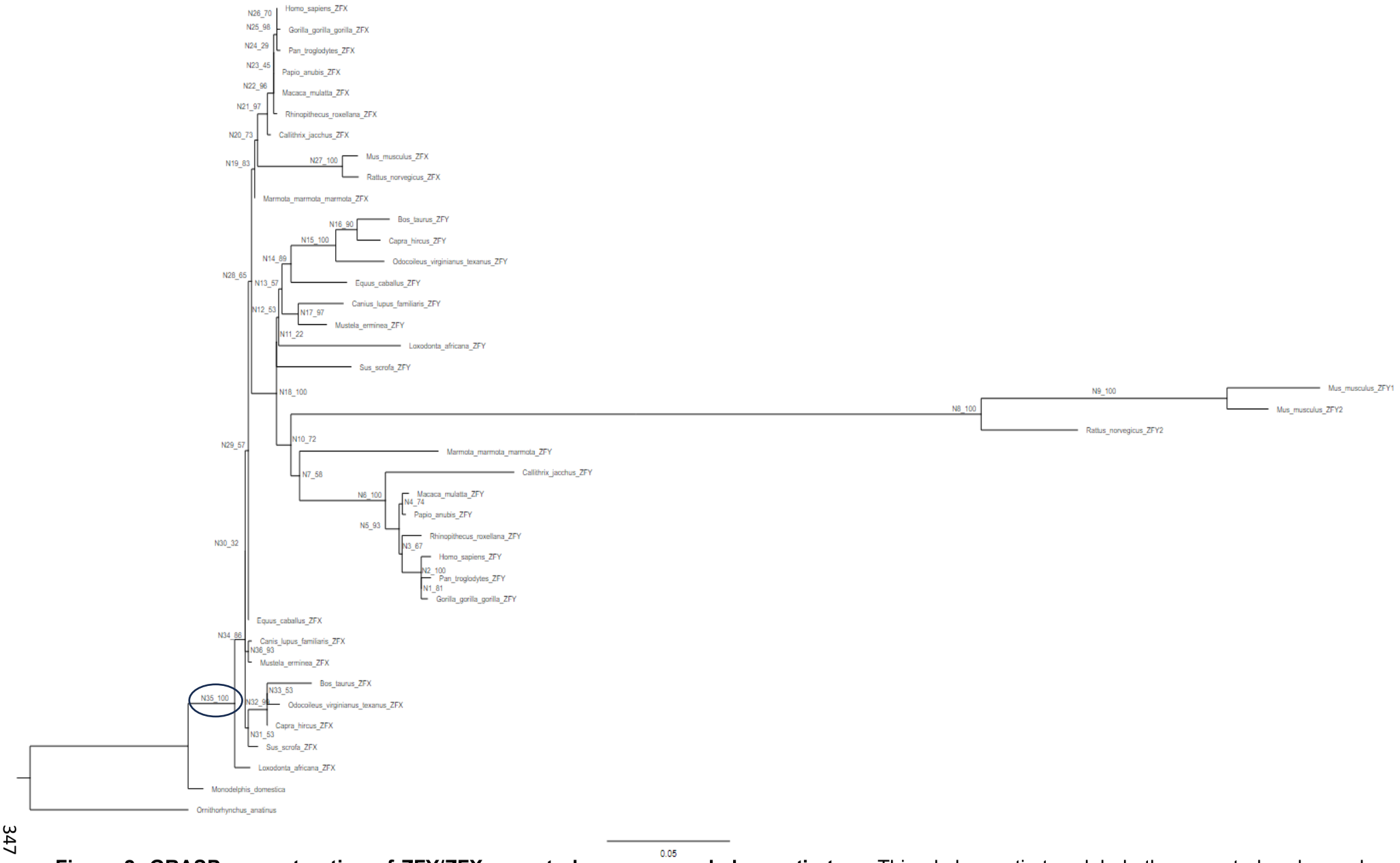


Figure 2: GRASP reconstruction of ZFY/ZFX ancestral sequences phylogenetic tree. This phylogenetic tree labels the ancestral nodes and provides weighted support for each branch. This tree was used to obtain the node associated with the first ZFY/ZFX ancestor sequence, labelled N35_100. N35_100 means the node number is 35 and the node is 100% supported. The scale bar denotes the branch length corresponding to 0.05 genetic change.

Figure 3: GRASP reconstruction of ZFY ancestral sequences phylogenetic tree. This phylogenetic tree labels the ancestral nodes and provides weighted support for each branch. This tree was used to obtain the nodes associated with the divergence of ZFY at lineage breaks. The scale bar denotes the branch length corresponding to 0.05 genetic change. Outlined is the first ZFY ancestor after the marsupial/eutherian split labelled N18_100.

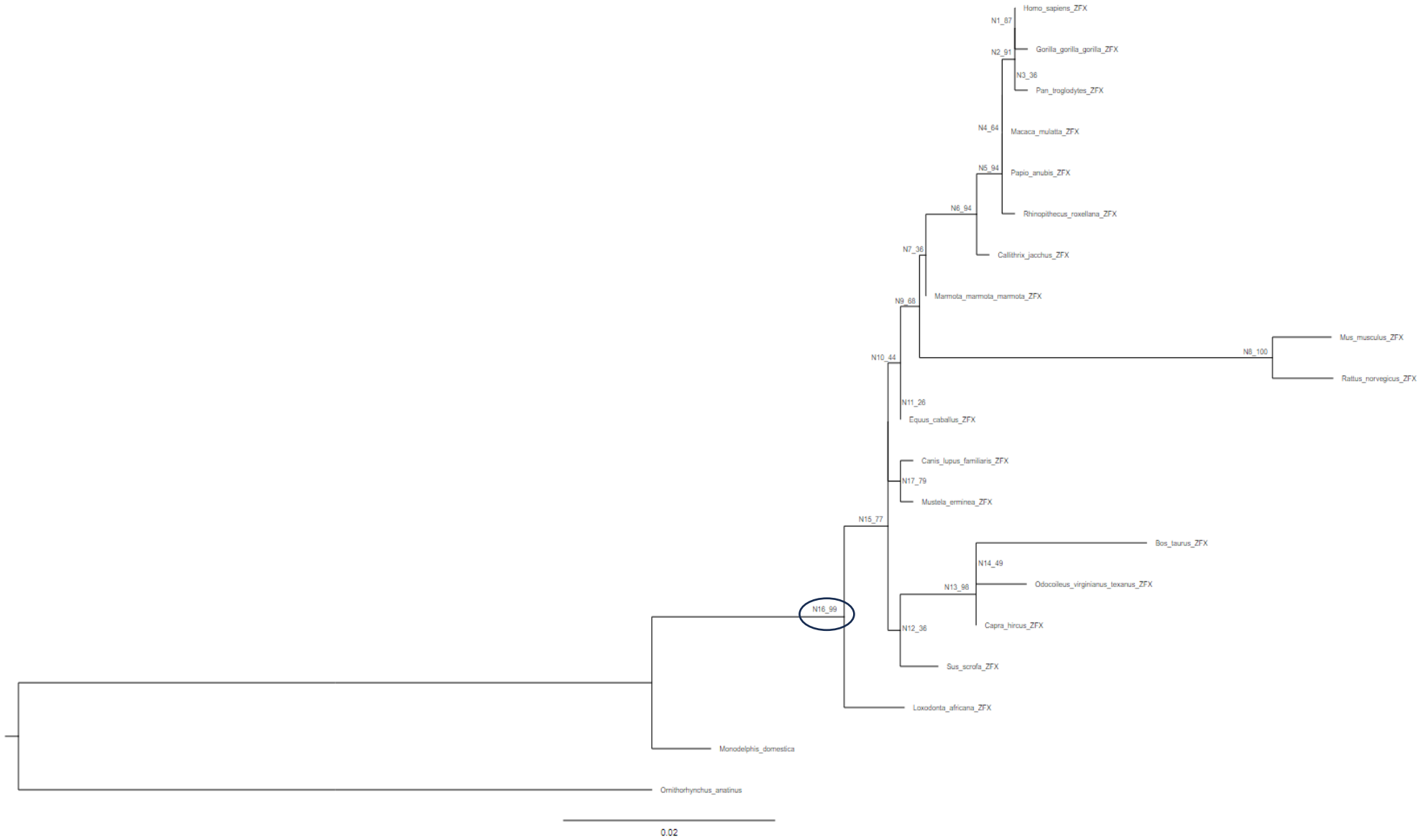


Figure 4: GRASP reconstruction of ZFX ancestral sequences phylogenetic tree. This phylogenetic tree labels the ancestral nodes and provides weighted support for each branch. This tree was used to obtain the nodes associated with the divergence of ZFX at lineage breaks. The scale bar denotes the branch length corresponding to 0.02 genetic change. Outlined is the first ZFX ancestor after the marsupial/eutherian split labelled N19_99.

Table 6: Reporter nucleotide sequence verifying the insertion of the *ZFY* exon into the GFP reporter system. Black = partial vector backbone including split GFP and luciferase intron. The XmaI and XhoI enzyme sites used for cloning are underlined. Blue = *ZFY* intronic sequence flanking the cassette exon. Red = *ZFY* cassette exon (second coding exon). Sequence 1 contains the reference human *ZFY* sequence whilst in sequence 2, SNPs are introduced (highlighted with bold underline) to remove the splice donor and acceptor site and any other nearly potential alternative donor/acceptor sites.

Description	Sequence
Sequence 1: *ZFY* exon 2 with flanking introns	AGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCC ATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAA CGGGACTTTCCAAAAGTTCGTAACAATTCCGCCCCCAT GACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTC TATATAAGCAGAGCTGGTTTAGTGAACCGTCAGATCCGC TAGCGCTACCGGTCGCCACCATGGTGAGCAAGGGCGA GGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAG CTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGT CCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGC TGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCC CGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTAC GGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGA AGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGG CTACGTCCAGGTAAGTATCAACGCGTTACAAGACAGGT TTAAGGAGACCAATAGAAACTCCCGGGggttttttttatttttctttttct tgttagttgattttatatttctttttttactttgtatttctttttcttgttagttgattagggttgagta aaataaatttttatgttaaaaatttgtttgttaagtgaatcaaaagtggattcattctctt atttattgtctcaatgtttgaggaacaccaaggacatagtctgttttcccagaagaa attaaataaaaactatatttagtcatgtatacttacctgagatttgtcattttaatttttatt ctttaaggagctgatgctacacacatggatggtgatcagattgttgtggaaataca agaagcagtttttgtttctaatattgtggattctgacataactgtgcataactttgttcct gatgacccagactcagttgtaatccaagatgttgttgaagatgttgtcatagagga ggatgttcagtgctcagatatcttagaagaggcagatgtatctgaaaatgtcatca ttcctgagcaagtgctggactcagatgtaactgaagaagtttctttaccacactgc acagtcccagatgatgttttagcttctgacattacttcaacctcaatgtctatgccag aacatgttttaacgagtgaatccatgcatgtgtgtgacattggacatgttgaacata tggtgcatgatagtgtagtggaagcagaaatcattactgatcctctgacgagtgac atagtttcagaagaagtattggtagcagactgtgcccctgaagcagtcatagatg ccagcgggatctcagtggaccagcaagataatgacaaagccagctgtgagga ctacctaatgatttcgtgtaagtcatggggtacagtgattttaagcaatgtttttgaaa cctgtgttttctaacataaataatggtgaaattttattgagtttaacttctgtaaagttaa agtagaaatgataaacctacatgcattgttgcttagtttgcatcccagaagataaa ccgtaattaagtgacatggtgcagaagagagttatttctatattaggatagaacat aaggatgaaagaataagaaatgaaaataacagaattgagtcttatgaaatcat acaaagtatgatCTCGAGTAGGCTAGCCTATTGGTCTTACTG ACATCCACTTTGCCTTTCTCTCCACAGGAGCGCACCAT CTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCC GAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCA TCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAA CATCCTGGGGCACAAGCTGGAGTACAACTACAACAGC CACAACGTCTATATCATGGCCGACAAGCAGAAGAACGG CATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGG ACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAA CACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGA CAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAA GACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGG

	AGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGA CGAGCTGTACAAGTCCGGACTCAGATCTCGACAGAGC CATGGCTTCCCGCCGGAGGTGGAGGAGCAGGATGATG GCACGCTGCCCATGTCTTGTGCCCAGGAGAGCGGGAT GGACCGTCACCCTGCAGCCTGTGCTTCTGCTAGGATC AATGTGTAGATTTATTTAATTAATTAATTTGGGATCCACC GGATCTAGATAACTGATCATAATCAGCCATACCACATTTG TAG
Sequence 2: *ZFY* exon 2 with flanking introns (contains small deletions to remove the splice donor and acceptor site)	AGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCC ATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAA CGGGACTTTCCAAAAGTTCGTAACAATTCCGCCCCCAT GACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTC TATATAAGCAGAGCTGGTTTAGTGAACCGTCAGATCCGC TAGCGCTACCGGTCGCCACCATGGTGAGCAAGGGCGA GGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAG CTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGT CCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGC TGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCC CGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTAC GGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGA AGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGG CTACGTCCAGGTAAGTATCAACGCGTTACAAGACAGGT TTAAGGAGACCAATAGAAACTCCCGGGggttttttttatttttctttttct tgttagttgattttatatttctttttttactttgtatttctttttcttgttagttgattagggttgagta aaataaatttttatgttaaaaatttgtttgttaagtgaatcaaaagtggattcattctctt atttattgtctcaatgtttgaggaacaccaaggacatagtctgttttcccagaagaa attaaataaaaactatatttagtcatgtatacttacctgagatttgtcattttaatttttatt cttta**TC**g**T**gctgatgctacacacatggatggtgatcagattgttgtggaaatac aagaagcagtttttgtttctaatattgtggattctgacataactgtgcataactttgttcc tgatgacccagactcagttgtaatccaagatgttgttgaagatgttgtcatagagg aggatgttcagtgctcagatatcttagaagaggcagatgtatctgaaaatgtcatc attcctgagcaagtgctggactcagatgtaactgaagaagtttctttaccacactg cacagtcccagatgatgttttagcttctgacattacttcaacctcaatgtctatgcca gaacatgttttaacgagtgaatccatgcatgtgtgtgacattggacatgttgaacat atggtgcatgatagtgtagtggaagcagaaatcattactgatcctctgacgagtg acatagtttcagaagaagtattggtagcagactgtgcccctgaagcagtcataga tgccagcgggatctcagtggaccagcaagataatgacaaagccagctgtgag gactacctaatgatttc**C**t**CA**aagtcatggggtacagtgattttaagcaatgttttt gaaacctgtgttttctaacataaataatggtgaaattttattgagtttaacttctgtaaa gttaaagtagaaatgataaacctacatgcattgttgcttagtttgcatcccagaaga taaaccgtaattaagtgacatggtgcagaagagagttatttctatattaggataga acataaggatgaaagaataagaaatgaaaataacagaattgagtcttatgaaa tcatacaaagtatgatCTCGAGTAGGCTAGCCTATTGGTCTTAC TGACATCCACTTTGCCTTTCTCTCCACAGGAGCGCACC ATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGC CGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCG CATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGC AACATCCTGGGGCACAAGCTGGAGTACAACTACAACAG CCACAACGTCTATATCATGGCCGACAAGCAGAAGAACG GCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAG GACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGA ACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGA CAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAA GACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGG AGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGA CGAGCTGTACAAGTCCGGACTCAGATCTCGACAGAGC

	CATGGCTTCCCGCCGGAGGTGGAGGAGCAGGATGATG GCACGCTGCCCATGTCTTGTGCCCAGGAGAGCGGGAT GGACCGTCACCCTGCAGCCTGTGCTTCTGCTAGGATC AATGTGTAGATTTATTTAATTAATTAATTTGGGATCCACC GGATCTAGATAACTGATCATAATCAGCCATACCACATTTG TAG

Human *ZFY* Clones

Clones were synthetically made for the experiments used in Chapter 4 and Chapter 5. Chapter 4 used clones that contained the complete open reading frame for the major splice form of human *ZFY*, and the complete open reading from for the minor short splice form of human *ZFY* tagged with either HA or eGFP (Clones 1, 2, 3 and 4). Chapter 5 used a pET-15b vector with the truncated open reading frame for the major splice form and short minor splice form of human *ZFY* with the terminal exon removed which codes for the DNA binding domain cloned in (Clones 5 and 6).

Clone 1: Sequence A, cloned into vector pcDNA3.1+N-HA using XhoI / XbaI

Clone 2: Sequence A, cloned into vector pcDNA3.1+N-GFP using XhoI / XbaI

Clone 3: Sequence B, cloned into vector pcDNA3.1+N-HA using XhoI / XbaI

Clone 4: Sequence B, cloned into vector pcDNA3.1+N-GFP using XhoI / XbaI

Clone 5: Sequence C, cloned into vector pET-15b using XhoI / BlpI

Clone 6: Sequence D, cloned into vector pET-15b using XhoI / BlpI

See below for the inserted sequences and note colour coding of sites used for cloning:

AsiSI, XhoI, XbaI, BlpI

SEQUENCE A

The complete open reading frame for the major splice form of human *ZFY* (NM_003411.4)

>h*ZFY*_full

GCGATCGCCCTCGAG
ATGGATGAAG ATGAATTTGA ATTGCAGCCA CAAGAGCCAA ACTCATTTTT TGATGGAATA
GGAGCTGATG CTACACACAT GGATGGTGAT CAGATTGTTG TGGAAATACA AGAAGCAGTT
TTTGTTTCTA ATATTGTGGA TTCTGACATA ACTGTGCATA ACTTTGTTCC TGATGACCCA
GACTCAGTTG TAATCCAAGA TGTTGTTGAA GATGTTGTCA TAGAGGAGGA TGTTCAGTGC
TCAGATATCT TAGAAGAGGC AGATGTATCT GAAAATGTCA TCATTCCTGA GCAAGTGCTG
GACTCAGATG TAACTGAAGA AGTTTCTTTA CCACACTGCA CAGTCCCAGA TGATGTTTTA
GCTTCTGACA TTACTTCAAC CTCAATGTCT ATGCCAGAAC ATGTTTTAAC GAGTGAATCC
ATGCATGTGT GTGACATTGG ACATGTTGAA CATATGGTGC ATGATAGTGT AGTGGAAGCA
GAAATCATTA CTGATCCTCT GACGAGTGAC ATAGTTTCAG AAGAAGTATT GGTAGCAGAC
TGTGCCCCTG AAGCAGTCAT AGATGCCAGC GGGATCTCAG TGGACCAGCA AGATAATGAC
AAAGCCAGCT GTGAGGACTA CCTAATGATT TCGTTGGATG ATGCTGGCAA AATAGAACAT
GATGGTTCCA CTGGAGTGAC CATCGATGCA GAATCAGAAA TGGATCCTTG TAAAGTGGAT
AGCACTTGTC CTGAAGTCAT CAAGGTGTAC ATTTTTAAAG CTGACCCTGG AGAAGATGAC
TTAGGTGGAA CTGTAGACAT TGTGGAGAGT GAACCTGAAA ATGATCATGG AGTTGAACTA
CTTGATCAGA ACAGCAGTAT TCGTGTTCCC AGGGAAAAGA TGGTTTATAT GACTGTCAAT
GACTCTCAAC AAGAAGATGA AGATTTAAAT GTTGCTGAAA TTGCTGATGA AGTTTATATG

GAAGTGATCG TAGGAGAGGA GGATGCTGCT GTTGCAGCAG CAGCAGCTGC TGTGCATGAG
CAGCAAATTG ATGAGGATGA AATGAAAACC TTCGTACCAA TTGCATGGGC AGCAGCTTAT
GGTAATAATT CTGATGGAAT TGAAAACCGG AATGGCACTG CAAGTGCCCT CTTGCACATA
GATGAGTCTG CTGGCCTTGG CAGACTGGCT AAACAGAAAC CAAAGAAAAA GAGAAGACCT
GATTCCAGGC AGTACCAAAC AGCAATAATT ATTGGCCCTG ATGGTCATCC TTTGACTGTC
TATCCTTGCA TGATTTGTGG GAAGAAGTTT AAGTCGAGGG GTTTTTTGAA AAGACACATG
AAAAACCATC CTGAACACCT TGCCAAGAAG AAGTACCACT GTACTGACTG TGATTACACT
ACCAATAAGA AGATAAGTTT ACATAACCAC CTGGAGAGCC ACAAGCTGAC CAGCAAGGCA
GAGAAGGCCA TTGAATGTGA TGAGTGTGGG AAGCATTTTT CTCATGCAGG GGCTTTGTTT
ACTCACAAAA TGGTGCATAA GGAAAAAGGG GCCAACAAAA TGCACAAGTG TAAATTCTGT
GAATATGAGA CAGCTGAACA GGGGTTATTG AATCGCCACC TCTTGGCAGT CCACAGCAAG
AACTTTCCTC ATATTTGTGT GGAGTGTGGT AAAGGTTTCC GACACCCGTC GGAACTGAGA
AAGCACATGC GAATCCATAC CGGCGAGAAG CCATACCAAT GCCAGTACTG TGAATATAGG
TCTGCAGACT CTTCTAACTT GAAAACACAT ATAAAAACAA AGCATAGTAA AGAGATGCCA
TTCAAGTGTG ACATTTGTCT TCTGACTTTC TCAGATACCA AAGAAGTGCA GCAACATACT
CTTGTCCACC AAGAAAGCAA AACACATCAG TGTTTGCATT GCGACCACAA GAGTTCAAAC
TCAAGTGATT TGAAACGACA TGTAATTTCA GTTCATACGA AAGACTATCC TCATAAGTGT
GAGATGTGCG AGAAAGGCTT TCACAGGCCT TCAGAACTTA AGAAACATGT GGCTGTCCAC
AAAGGTAAAA AAATGCACCA ATGTAGACAT TGTGACTTTA AGATTGCAGA CCCATTTGTT
CTAAGTCGCC ATATTCTCTC AGTTCACACA AAGGATCTTC CATTTAGGTG TAAGAGATGT
AGAAAGGGAT TTAGGCAACA AAATGAGCTT AAAAAGCATA TGAAGACACA CAGTGGCAGG
AAAGTATATC AGTGTGAGTA CTGTGAGTAT AGCACTACAG ATGCCTCAGG CTTTAAACGG
CACGTTATTT CCATTCATAC AAAAGACTAT CCTCATCGGT GTGAGTACTG CAAGAAAGGC
TTCCGAAGAC CTTCAGAAAA GAACCAGCAC ATAATGAGAC ACCATAAAGA AGTTGGTCTG
CCCTAA
TCTAGAGCTGAGC

SEQUENCE B

The complete open reading frame for the minor, short splice form of human *ZFY* (NM_001145276.2), which omits the second coding exon.

>h*ZFY*_short

GCGATCGCCCTCGAG
ATGGATGAAG ATGAATTTGA ATTGCAGCCA CAAGAGCCAA ACTCATTTTT TGATGGAATA
GTGGATGATG CTGGCAAAAT AGAACATGAT GGTTCCACTG GAGTGACCAT CGATGCAGAA
TCAGAAATGG ATCCTTGTAA AGTGGATAGC ACTTGTCCTG AAGTCATCAA GGTGTACATT
TTTAAAGCTG ACCCTGGAGA AGATGACTTA GGTGGAACTG TAGACATTGT GGAGAGTGAA
CCTGAAAATG ATCATGGAGT TGAACTACTT GATCAGAACA GCAGTATTCG TGTTCCCAGG
GAAAAGATGG TTTATATGAC TGTCAATGAC TCTCAACAAG AAGATGAAGA TTTAAATGTT
GCTGAAATTG CTGATGAAGT TTATATGGAA GTGATCGTAG GAGAGGAGGA TGCTGCTGTT
GCAGCAGCAG CAGCTGCTGT GCATGAGCAG CAAATTGATG AGGATGAAAT GAAAACCTTC
GTACCAATTG CATGGGCAGC AGCTTATGGT AATAATTCTG ATGGAATTGA AAACCGGAAT
GGCACTGCAA GTGCCCTCTT GCACATAGAT GAGTCTGCTG GCCTTGGCAG ACTGGCTAAA
CAGAAACCAA AGAAAAAGAG AAGACCTGAT TCCAGGCAGT ACCAAACAGC AATAATTATT
GGCCCTGATG GTCATCCTTT GACTGTCTAT CCTTGCATGA TTTGTGGGAA GAAGTTTAAG
TCGAGGGGTT TTTTGAAAAG ACACATGAAA AACCATCCTG AACACCTTGC CAAGAAGAAG
TACCACTGTA CTGACTGTGA TTACACTACC AATAAGAAGA TAAGTTTACA TAACCACCTG
GAGAGCCACA AGCTGACCAG CAAGGCAGAG AAGGCCATTG AATGTGATGA GTGTGGGAAG
CATTTTTCTC ATGCAGGGGC TTTGTTTACT CACAAAATGG TGCATAAGGA AAAAGGGGCC
AACAAAATGC ACAAGTGTAA ATTCTGTGAA TATGAGACAG CTGAACAGGG GTTATTGAAT
CGCCACCTCT TGGCAGTCCA CAGCAAGAAC TTTCCTCATA TTTGTGTGGA GTGTGGTAAA
GGTTTCCGAC ACCCGTCGGA ACTGAGAAAG CACATGCGAA TCCATACCGG CGAGAAGCCA
TACCAATGCC AGTACTGTGA ATATAGGTCT GCAGACTCTT CTAACTTGAA AACACATATA
AAAACAAAGC ATAGTAAAGA GATGCCATTC AAGTGTGACA TTTGTCTTCT GACTTTCTCA

GATACCAAAG AAGTGCAGCA ACATACTCTT GTCCACCAAG AAAGCAAAAC ACATCAGTGT
TTGCATTGCG ACCACAAGAG TTCAAACTCA AGTGATTTGA AACGACATGT AATTTCAGTT
CATACGAAAG ACTATCCTCA TAAGTGTGAG ATGTGCGAGA AAGGCTTTCA CAGGCCTTCA
GAACTTAAGA AACATGTGGC TGTCCACAAA GGTAAAAAAA TGCACCAATG TAGACATTGT
GACTTTAAGA TTGCAGACCC ATTTGTTCTA AGTCGCCATA TTCTCTCAGT TCACACAAAG
GATCTTCCAT TTAGGTGTAA GAGATGTAGA AAGGGATTTA GGCAACAAAA TGAGCTTAAA
AAGCATATGA AGACACACAG TGGCAGGAAA GTATATCAGT GTGAGTACTG TGAGTATAGC
ACTACAGATG CCTCAGGCTT TAAACGGCAC GTTATTTCCA TTCATACAAA AGACTATCCT
CATCGGTGTG AGTACTGCAA GAAAGGCTTC CGAAGACCTT CAGAAAAGAA CCAGCACATA
ATGAGACACC ATAAAGAAGT TGGTCTGCCC TAA
TCTAGAGCTGAGC

SEQUENCE C

The truncated open reading frame for the major splice form of human *ZFY* (NM_003411.4), with the terminal exon that codes for the DNA binding domain removed.

>h*ZFY*_full_noDBD
GCGATCGCCCTCGAG
ATGGATGAAG ATGAATTTGA ATTGCAGCCA CAAGAGCCAA ACTCATTTTT TGATGGAATA
GGAGCTGATG CTACACACAT GGATGGTGAT CAGATTGTTG TGGAAATACA AGAAGCAGTT
TTTGTTTCTA ATATTGTGGA TTCTGACATA ACTGTGCATA ACTTTGTTCC TGATGACCCA
GACTCAGTTG TAATCCAAGA TGTTGTTGAA GATGTTGTCA TAGAGGAGGA TGTTCAGTGC
TCAGATATCT TAGAAGAGGC AGATGTATCT GAAAATGTCA TCATTCCTGA GCAAGTGCTG
GACTCAGATG TAACTGAAGA AGTTTCTTTA CCACACTGCA CAGTCCCAGA TGATGTTTTA
GCTTCTGACA TTACTTCAAC CTCAATGTCT ATGCCAGAAC ATGTTTTAAC GAGTGAATCC
ATGCATGTGT GTGACATTGG ACATGTTGAA CATATGGTGC ATGATAGTGT AGTGGAAGCA
GAAATCATTA CTGATCCTCT GACGAGTGAC ATAGTTTCAG AAGAAGTATT GGTAGCAGAC
TGTGCCCCTG AAGCAGTCAT AGATGCCAGC GGGATCTCAG TGGACCAGCA AGATAATGAC
AAAGCCAGCT GTGAGGACTA CCTAATGATT TCGTTGGATG ATGCTGGCAA AATAGAACAT
GATGGTTCCA CTGGAGTGAC CATCGATGCA GAATCAGAAA TGGATCCTTG TAAAGTGGAT
AGCACTTGTC CTGAAGTCAT CAAGGTGTAC ATTTTTAAAG CTGACCCTGG AGAAGATGAC
TTAGGTGGAA CTGTAGACAT TGTGGAGAGT GAACCTGAAA ATGATCATGG AGTTGAACTA
CTTGATCAGA ACAGCAGTAT TCGTGTTCCC AGGGAAAAGA TGGTTTATAT GACTGTCAAT
GACTCTCAAC AAGAAGATGA AGATTTAAAT GTTGCTGAAA TTGCTGATGA AGTTTATATG
GAAGTGATCG TAGGAGAGGA GGATGCTGCT GTTGCAGCAG CAGCAGCTGC TGTGCATGAG
CAGCAAATTG ATGAGGATGA AATGAAAACC TTCGTACCAA TTGCATGGGC AGCAGCTTAT
GGTAATAATT CTGATGGAAT TGAAAACCGG AATGGCACTG CAAGTGCCCT CTTGCACATA
GATGAGTCTG CTGGCCTTGG CAGACTGGCT AAACAGAAAC CAAAGAAAAA GAGAAGACCT
GATTCCAGGC AGTACCAAAC ATAA
TCTAGAGCTGAGC

SEQUENCE D

The truncated open reading frame for the minor splice form of human *ZFY* (NM_003411.4), with the terminal exon that codes for the DNA binding domain removed.

>h*ZFY*_short_noDBD
GCGATCGCCCTCGAG
ATGGATGAAG ATGAATTTGA ATTGCAGCCA CAAGAGCCAA ACTCATTTTT TGATGGAATA
GTGGATGATG CTGGCAAAAT AGAACATGAT GGTTCCACTG GAGTGACCAT CGATGCAGAA
TCAGAAATGG ATCCTTGTAA AGTGGATAGC ACTTGTCCTG AAGTCATCAA GGTGTACATT
TTTAAAGCTG ACCCTGGAGA AGATGACTTA GGTGGAACTG TAGACATTGT GGAGAGTGAA
CCTGAAAATG ATCATGGAGT TGAACTACTT GATCAGAACA GCAGTATTCG TGTTCCCAGG
GAAAAGATGG TTTATATGAC TGTCAATGAC TCTCAACAAG AAGATGAAGA TTTAAATGTT

GCTGAAATTG CTGATGAAGT TTATATGGAA GTGATCGTAG GAGAGGAGGA TGCTGCTGTT
GCAGCAGCAG CAGCTGCTGT GCATGAGCAG CAAATTGATG AGGATGAAAT GAAAACCTTC
GTACCAATTG CATGGGCAGC AGCTTATGGT AATAATTCTG ATGGAATTGA AAACCGGAAT
GGCACTGCAA GTGCCCTCTT GCACATAGAT GAGTCTGCTG GCCTTGGCAG ACTGGCTAAA
CAGAAACCAA AGAAAAAGAG AAGACCTGAT TCCAGGCAGT ACCAAACATA A
TCTAGAGCTGAGC

Table 7: Enrichment pathway analysis performed as part of Chapter 4. List of potentially interesting pathways associated with the *ZFY* variants, and the genes identified within the dataset associated with the pathways. FDR = False Discovery Rate**.**

	Pathway	p-value	FDR	Genes
Enriched by both *ZFY* Variants	Degradation of Extracellular Matrix	$4.24e^{-06}$	0.004	ACAN, ADAMTS8, CAPN5, COL13A1, COL19A1, COL26A1, COL5A1, COL9A2, CTSV, HSPG2, SCUBE1, TMPRSS6, ADAM15, BCAN, CAPNS2, COL16A1, COL1A1, COL4A2, COL6A1, COL9A3, ELN, MMP15, TIMP2, TPSAB1, ADAMTS4, CAPN1, CDH1, COL18A1, COL23A1, COL4A4, COL6A2, CTSL, FDN3, MMP25, TLL2, TPSB2
	Collagen Degradation	2.41e-05	0.016	COL13A1, COL19A1, COL26A1, COL5A1, COL9A2, CTSV, COL16A1, COL1A1, COL4A2, COL6A1, COL9A3, MMP15, COL18A1, COL23A1, COL4A4, COL6A2, CTSL, TMPRSS6
	Collagen Chain Trimerization	$1.75e^{-06}$	0.003	COL13A1, COL19A1, COL26A1, COL5A1, COL9A2, COL16A1, COL1A1, COL4A2, COL6A1, COL9A3, COL18A1, COL23A1, COL4A4, COL6A2
Enriched by *ZFYS* only	ERBB2 Regulates Cell Motility	$9.54e^{-07}$	$4.65e^{-04}$	ERBB4, NRG1
	ERBB2 Activates PTK6 Signalling	$2.67e^{-06}$	$4.65e^{-04}$	ERBB4, NRG1

	PI3K Events in ERBB2 Signalling	$3.47e^{-06}$	$4.65e^{-04}$	ERBB4, NRG1
Enriched by both *ZFY* Variants and *ZFX*	*WNT* Ligand Biogenesis and Trafficking	$3.28e^{-07}$	$3.72e^{-04}$	*WNT*11, *WNT*5B, *WNT*9A, *WNT*3A, *WNT7*A, *WNT*4, *WNT7*B
	Voltage Gated Potassium Channels	$1.48e^{-06}$	$8.41e^{-04}$	KCNA2, KCNB1, KCNQ1, KCNA3, KCNH2, KCNA6, KCNH6

Table 8: List of the 55 potential protein interactors obtained from n=1 proteomics following pull-down assay. No filtering has been performed on these genes. It should be noted that all potential interactors in this table are for *ZFYS* and *ZFYL*.

Description	Gene Name
Poly [ADP-ribose] polymerase 1	PARP1
Nucleolin	NCL
Heat shock cognate 71 kDa protein	HSPA8
Heat shock 70 kDa protein 1A	HSPA1A
60S acidic ribosomal protein P0	RPLP0
40S ribosomal protein S3	RPS3
Heterogeneous nuclear ribonucleoproteins C1/C2	HNRNPC
60S ribosomal protein L18	RPL18
ATP synthase subunit alpha_ mitochondrial	ATP5F1A
Actin_ cytoplasmic 1	ACTB
L-lactate dehydrogenase B chain	LDHB
Probable 28S rRNA (cytosine(4447)-C(5))-methyltransferase	NOP2
Elongation factor 1-alpha 1	EEF1A1
Tubulin beta chain	TUBB
Polyubiquitin-B	UBB
Heterogeneous nuclear ribonucleoprotein U	HNRNPU
Heterogeneous nuclear ribonucleoprotein R	HNRNPR
Heterogeneous nuclear ribonucleoproteins A2/B1	HNRNPA2B1
60S ribosomal protein L7	RPL7
60S ribosomal protein L4	RPL4
60S ribosomal protein L12	RPL12
Clathrin heavy chain	CLTC
40S ribosomal protein S8	RPS8
Heterogeneous nuclear ribonucleoprotein H	HNRPH1
Histone H4	H4C1
60S ribosomal protein L14	RPL14
60S ribosomal protein L23a	RPL23A
ATP-dependent RNA helicase DDX3X	DDX3X
ATP-dependent RNA helicase DDX50	DDX50

Tubulin alpha-3C chain	TUBA3C
Y-box-binding protein 1	YBX1
Histone H2A type 1-B/E	H2AC4
Nucleophosmin	NPM1
Heterogeneous nuclear ribonucleoprotein K	HNRNPK
60S ribosomal protein L29	RPL29
116 kDa U5 small nuclear ribonucleoprotein component	EFTUD2
60S ribosomal protein L11	RPL11
Interleukin enhancer-binding factor 3	ILF3
28S ribosomal protein S29_ mitochondrial	DAP3
Insulin-like growth factor 2 mRNA-binding protein 1	IGF2BP1
Tubulin alpha-1B chain	TUBA1B
Y-box-binding protein 3	YBX3
Pyruvate kinase	PKM
Eukaryotic translation initiation factor 2 alpha kinase 3	EIF2AK3
Ribosomal protein L19	RPL19
Spliceosome RNA helicase DDX39B (Fragment)	DDX39B
Vimentin (Fragment)	VIM
Heterogeneous nuclear ribonucleoprotein D	HNRNPD
Thyroxine 5-deiodinase	DIO3
Glyceraldehyde-3-phosphate dehydrogenase	GAPDH
Calmin	CLMN
Polyadenylate-binding protein	PABPC1
Galactose-3-O-sulfotransferase 4	GAL3ST4
Cytoskeleton-associated protein 5	CKAP5
ATP-dependent RNA helicase DHX15	DHX15

Table 9: List of 61 potential protein interactors obtained from n=2 proteomics following pull-down assay. No filtering has been performed on these genes. It should be noted that all potential interactors in this table are for *ZFYS* and *ZFYL*.

Description	Gene Name
Poly [ADP-ribose] polymerase 1	PARP1
Histone H4	H4C1
60S ribosomal protein L7	RPL7
Histone H1.4	H1-4
Heat shock cognate 71 kDa protein	HSPA8
Nucleolin	NCL
Zinc finger CCCH domain-containing protein 18	ZCH18
Ribosomal protein L18	RPL18
L-lactate dehydrogenase A chain	LDHA
60S ribosomal protein L6	RPL6
Heat shock 70 kDa protein 1B	HSPA1B
Ninein	NIN

Adenosylhomocysteinase	AHCY
L-lactate dehydrogenase B chain	LDHB
Heterogeneous nuclear ribonucleoproteins A2/B1	HNRNPA2B1
ATP-dependent RNA helicase DDX3X	DDX3X
Heterogeneous nuclear ribonucleoproteins C1/C2	HNRNPC
Tubulin alpha-3C chain	TUBA3C
Probable 28S rRNA (cytosine(4447)-C(5))-methyltransferase	NOP2
Histidine--tRNA ligase_ mitochondrial	HARS2
Polyubiquitin-B	UBB
Heterogeneous nuclear ribonucleoprotein U (Fragment)	HNRNPU
ATP synthase subunit alpha_ mitochondrial	ATP5F1A
Phosphatidylinositol 3_4_5-trisphosphate 5-phosphatase 2	INPPL1
40S ribosomal protein S2	RPS2
ELAV-like protein 1	ELAVL1
28S ribosomal protein S29_ mitochondrial	DAP3
Family with sequence similarity 184 member A	FAM184A
60S ribosomal protein L8 (Fragment)	RPL8
Heat shock protein HSP 90-alpha	HSP90AA1
60S ribosomal protein L11	RPL11
Tubulin beta chain	TUBB
Histone H2B type 1-J	H2BC11
Actin beta (Fragment)	ACTB
Voltage dependent anion channel 2 (Fragment)	VDAC2
Coiled-coil domain-containing protein 113	CCDC113
GTP-binding nuclear protein Ran	RAN
ATP-dependent RNA helicase	EIF4A2
Prohibitin	PHB2
60S ribosomal protein L14	RPL14
RNA helicase	DDX17
Probable ATP-dependent RNA helicase DDX5	DDX5
Small ubiquitin like modifier 3	SUMO3
DNA helicase	CHD4
Ribosomal protein lateral stalk subunit P0	RPLP0
Vimentin (Fragment)	VIM
Ankyrin repeat domain 20 family member A1 (Fragment)	ANKRD20A1
Heterogeneous nuclear ribonucleoprotein R (Fragment)	HNRNPR
Mitochondrial ribosomal protein S7	MRPS7
40S ribosomal protein S6	RPS6
60S ribosomal protein L35	RPL35
40S ribosomal protein S8	RPS8
60S ribosomal protein L27a	RPL27A
60S ribosomal protein L29	RPL29

T-complex protein 1 subunit alpha	TCP1
Eukaryotic initiation factor 4A-III	EIF4A3
Protein transport protein Sec31B	SEC31B
NOP56 ribonucleoprotein	NOP56
60S ribosomal protein L13a	RPL13A
Ribosomal protein L5 (Fragment)	RPL5
Synaptotagmin like 2	SYTL2

Table 10: List of the 42 potential interactors collected from n=2. These genes meet the criterial of having at least 2 unique peptide hits. It should be noted that all potential interactors in this table are for *ZFYS* and *ZFYL*.

Description	Gene Name
Poly [ADP-ribose] polymerase 1	PARP1
Histone H4	H4C1
60S ribosomal protein L7	RPL7
Histone H1.4	H1-4
Heat shock cognate 71 kDa protein	HSPA8
Nucleolin	NCL
Zinc finger CCCH domain-containing protein 18	ZCH18
Ribosomal protein L18	RPL18
L-lactate dehydrogenase A chain	LDHA
60S ribosomal protein L6	RPL6
Heat shock 70 kDa protein 1B	HSPA1B
Ninein	NIN
Adenosylhomocysteinase	AHCY
L-lactate dehydrogenase B chain	LDHB
Heterogeneous nuclear ribonucleoproteins A2/B1	HNRNPA2B1
ATP-dependent RNA helicase DDX3X	DDX3X
Heterogeneous nuclear ribonucleoproteins C1/C2	HNRNPC
Tubulin alpha-3C chain	TUBA3C
Probable 28S rRNA (cytosine(4447)-C(5))-methyltransferase	NOP2
Histidine--tRNA ligase_ mitochondrial	HARS2
Polyubiquitin-B	UBB
Heterogeneous nuclear ribonucleoprotein U (Fragment)	HNRNPU
ATP synthase subunit alpha_ mitochondrial	ATP5F1A
Phosphatidylinositol 3_4_5-trisphosphate 5-phosphatase 2	INPPL1
40S ribosomal protein S2	RPS2
ELAV-like protein 1	ELAVL1
28S ribosomal protein S29_ mitochondrial	DAP3
Family with sequence similarity 184 member A OS=Homo sapiens OX=9606 GN=FAM184A PE=1 SV=1	FAM184A
60S ribosomal protein L8 (Fragment)	
Heat shock protein HSP 90-alpha	HSP90AA1

60S ribosomal protein L11	RPL11
Tubulin beta chain	TUBB
Histone H2B type 1-J	H2BC11
Actin beta (Fragment)	ACTB
Voltage dependent anion channel 2 (Fragment)	VDAC2
Coiled-coil domain-containing protein 113	CCDC113
GTP-binding nuclear protein Ran	RAN
ATP-dependent RNA helicase	EIF4A2
Prohibitin	PHB2
60S ribosomal protein L14	RPL14
RNA helicase	DDX17
Probable ATP-dependent RNA helicase DDX5	DDX5

Table 11: Proteomics duplicated pull-down hits. Using n. of 2, direct targets of *ZFY* were identified across both sample (*ZFYS* and *ZFYL*). The hits in this table were identified in both samples and have a unique peptide count greater than or equal to 2.

		Sample 1			Sample 2		
Gene Name	**Gene Symbol**	**Peptide Count**	**Unique Peptide**	**Confidence**	**Peptide Count**	**Unique Peptide**	**Confidence**
Heat Shock cognate 71kDa protein	HSPA8	12	10	74.6029	8	7	48.3901
Nucleolin	NCL	11	11	59.655	6	6	23.1967
Poly [ADP-Ribose] polymerase 1	PARP1	12	12	59.655	10	10	57.349
Tubulin beta chain	TUBB	10	4	52.9523	2	2	8.2613
Tubulin alpha-3C chain	TUBA3C	6	2	31.0904	4	4	22.8453
Heterogenous nuclear ribonucleoprotein C1/C2	HNRNPC	5	5	27.8964	4	4	19.5026
ATP synthase subunit alpha_mitochondrial	ATP5F1A	5	5	26.8599	3	3	12.7583
Polyubiquitin-B	UBB	5	4	25.6965	4	4	11.8271
L-lactate dehydrogenase B chain	LDHB	5	5	22.6404	5	4	26.2335
Probable 28s rRNA	NOP2	5	5	21.4783	3	3	11.9395
60S ribosomal protein L7	RPL7	4	4	20.146	8	8	49.4532
Histone H4	H4C1	3	3	17.339	9	9	46.2052
ATP-dependent RNA helicase	DDX3X	3	3	13.6898	4	4	15.7706
60S ribosomal protein L11	RPL11	2	2	9.2241	2	2	10.3532
28S ribosomal protein S29	DAP3	2	2	7.5932	2	2	9.3076
Heterogenous nuclear ribonucleoprotein A2/B1	HNRNPA2B1	4	4	20.9158	4	4	19.5026
Heterogenous nuclear ribonucleoprotein U	HNRNPU	4	4	24.3827	3	3	16.4796
Ribosomal protein L18	RPL18	5	5	27.8452	5	5	28.8642
60S ribosomal protein L14	RPL14	4	4	20.1184	2	2	16.9169