Kent Academic Repository

Martell, Henry (2018) *Analysis of Genetic Variation in Humans and Other Species.* Doctor of Philosophy (PhD) thesis, University of Kent,.

Downloaded from

https://kar.kent.ac.uk/76218/ The University of Kent's Academic Repository KAR

The version of record is available from

This document version

UNSPECIFIED

DOI for this version

Licence for this version

UNSPECIFIED

Additional information

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in *Title of Journal* , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

Analysis of Genetic Variation in Humans and Other Species

A PhD Thesis for the Degree of

Doctor of Philosophy

in

Computational Biology

Faculty of Sciences, School of Biosciences,

University of Kent

Henry James Martell

2018

Declaration

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

Name: Henry James Martell

Date: 19th September 2018

....

Abstract

Recent advances in sequencing technologies have led to the generation of vast amounts genetic variation data for many species, including humans, with advances in our understanding of disease now limited by the speed at which these data can be analysed. This thesis focuses on the analysis of genetic variation at multiple levels. First, in the human disease cystinuria, an inherited form of kidney stones. Genetic variants previously associated with cystinuria were characterised using a series of computational methods, identifying key functional features of these mutations. Predictions of disease severity for a cohort of 74 cystinuria patients were then made based on the genotypes of each individual. When compared to clinical outcomes, these predictions demonstrate the potential for computational methods in delivering precision medicine to cystinuria patients.

Second, a genome-wide analysis of variant combinations in individual human genomes identified combinations of variants protein-wide, within close spatial proximity in the 3-dimensional structures of proteins, and in protein-protein interface sites. The vast majority of computational methods for analysing genetic variation consider only one variant at a time. This work highlights the importance of analysing the combined effects of variants, which will be a key challenge in the future of computational biology and precision medicine.

Finally, two different analyses of ebolaviruses were performed. The first study focused on human pathogenicity of ebolaviruses, a critical challenge in epidemiology. This study identified a set of key variants that differentiate human pathogenic and non-pathogenic ebolaviruses. The second study focused on the evolution of the *Ebola virus* genome, the most common causative species of human ebolavirus outbreaks. *Ebola virus* genome evolution was analysed over time since its identification in 1976, and over the course of the 2013-2016 West Africa outbreak. A strong bias for transition mutations was identified, with suspected mutational pressure from host APOBEC and ADAR enzymes.

Acknowledgements

I would like to thank my primary supervisor Mark Wass who has helped me immensely in all of the work that I have done during my PhD, as well as given me countless fantastic opportunities, and always been a wonderful person to work for. I would like to thank my secondary supervisor Darren Griffin, who has given me many opportunities inside and outside of science, and has always encouraged me and offered advice. I would also like to thank Martin Michaelis, Tim Fenton, and Peter Ellis, all of whom have given their time and expertise to me during my PhD.

I would like to thank the BBSRC, who funded my PhD, along with all of my collaborators at JSR Genetics and Topigs-Norsvin, especially Grant Walling and Egbert Knol, and finally Barbara Harlizius, Maren van Son, and Marcos Lopes for our regular discussions over the last few years.

I would like to thank all of the members of the Wass lab, past and present, especially Morena Pappalardo, Miguel Juliá-Molina, Stefani Dritsa, and Magdalena Antczak, who have made the lab a great place to work and helped me at countless times in my PhD. I would like to especially thank Ian Reddin for being one of the funniest people I have ever met, particularly necessary when we were learning Perl together. I would like to thank Jake McGreig & Helen Grimsley for keeping me sane in my final year with endless racquet sports. I would like to thank all of the members of the Griffin lab, past and present, but especially Becky O’Connor who mentored me when I started and has been a close friend for the last four years.

I would like to thank my mum and dad, whom I owe everything to, and who have supported me throughout my PhD. I would like to thank my siblings Rebecca, Emma, and Thomas for their love and support, and especially Thomas for teaching me to love science. I would also like to thank all of my friends, the glorious Mrs Susan, and literally every dog ever.

Most of all, I would like to thank Joanna Bird, my partner in everything, who has been there for me every step of my PhD, who always believed in me, who kept me calm during the heights of PhD stress, and who always kept me smiling.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AUC	Area Under the Curve
BDBV	*Bundibugyo virus*
bp	Base Pairs
CAFA	Critical Assessment of Function Annotation
CAGI	Critical Assessment of Genome Interpretation
CAPRI	Critical Assessment of Prediction of Interactions
CASP	Critical Assessment of Protein Structure Prediction
CNV	Copy Number Variant
cryoEM	Single-particle cryo electron microscopy
DCA	Direct Coupling Analysis
DDD	Deciphering Developmental Disorders
DNA	Deoxyribonucleic Acid
EBOV	*Ebola virus*
ESP	Exome Sequencing Project
EVD	Ebolavirus Disease
ExAC	Exome Aggregation Consortium
FNR	False Negative Rate
FPR	False Positive Rate
GATK	Genome Analysis Toolkit
Gb	Gigabases
GnomAD	Genome Aggregation Database
HIV	Human Immunodeficiency Virus
HMM	Hidden Markov Model
IFN	Interferon
Indels	Insertions and deletions
Kb	Kilobases
kDa	Kilodalton (1,000 daltons)
mRNA	Messenger RNA
MNV	Multinucleotide Variant
MSA	Multiple Sequence Alignment
NCBI	National Center for Biotechnology Information

NMR	Nuclear Magnetic Resonance
nsSNV	Non-Synonymous Single Nucleotide Variant
PPARG	Peroxisome Proliferator-Activated Receptor Gamma
RNA	Ribonucleic Acid
RESTV	*Reston virus*
SDP	Specificity Determining Position
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
SNV	Single Nucleotide Variant
sSNV	Synonymous Single Nucleotide Variant
SUDV	*Sudan virus*
TAFV	*Tai Forest virus*
TCGA	The Cancer Genome Atlas
TNR	True Negative Rate
TPR	True Positive Rate
USA	United States of America
VCF	Variant Call Format
VEP	The Ensembl Variant Effect Predictor
Vif	Viral infectivity factor
WES	Whole Exome Sequencing

Chapter 1: Introduction

This thesis focuses on deciphering functional impacts of genetic variation. This broad aim has been applied to the study of human genetic variation, both at a global proteome level and to the study of a specific human disease, cystinuria, and to the study of ebolaviruses. This thesis is structured as a series of papers, one focused on analysis of genetic variants associated with the disease cystinuria, one on identifying coevolution within human genomes, one on the molecular determinants of human pathogenicity in ebolaviruses, and one on mutational bias in *Ebola virus* genome evolution. Additionally, one paper describing a new method for constructing chromosome-level genome assemblies, with proof of principle construction and analysis of the pigeon and peregrine falcon genomes, and another paper using quantitative genetics and next-generation sequencing data to identify functional variants affecting phenotypic traits in pigs are included as appendices (Appendix 5 and Appendix 6, respectively).

1.1 Genetic Variation

The phenotype of an organism is determined by its genome, the sum total of its genetic content. Environmental factors being equal, phenotypic differences between organisms are determined by genetic variation – large and small differences between genomes. Genetic variation is critical for life, allowing for the advent of new functions and adaptation to changing environments at the species level, but within one individual organism genetic variation can be the difference between health and disease. Due to recent advances in molecular biology, there is an unprecedented amount of genetic variation data available today. This big biological data was unimaginable even 20 years ago and it continues to grow rapidly. One key aim of computational biology is to understand how genetic variation observed in nature causes phenotypic differences, and to build predictive models that can glean actionable insights from big biological data.

1.1.1 Single Nucleotide Variants

A single nucleotide variant (SNV) is the change of a single nucleotide in a genome, and is the smallest and most basic type of genetic variation. If an SNV occurs commonly in a population (>1% frequency), it is typically referred to as a single nucleotide polymorphism (SNP). The human reference genome is roughly 3.1Gb in length, and so an SNV corresponds to ~0.0000000003% difference between an otherwise identical genome. However, such small differences are not trivial, with many human diseases caused by SNVs. Sickle cell disease is associated with various health complications in humans, including episodes of acute illness, progressive organ damage, and reduced lifespan. The most common form of sickle cell disease is caused by an SNV that results in a mutation from glutamic acid to valine in the β-globin gene (Ingram 1959; OMIM 2017b).

The SNV causing sickle cell disease is an extreme example of an SNV with a large phenotypic effect, but it demonstrates the importance of the sequence context of mutations. Not all regions of the genome are equally important to the fitness of an organism. SNVs in protein-coding regions of the genome, like the SNV that causes sickle cell disease, have the potential to change the sequences of the encoded proteins. An SNV that changes the amino acid sequence of a protein is termed a non-synonymous SNV (nsSNV). An SNV in a coding region that preserves the amino acid sequence of the protein, due to the degeneracy in the genetic code, is termed a synonymous SNV (sSNV).

sSNVs generally have a smaller impact on protein function, but their effects are not negligible. sSNVs can affect mRNA splicing, mRNA stability, and protein folding via changes in translation speed (Sauna & Kimchi-sarfaty 2011), and are implicated in human diseases, such as cancer (Supek et al. 2014). The Multidrug Resistance 1 gene encodes the product P-glycoprotein, which is an ATP-dependent efflux pump, genetic alterations to which are associated with resistance to multiple drugs in cancer (Gottesman et al. 2002). A common haplotype (a combination of variants inherited together) in P-glycoprotein, that results in differences in substrate specificity and inhibitor interaction, contains both nsSNVs and sSNVs (Kimchi-Sarfaty et al. 2007). Crucially, the nsSNVs on their own do not cause the observed differences P-glycoprotein, which are only observed when the sSNV is also present (Kimchi-Sarfaty et al. 2007) . In summary, functional effects are rarer for sSNVs compared to nsSNVs, but sSNVs can play an important role in the phenotype of an organism and their historic designation as 'silent' is now known to be a misnomer. However, computational prediction of the effects of sSNVs remains more difficult than for nsSNVs.

Roughly 1.22% of the human genome is composed of protein-coding exons (Dunham et al. 2012), which are the regions of the genome where nsSNVs and sSNVs occur. The remaining ~99% of the human genome is the non-coding portion. For all variants currently associated with disease, >80% are within non-coding regions of the genome (Dunham et al. 2012). Diagnostic rates of rare childhood diseases using whole exome sequencing (WES), which is unable to detect many types of non-coding variation (Zappala & Montgomery 2016), ranges depending on the disease. Most diseases have diagnostic rates well below 100%, such as syndromic congenital heart disease, which has a diagnostic rate of 9.7% using whole exome sequencing (Zappala & Montgomery 2016; Wright et al. 2018). There are some exceptions where WES performs much better, such as osteogenesis imperfecta (100%) (Wright et al. 2018), but for cases where WES data is unable to make a diagnosis it is assumed that non-coding variation is causing the observed disease phenotype (Wright et al. 2018).

Non-coding regions of the genome are typically under less evolutionary constraint than protein-coding regions of the genome (Dunham et al. 2012), but genetic variation in non-coding regions can still be damaging. Changes to promoters, enhancers, splice sites, and genome packaging and organisation can all have profound impacts on the phenotype of an organism. While progress has been made in understanding the roles of different non-coding regions (Feingold et al. 2004; Birney et al. 2007; Dunham et al. 2012), the impact of genetic variation within these sites remains more difficult to predict compared to protein-coding variation. Typically, methods use information about interspecies and intraspecies conservation of non-coding regions, as well as functional annotations of non-coding regions (Dunham et al. 2012), to train machine learning methods to identify potentially functional non-coding genetic variation (Ritchie et al. 2014; Fu et al. 2014; Zhou & Troyanskaya 2015; Shihab et al. 2015; Ionita-Laza et al. 2016; Smedley et al. 2016). These methods enable the prioritisation of non-coding variants, which are most likely to have functional effects, but much progress is needed to improve predictive accuracy.

1.1.2 Structural Variants

Genetic variation also occurs at a structural level and there are many different types of structural variation, ranging from single nucleotide to whole genome aberrations. Insertions and deletions (indels) are the addition or removal of bases to an organism's genome, and these can be ≥1 base in size. Indels within protein-coding regions of the genome can be either in-frame or out-of-frame, depending on whether they conserve the open-reading frame of the gene. Out-of-frame indels are typically more damaging, as they will likely change all encoded amino acids downstream of the indel, and can lead to early stop codons within exons. Indels in non-coding regions of the genome can also potentially disrupt functional non-coding elements by changing their sequence.

Other types of structural variants often affect much larger regions of the genome compared to indels. Copy number variants (CNVs) are differences in the number of copies of a region of a genome compared to the reference genome. CNVs can be small, such as CNVs of the CAG trinucleotide repeat observed in Huntington's disease (OMIM 2018a), or span larger genome regions, including rare cases of CNVs spanning entire genes, such as human amylase genes (Perry et al. 2007).

Structural rearrangements also contribute to phenotypic diversity. These include inversions, where the orientation of a genome region is reversed, and translocations, where a portion of a chromosome is transferred to another chromosome. Translocations can be reciprocal, where material is transferred between two non-homologous chromosomes, or non-reciprocal, where material is transferred from one non-homologous chromosome to another. Translocations are a common causative agent in cancer, often resulting in gene fusions that lead to aberrant functions. Many cancers are characterised by specific, recurrent translocations, such as fusions of the EWSRI gene in Ewing Sarcoma (Riggi et al. 2007).

At the species level, chromosomal rearrangements often delineate key evolutionary events. For example, avian genomes are uniquely organised and almost all extant birds have a characteristic karyotype containing a large number of microchromosomes – which are chromosomes <20Mb in size, smaller than the smallest human chromosome (chromosome 21 - ~47Mb) (Nishida et al. 2008). The chicken genome is a typical avian genome, with 39 pairs of chromosomes, 33 of which are microchromosomes. A less typical avian genome is that of the Peregrine falcon, in which over the course of evolution many of its microchromosomes have fused back together (Damas et al. 2017). Both inter- and intra-chromosomal rearrangements have played a key role in the evolution of modern birds, and chromosomal evolution of the Peregrine falcon was one focus of the work presented in Appendix 5 of this thesis (Damas et al. 2017).

Variation can also occur at the level of entire chromosomes, with multiple human diseases caused by non-disjunction of chromosomes during cell replication, resulting in an additional copy of a chromosome (aneuploidy), such as in Down's syndrome where individuals have an extra copy of chromosome 21 (OMIM 2017a). At the species level, variation in chromosome number has led to the evolution of polyploid genomes in many plants – genomes with more than 2 copies of each chromosome (Adams & Wendel 2005; Comai 2005). Many species have also evolved what are referred to as B chromosomes – these are chromosomes which are not essential for the survival of an organism, but are present in some individuals within populations and provide an additional source of genetic variation (Houben 2017).

1.2 Human Genetic Variation

1.2.1 The Human Genome Project

Over the course of the 20th century, the study of genetics progressed from Mendel's fundamental laws of heredity to a working knowledge of the molecular basis by which cells read and store the information that encodes life, as wells as the tools with which to manipulate these biological processes (Lander et al. 2001). The study of human genetics was revolutionised again in 2001, with the publication of the first draft of the human genome (Lander et al. 2001). This was followed by a more complete version published in 2004 (Collins et al. 2004), and although the human genome remains unfinished, with some highly repetitive regions proving hard to sequence and assemble correctly, many improvements have been made since the original publication of the human genome, up until the most recent genome version (GRCh38/HG38).

The end result of the Human Genome Project was the first near-complete genome of a vertebrate, detailing the sequences of the 23 pairs of chromosomes that make up the human genome, and providing researchers with a tool for more precise genetic analyses than previously possible. One of the most notable findings from the project was that the human genome only encodes ~20,000-25,000 proteins, far fewer than was predicted at the time (Collins et al. 2004). This finding led to a fundamental shift in the perception of biological complexity and how it is achieved.

The Human Genome Project took roughly 13 years to complete, cost billions of dollars, and required the combined efforts of multiple international research groups, but just over a decade after the completion of the Human Genome Project the sequencing of a single human genome is comparatively trivial.

1.2.1 Genetic Variation in the Post-Genome Era

In the early 2000s the cost of sequencing a genome was still prohibitively high for large-scale projects (Figure 1.1). Many projects studying genetic variation during this time focused on the use of genotyping arrays, an older and cheaper technology compared to genome sequencing, but which provides lower-density data.

The International HapMap project was launched in 2002 with the aim of uncovering common patterns of genetic variation and producing a haplotype map of the human genome – sets of variants inherited together on single chromosomes or in a chromosome region (The International HapMap Consortium 2003). The HapMap project proceeded through three main phases. Phase One was completed in 2005 and reported complete genotypes for 269 samples from four populations, containing 1,007,329 high-quality variants (Belmont et al. 2005). Phase Two was completed in 2007, and improved on the Phase One HapMap result with >3.1 million SNPs genotyped for 270 individuals from four populations.

The third and final HapMap Phase (HapMap 3) was completed in 2010, and combined the analysis of common and rare variants. A total of 1,184 individuals were genotyped for 1.6 million common variants, and then a subset of 692 of these individuals had select genome regions sequenced (Altshuler et al. 2010). In addition to providing insight in to common inheritance patterns of human genetic variation, as the first large-scale analysis of human haplotypes, the HapMap Project data has also been utilised in many genome-wide association studies to identify phenotypic traits associated with different genotypes (Frazer et al. 2007).

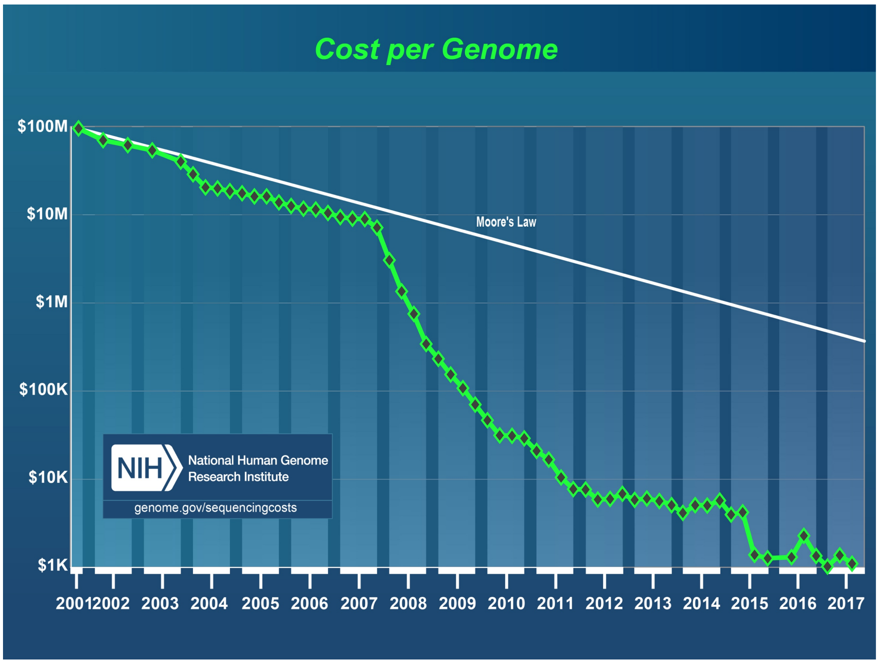


Figure 1.1: The cost of sequencing a genome over time. This figure is reproduced with permission from the National Human Genome Research Institute (Wetterstrand 2018).

With so much genetic variation data being generated, and with many large-scale sequencing projects on the horizon, a multitude of genetic variation databases have been established to store the data and make it publicly available. The most general genetic variation database, dbSNP, hosts variation information for multiple species (53 in total) and was established by the National Center for Biotechnology Information (NCBI) (Sherry 2001). The current dbSNP release (build 151, last updated 22/03/2018) contains 660,773,127 reference SNPs and 1,803,563,957 constituent submitted SNPs (https://www.ncbi.nlm.nih.gov/projects/SNP/). NCBI also produces a sister database to dbSNP called dbVar, which is focused on structural variants and complements dbSNP (which is focused on SNVs). The current release of dbVar (last updated 10/06/2018) contains 34.6 million variant calls (https://www.ncbi.nlm.nih.gov/dbvar/).

Other databases of genetic variation are more specialised than dbSNP. Both Humsavar (Bateman et al. 2015) and ClinVar (Harrison et al. 2016) are genetic variation databases focused on clinically relevant human genetic variants. Humsavar (07/2018 release) contains a total of 77,936 variants, with 30,210 associated with disease, 39,959 natural polymorphisms, and 7,767 unclassified variants. ClinVar contains a total of 701,674 variants, 684,317 of which have an associated clinical interpretation. Humsavar and ClinVar are both important tools in interpreting genetic variation data from new samples, combined with tools like Online Mendelian Inheritance in Man (OMIM), which focuses on the relationship between genes and disease phenotypes as well as specific disease variants (OMIM 2018b). VariBench is another specialised database of genetic variants that contains a mixture of disease causing variants and natural variants, but was specifically designed for benchmarking methods for predicting the pathogenicity of variants (Sasidharan Nair & Vihinen 2012).

Since the first draft of the human genome in 2001, the cost of sequencing a genome has dropped dramatically (Figure 1.1). The human genome project itself finished two years earlier than planned and also under budget, due to advances in sequencing technologies made during the project (Collins et al. 2004). After the human genome project was completed, these sequencing advances were leveraged to generate additional human genomes and the extent of variation between two human genomes could now be investigated in more detail than had previously been possible.

Initial comparisons of entire human genomes showed that there could be as many as 4.1 million DNA variants compared to the reference genome: 3,213401 SNVs, 53,823 substitutions between 2-206bp in size, 292,102 heterozygous indels (1-571bp in size), 559,473 homozygous indels (1-82,711bp in size), 90 inversions, and a number of segmental duplications and CNVs (Levy et al. 2007). SNVs were found to be the most numerous type of variant, accounting for the majority of variants between genomes (78%), but 74% of variant bases were from larger non-SNV changes (Levy et al. 2007).

While various sequencing projects were happening concurrently after the completion of the human reference genome, the successor to the Human Genome Project in terms of scale and ambition was the 1,000 Genomes Project.

1.2.2 The 1,000 Genomes Project

The 1,000 Genomes Project began in 2008 with a pilot study, which was followed by three project phases, the last of which ended in 2015 (1000 Genomes Project Consortium 2015). The aim of the project was to produce a comprehensive reference of common genetic variation in humans, defined as variants occurring in >1% of individuals, which could then be used to advance many areas of biological research. The project used a combination of whole-genome sequencing (at low coverage), deep exome sequencing, and dense microarray genotyping to reconstruct the genomes of 2,504 individual humans and identified >88 million variants. Of these, 84.7 million were SNVs, 3.6 million were short indels, and 60,000 were larger structural variants (1000 Genomes Project Consortium 2015).

The majority of the variants identified in each individual genome occur in non-coding regions, and of these 459,000-565,000 per individual overlap with known functional non-coding regions. Protein-coding variants are much rarer but still occur fairly frequently. Each genome contains 149-182 protein truncating variants and 10,000-12,000 variants that alter the sequence of the encoded polypeptides (1000 Genomes Project Consortium 2015). The majority of the variants identified across individuals are rare (frequency <0.5%), but the majority of variants observed within a single human genome were found to be common, with only 1-4% of variants in an individual having a frequency <0.5%.

This set of 2,504 genomes was sampled from 26 different human populations (Figure 1.2), with the design of the study and the population sampling informed by previous projects, such as the HapMap Project (Altshuler et al. 2010). These data provide a reference for >99% of genetic variants in these 26 populations that occur in >1% of individuals (1000 Genomes Project Consortium 2015). The typical human genome was found to differ from the human reference genome at 4.1 million to 5.0 million sites. As previously found, the majority of variants are SNVs but the majority of variant bases are from larger types of structural variation, with the typical genome having 2,100-2,500 structural variants (Levy et al. 2007; 1000 Genomes Project Consortium 2015).

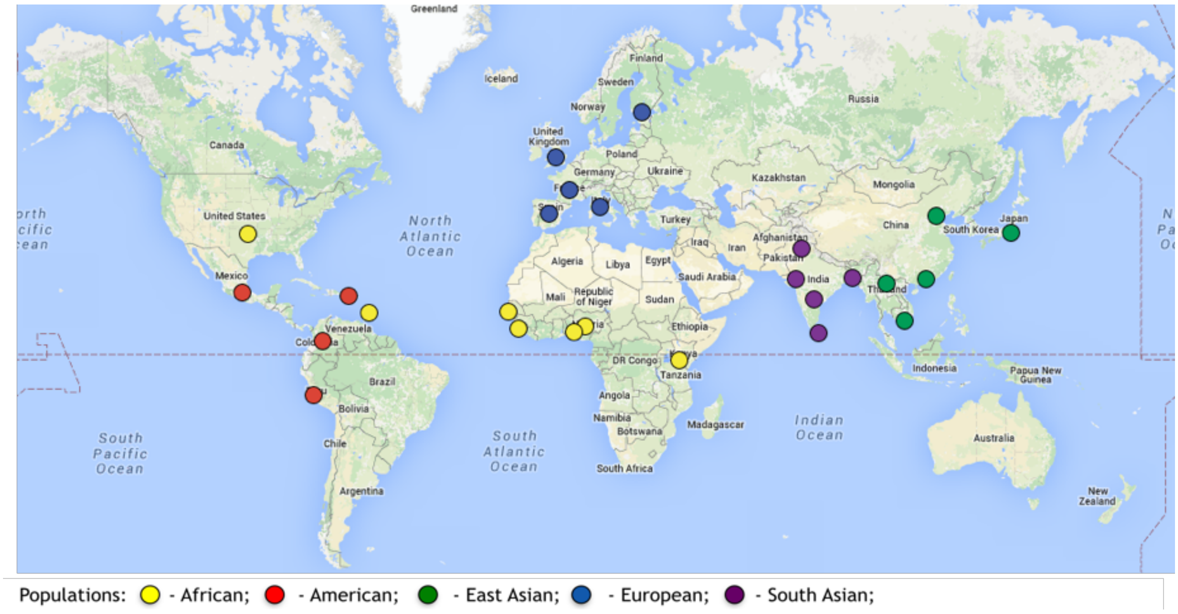


Figure 1.2: Populations sampled for the 1,000 Genomes Project. This figure is reproduced with permission from the International Genome Sample Resource (International Genome Sample Resource 2018).

The numbers of observed variant sites compared to the reference genome varied considerably between the populations sampled, with the African super population having the highest number of variant sites. Individual variants also occurred at highly different frequencies between different populations used in the 1,000 Genomes Project (Figure 1.2), with 762,000 variants that were found to be rare across populations (<0.5% frequency) but common in at least one population sampled (>5% frequency) (1000 Genomes Project Consortium 2015).

Each of the 2,504 genomes sequenced in the 1,000 Genomes Project was also phased to high-quality. This means that for each genome it is known which variants correspond to which copy of a chromosome, and therefore which combinations of variants were inherited together, and for protein-coding variants which variants occur in the same copies of proteins. The high-quality, size and public availability of the 1,000 Genomes Project data have made it a crucial resource for studies of genetic variation, even as other studies have eclipsed it in terms of sample size.

1.2.3 Rapid Human Genome Sequence Growth

Since the completion of the 1,000 genomes project the pace of genome sequencing has further increased, with sequencing projects targeting ever larger sample sizes. The Exome Sequencing Project (ESP) began shortly after the 1,000 Genomes Project in 2009 and was completed before phase 3 of the 1,000 Genomes Project. Its focus was in identifying rare, likely functional protein-coding variants associated with heart, lung, and blood diseases (Fu et al. 2013). The ESP exceeded the 1,000 Genomes Project in terms of sample size, with 6,515 samples, but was more limited due to only sequencing the exomes of the samples and also due to sampling a smaller set of populations.

Despite these limitations, the ESP has been a valuable resource for quantifying protein-coding variation, for associating >70 different traits associated with heart, lung, and blood diseases, and in the planning of future large-scale sequencing projects (Auer et al. 2016). One interesting finding from the ESP data was the estimation that 73% of all protein-coding SNVs and 86% of SNVs predicted to be deleterious have arisen within the last 5,000-10,000 years. This indicates that not enough time has elapsed for these mutations to be removed from human populations via purifying selection, and highlights the burden of recently occurring deleterious mutations segregating in human populations (Fu et al. 2013).

The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/) aims to catalogue the key genomic changes for many types of cancer, and is a collection of matched tumour and normal tissue samples from 11,000 patients, across 33 different tumour types. The project produced 2.5 petabytes of data, combining data from RNA sequencing, MicroRNA sequencing, DNA sequencing, SNP arrays, DNA methylation arrays, and protein expression arrays (Tomczak et al. 2015). TCGA data have been used to identify many key genomic changes associated with different types of cancer, changed the way that tumours are classified by identifying many sub-types of tumours, and advanced therapeutic options for cancers by identifying targetable genomic changes for current treatments and also identifying targets for future therapy development (Tomczak et al. 2015).

Similar to TCGA, the Deciphering Developmental Disorders (DDD) Study is a large-scale sequencing project aimed at understanding disease phenotypes, but the DDD study is focused on the analysis of human developmental disorders. The DDD study began recruiting patients in 2011 and has since recruited thousands of children with developmental disorders, and there are 120 published studies using the DDD Study to date (as of 29/08/2018).

In the latest DDD study publication, an analysis of 4,293 patients with severe, undiagnosed developmental disorders was reported, with exome sequencing performed for the affected children and also their parents (Deciphering Developmental Disorders Study 2017). Utilising the data from affected individuals as well as their parents, this work focused on the prevalence of *de novo* mutations in developmental disorders, and estimated that 42% of their cohort carry pathogenic *de novo* mutations in coding regions. From these data, and given current global demographics, it was estimated that 400,000 children worldwide are born per year with developmental disorders caused by *de novo* mutations (Deciphering Developmental Disorders Study 2017). The DDD study has also helped identify 12 novel genes associated with developmental disorders. These new genes have increased the diagnosis rate of the children with developmental disorders in the study by 10% (from 28% to 31%) (The Deciphering Developmental Disorders Study 2015).

Many of these large sequencing projects have produced vast amounts of genetic variation data, but their true potential is often limited by the difficulty in comparing data between studies. Many studies use different sequencing approaches, and almost all use different methods to call and filter the raw sequencing data. To try and maximise the utility of the sequencing data available a project was started to collate multiple sequencing efforts and standardise the protocols used to call and filter variants. This became known as the Exome Aggregation Consortium (ExAC).

In 2016, ExAC published its analysis of a set of 60,706 high-quality exomes, all of which were generated using a standardised protocol for calling variants (Monkol Lek et al. 2016). Across the set of exomes, 10,195,872 unique variants were identified, and 7,404,909 were kept after filtering – this corresponds to one variant for every eight base pairs in the exome (Monkol Lek et al. 2016). The majority of the variants identified are rare variants, with 99% having a frequency <1%, and 54% occurring in only one sample in the dataset. The sample size of ExAC enabled the detection of rare variants on a scale not previously possible, with 72% of the variants identified not present in either the 1,000 Genomes dataset or the ESP dataset (Monkol Lek et al. 2016). The size of the ExAC dataset has also identified that 7.9% of high-quality sites analysed are multiallelic (>1 variant observed at the same position in different individuals), far more than the 0.48% identified from the 1,000 Genomes Project or the 0.43% from the ESP (Monkol Lek et al. 2016).

Mutational recurrence is the phenomenon of identical mutations occurring independently in different individuals. Another key finding from the ExAC dataset was the identification of high-rates of mutational recurrence, with 43% of *de novo* mutations previously identified from parent-offspring trio studies also identified in the ExAC dataset, implicating recurrence of variants (Monkol Lek et al. 2016).

The ExAC data set has also highlighted the importance of haplotype phasing to avoid incorrect annotation of multinucleotide variants (MNVs), which are cases where multiple nucleotide substitution events have occurred in the same codon of a protein-coding gene. Haplotype phasing is the assignment of variants to copies of chromosomes, thereby identifying which variants are inherited as haplotypes (variants on the same copies of chromosomes that are inherited together), and without haplotype phasing it is not possible to correctly annotate MNVs. For example, if an individual has two nucleotide variants that both occur in the same codon, these may occur on different copies of that individual's chromosomes, resulting in each copy of the encoded protein having a single variant (which could be synonymous or non-synonymous). Alternatively, these two nucleotide variants may have occurred on the same chromosome copy, which would result in a different codon than either of the single variants alone (this final codon could also be a synonymous or non-synonymous change compared to the reference genome).

In the ExAC data set, there were 5,495 MNVs identified, with on average 23 MNVs per sample. In each of these cases, analysis of the individual SNVs, without haplotype phasing, would have led to the incorrect interpretation of the variants. Of these MNVs, there were 647 examples where an MNV would rescue a protein-truncating variant caused by one of the constituent SNVs, and 131 examples where the MNV resulted in a protein-truncating variant not caused by the constituent SNVs alone (Monkol Lek et al. 2016).

The ExAC data resource was released in an aggregated format, which was necessary to maintain the privacy of the individuals involved, but also limits the information available from ExAC and therefore its utility. For example, the work presented in Chapter 3 of this thesis requires knowledge of all of the variants that each individual sample has, which is possible using the 1,000 Genomes Project data. However, with the ExAC data set it is only possible to know how frequently each variant occurs in the population. Nevertheless, such a large reference database of genetic variation has proved a useful tool for the analysis of new sample data. For rare variants and rare diseases, variants that occur above a certain frequency in ExAC can be ruled out as unlikely to be the causative variants, with variants occurring at >1% frequency often filtered out in variant prioritisation (Monkol Lek et al. 2016). However, not all of the individuals within ExAC can be considered ‘healthy’, as some samples are taken from disease cohort studies, which must be taken in to account if using ExAC data to study diseases that a subset of ExAC samples have (Monkol Lek et al. 2016).

The scale of the ExAC dataset has already changed the way that genetic variation in humans is viewed. In prion disease, comparison of the genomes of prion disease sufferers with the large control population of ExAC showed that many of the genetic variants previously associated with prion disease occur far too frequently in the ExAC population to be causative of the disease. Some of the previously associated variants were 30 times more common in ExAC than expected based on the prevalence of prion disease, and these variants were either previously falsely associated with the disease or their penetrance depends on other genetic factors (Minikel et al. 2016). Similar trends have been observed for mutations previously associated with cardiomyopathy, with multiple mutations occurring far more frequently in the ExAC population than expected based on the prevalence of the disease (Walsh et al. 2017). We observed a similar pattern in the work presented in Chapter 2 – many cystinuria associated variants were found to have an allele frequency ~30% in ExAC, despite the worldwide incidence of cystinuria being 1 in 7,000 (Martell et al. 2017).

With the ever-increasing pace of genome sequencing, the ExAC database has already been replaced by the Genome Aggregation Database (GnomAD) (Karczewski et al. 2017). The GnomAD database not only contains all of the exomes of ExAC, but an additional 62,430 exomes, and now its first whole genomes, with 15,496 full human genomes in the current GnomAD release, drawing data from 47 individual sequencing projects (Karczewski et al. 2017)

1.2.4 Ongoing and Future Genome Sequencing Projects

The pace of genome sequencing continues to advance, with many ongoing private and public sequencing projects. One of the largest ongoing projects is the 100,000 Genomes Project run by Genomics England, a company started and run by the Department of Health, England, in collaboration with the National Health Service (NHS), Public Health England, and Health Education England (Davies 2017). The project began in 2012, with the aim of sequencing 100,000 whole human genomes, with a particular focus on identifying the underlying causes of multiple types of cancer and rare diseases. There are also a series of smaller projects within the 100,000 Genomes Project, such as the analysis of infectious diseases, with 3,000 multidrug resistant tuberculosis organisms already sequenced. The large sample cohort combined with the detailed medical and family records available to the 100,000 Genomes Project make it a powerful data set for the advancement of precision medicine (Davies 2017).

The Personal Genomes Project was launched in 2005 with the aim of recruiting 100,000 volunteers to have their complete genomes published as well as their medical records. The project began exclusively in the United States at Harvard, but has since expanded to a global network of individual projects, with local projects in the United Kingdom, Canada, Austria, and the People's Republic of China. The aim of the Personal Genomes Project is to advance personal genomics and precision medicine by making this large collection of genotype and phenotype data freely available (Church 2005).

The UK Biobank is another large-scale, open-data precision medicine project, with a focus on the diseases of middle and old age. The study has recruited 500,000 participants aged between 40-69 years old, all of whom have consented to have a wide range of phenotypic and genotypic measurements taken, as well as longitudinal follow-up (Sudlow et al. 2015). Initial genotypic measurements have been limited to genome-wide genotyping, but whole exome sequencing of samples has now been approved in an industrial partnership with GlaxoSmithKline and Regeneron (UK Biobank 2018).

In 2015, President Obama announced the United States' Precision Medicine Initiative, a $215 million investment in precision medicine, with a particular focus on precision medicine in cancer treatment. This project became the 'All of Us' initiative, which aims to recruit one million participants and spearhead precision medicine advancement in the United States of America (USA) (National Institutes of Health 2018).

The Resilience Project is another large-scale precision medicine project in the USA, which is being coordinated by the Icahn Institute for Genomics. This project focuses on identifying individuals that carry known disease-causing variants but who show no clinical manifestation of the disease. The potential of the project was demonstrated in an analysis of 589,306 genomes that identified individuals with variants known to cause severe childhood mendelian disease who have reached adulthood with no manifestation of the disease (Chen et al. 2016). Such 'resilient' individuals are believed to carry additional variants that buffer the effects of the disease-causing variants, the identification of which could help in the development of new treatment options. The resilience project is also intriguing, because some of the data used in the project comes from 23andme, a private company that offers direct-to-consumer genetic testing. Such genetic testing products began primarily to report the ancestry and health-risks of customers, but the data generated is now being used in primary medical research (Chen et al. 2016).

Much of the current boom in genomics was fuelled by the development of next-generation sequencing technologies that were much faster than traditional Sanger sequencing, but which produced shorter reads. These short-read technologies are now referred to as second-generation sequencing technologies. Future sequencing projects will be able to leverage the advances made in third-generation sequencing technologies. Nanopore sequencing is one third-generation sequencing method, and it has already been used as a platform for rapid sequencing of virus genomes during the 2013-2016 West Africa *Ebola virus* outbreak, with the portability of the MinION platform being a major advantage over other platforms (Quick et al. 2016; Hoenen et al. 2016).

Other fast, long-read technologies have also made recent advances, such as PacBio SMRT sequencing. SMRT sequencing is able to produce much longer reads than second-generation sequencing technologies, with an average read length of 15,000 bp and some reads as long as 100,000 bp, with most second-generation sequencing technologies having a maximum read length between 150-400 bp. SMRT sequencing also has the advantage that its error rate is uniform across sequences, unlike many second-generation sequencing technologies which have large error biases (Roberts et al. 2013).

It has long been known that the organisation of an organism’s genome has a large functional impact. High-throughput detection of large-scale organisation of genomes is now becoming possible with advances in chromosome conformation capture technologies. These methods have applications both in the assembly of genomes (Mascher et al. 2017; Moll et al. 2017), and also in the study of variation in genome organisation, such as analysis of chromosomal rearrangements in cancer (Harewood et al. 2017).

1.3 Non-Human Genetic Variation

The genomics revolution has not been confined to the study of humans, with some animals having their genome sequenced before the human genome, such as *Caenorhabditis elegans* in 1998 (The C. elegans Sequencing Consortium 1998). Not long after the publication of the human genome in 2001, multiple animal genomes had already been sequenced. The initial focus was on animals of research or agricultural relevance, with the mouse genome published in 2002 (Waterston et al. 2002), followed by the rat genome in 2004 (Gibbs et al. 2004), the chicken genome in 2004 (International Chicken Genome Sequencing Consortium 2004), the cow genome in 2009 (The Bovine Genome Sequencing and Analysis Consortium 2009), and the pig genome in 2012 (Groenen et al. 2012). As of 03/09/2018 there are a total of 38,734 published genomes in the NCBI genomes database, including 6,504 eukaryotes, 156,689 prokaryotes, 18,749 viruses, 13,483 plasmids, and 12,079 organelles (https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/).

The number of non-human genome sequences is continuing to grow, with many ongoing large-scale projects. The Genome 10K Project aims to sequence the genome of at least one individual in each vertebrate genus – around 10,000 genomes (Haussler et al. 2009), the 5,000 Insect Genome Project aims to sequence the genomes of 5,000 insects and other arthropods (Robinson et al. 2011), the Bird 10K project aims to produce draft genomes for all extant species of birds – around 10,500 genomes (Zhang 2015), and the Earth BioGenome Project aims to sequence, catalogue, and characterise the genomes of all of Earth's eukaryotic biodiversity (Lewin et al. 2018).

One drawback of the majority of these *de novo* sequencing projects is that they often produce fragmented genome assemblies. The aim of sequencing and assembling a new reference genome for a species is to produce one contiguous sequence block per chromosome. This is often not possible due to the short-read lengths of many of the second-generation technologies that have enabled such large-scale sequencing projects. This means that the utility of the genomes produced is limited, as much of the functional information contained in the organisation of a species' genome is lost. This was the problem we sought to ameliorate in the paper presented in Appendix 5, which describes a new method combining computational and experimental methods to upgrade fragmented genome assemblies to chromosome-level. The potential of this new method is demonstrated with upgraded assemblies of both the pigeon and peregrine falcon genomes. Analyses of these new genome assemblies also provided insight in to the evolution of the unique avian karyotype, see Appendix 5 (Damas et al. 2017).

The generation of non-human genomes has clear impacts for the respective research fields of each organism, but their use in comparative genomics analyses can also advance our understanding of human genetic variation. High-quality non-human genomes can be used to identify conserved regions in genomes and help to identify regions of unrecognised importance, with inter-species sequence conservation a key feature used in many tools for predicting the effects of genetic variants.

1.4 Genetic Variation Analysis Tools

1.4.1 Variant Calling and Filtering

The rapid advances in sequencing technologies and subsequent growth in sequencing data have necessitated methods for their analysis which are fast and scalable. These methods developed rapidly and divergently, but in recent years there has been a drive to standardise approaches to the processing of sequencing data, such as the Genome Analysis Toolkit (GATK) developed by the Broad Institute (McKenna et al. 2010). The following section describes the best practices for the analysis of DNA sequencing data defined by GATK.

Modern sequencing technologies produce vast amounts of data in the form of reads – small sequence fragments of the sample being analysed. These reads are stored in FASTQ files, which store both the sequence of the read and also individual base quality scores for the read. Sequencing reads can then be aligned to a reference genome, and a SAM/BAM file can be produced that stores the read, its quality, and where it aligns to the reference genome (SAM files are uncompressed whereas BAM files are compressed and indexed, but both files store the same information). SAM/BAM files are a “platform-agnostic” file format, and once produced allow for the downstream analysis of sequence data from any sequencing platform (McKenna et al. 2010).

The next step is variant calling to identify differences between the sequenced sample and the reference genome. This step involves quality filtering of the sequence reads, either with hard-filtering, for example using thresholds for Phred scores (which describe the probability that a base in a sequence read has been called correctly), or more recent methods that use machine learning approaches, such as random forest classifiers, or with a combination of hard-filtering and machine learning, as with the GnomAD variant filtering protocol (Monkol Lek et al. 2016; Karczewski et al. 2017).

There are a large number of tools available to perform variant calling. One recent paper compared eight frequently used variant calling tools: GATK HaplotypeCaller, Platypus, VarScan, LoFreq, FreeBayes, SNVer, SAMtools and VarDict. These tools will each produce different call sets for the same input data, with each method having trade-off between sensitivity and precision of calling variants (Sandmann et al. 2017).

Variant calling remains a tricky problem, partially because of variation between pipelines used in different studies, but also because of reproducibility problems even when using the same calling pipeline. Using a set of 200 samples all taken from the same individual, variant calling and standardised variant filtering was performed to create 200 call sets for the same sample, filtering out low-complexity regions such as centromeres, telomeres and repetitive regions (~70% of the genome remains after this filtering). Across these samples, 97.3% of the region could be sequenced with high reproducibility (2,157Mb), but 184Mb of this region had acceptable reproducibility rates >90%, and 184Mb had reproducibility rates <90% (Telenti et al. 2016).

Regardless of the method used, this variant calling step produces a VCF (Variant Call Format) file that describes all of the differences between the sample and the reference genome. After variant calling additional filters may be applied to filter out common variants, or those with known clinical associations (Bateman et al. 2015; Monkol Lek et al. 2016; Landrum et al. 2018).

1.4.2 Variant Annotation

Once a high-quality variant call set has been produced, the next step is to annotate the variants. The call set simply provides coordinates for where in the genome a variant has occurred compared to the reference genome used. The purpose of variant annotation is to add context and functional information to these coordinates. For example, whether a variant is in a gene or in an intergenic region of the genome, and if in a protein coding gene whether it changes the sequence of the encoded protein.

The problem of variant annotation is not as simple as it first appears. Many tools are available for variant annotation, with two of the most commonly used ones being Annovar (Wang et al. 2010) and the Ensembl Variant Effect Predictor (VEP) (McLaren et al. 2016). Each of these tools is able to take a VCF file as input and annotate each of the variants contained, determining which variants occur in non-coding regions, which variants occur in coding regions, and what their likely functional effects are. However, the annotation agreement between these tools has been shown to be quite poor. For exonic variants, the agreement between Annovar and VEP has been measured at only 87%, with only 65% agreement for putative loss-of-function mutations (McCarthy et al. 2014). The choice of transcript set can also have a large effect on the annotations produced by Annovar and VEP, with loss-of-function variants having only 44% annotation agreement when comparing the two commonly used transcript sets of Ensembl and RefSeq (McCarthy et al. 2014).

The discrepancies in variant annotation between methods and transcript sets highlights the need for standardised methods for variant annotation, and also highlights the importance of efforts like ExAC in producing large data sets that are comparable across samples due to being processed with a single pipeline (Monkol Lek et al. 2016).

As discussed earlier, within the ExAC dataset there were 5,945 cases of MNVs, where the final impact of the variants would be incorrectly annotated without haplotype phasing and consideration of the combined variant effects (Monkol Lek et al. 2016). Even with a correctly phased variant call set, many annotation tools do not consider the effects of MNVs and will annotate each constituent SNV separately. Annovar does not currently interpret MNVs in a combined manner, and VEP has only recently implemented its Haplosaurus tool for annotating phased variants. MAC is a pipeline tool that can identify MNVs and correct annotations from Annovar and VEP outputs, but it requires input of the raw sequencing data (Wei et al. 2015). This may not be possible in some cases, such as public data sets that only release the variant call sets and not the raw sequencing data. Correction for MNVs was an important step of the work presented in Chapter 3 of this thesis, in order to correctly determine variant combinations within individual human genomes.

1.4.3 Computational Pathogenicity Prediction

After an annotated VCF file has been produced, the next step is to prioritise the variants identified. This step will depend greatly on the aim of the analysis, for example if trying to identify functional variants associated with heart disease one might filter for variants in known heart disease-associated genes. However, many analyses involve traits where the associated genes are unknown, and therefore the goal is often to identify variants that are most likely to have large functional impacts.

Despite advances in high-throughput mutational scanning methods (Gasperini et al. 2016; Fowler & Fields 2014), which can experimentally test the effects of thousands of mutations on certain protein features in parallel, the scale of modern variation data means that experimental characterisation of each variant identified is impractical. Instead, a large number of computational tools have been developed to plug this gap and try to predict the pathogenicity of variants (Figure 1.3).

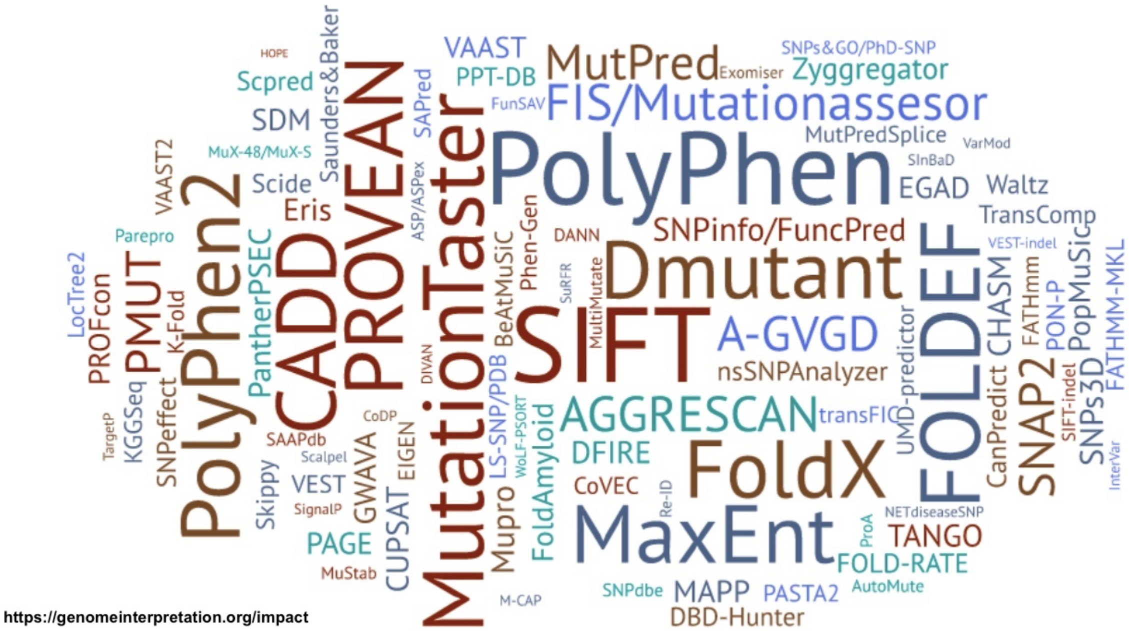


Figure 1.3: Computational tools for predicting pathogenicity of genetic variants. The size of the name is scaled to the logarithm of the number of citations. This figure is reproduced with permission from Hoskins et el. (Hoskins et al. 2017).

The aim of these computational prediction tools is to classify a given variant as either pathogenic or neutral, in order to broadly classify and prioritise the overwhelming number of genetic variants identified in large-scale sequencing projects. Many of these methods rely on similar principles, and largely differ in their implementation (typically using machine learning approaches).

One of the most commonly used tools for predicting variant pathogenicity has been SIFT, which takes a fairly simplistic approach of considering conservation of positions in multiple sequence alignments (Ng & Henikoff 2003; Kumar et al. 2009; Sim et al. 2012). The program takes the input sequence of the user and identifies related sequences from a database of protein sequences. SIFT then builds an alignment of the related sequences, and calculates the probability of each of the 20 amino acids to occur at a given position. SIFT then provides a probability that the observed variant will affect the function of the protein (SIFT score), with variants at highly conserved positions being more likely to affect protein function.

PolyPhen2 (Adzhubei et al. 2010) is another commonly used tool that also uses information about evolutionary conservation in its prediction. However, it also uses the basic physiochemical properties of the wild type and mutant amino acids, and considers many sequence features as annotated by UniProt. These sequence features include: residues forming bonds (e.g. disulphide bonds), active site residues, binding sites (e.g. metal binding sites), and residues that undergo lipid or sugar modifications. PolyPhen2 also uses information from known 3D structures of proteins, for example identification of mutations that would disrupt hydrophobic cores, and it also uses DSSP to analyse protein secondary structure, solvent accessibility, and phi-psi dihedral angles (Joosten et al. 2011).

PolyPhen2 implements machine learning (naive Bayes classifier) to integrate all of these features and classify variants, and has been trained on two different datasets: HumVar and HumDiv, with each of these training sets having its own advantages. HumVar is composed of 13,032 human disease-causing variants (annotated in UniProt) as well as 8,946 nsSNVs without disease annotation (which are regarded as non-damaging by PolyPhen2), and is better suited to classifying variants with drastic effects, such as those that cause Mendelian diseases. HumDiv is composed of 3,155 variants annotated as causing Mendelian diseases in UniProt as well as 6,321 sequence differences between human proteins and closely related homologs, and is better suited to classifying variants involved in complex phenotypes (Adzhubei et al. 2010).

Many other predictions also leverage similar features to SIFT and PolyPhen2, but combine these features with their own unique features. SuSPect utilises network centrality in protein-protein interaction networks as an important feature in its prediction method (Yates et al. 2014). FATHMM uses hidden Markov models (HMMs), as well as homologous proteins and protein domain conservation to weight how tolerant certain proteins/conserved domains are to variants (Hashem A. Shihab et al. 2013). MutationAssessor also uses homologues and domain conservation in its predictions, but implements a sequence entropy metric as a measure of position-specific conservation, and compares the entropy of the position with the wild type and variant amino acids (Reva et al. 2011). MutationAssessor also identifies sequence subfamilies within homologues to determine putative function specificity residues, and this information is used to predict variants that cause switch of function, as opposed to either loss or gain of function. MutationAssessor was used to analyse a set of cancer-relevant variants from COSMIC, and of the 3,631 variants in functional regions/binding sites it estimated that at least 5% result in switch of function (Reva et al. 2011).

Foldx (Schymkowitz et al. 2005) and mCSM (Pires et al. 2014) are two more methods for predicting the effects of genetic variants, but specialise in predicting the effects of the mutations on the 3-dimensional structures of proteins, and as such both methods require a protein structure in order to make their predictions. Foldx is a tool for calculating the free energy of a macromolecule, and can be used to compare the free energy of a wild type protein to the free energy of the same protein with one or more variant (Schymkowitz et al. 2005). This can be used to identify variants that will have destabilising or stabilising effects on proteins, both of which can be damaging.

mCSM is a newer method that uses graph-based signatures to encode distances between atoms and environments around individual residues, e.g. hydrophobic environments. The properties of the wild type and variant type amino acids are then considered within the context of the environment, e.g. a hydrophobic wild type amino acid in a hydrophobic environment mutated to a polar type variant amino acid. mCSM then uses machine learning to interpret these graph-based signatures and predict the effects of mutations on protein stability, protein-protein interactions, and protein-DNA interactions (Pires et al. 2014).

Molecular dynamics analyses are another type of prediction method that utilise protein structures. These methods focus on considering the inherent flexibility and motion of proteins, by attempting to predict the movement of individual atoms in structures over time. These methods are computationally expensive, especially when analysing structural fluctuations over biologically relevant timescales, but have shown potential in assessing the impacts of variants on protein flexibility and stability (Zimmermann et al. 2017; De Vivo et al. 2016).

Some methods focus on combining the predictions of other tools to make an aggregate prediction, such as CADD (Kircher et al. 2014), Condel (González-Pérez & López-Bigas 2011), and REVEL (Ioannidis et al. 2016). Each of these three tools integrates the predictions of other tools, including SIFT and PolyPhen, to compute a combined prediction. However, each tool uses slightly different input predictions as well as different machine learning algorithms for the final predictions, with CADD using a support vector machine, Condel using weighted average scores, and REVEL using a random forest classifier (González-Pérez & López-Bigas 2011; Kircher et al. 2014; Ioannidis et al. 2016).

In benchmarking, each of these three combined prediction tools report improved performance compared to using the constituent predictions on their own. However, each new prediction tool that is published reports an improved predictive ability compared to existing tools, and the performance of each tool is heavily influenced by the set of variants used for benchmarking. Independent comparisons of variant prediction tools have shown that different tools have low levels of agreement about the pathogenicity of the same variants (Chun & Fay 2009; Walters-Sen et al. 2015), and also low levels of agreement with known consequences of variants (Miosge et al. 2015). Additionally, the reported accuracies of the prediction methods in their initial publications have been shown to be likely over-estimated when independently compared (Grimm et al. 2015; Mahmood et al. 2017).

We saw a similarly low level of agreement between different pathogenicity predictors in the work presented in Chapter 2. We implemented six commonly used pathogenicity predictors (SIFT, PolyPhen2, MutationAssessor, FATHMM, Condel, and CADD) to predict the effects of cystinuria associated mutations. Across the set of mutations, these tools had between 32.9-34.5% agreement in their predictions, depending on the affected gene. We also observed incomplete agreement of predictions with experimental characterisation of a small set of cystinuria associated mutations, see Chapter 2 (Martell et al. 2017).

The difficulties in comparing computational pathogenicity predictors has led to the drive for more independent comparisons of tools, and to the formation of the Critical Assessment of Genome Interpretation (CAGI) community experiment

(Hoskins et al. 2017). The aim of CAGI is to objectively compare the abilities of tools for interpreting genetic variants, similar to the Critical Assessment of Protein Structure Prediction (CASP) for the assessment of protein structure prediction (Moult et al. 2018), to the Critical Assessment of Function Annotation (CAFA) for the assessment of protein function prediction (Jiang et al. 2016), and to the Critical Assessment of Prediction of Interactions (CAPRI) project for the assessment of prediction of protein-protein interactions (Lensink et al. 2017).

Objective comparison of the performance of these missense variant prediction tools is difficult. Traditional benchmarking of the tools relies on the use of a large set of variants for which the effects are known, i.e. pathogenic or benign. The source of this functional information depends on the variant set, with some benchmarking sets sourced from reported clinical information about variants and others sourced from experimental testing of the effects of variants. Tools are then assessed on their ability to correctly predict the functional effects of the variants. However, no single metric can adequately describe the utility of a prediction tool, and therefore predictions tools are assessed over a range of metrics.

Common metrics include: the true positive rate (the proportion of deleterious variants that are correctly predicted to be deleterious; TPR), the true negative rate (the proportion of benign variants that are correctly predicted to be benign; TNR), the false positive rate (the proportion of benign variants are falsely predicted to be deleterious; FPR), and the false negative rate (the proportion of deleterious variants that are incorrectly predicted to be benign; FNR). Another method frequently used to directly compare prediction methods is to plot a Receiver Operating Characteristic (ROC) curve, which reflects the TPR and the FPR at different pathogenicity thresholds for an individual method. For each method the area under the curve (AUC) is calculated to reduce the complexity of the ROC curve. The AUC is a single value and the higher the AUC the better the performance of the prediction tool.

Recent work has attempted to comprehensively and independently assess the performance of a set of 23 commonly used missense variant prediction methods (Li et al. 2018). Each of the 23 tools was used to predict the effects of missense variants for 3 different benchmark variant sets: i) human somatic variants from ClinVar ii) human somatic variants from the IARC TP53 database and the ICGC database (two cancer focused mutation databases) iii) a set of experimentally evaluated Peroxisome Proliferator-Activated Receptor Gamma (PPARG) variants. The predictions of the different tools were then compared for 12 different performance measures, including TPR, TNR, FPR, FNR, and AUC. For the ClinVar variant set, a large amount of variation in performance was seen across the methods: TPR (50.52%-96.07%), TNR (34.8%-89.95%), FPR (10.05%-65.20%), FNR (3.93%-45.99%), AUC (0.610-0.929). Each of the 23 tools performs better for certain metrics, e.g. some tools have relatively high TPR but relatively low TNR. Crucially, the 23 tools also vary in their performance between the three benchmarking datasets, demonstrating the need for researchers and clinicians to choose not the best overall prediction tool but the best prediction tool for a given problem (Li et al. 2018).

The effects of mutations on splicing are another important aspect of pathogenicity prediction, as variants that alter splicing frequently have much larger effects on protein function compared to SNVs that change one amino acid. Both Annovar and VEP will annotate variants as likely to affect splicing, but prediction of splice variants shows the highest degree of annotation discrepancy between the variant annotation methods (McCarthy et al. 2014).

As previously discussed, prediction of the effects of non-coding variants remains more difficult than predicting the effects of protein-coding variants, with existing methods focused on combining functional annotations of non-coding regions, evolutionary conservation, and machine learning to prioritise non-coding variants (Ritchie et al. 2014; Fu et al. 2014; Zhou & Troyanskaya 2015; Shihab et al. 2015; Ionita-Laza et al. 2016; Smedley et al. 2016). These methods are likely to improve as annotations of non-coding regions of genomes improve. Progress has been made on the annotation of functional non-coding regions in the human genome (Dunham et al. 2012), but annotation of functional non-coding regions in other non-human reference genomes is still poor (Andersson et al. 2015).

With non-coding variants known to be the functional variants in a number of human diseases it is essential that methods to predict their effects are improved (Zhang & Lupski 2015). In the future it will also be essential to integrate whole-genome sequencing data with RNA sequencing and proteomics data, in order to directly study the effects of genetic variation at the RNA level and on protein expression. These data can then be compared to reference expression data sets, such as the Expression Atlas, which has integrated and reanalysed expression data from thousands of individual studies (Petryszak et al. 2016). This will enable the identification of variants that deregulate RNA expression leading to aberrant functions and disease.

Finally, one of the biggest challenges in predicting the effects of variants is predicting the combined effects of variants. Almost all current prediction methods consider the effects of variants in isolation, essentially predicting the effect of a single difference compared to the reference genome. However, each individual human genome contains 4.1-5 million variants compared to the reference genome, and therefore to achieve more accurate predictions the combined effects of variants must be considered.

EVmutation (Hopf et al. 2017) is a new tool that aims to predict the effects of variants in the context of other variants. This method considers coevolution and covariation between residue positions, taking in to account dependencies between residues (see Section 1.6.4 for a summary of protein residue coevolution). For example, one variant may be damaging on its own, but when it occurs with one or more other variants it is no longer damaging. Using this context-dependent approach, EVmutation demonstrated improved performance compared to methods that consider variants in isolation (Hopf et al. 2017). Approaches that consider the effects of multiple variants together will be essential in the future, especially in the field of precision medicine, and is one of the topics covered in Chapter 3 of this thesis.

1.4.4 Structural Analysis of Protein Variants

The functional effect of a protein sequence variant is often determined by its effect on the final 3-dimensional structure of the protein. Manual analysis of variants within protein structures, or automated predictors like Foldx, mCSM and molecular dynamics approaches (Schymkowitz et al. 2005; Pires et al. 2014; De Vivo et al. 2016; Zimmermann et al. 2017), can be used to identify such functional impacts of variants on proteins. However, each of these approaches require a high-quality protein structure.

The requirement of a protein structure in order to perform functional analyses of genetic variants is a limiting factor, due to the sequence-structure gap – the fact that there are a large number of sequenced proteins but only a comparatively small number of experimentally determined structures. In the current release of UniProt (07/2018) there are 120,243,849 distinct proteins (Bateman et al. 2015). However, in the current release of the PDB (09/2018) there are only 133,749 protein structures (Berman et al. 2000). Filtering these structures for only representative structures (<95% sequence identity), results in a set of only 52,868 structures, demonstrating the redundancy of proteins even within the relatively small set of experimentally resolved protein structures.

This relative paucity of protein structures is caused by the difficulty of experimental protein structure determination compared to the experimental determination of a protein's sequence. The three main techniques for experimental structure determination are X-ray crystallography (120,564 structures in the PDB), Nuclear Magnetic Resonance (NMR) spectroscopy (10,826 structures in the PDB), and electron microscopy (1,717 structures in the PDB). Each of these techniques has its own advantages and disadvantages.

X-ray crystallography can produce highly accurate atomic resolution structures, but requires the protein of interest to be crystallised. The crystallisation of a protein is non-trivial and for many proteins not currently possible. Crystallisation also results in a static representation of the protein that does not represent the fluidity in native protein structures. Nevertheless, X-ray crystallography has been the leading technology of protein structure determination, as demonstrated by the large number of X-ray solved structures in the PDB (Shi 2014).

NMR spectroscopy is also capable of producing high-resolution structures of proteins, and has the added advantage of the ability to interrogate the dynamics of protein folding and structural fluctuations over time. The main weakness in NMR spectroscopy is the difficulty in determining the structures of large proteins. For proteins >25 kDa, more advanced NMR methodologies are required, and the resolution of the structures produced can be limited (Sugiki et al. 2017).

Single-particle cryo electron microscopy (cryoEM) is a less mature technology than both X-ray crystallography and NMR spectroscopy, but has shown great potential especially for the determination of the structures of large macromolecular complexes (Zhou 2011; Shi 2014). In the future it will be beneficial to combine the benefits of X-ray, NMR, and cryoEM techniques, both in choosing the most appropriate technique on an individual protein basis and also in multi-technique approaches for solving complex structures. In the current release of the PDB (09/2018) there are already 115 structures produced using a combination of experimental techniques.

In order to overcome this lack of experimental structures many computational methods have been developed to predict the structures of proteins. The current gold standard approaches for predicting protein structures are homology-based predictors. These methods leverage the fact that homologues proteins have high structural similarity, with protein structure being more conserved than protein sequence (Chothia & Lesk 1986; Koonin et al. 2002). Additionally, it is believed that there are only around 1,000-10,000 unique protein folds present in nature (Koonin et al. 2002; Kelly et al. 2015). Therefore, for a protein without an experimental structure, a homologous protein can be identified that does have an experimental structure and this experimental structure can be used as a template to predict the structure of the protein without experimental data.

There are now a large number of homology-based structure prediction tools available, many of which are assessed as part of CASP, which is a community effort to objectively assess the accuracy of different protein structure prediction tools (Moult et al. 2018). Two world-leading homology-based structure prediction tools are Phyre2 (Kelly et al. 2015) and I-TASSER (Yang et al. 2014), both of which receive thousands of submissions every year. If a homology template can be identified, these tools can produce high-quality structural models of either complete proteins or individual domains.

For some proteins homology modelling is not possible due to a lack of available high-quality template structures. Such cases necessitate the development of non-homology-based methods, often termed template-free or *de novo* structure prediction methods. These *de novo* prediction methods are less accurate compared to template-based modelling, and despite rapid improvement in recent years much work is still required to improve accuracy and performance, especially for larger proteins (Moult et al. 2016; Moult et al. 2018).

Predicting the structure of a protein without a template structure to guide the prediction is far more difficult, due to the large number of possible conformations of a given protein sequence. To compare all possible conformations would be computationally infeasible, therefore current *de novo* structure prediction methods focus on reducing this search space of conformations. One approach to reduce the search space of conformations is to use large libraries of fragments of known protein structures and model these conformations on the unknown protein to identify local conformations with the lowest overall energy. This general approach has been implemented in a number of *de novo* structure prediction tools (Bhattacharya et al. 2016; de Oliveira et al. 2015; De Oliveira et al. 2018; Gront et al. 2011; Shen et al. 2013; Trevizani et al. 2017; Wang et al. 2017).

An improvement on this general fragment-based approach has recently been demonstrated by integrating information about coevolving residues in protein sequences (De Oliveira & Deane 2018). This method is based on the observation that coevolving residues in proteins frequently occur close in space within the 3-dimensional structure of the protein (Buslje et al. 2009; Marks et al. 2011; Morcos et al. 2011; Jeong & Kim 2012; Kamisetty et al. 2013; Ekeberg et al. 2013; Schneider & Brock 2014; Seemayer et al. 2014; Kaján et al. 2014; Jones et al. 2015; Adhikari & Cheng 2016; Ovchinnikov et al. 2017).

In Chapter 3 of this thesis, we used a combination of experimental and homology modelled structures to investigate coevolutionary relationships between variants within individual human genomes. This combination of experimental and predicted structures enabled the analysis of a much larger set of proteins and variants than if experimental structures alone had been used.

1.5 Cystinuria

Cystinuria is a rare human disease, caused by mutations in two genes: SLC3A1 and SLC7A9, which encode the proteins rBAT and b(0+)AT, respectively (Thomas et al. 2014). These two proteins form a membrane transporter in the kidney and the intestinal tract that transports dibasic amino acids from the urine in to the blood, with the exchange of a neutral amino acid in the opposite direction (antiporter activity). This transporter is crucial in controlling the levels of these molecules in the urine (Figure 1.4).

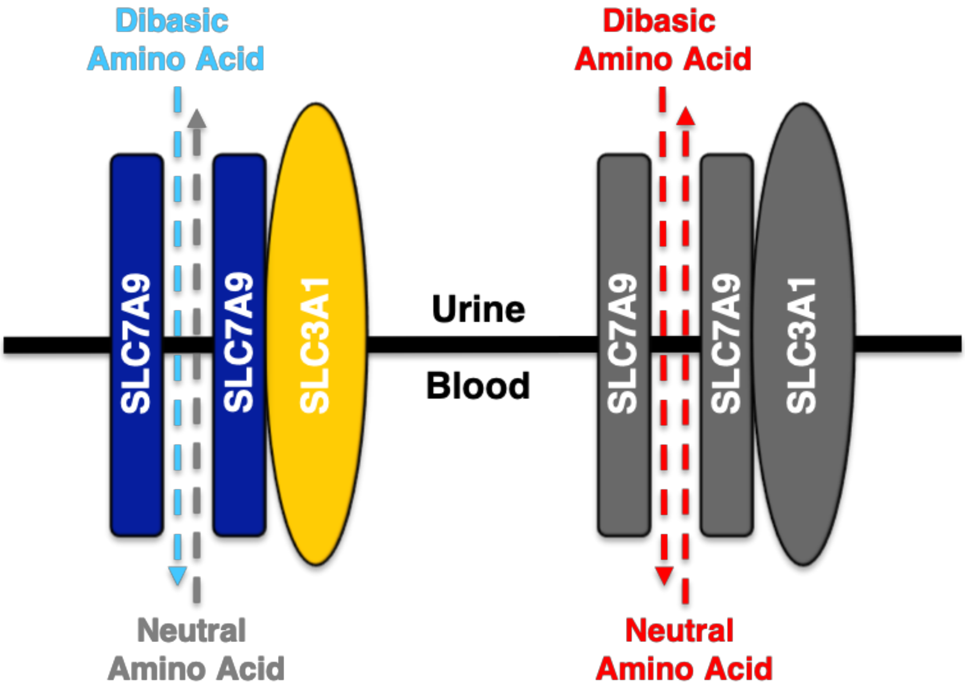


Figure 1.4: Transportation of dibasic amino acids in the kidney, in exchange for neutral amino acids, by the gene products of SLC3A1 and SLC7A9 (left) and disrupted transportation by a defective version of the transporter (right).

Defective transport of these dibasic amino acids from the urine in to the blood, caused by mutations in SLC3A1 and SLC7A9, results in their build up in the urine (Figure 1.4). One of the dibasic amino acids transported by this protein complex is cystine, which is two cysteine molecules linked together by a disulphide bond. Cystine is relatively insoluble, and high concentrations in the urine lead to it crystallising, resulting in kidney stones (Wong et al. 2015).

The incidence of kidney stones in the United States is ~9%, and is on the rise (Zee et al. 2017). Cystinuria is a rare subtype of kidney stones, with a worldwide prevalence of 1 in 7,000 (Barbosa et al. 2012). However, the incidence of cystinuria is highly variable between different human populations, with incidence as high as 1 in 2,500 in Libyan Jews (Chillarón et al. 2010), and as low as 1 in 100,000 in Sweden (Harnevik et al. 2003).

The clinical presentation of cystinuria is highly variable. Patients with a mild form of the disease may have a single stone episode in their lifetime. Whereas, patients with a severe form of the disease may suffer from chronic kidney disease, frequent stone complications and hospitalisation, and sometimes nephrectomy (Thomas et al. 2014; Wong et al. 2015). Successful treatment for patients with a mild form of the disease can sometimes be achieved through increase in fluid intake and changes in diet (Zee et al. 2017). However, treatment options for patients with a severe form of cystinuria are much more limited, with many of the prescribed drugs having a limited impact on stone episodes, and can cause adverse side effects (Zee et al. 2017). A recent study in an SLC3A1-/- mouse model provided hope that the dietary supplement alpha-lipoic acid could be used as a treatment option for cystinuria, with it seemingly increasing the solubility of cystine in the urine (Zee et al. 2017). Alpha-lipoic acid is currently undergoing phase 2 clinical trials as a treatment for cystinuria (ClinicalTrials.gov identifier: NCT02910531, https://clinicaltrials.gov/ct2/show/record/NCT02910531).

The work presented in Chapter 2 of this thesis has two main aspects. First, a comprehensive analysis of every variant previously associated with cystinuria was performed. Each variant was characterised by its location in the protein structure, its effect on ligand binding sites, its frequency in the ExAC population, its conservation in homologous proteins, its effect on structural stability, and finally its predicted pathogenicity from a range of pathogenicity predictors. Second, we analysed a cohort of 74 cystinuria patients, and predicted each sample's disease severity based on the variants they carry in the genes SLC3A1 and SLC7A9. We show that this approach has the potential to identify high-risk cystinuria patients, and help to direct treatments options in order to provide precision medicine.

1.6 Coevolution

Coevolution occurs when the evolution of two or more entities happens in a reciprocal manner, with each entity affecting the evolution of the other. Coevolution is a key principle in evolutionary theory that governs many biological relationships and occurs at many different scales: the species level, the protein level, and the protein residue level.

1.6.1 Coevolutionary Relationships

Coevolving species, proteins, and protein residues are said to have a coevolutionary relationship. This relationship can be categorised in to broad groups, based on whether the effects of the relationship are beneficial, negative, or neutral for each partner in the relationship.

The six main types of coevolutionary relationships at the species level are shown in Figure 1.5. These relationships are as follows:

1. Mutualism – both partner species gain a benefit from the relationship, e.g. the relationship between a flower and its pollinators. This relationship can be facultative, where each species gains a benefit from the relationship but can survive on their own, or obligative where each species requires the relationship in order to survive.
2. Parasitism – one species gains a benefit from the relationship, but the relationship is damaging for the other species, e.g. the relationship between a virus and its host

3. Commensalism – one species gains a benefit from the relationship, with little or no effect on the other species, e.g. the relationship between a human and its gut microflora
4. Neutralism – the impact on each species is neutral, e.g. two species that do not interact in nature or interact only indirectly such as rabbits and deer
5. Amensalism – one species is damaged by the relationship, with little or no effect on the other species, e.g. the secretion of penicillin by *Penicillium* which is damaging to many bacteria
6. Competition – one species gains a benefit from the relationship, with a negative effect on the other species, e.g. the predator-prey relationship between a lion and a zebra

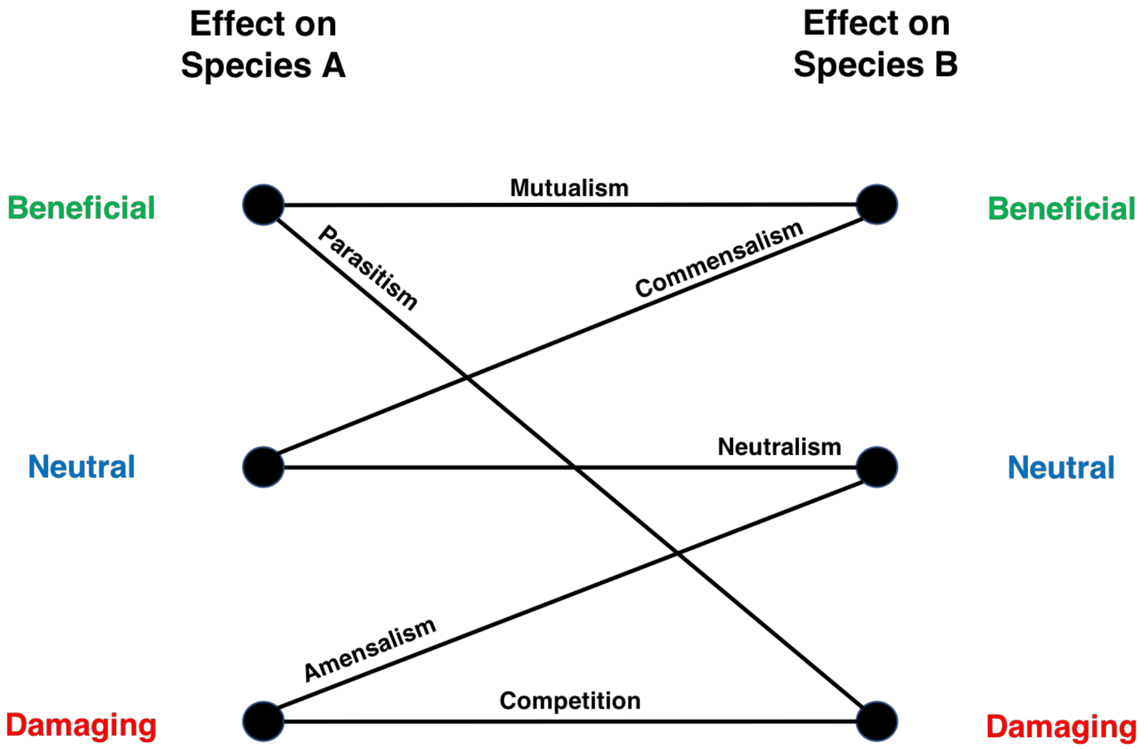


Figure 1.5: Summary of the different types of coevolutionary relationships at the species level. In this figure each black line represents a type of relationship and the line links the effect on each partner of the relationship, e.g. in Competition the relationship is damaging for both species. This figure was created for this thesis but is based on Ian Alexander's work (available here: https://en.wikipedia.org/wiki/Symbiosis#/media/File:Symbiotic_relationships_diagram.svg).

1.6.2 Coevolutionary Relationships Observed in Nature

An example of a mutualistic relationship at the species level is the relationship between hummingbirds and the flowers they pollinate (Kay et al. 2005). The sword-billed hummingbird (*Ensifera ensifera*) is a particularly extreme example, with the bird having evolved a beak longer than its body in order to pollinate the *Passiflora mixta* flower, which has a very deep lumen (Figure 1.6). In this relationship, the bird gains food from the plant, and the plant gains the ability to pollinate from the bird. Therefore, both partners benefit and this coevolutionary relationship is mutualistic.

Parasitic and competitive coevolutionary relationships frequently lead to evolutionary arms races, in which each partner is under selective pressure to outcompete the other. For example, in humans and other mammals the family of APOBEC genes encode a series of cytidine deaminase enzymes, which function as a defence mechanism against viruses by inducing hyper-mutation. To combat this, some viruses have evolved their own mechanisms to inhibit APOBEC enzymes and evade this defence mechanism, such as Viral infectivity factor (Vif) encoded by Human Immunodeficiency Virus (HIV). Vif is able to inhibit APOBEC enzymes and allow HIV to infect human cells without being hyper-mutated (Rose et al. 2004). In this example there is a parasitic coevolutionary relationship between HIV and humans, and underpinning this is a competitive evolutionary relationship at the protein level between human APOBEC enzymes and HIV Vif.

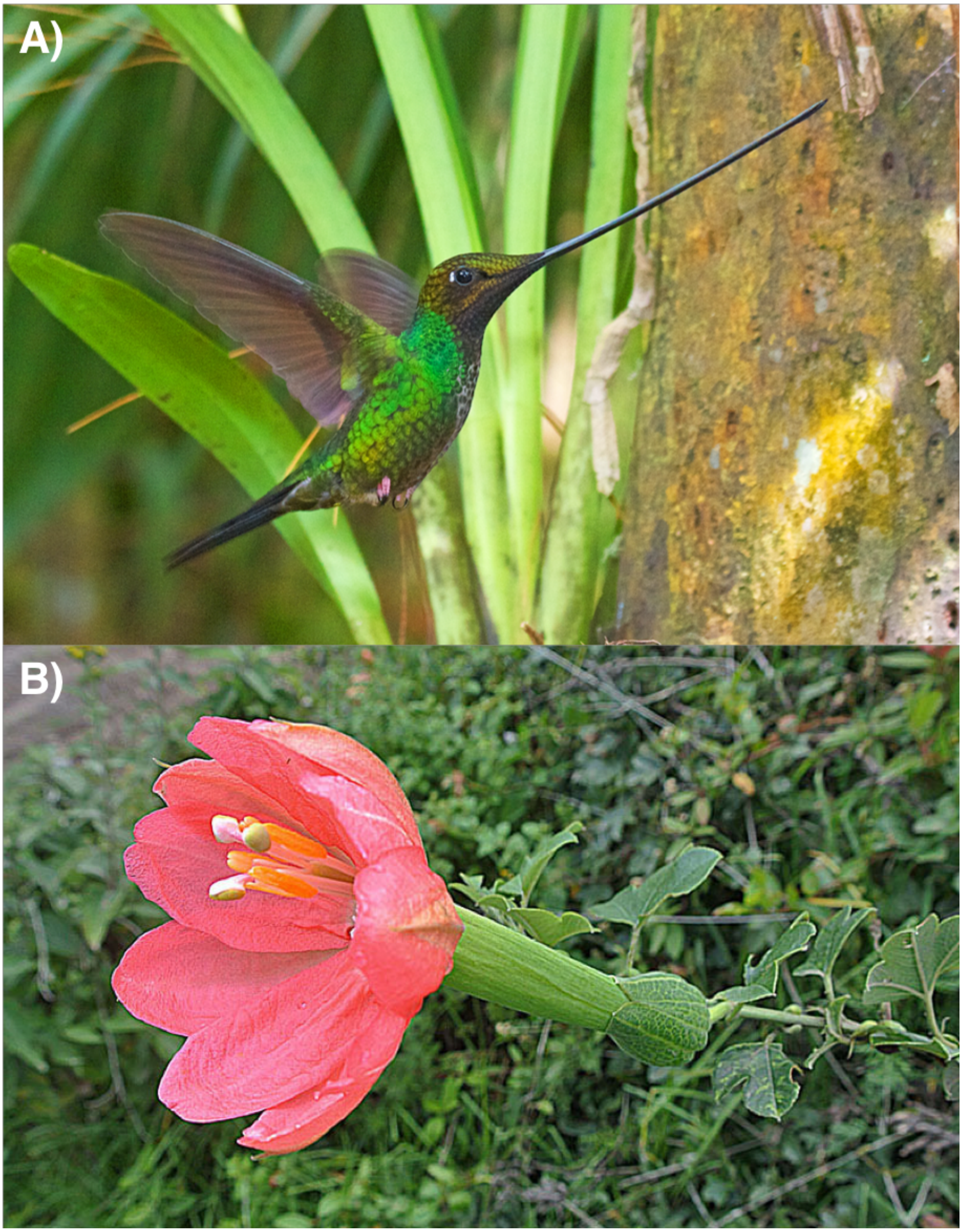


Figure 1.6: Mutualistic coevolution at the species level, between the sword-billed hummingbird (*Ensifera ensifera*) and the *Passiflora mixta* flower. **A)** The sword-billed hummingbird. Photo is used with permission from Joseph C Boone (www.calpoly.edu/~jboone/). **B)** The *Passiflora mixta* flower. Photo is used with permission from Dick Culbert (http://www.dixpix.ca/index.html), and has been horizontally flipped to align with the photo of the hummingbird.

1.6.3 Protein Coevolution

Coevolutionary relationships at the protein level represent functional relationships within the environment of the cell, where the evolution of one protein affects the evolution of another protein. Groups of proteins within a single organism may coevolve (intra-species), with proteins involved in similar biological processes or those that interact physically having shared selection pressures. Inter-species protein coevolution is the reciprocal evolution of proteins between multiple organisms, commonly seen in host-parasite interactions, such as the coevolutionary relationship of human APOBEC enzymes and HIV Vif described in Section 1.6.2.

The study of coevolution at the protein level is more difficult than at the species level and requires a large amount of data in order to avoid false associations. The successful identification of coevolving proteins has practical applications, for example in predicting proteins that interact within cells (Juan et al. 2008). A common feature of interacting proteins is the similarities between their phylogenetic trees. The underlying reason for the observed tree similarity is believed to be explained by either a coevolutionary relationship that results in compensatory changes between the two proteins, or that it is an indirect consequence of two proteins that are involved in the same process (Juan et al. 2008).

There are two main approaches for identifying coevolving proteins that have been iterated on since their inception. First, comparison of the absence or presence of proteins or protein families can help identify coevolving proteins, with proteins that interact functionally or physically having been shown to have similar species distributions (Pellegrini et al. 1999; Marcotte et al. 1999). This approach is implemented in the STRING database as a metric for predicting protein-protein interactions (Szklarczyk et al. 2015), and can identify extreme cases of protein coevolution, where proteins are coevolving not just certain features but their very existence (Juan et al. 2008). Second, comparisons of the shapes of protein phylogenetic trees has shown that interacting and functionally-related proteins tend to have similar phylogenetic tree shapes (Pazos & Valencia 2001; Pagès et al. 1997; Goh et al. 2000; Fryxell 1996). Methods such as MirrorTree can be used to exploit this pattern and compare the shapes of protein phylogenetic trees to detect more subtle coevolutionary relationships between proteins (Ochoa & Pazos 2010).

1.6.4 Protein Residue Coevolution

Protein residue coevolution involves epistatic interactions between residues in proteins, where the amino acid at one position influences the amino acid at another position. Coevolution at the residue level in proteins can occur within a single protein (intra-protein residue coevolution), for example between two residues in a protein chain that are close in 3-dimensional space (Figure 1.7.A), or between two different proteins of the same species or different species (inter-protein residue coevolution), for example between residues in interface regions of two physically interacting proteins (Figure 1.7.B).

An example of intra-protein residue coevolution observed in nature is between residues 20 and 69 in β-haemoglobin. In human β-haemoglobin the wild type amino acid at position 20 is valine (Figure 1.7.C), and mutation to glutamic acid results in the disease phenotype erythrocytosis (dbSNP identifier: rs33918474) (Sherry 2001). Conversely, in horse β-haemoglobin the wild type amino acid at position 20 is glutamic acid (Figure 1.7.C). However, in horse β-haemoglobin there is a histidine residue at position 69, this differs from human β-haemoglobin, which has a glycine at position 69. Glutamic acid is observed at position 20 in many other species and results in healthy phenotypes, but it is always combined with histidine at position 69.

This is an example of a Dobzhansky-Muller incompatibility – one amino acid is not present without the other in a species due to the reduction in fitness caused by having only one (Kondrashov et al. 2002). In human β-haemoglobin, Val20 can form a van der Waals interaction with Gly69, and in horse β-haemoglobin Glu20 can form a hydrogen bond with His69, but each interaction is lost if only one amino acid in the pair is changed. These types of context-dependent effects of variants between species are termed Compensated Pathogenic Deviations (CPDs), where the wild type amino acid in one species can cause disease in the other species (Kimura 1985). CPDs are well studied between species, but less so within species.

The work in Chapter 3 of this thesis explores potential compensatory variants within individual human genomes.

The general approach for identifying coevolving residues is to search for covariation at positions in multiple sequence alignments and detect positions with correlated variants. Direct Coupling Analysis (DCA) describes a set of statistical methods used to determine the strength of the relationships between residues and identify those with a likely direct functional relationship. These methods have also been shown to be useful in protein structure prediction, as coevolving residues have a propensity to be close in 3-dimensional space, see Section 1.4.4 (Buslje et al. 2009; Marks et al. 2011; Morcos et al. 2011; Jeong & Kim 2012; Kamisetty et al. 2013; Ekeberg et al. 2013; Schneider & Brock 2014; Seemayer et al. 2014; Kaján et al. 2014; Jones et al. 2015; Adhikari & Cheng 2016; Ovchinnikov et al. 2017).

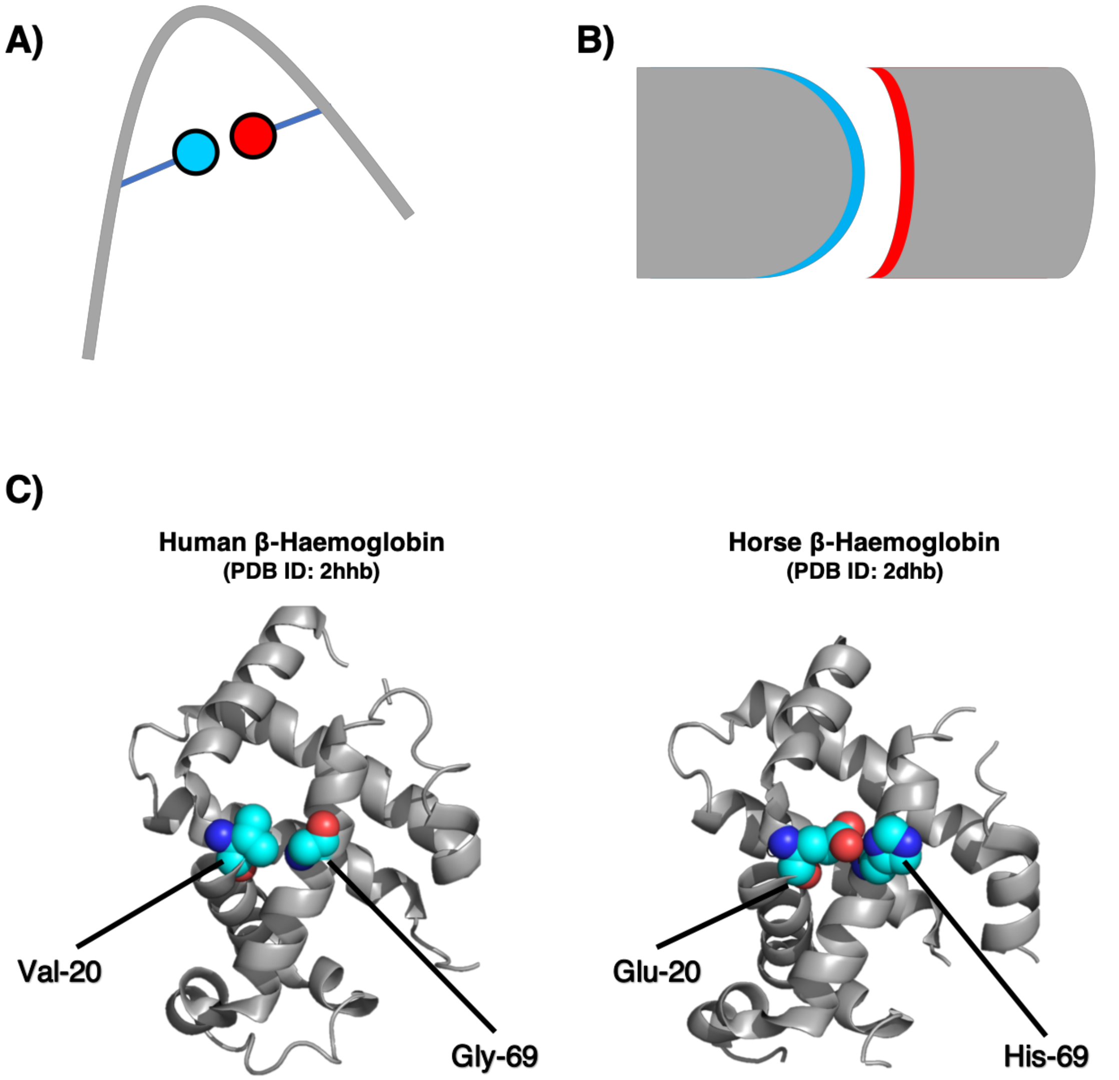


Figure 1.7: Examples of protein residue coevolution. **A)** An example of two coevolving residues within a protein chain that are in close proximity in the 3-dimensional space of the protein structure. **B)** An example of coevolving residues between two physically interacting proteins. **C)** An example of coevolving residues observed in nature within β-haemoglobin. The structures of human and horse haemoglobin are shown, each of which has a different pair of amino acids at positions 20 and 69 that have coevolved. In sub-figures **A)** and **B)** the proteins are coloured in grey and the coevolving residues/regions are shown in cyan and red. In sub-figure **C)** the chains are coloured in grey and shown in cartoon format, and residues 20 and 69 are shown in sphere format and coloured by atom.

1.7 Ebolaviruses

1.7.1 Ebolavirus Species

There are five species of ebolaviruses in the *Ebolavirus* genus. Kuhn et al. classify the five species of *Ebolavirus* as follows: *Bundibugyo ebolavirus* (type virus: *Bundibugyo virus*, BDBV), *Reston ebolavirus* (type virus: *Reston virus*, RESTV), *Sudan ebolavirus* (type virus: *Sudan virus*, SUDV), *Tai Forest ebolavirus* (type virus: *Tai Forest virus*, TAFV) and *Zaire ebolavirus* (type virus: *Ebola virus*, EBOV) (Kuhn et al. 2014).

While the reservoir species of ebolaviruses remains unclear, ebolaviruses do not normally circulate within humans (Leendertz 2016). However, since 1976, the first time ebolaviruses were observed to infect humans, there have been regular spill-over events of this zoonotic virus in to human populations (Table 1.1). *Ebola virus* has been the most common causative species, but each of the viruses has been observed to infect and cause disease in humans, except for *Reston virus* (CDC 2018; World Health Organization 2009).

Table 1.1: Ebolavirus outbreaks by species and by year. Outbreak data was taken from the CDC National Centre for Health Statistics (CDC 2018).

Species	Outbreak Years
Bundibugyo virus	2007, 2012
Ebola virus	1976, 1977, 1994, 1995, 1996, 2001, 2002, 2003, 2004, 2007, 2008, 2013-2016, 2017, 2018
Reston virus	-
Sudan virus	1976, 1979, 2000, 2011, 2012
Tai Forest virus	1994

1.7.2 Ebolavirus Pathogenicity

Reston virus is the only species in the *Ebolavirus* genus that is not known to be pathogenic to humans, despite being able to infect humans (Miranda et al. 1999; World Health Organization 2009). However, *Reston virus* is able to infect and cause disease in non-human primates (Miranda & Miranda 2011). Each of the other four *Ebolavirus* species are known to cause Ebolavirus Disease (EVD)

(World Health Organization 2009). EVD is a haemorrhagic fever, the initial symptoms of which include fever, chills, malaise, and myalgia. These symptoms are followed by multiple gastrointestinal, respiratory, vascular, and neurological symptoms, and at the peak of infection multiple haemorrhagic symptoms arise (Feldman & Geisbert 2011). The fatality rate of EVD has varied between Ebolavirus outbreaks, but has been observed as high as 90% (CDC 2018).

1.7.3 The 2013-2016 West Africa Ebolavirus Outbreak

The 2013-2016 West Africa *Ebola virus* outbreak was the largest to date, with 28,610 confirmed cases, and 11,308 deaths (CDC 2018). This outbreak primarily affected Sierra Leone, Guinea, and Liberia, though there were transmission chains that led to isolated cases in Italy, Mali, Nigeria, Senegal, Spain, and the United States of America (CDC 2018). This outbreak required a coordinated international effort to control, and highlighted the severe threat posed by ebolaviruses to public health. This threat has been reemphasised by the concurrent outbreak in the Democratic Republic of the Congo in 2014, and the two further outbreaks in the Democratic Republic of the Congo in 2018 (CDC 2018; Rimmer 2018).

1.7.4 The Ebolavirus Genome and Life Cycle

Ebolaviruses are negative, single-stranded RNA viruses, with a genome size ~19 kilobases (kb) (Messaoudi et al. 2015). The ebolavirus life cycle (Figure 1.8) follows six major steps (Messaoudi et al. 2015), these are:

1. The virus attaches to the host cell via the ebolavirus surface protein glycoprotein
2. The virus enters the cell via macropinocytosis
3. After acidification of the endosome, and cleavage of glycoprotein by cellular proteases cathepsins B and L, glycoprotein interacts with the host protein Niemann-Pick C1 (NPC1) triggering fusion of the endosomal and viral membranes. Although NPC1 is required in this step, ebolavirus glycoprotein may interact with other host receptors.
4. The viral ribonucleocapsid is then released in to the host cell cytoplasm, where it can undergo transcription and replication. Individual ebolavirus

genes are transcribed in to mRNA, which are then translated in to the seven ebolavirus proteins.

5. The RNA genome is used as a template to create a full-length complement, which is then itself used as a template to synthesise new negative-sense ebolavirus genomes.
6. New viral particles are assembled at the plasma membrane, incorporating the viral genome and viral proteins. This budding process is directed by the ebolavirus protein VP40.

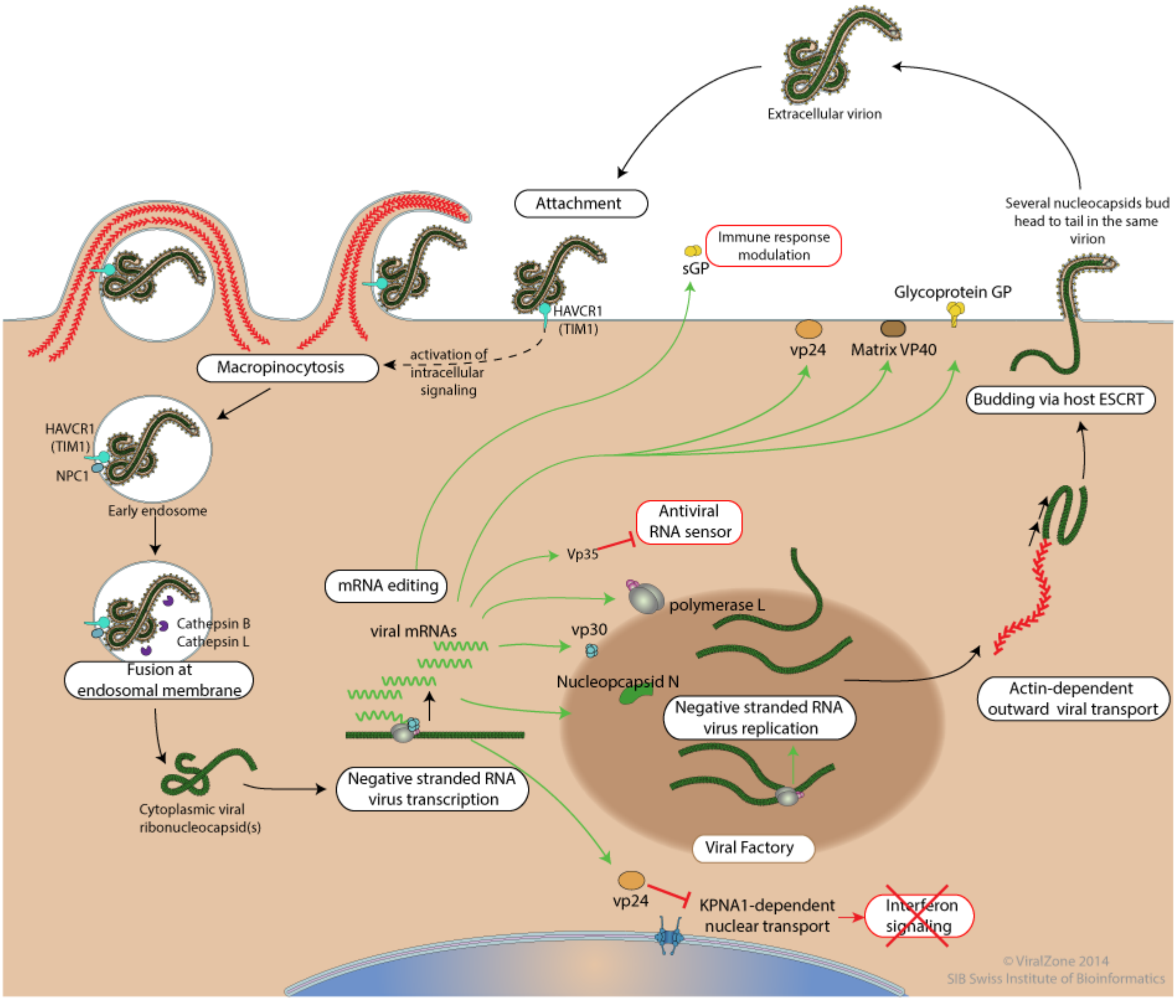

Extracellular virion
Attachment
Several nucleocapsids bud head to tail in the same virion
Immune response modulation
sGP
Glycoprotein GP
HAVCR1 (TIM1)
vp24
Matrix VP40
Macropinocytosis
activation of intracellular signaling
Budding via host ESCRT
HAVCR1 (TIM1)
NPC1
Early endosome
Antiviral RNA sensor
Vp35
mRNA editing
polymerase L
Cathepsin B
Cathepsin L
viral mRNAs
vp30
Fusion at endosomal membrane
Nucleopcapsid N
Actin-dependent outward viral transport
Negative stranded RNA virus replication
Negative stranded RNA virus transcription
Cytoplasmic viral ribonucleocapsid(s)
Viral Factory
vp24
KPNA1-dependent nuclear transport
Interferon signaling
© ViralZone 2014
SIB Swiss Institute of Bioinformatics

Figure 1.8: The replication cycle for Ebolaviruses within host cells. This figure is reproduced with permission from ViralZone - www.expasy.org/viralzone (Hulo et al. 2011).

1.7.5 Ebolavirus Proteins

The ebolavirus genome encodes seven different proteins: glycoprotein (GP), matrix protein (VP40), nucleoprotein (NP), VP35, VP40, VP24, and polymerase protein (L; Figure 1.9). As with many viruses, especially those that only encode a small number of proteins, most ebolavirus proteins perform multiple functions.

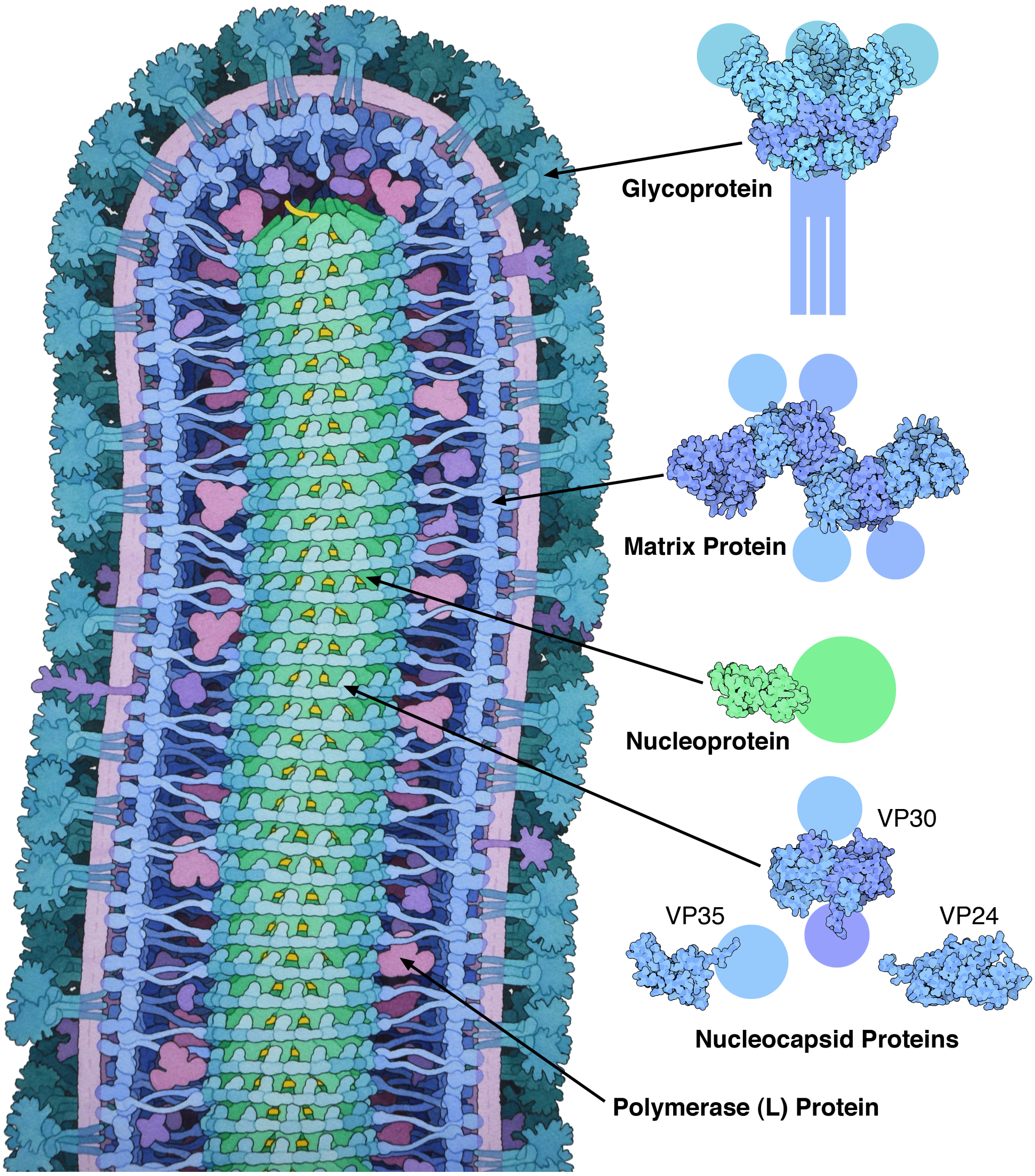


Figure 1.9: The seven proteins of the ebolavirus genome: Glycoprotein (GP), Matrix protein (VP40), Nucleoprotein (NP), VP35, VP40, VP24, and Polymerase protein (L). Spheres represent regions of the proteins not solved in the corresponding structures. This figure is reproduced with permission from the Protein Data Bank (Berman et al. 2000).

GP is the only surface protein encoded by ebolaviruses, and is primarily responsible for host-cell entry (Dutta et al. 2017). Full-length GP is not encoded directly by the Ebolavirus genome, and is the result of mRNA editing (Volchkov et al. 1995). The standard reading frame of GP results in the production of sGP. Two separate mRNA editing events result in the production of full-length GP and ssGP, but the roles of sGP and ssGP are still unclear (Mehedi et al. 2011; Miranda & Miranda 2011; Hoenen et al. 2015).

The L protein is an RNA polymerase, which is responsible for both replication of the ebolavirus genome and also transcription (Volchkov et al. 1999). L also forms a part of the nucleocapsid complex (with NP, VP35, and VP30), which is key for transcription, replication, and virus assembly (Dutta et al. 2017; Wan et al. 2017). NP also plays a role in virus replication, and encapsidates the RNA genome (with VP35 and VP24) (Huang et al. 2002).

The four Ebolavirus VP proteins (VP24, VP30, VP35, and VP40) each have multiple functions. VP35 and VP30 both form part of the nucleocapsid complex, and VP35 and VP24 both help encapsidate the RNA genome (Huang et al. 2002; Dutta et al. 2017; Wan et al. 2017). VP30 also plays a key role in regulating transcription of ebolavirus genes, acting as a transcription activator (Weik et al. 2002), that can itself be regulated by phosphorylation (Modrof et al. 2002). However, the mechanism by which VP30 activates transcription remains unclear (Dutta et al. 2017). VP40 is known to have a dimeric, a hexameric, and an octameric conformation (Bornholdt et al. 2013). The dimeric form is responsible for trafficking VP40 to the cellular membrane. VP40 in its hexameric conformation forms filamentous matrix structures and is essential for ebolavirus budding, see Figure 1.8 (Messaoudi et al. 2015). The octameric form of VP40 plays a separate role in regulating viral transcription (Bornholdt et al. 2013), which is critical during the life cycle of ebolaviruses (Hoenen et al. 2005).

VP35 and VP24 also play key roles in downregulating the host interferon (IFN) response. In humans, the IFN response is a key modulator of the adaptive immune system, and sets the body in to a defensive mode upon detection of an invading virus. For a virus to successfully infect a host it is important for it to downregulate the IFN response (Haller et al. 2006). Detection of double-stranded RNA (dsRNA) by host cells leads to upregulation of IFN signalling (Haller et al. 2006). One of the functions of VP35 is to bind dsRNA, thereby preventing its detection by the host, and preventing upregulation of IFN signalling (Basler et al. 2000; Basler et al. 2003; Prins et al. 2009).

VP24 also plays a role in downregulating IFN signalling, but acts via a different mechanism to VP35. Physiologically in human cells, accumulation of phosphorylated STAT1 in the nucleus upregulates IFN signalling. However, STAT1 must interact with karyopherin proteins in order to trigger nuclear import, as they do not contain a classical nuclear localisation signal, see Figure 1.10.A (Xu et al. 2014). Ebolavirus VP24 is able to bind directly to multiple human karyopherin alpha proteins (KPNA; KPNA1, KPNA5, and KPNA6) and also directly with STAT1, and in doing so prevent the accumulation of phosphorylated STAT1 in the nucleus (Figure 1.10.B). This prevents upregulation of IFN signalling and allows the virus to replicate without attenuation by the IFN response (Reid et al. 2006; Reid et al. 2007; Zhang et al. 2012; Xu et al. 2014).

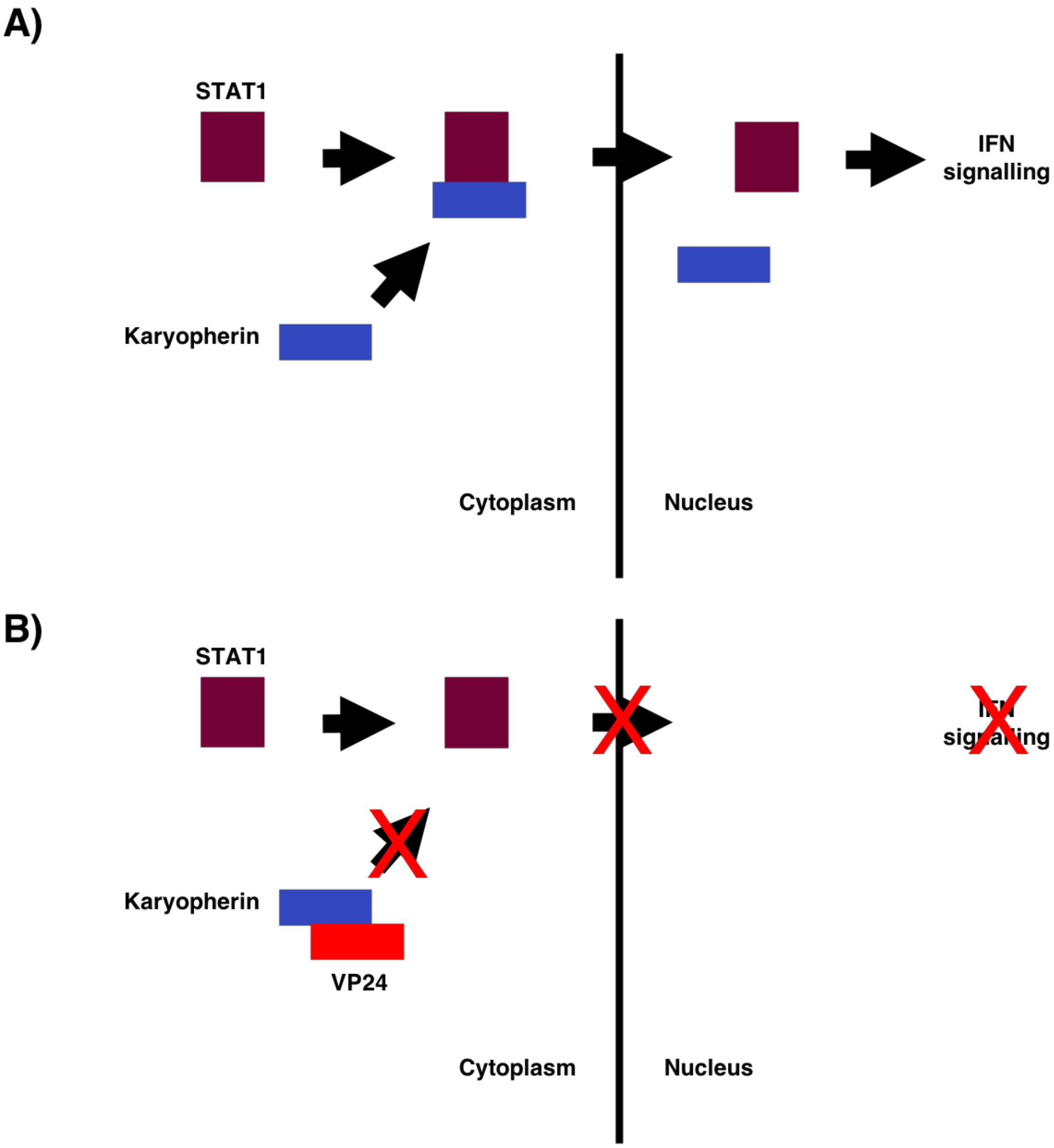


Figure 1.10: VP24 and IFN signalling. **A)** Interaction of human STAT1 and Karyopherin proteins to upregulate IFN signalling. **B)** Inhibition of STAT1 and Karyopherin protein interaction and prevention of upregulation of IFN signalling.

1.7.6 Specificity Determining Positions

Specificity Determining Positions (SDPs) are positions in protein sequences that are differentially conserved between groups. In enzyme families, SDPs often determine the substrate-specificity of an enzyme (Rausell et al. 2010). In the case of ebolaviruses, SDPs are the positions in the genome that are most likely to be causing the pathogenicity differences between the human pathogenic ebolavirus species and the human non-pathogenic species (Section 1.7.2).

Figure 1.11 shows an example of SDP determination within a protein sequence alignment. Here, there are two groups: pathogenic species (red) and non-pathogenic species (cyan). In the example multiple sequence alignment, columns 1 and 4 are both conserved across all samples and across both groups, whereas column 2 is variable between groups and also within groups. These positions are unlikely to determine functional differences between groups. Column 3 is an example of an SDP – the position is conserved within pathogenic samples (V) and also within non-pathogenic groups (K), but the amino acid conserved in each of the groups is not the same across groups (differentially conserved).

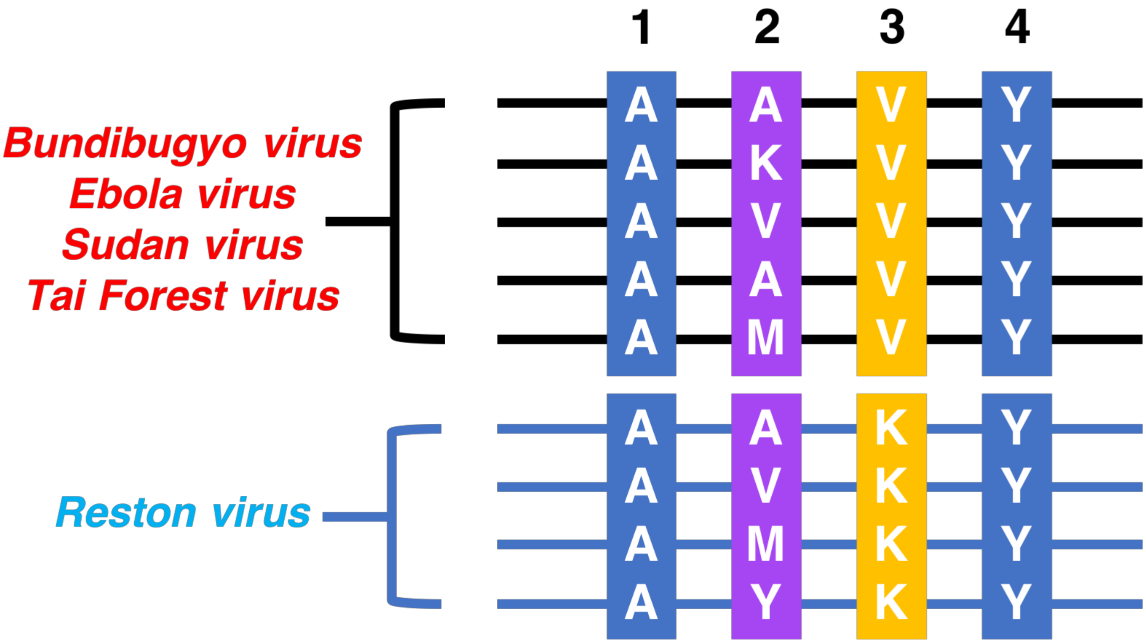


Figure 1.11: Determination of specificity determining positions. Ebolavirus species groupings are shown on the left, with pathogenic species in one group (red), and non-pathogenic species in the other group (cyan). Positions 1&4 in the sequence alignment are both conserved across the pathogenic and non-pathogenic groups. Position 2 is variable across and within the pathogenicity groups. Position 3 is differentially conserved between the groups, and would be classed as an SDP.

S3det is one method for determining SDPs, and is implemented in Chapter 4 of this thesis to determine SDPs associated with Ebolavirus pathogenicity. This tool is run on a protein multiple sequence alignment (MSA), which can be produced by a large number of tools, but in Chapter 4 the MSAs were produced using Clustal Omega (Fabian Sievers et al. 2011). S3det takes the MSA as input and uses Multiple Correspondence Analysis and the Wilcoxon test to identify significant sources of variation within the MSA. S3det can be run in a supervised or unsupervised manner. In the supervised manner, the groupings of the samples are provided to S3det by the user, whereas in the unsupervised mode S3det will attempt to determine the sample groupings itself using k-means clustering to define protein sub-families. In chapter 4, S3det was run in supervised mode, as the sequence groupings were known: pathogenic vs non-pathogenic. Positions in the MSA that segregate with the protein groupings are defined as SDPs (Rausell et al. 2010).

Chapters 4 and 5 of this thesis detail two different analyses of ebolaviruses. Chapter 4 focuses on the molecular determinants of ebolavirus pathogenicity, using an SDP approach at the protein level. This analysis represents an important update of previous work identifying SDPs associated with ebolavirus pathogenicity, with a much larger sample size than previous analyses. Chapter 5 focuses on the evolution of the *Ebola virus* genome at the nucleotide level, both since the first human outbreak in 1976, and also over the course of the 2013-2016 West Africa *Ebola virus* outbreak.

1.8 Organisation of this Thesis

This thesis broadly covers the analysis of genetic variation in humans and ebolaviruses, and with the addition of Appendices 5 and 6 the analysis of genetic variation in birds and in pigs.

Chapter 1 introduces the topics covered in this thesis, and describes the current state of genetic variation data generation, and the computational tools used in its analysis.

Chapter 2 describes an analysis of the rare human disease cystinuria, and was published under the title “Associating mutations causing cystinuria with disease severity with the aim of providing precision medicine” in BMC Genomics (Martell et al. 2017).

Chapter 3 describes an analysis of coevolution in human genomes, focusing on cooccurring variants within protein structures and within protein-protein interfaces. This paper is entitled “Identifying Coevolution Within the Human Genome” and is currently being prepared for submission.

Chapter 4 describes an analysis of ebolavirus pathogenicity, utilising the large number of ebolavirus genome sequences that have come from the 2013-2016 West Africa Outbreak. This paper is entitled “Investigating the Molecular Determinants of Ebolavirus Pathogenicity” and is currently being prepared for submission.

Chapter 5 describes an analysis of mutation patterns and biases over time within *Ebola virus* genomes. This paper is entitled “Transition-to-Transversion Bias in the Evolution of *Ebola virus*” and is currently being prepared for submission.

Chapter 6 summarises and discusses the topics covered in chapters 2-5, as well as the potential for future work building on the research presented, and the future challenges for variant interpretation and precision medicine.

Appendix 5 contains the published manuscript for the paper “Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set”, which has been published in *Genome Research* (Damas et al. 2017).

Appendix 6 contains the manuscript for the paper “A QTL for number of teats shows line specific effects on number of vertebrae in pigs: Bridging the gap between molecular and quantitative genetics”, which has been submitted to the journal *Frontiers in Genetics* and is awaiting review.

Chapter 2: Associating Mutations Causing Cystinuria with Disease Severity with the Aim of Providing Precision Medicine

Martell, H.J. et al., 2017. “Associating mutations causing cystinuria with disease severity with the aim of providing precision medicine”. *BMC Genomics*, 18(Suppl 5).

My contribution to the work was as follows:

1. Helped devise the project with Mark Wass, Kay Thomas, and Kathie Wong.
2. Wrote all of the scripts for processing the genetic variation data and the clinical data
3. Performed the protein analyses, with Juan Martin and Mark Wass
4. Performed the majority of the statistical analyses, with Kathie Wong and Ziyan Kassam
5. Produced all figures and tables, except for Figures 2.3 and 2.4 which were produced by Mark Wass
6. Wrote the manuscript with Mark Wass, with contributions from the other authors

2.1 Abstract

Cystinuria is an inherited disease that results in the formation of cystine stones in the kidney, which can have serious health complications. Two genes (SLC7A9 and SLC3A1) that form an amino acid transporter are known to be responsible for the disease. Variants that cause the disease disrupt amino acid transport across the cell membrane, leading to the build-up of relatively insoluble cystine, resulting in formation of stones. Assessing the effects of each mutation is critical in order to provide tailored treatment options for patients. We used various computational methods to assess the effects of cystinuria associated mutations, utilising information on protein function, evolutionary conservation and natural population variation of the two genes. We also analysed the ability of some methods to predict the phenotypes of individuals with cystinuria, based on their genotypes, and compared this to clinical data. Using a literature search, we collated a set of 94 SLC3A1 and 58 SLC7A9 point mutations known to be associated with cystinuria. There are differences in sequence location, evolutionary conservation, allele frequency, and predicted effect on protein function between these mutations and other genetic variants of the same genes that occur in a large population. Structural analysis considered how these mutations might lead to cystinuria. For SLC7A9, many mutations swap hydrophobic amino acids for charged amino acids or vice versa, while others affect known functional sites. For SLC3A1, functional information is currently insufficient to make confident predictions but mutations often result in the loss of hydrogen bonds and largely appear to affect protein stability. Finally, we showed that computational predictions of mutation severity were significantly correlated with the disease phenotypes of patients from a clinical study, despite different methods disagreeing for some of their predictions. The results of this study are promising and highlight the areas of research which must now be pursued to better understand how mutations in SLC3A1 and SLC7A9 cause cystinuria. The application of our approach to a larger data set is essential, but we have shown that computational methods could play an important role in designing more effective personalised treatment options for patients with cystinuria.

2.2 Introduction

Cystinuria is an inherited disorder resulting in urinary dibasic aminoaciduria (Thomas et al. 2014). The clinical presentation is varied; ranging from some patients having stone episodes every few months to other patients having only one stone in their lifetime. It is primarily caused by mutations in two genes; SLC3A1 encodes the neutral and basic amino acid transport protein (rBAT) and SLC7A9 encodes the light chain b amino acid transporter b(0+)AT (Calonge et al. 1994; Fernández et al. 2002). These two proteins form a dimer linked by a disulphide bridge (Wagner et al. 2001). b(0+)AT contains 12 transmembrane helices that form the channel through which dibasic amino acids (cystine, lysine, arginine and ornithine) are transported into the cell with the exchange of neutral amino acids. rBAT has a single transmembrane domain and a large extracellular domain. There is evidence to suggest that the extracellular glycosidase domain has a role in cystine transport but not the other dibasic amino acids (Lundgren et al. 1995). rBAT also requires chaperones to fold correctly and some mutations have been linked with incorrect folding of the protein and/or trafficking to the plasma membrane (Franca et al. 2005).

Experimental studies suggest that rBAT may function as an activator of b(0+)AT (Fernández et al. 2002; Palacín et al. 2001) but the functional role of rBAT remains unclear, although it is required for efficient transport to occur. Mutations in either of these two genes can result in defective transport of dibasic amino acids across the renal tubular membrane and intestine (Feliubadaló et al. 1999; Chairoungdua et al. 1999). In the kidneys, this results in cystine accumulating in the urine and forming stones.

SLC3A1 mutations are inherited in an autosomal recessive pattern whilst mutations in SLC7A9 can be regarded as inherited in an autosomal dominant pattern with incomplete penetrance (Eggermann et al. 2012). In SLC3A1, mutations in both alleles of the gene are required for disease presentation. In SLC7A9, some patients only have one mutation in one allele and can form cystine stones (Strologo Dello et al. 2002).

Many mutations have been identified in both SLC3A1 and SLC7A9 in individuals with cystinuria (Chillarón et al. 2010; Bisceglia et al. 1997). Frame shift, deletion, duplication, splice site and nonsense mutations typically result in large effects on the encoded protein and therefore its protein structure or function. Most mutations described in Cystinuria however, are missense mutations resulting in the change of a single amino acid in the protein. The effect of a missense mutation can range from having no effect on protein function to rendering it non-functional. For many of the missense mutations in SLC3A1 or SLC7A9, without further analysis it is not clear what effect they have on protein function and how they lead to disease presentation.

The sequencing of many people has demonstrated that each individual has between 4 and 5 million genetic variants compared to the reference human genome (1000 Genomes Project Consortium 2015). Some of these variants will cause disease or increase the risk of disease, however it is difficult from this large set of variants to identify those that are most likely to have a phenotypic effect and may have a role in disease. As a result many computational methods have been developed to predict if a genetic variant is likely to be deleterious (reviewed in (Bromberg 2013)). These methods largely focus on the analysis of non-synonymous single nucleotide variants (nsSNVs) and use many different features from sequence conservation to structural and functional information. Methods include SIFT (Ng & Henikoff 2003; Sim et al. 2012), PolyPhen2 (Adzhubei et al. 2010), SuSPect (Yates et al. 2014), VarMod (Pappalardo & Wass 2014), SNAP (Hecht et al. 2015), MutationAssesor (Reva et al. 2011), FATHMM (H A Shihab et al. 2013), CADD (Kircher et al. 2014) and Condel (González-Pérez & López-Bigas 2011).

We recently proposed that protein structural modelling and analysis of mutations present in cystinuria could be used to further our understanding of how mutations alter the function of the transporter and the extent of functional effect caused by each mutation (Wong et al. 2016). Here we perform an extensive literature survey of clinical studies to identify the range of different mutations associated with cystinuria. A structural analysis of all the identified single point mutations present in rBAT and b(0+)AT is performed to investigate the effect of the mutations on the transporter structure and function. We also compare these mutations that have been reported to cause cystinuria with the natural variation of SLC3A1 and SLC7A9 present in the large population study of genetic variation ExAC (Monkol Lek et al. 2016). Finally, the ability of automated predictors to assess the effect of cystinuria associated mutations is considered using clinical data from a cohort of 74 patients (Wong et al. 2015).

2.3 Methods

2.3.1 Cystinuria Literature Survey

A literature search was performed to identify all clinical studies of cystinuria patients. PubMed was searched with the terms “cystinuria” “cystinuria mutation” and “SLC3A1” and “SLC7A9”. Papers were first filtered on the basis of being original articles or reviews, with reviews discarded. Further filtering was performed by reading the abstracts of all papers to check for relevance. Those that were assessed to be relevant were then read fully and any relevant data extracted.

2.3.2 ExAC Data

Genetic variation data was downloaded from the Exome Aggregation Consortium (ExAC) browser, for both SLC3A1 and SLC7A9, on 28/10/2016 (Monkol Lek et al. 2016). This data set was then filtered to contain only variants that affect canonical transcripts, and then further filtered to contain only non-synonymous single nucleotide variants (nsSNVs). This resulted in a set of 318 and 144 nsSNVs not known to be associated with cystinuria for SLC3A1 and SLC7A9, respectively. The ExAC data was used to determine the allele frequencies of the variants identified by the literature search to have a role in cystinuria. In addition, all variants present in ExAC that were not identified to have a role in cystinuria form the SLC3A1 and SLC7A9 ExAC only variant sets.

2.3.3 Structural Modelling and Analysis

The protein structures of rBAT and b(0+)AT were modelled using the Phyre2 web server (Kelley et al. 2015), with default parameters. rBAT was modelled on the PDB structure 1UOK, chain A, with 78% coverage (alignment: residues 115-651) and 100% confidence. b(0+)AT was modelled on the PDB structure 4DJI, chain A, with 94% coverage (alignment: residues 28-487) and 100% confidence. Functional sites of the protein were modelled using multiple methods. The ligand binding sites including the amino acid, sugar and calcium binding sites were modelled using 3DLigandSite (Wass & Sternberg 2009; Wass et al. 2010) and firestar (Lopez et al. 2011). Protein stability predictions were made using mCSM (Pires et al. 2014) for all nsSNVs.

Residue conservation in rBAT and b(0+)AT was calculated using the following approach. Homologues were identified using BLAST (Altschul et al. 1990) to search the UniProtKB with default parameters (Bateman et al. 2015). A multiple sequence alignment and a phylogenetic tree were generated for each of these sets of homologues using Clustal Omega with default parameters (McWilliam et al. 2013; F Sievers et al. 2011). For each gene, the multiple sequence alignment, phylogenetic tree, and phyre2 structural model were submitted to ConSurf (Ashkenazy et al. 2016), which was run using the Bayesian prediction method with all other parameters set to default. The proteins used for the alignment, and the species that they come from, are shown in Appendix 1 Tables 1 & 2. For the 2 proteins, there were 158 common species, 87 species unique to the alignment of rBAT, and 45 species unique to the b(0+)AT alignment. This shows that the majority of the species used in the two alignments are the same, and there is not a large difference in the species distributions of the two alignments. This means that reasonable comparisons of conservation between the two proteins can be made.

2.3.4 Clinical Data

Phenotypic data associated with cystinuria was available for a cohort of 74 patients in the UK (Wong et al. 2015), consisting of 41 patients with mutations in SLC3A1, 32 in SLC7A9 and one patient without a mutation in either SLC3A1 or SLC7A9. Available phenotypic data included, urinary dibasic amino acids levels for cysteine, ornithine, arginine and lysine, the age of disease presentation, and the number of stone episodes and number of interventions over a three-year period. Two of the patients in this cohort were removed from this study as they had mutations in both SLC3A1 and SLC7A9.

2.3.5 Automated Prediction of the Effect of Mutations

The mutations found in the literature search were submitted to SIFT (Sim et al. 2012), PolyPhen2 (Adzhubei et al. 2010) (using both predictive models HumDiv and HumVar), MutationAssessor (Reva et al. 2011), FATHMM (H A Shihab et al. 2013), Condel (Gonzalez-Perez & Lopez-Bigas 2011) and CADD (Kircher et al. 2014) using default settings. Condel and CADD differ from the other prediction methods in that they integrate predictions from individual methods to create an overall prediction. For example, Condel uses predictions from PolyPhen2, SIFT, Mutation Assessor, and FATHMM to make predictions.

The different methods make predictions in different categories, SIFT only predicts two categories “Tolerated” and “Damaging”, as do FATHMM, CADD, and Condel (“neutral” and “damaging”), while PolyPhen2 categorises mutations into three categories “benign”, “possibly damaging” and “probably damaging” and MutationAssessor predicts four categories (“neutral”, “low”, “medium”, and “high”). PolyPhen-2 is available using two different training models, HumDiv and HumVar. These two models agree for 47 of 58 mutations in b(0+)AT and 81 of 94 mutations in rBAT. As we want to distinguish between mildly deleterious and more severe mutations, the remaining analyses consider only the results using the HumVar training model. However, as shown above the two models give similar results.

2.3.6 Grouping Patients by Mutation Severity and Comparison of Phenotypes

The different categories of the prediction methods make analysis of the overlap of agreement between the methods difficult to assess. Therefore, the prediction scores made by each of the automated methods were classified into two groups, either mild or severe and these two groups were then associated with a score: mild = 1 and severe = 2. Multiple thresholds for grouping mutations were tested for each of the prediction methods (see Appendix 1 Tables 3 & 4), starting from the recommended threshold for separating deleterious and neutral mutations for the specific method (see Table 2.1). The stringency of the threshold was then increased incrementally to separate the high confidence predictions from the medium confidence predictions. Thresholding above the cut-offs for deleterious vs neutral variants was necessary because these methods are designed to predict even mildly deleterious variants as deleterious, and we want to separate mild from severe mutations.

Frameshift, deletion, splice site and nonsense mutations are typically likely to have a significant effect on protein function and were therefore all assigned scores of 2.

This may represent a simple scoring scheme but given the different categories and scoring scales of the different methods it appeared to be the most appropriate. Using the mutation scores from the predictive methods, each patient was assigned an overall severity score for the mutations that they have. As SLC7A9 mutations show dominant inheritance with incomplete penetrance, for patients with SLC7A9 mutations, patient scores were the total of the scores for the individual mutations in each allele. This results in scores ranging from one (only one mild mutation present) to four (patient has two mutations classified as severe, one in each allele).

Table 2.1: Mutation severity prediction score thresholds used for each method, based on stabilisation of group numbers above the recommended deleterious threshold.

Method	Standard Deleterious Threshold	Threshold for Mutation Severity Score of 1	Threshold for Mutation Severity Score of 2
SIFT	Score < 0.05	Score > 0.025	Score ≤ 0.025
PolyPhen2	Score ≥ 0.5	Score < 0.80	Score ≥ 0.80
MutationAssessor	Score > 1.9	Score < 2.7	Score ≥ 2.7
FATHMM	Score ≤ -1.5	Score ≥ -8.5	Score < -8.5
Condel	Score > 0.522	Score ≤ 0.672	Score > 0.672
CADD	Score ≥ 15	Score < 27.5	Score ≥ 27.5

A similar approach was taken for SLC3A1, but inheritance of cystinuria from SLC3A1 mutations is autosomal recessive, therefore mutations are required in both alleles to have the disease. For each patient, it was considered that the allele with the worst mutation would not be expressed while the other allele would be. For example, an individual with two mutations scored at 1, would have an overall score of 1, as would a patient with one mutation scored at 1 and the other at 2. Finally, an individual with two severe mutations would score 2 overall. This strategy is valid for our dataset, because no individual has more than one mutation in a single allele of SLC3A1.

The properties of each set of data were compared using the Wilcoxon rank sum test to find any statistically significant differences between the groups. All p-values were corrected for multiple testing using the Bonferroni method. Statistical figures were produced using the R statistical package, version 3.2.1 (R Core Team 2015). Additionally, plots with axes gaps were produced using the R package ‘plotrix’ (Lemon 2006).

2.4 Results

A literature search identified 52 articles, consisting of 49 original articles and three reviews. All of the original articles were deemed relevant from the abstract and read in full to extract data. From the clinical studies we identified a total of 94 SLC3A1 and 58 SLC7A9 cystinuria associated point mutations.

The 94 unique nsSNVs in SLC3A1 affect 81 different amino acid positions as 10 residues have two variant amino acids and residue p.Arg365 has four different variant amino acids present. For SLC7A9 the 58 nsSNVs affect 55 different amino acid positions with only residues 105, 195 and 333 having two different variant amino acids.

2.4.1 Initial Comparison of Cystinuria Associated Mutations with Variation Present in a Large Population

The ExAC resource (M Lek et al. 2016) provides access to the variant frequencies from over 60,000 individuals. We identified all variants present in SLC7A9 and SLC3A1 (Appendix 1 Tables 5 & 6) to investigate the variation present in a large population of individuals and compare variants/mutations associated with cystinuria and those not associated with the disease. Worldwide prevalence of cystinuria is estimated at 1 in 7,000 (Barbosa et al. 2012), though variation by geographical location is large (1 in 100,000 in Sweden (Harnevik et al. 2003), and 1 in 2500 in Libyan Jews (Chillarón et al. 2010). Using the worldwide prevalence, in the ExAC set of just over 60,000 individuals, we would therefore expect approximately nine individuals to have the disease.

The vast majority of cystinuria associated variants in both SLC7A9 and SLC3A1 occur very rarely with an allele frequency of less than 0.01% and many are not present in the ExAC dataset (allele frequency of 0% - Figure 2.1.C). This suggests that these variants are under purifying selection and that these mutations are deleterious. A few disease-associated variants have much higher frequencies (between 0.27%–31%). The cystinuria associated variant p.Val142Ala in SLC7A9 has an allele frequency of 31% indicating that it regularly occurs in individuals. Given the high frequency of this variant it is likely that it has a limited effect on

SLC7A9 function as cystinuria is a rare disease, this is reinforced by the low evolutionary conservation of residue 142 in the protein (ConSurf score of 1). Many non-disease associated variants in SLC7A9 and SLC3A1 also have low frequency (<0.01%; Figure 2.1.C). For SLC7A9 there are a few variants with higher frequencies, and a considerable number more for SLC3A1. This demonstrates that there is limited variation in SLC3A1 and SLC7A9 in the population. The higher frequency of variants in SLC3A1 may reflect that it has autosomal recessive inheritance, whereas SLC7A9 inheritance is autosomal dominant with incomplete penetrance. Thus, a single SLC7A9 allele can result in cystinuria.

At the protein level there appears to be some clustering of the rBAT cystinuria associated point mutations in sequence (Figure 2.1.B). For example, there are some mutated positions that occur in stretches throughout the protein (including 121–124, 253–256, 480–482, 552–568). While some ExAC variation occurs in the same sequence regions, there appears to be less clustering of these variants and ExAC only variants largely occur in parts of the protein sequence where cystinuria associated mutations are not present (Figure 2.1.B).

For b(0+)AT, unlike rBAT, there is less evidence of clustering of the amino acids that are mutated, with no runs of residues being mutated and only a few examples of adjacent residues being mutated (Figure 2.1.A). Again, there is limited overlap with ExAC variation data (Figure 2.1.A).

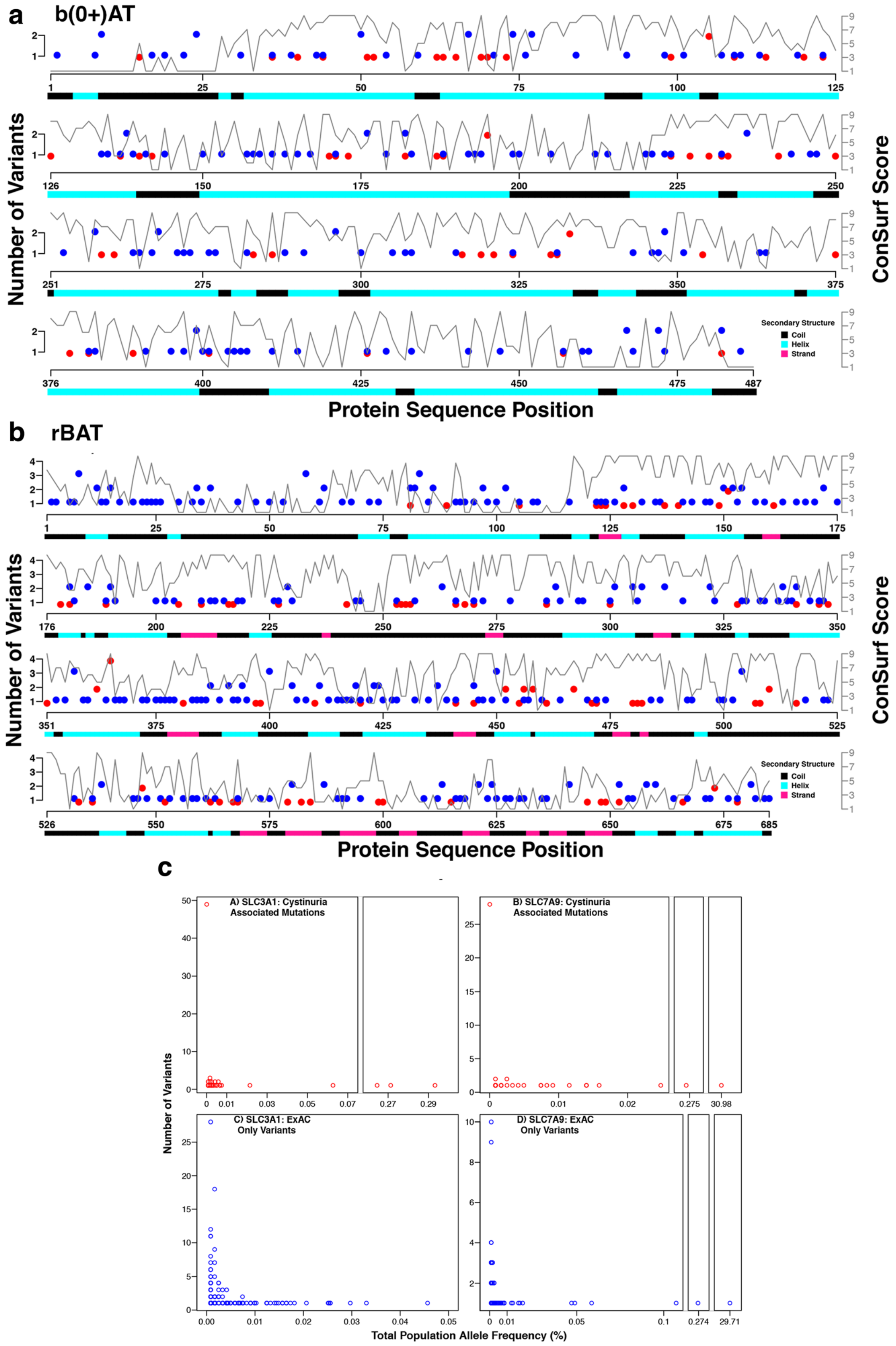


Figure 2.1: (see legend on next page)

Figure 2.1: Mutations present in b(0+)AT (SLC7A9) and rBAT (SLC3A1) in patients with cystinuria. Plots of the sequence of **A)** b(0+)AT and **B)** rBAT. For each protein the location of cystinuria associated mutations is shown (red circles) with the position of variants present in ExAC (blue circles). The conservation score is shown (grey line with values ranging from 1 to 9). The lower bar shows the protein secondary structure. **C)** Total population allele frequencies based on the ExAC data set. Each point represents an allele frequency. Multiple variants may have the same allele frequency, and the number of variants with the specific allele frequency is represented by the position of the point on the Y axis. The individual plots correspond to the four different sets of variants. **C.A)** Variants of SLC3A1 reported to be associated with cystinuria. **C.B)** Variants of SLC7A9 reported to be associated with cystinuria. **C.C)** Variants of SLC3A1 not reported to be associated with cystinuria but present in ExAC. **C.D)** Variants of SLC7A9 not reported to be associated with cystinuria but present in ExAC.

The conservation of each variant position was calculated using ConSurf (see Methods). The conservation scores range from 1 to 9, with 1 being the most variable and 9 the most conserved. For both genes, there is a clear difference in the distributions between the mutations known to be associated with cystinuria and the variants only found in ExAC (p = 1.69e-10 for SLC3A1, and p = 2.078e-06 for SLC7A9, Wilcoxon rank sum test) (Figure 2.2). The cystinuria associated mutations of both genes are predominantly at positions with high ConSurf scores, suggesting that these positions are of high importance to the function of the protein. This skew is larger for SLC7A9, where ~80% of the mutations have ConSurf scores between 6 and 9 (Figures 2.1.A-B and 2.2.A-B). Conversely, for both genes, a large number of variants that are not known to be associated with cystinuria have ConSurf scores of 1 (Figures 2.1.A-B and 2.2.C-D), suggesting that the functional roles of these positions are minimal. This agrees with these variants being neutral, as such positions are less likely to have an effect on protein structure or function. Around 40% of the positions of ExAC only variants have ConSurf scores between 6 and 9, it is possible that they may have some effect upon protein function or that the variants observed conserve the property of the wild type amino acid more so than the cystinuria associated mutations. We did not observe a correlation between ExAC allele frequency and ConSurf conservation score (Appendix 1 Figure 1).

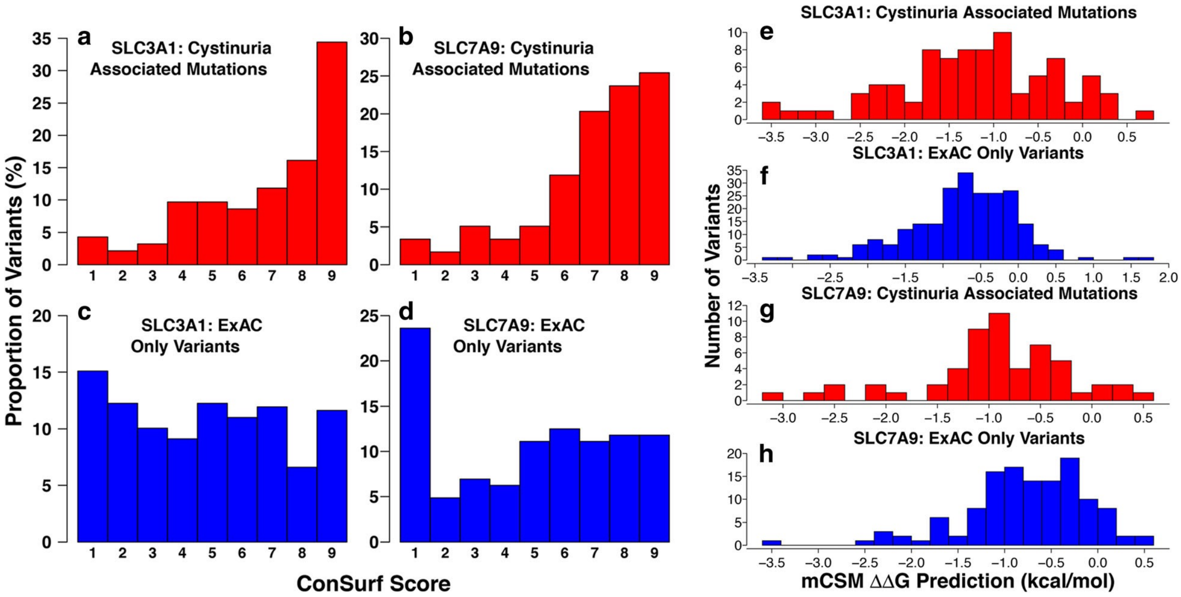


Figure 2.2: Conservation of nsSNVs in SLC7A9 and SLC3A1 and their predicted effect on protein stability. **A-D)** Distribution of ConSurf conservation scores for nsSNVs in SLC7A9 and SLC3A1 that are either i) present in individuals with cystinuria (**A** and **B**) or ii) present in the ExAC dataset (**C** and **D**). ConSurf scores vary between 1 and 9, with 9 being highly conserved and 1 being not conserved. **E-H)**. Effect of nsSNV on protein stability predicted by mCSM. mCSM predicts the change in Gibbs free energy (kcal/mol) negative values indicate destabilisation and positive values stabilisation.

2.4.2 Protein Structural Modelling

To investigate where in the protein structure the cystinuria associated variants occur and to analyse the effect they may have on protein structure and function, protein structural modelling of the protein was performed. Phyre2 (Kelly et al. 2015) generated high confidence structural models of both b(0+)AT and rBAT (Figure 2.3). For rBAT, the extracellular alpha amylase-like domain was modelled using the structure of Bacillus Cereus oligo-1,6-glucosidase (Watanabe et al. 1997) as a template (PDB code: 1uok). The structure of a glutamate and γ-aminobutyric acid antiporter (Ma et al. 2012) (PDB code: 4DJI) was used as the template structure for modelling b(0+)AT.

The b(0+)AT protein transports dibasic amino acids into the cell in exchange for neutral amino acids. There are therefore two sites for amino acid binding, one on each side of the transporter. The outward facing binding site was modelled using 3DLigandSite and firestar using the Arginine bound to another related APC transport (AdiC, PDB code: 3OBM) (Watanabe et al. 1997). To model the inward facing conformation, the putative binding site identified for ApcT (another member of the APC transporter family) was mapped onto our model (Kowalczyk et al. 2011) (Figure 2.3.A).

Studies have proposed that Lys158 in ApcT has a role equivalent to sodium in sodium dependent transporters (Shaffer et al. 2009). This lysine is conserved in b(0+)AT (Lys184) and three of the four residues coordinating with it are also conserved (Gly41, Ile44, Ser312).

Overall potential functional residues identified in b(0 +)AT for amino acid binding were: Ile38, Thr42, Ile43, Ser46, Gly47, Val50, Thr91, Lys92, Leu117, Lys121, Ser124, Ile128, Trp230, Ala231, Tyr232, Ile371. As the mechanism of transport is not clearly understood there are likely to be further residues that are functionally important that have not been identified here.

In rBAT the residues predicted to have a potential functional role in the alpha amylase domain for sugar binding were: Asp172, Tyr175, His215, Val258, Tyr259, Phe278, Met279, Gln282, Ser312, Asp314, Ala315, Phe318, Glu384, Asp449 (Figure 2.3.B). Additionally, Asp133, Asn135, Asp137, Asn139, Asp141 are predicted to bind calcium (Figure 2.3.B). However, it is not clear if rBAT binds sugar molecules or if it has an alpha amylase enzyme activity.

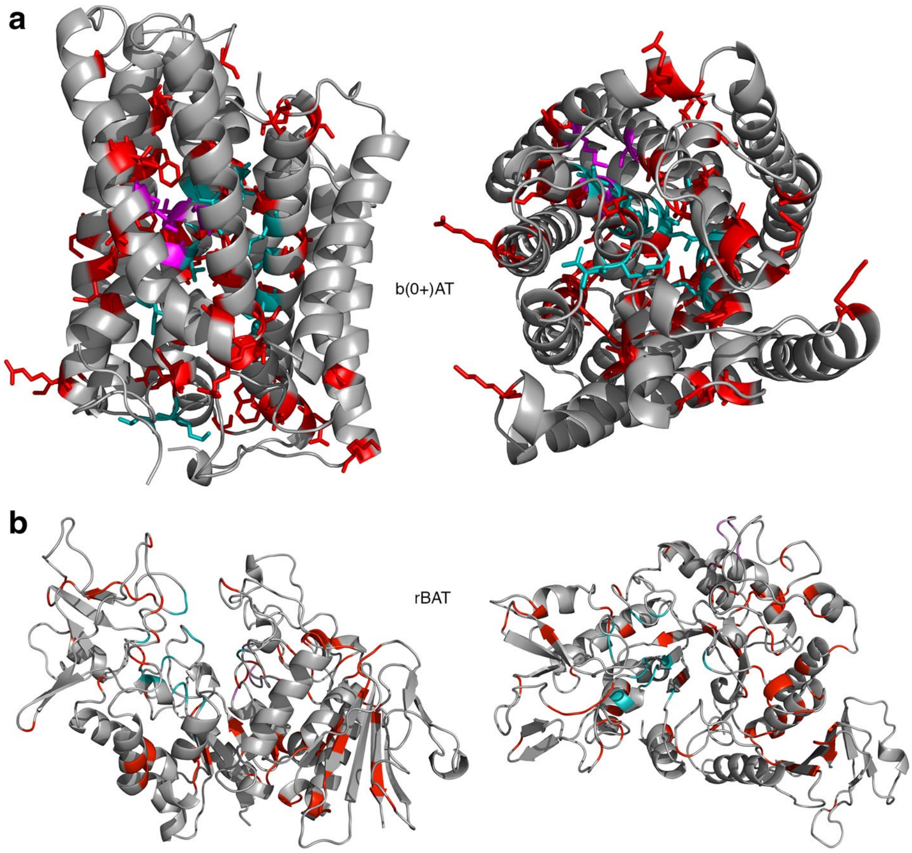


Figure 2.3: Structural models of rBAT and b(0+)AT. For both proteins cystinuria associated mutations are coloured red. **A)** Model of b(0+)AT. Residues modelled to contact the transported amino acids are coloured cyan. The conserved p.Lys184 and residue coordinating with it are coloured magenta. **B)** Model of rBAT. The modelled sugar binding site residues are coloured cyan and the predicted calcium binding site is magenta.

2.4.3 Structural Analysis of Mutations in b(0+)AT

Initial analysis of the substitutions that occur in b(0+)AT shows that for only 22 of the 58 point mutations the type of amino acid is not changed (Table 2.2). The majority of residues that are mutated are hydrophobic and for more than half of the changes (25 of 45) the mutated residue is polar or charged. This shows that mutations are regularly introducing charge into the protein.

The likely effects of the mutations fall into a few categories (full analysis details in Appendix 1 Table 7). Firstly, some mutations alter residues with a functional role (e.g. ligand binding) or they are located close to functional sites. Secondly, some mutations seem likely to alter protein conformation as they either introduce charge or change the size/shape of the sidechain (often in buried or densely packed regions of the protein). Finally, some mutations are located on the protein surface and they could affect the interaction with the membrane or with rBAT.

Table 2.2: Type of amino acid change for mutations in b(0+)AT.

	Mutation Amino Acid Type				
Original Amino Acid Type	Hydrophobic	Polar	Positive	Negative	Total
Hydrophobic	20	10	12	3	45
Polar	5	0	1	1	7
Positive	2	1	1	1	5
Negative	0	0	0	1	1

There are a set of mutations close to the functional residue Lys184, which is likely to function in an equivalent way to sodium in sodium dependent transporters. One of the residues thought to coordinate with Lys184, Ile44, is mutated to Thr (Figure 2.4.A). Additionally, in the same area there are the mutations p.Ile36Asn, p.Val40Met, p.Ala182Thr, p.Ile187Phe, p.Val188Met and p.Pro261Leu (Figure 2.4.A). Many of these mutations seem fairly conservative, and suggest that minor changes to the conformation of the protein, through altered packing of sidechains may be sufficient to alter function. This may be particularly relevant as helix 1 (containing p.Ile36Asn, p.Val40Met, p.Ile44Thr) is thought to undergo conformational change during transport and the other side of the helix contains multiple residues that are likely to have a role in binding the transported amino acids (Figure 2.4.A – cyan coloured residues).

Other mutations are close to the residues likely to have a functional role in transporting the amino acids. One of these functional residues, Trp230, is mutated to Arg (Figure 2.4.B) and there are multiple other mutations in the same area that are close to functional residues (Figure 2.4.B-D).

b(0+)AT contains many hydrophobic amino acids, which are often tightly packed. In multiple examples, a smaller hydrophobic is replaced by either a polar/charged amino acid or a larger hydrophobic (for example p.Gly319Arg - Figure 2.4.C).

A final group of mutations may affect the interactions of the protein with the lipid bilayer and its stability. Most of these mutations either introduce (p. Tyr99His, p.Ala109Thr, p.Cys137Arg, p.Phe140Ser, p.Gly195Arg, p.Tyr457His), remove (p.Arg171Trp, p.Asp333Trp) or alter charge (p.Arg250Lys, p.lys401Arg) mainly at the end of helices on the protein surface near the end of the membrane (examples shown in Figure 2.4.D). Another possible impact of mutations on the protein surface of b(0+)AT (and also rBAT) is that they interfere with the dimerization of the two proteins. However, little is known about how these two proteins interact and what residues are involved in the interaction, so predictions of the impact on dimerization were not possible.

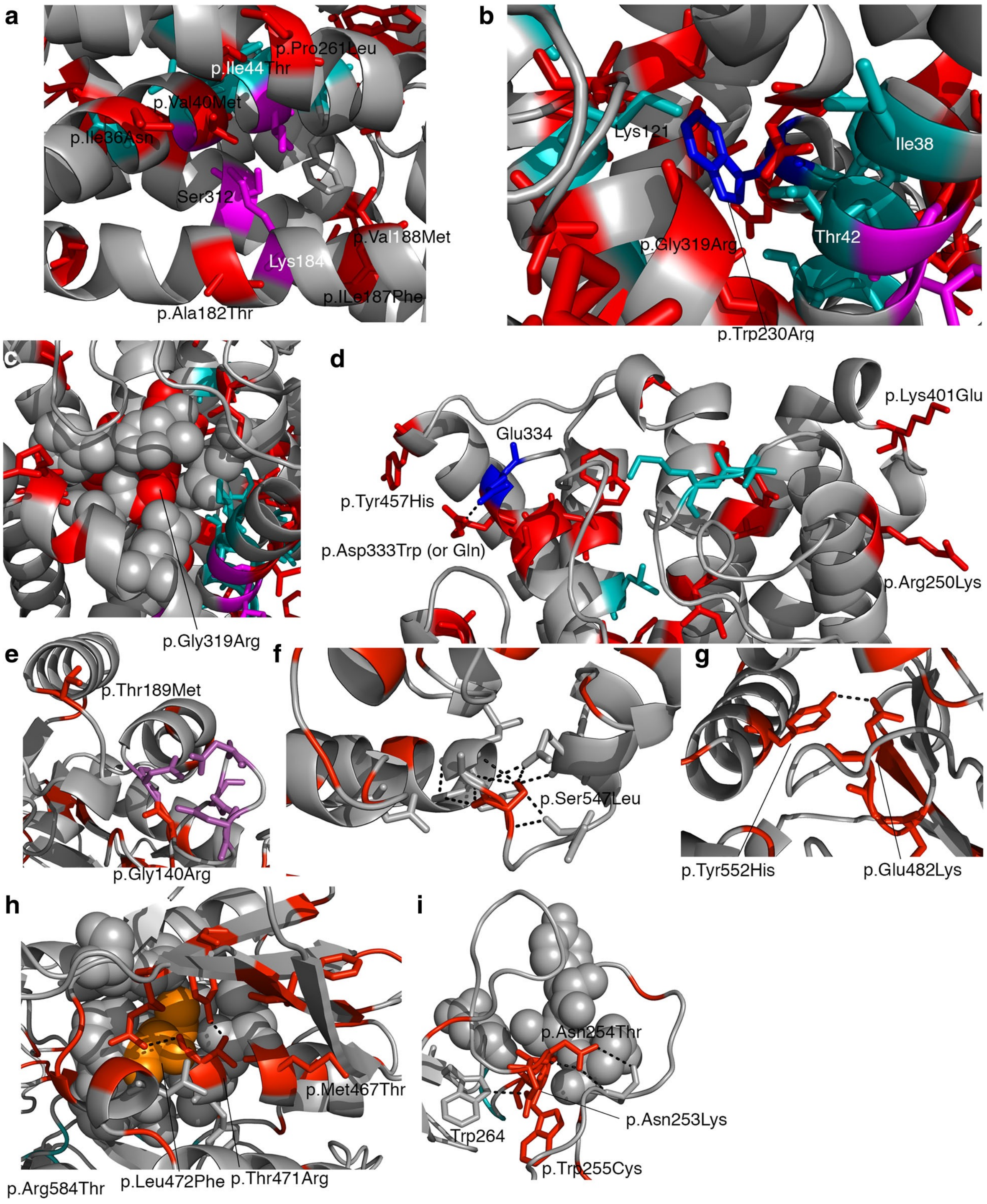


Figure 2.4: (see legend on next page)

Figure 2.4: Mutations in b(0+)AT and rBAT. In all images the mutated residues are displayed as red sticks in their wild type format. Predicted functional residues are coloured cyan. Hydrogen bonds are shown as dashed black lines. Images **A-D)** refer to b(0+)AT and images **E-I)** refer to rBAT. **A)** p.Ile187Phe and p.Ala182Thr mutations are adjacent to p.Lys184 which is thought to play a role equivalent to sodium in sodium dependent transporters. **B)** p.Trp230Arg (coloured blue) is adjacent to multiple functional residues. **C)** The mutation p.Gly319Arg occurs in a buried region (p.Gly319 shown in red spheres). **D)** Mutations close to the end of transmembrane helices may reduce stability in the membrane. **E)** Mutations occurring close to the predicted calcium binding site in rBAT. **F)** Mutation p.Ser547Leu will remove hydrogen bonding. **G)** p.Tyr552His and p.Glu482Lys as wild type form a hydrogen bond, it is not clear if this will be retained upon mutation. **H)** Multiple mutations present in a single region. p.Leu472Phe (orange spheres) will result in increased size in well packed area. Other mutations will remove hydrogen bonding. **I)** Mutations occur in residues 253–256.

2.4.4 Structural Analysis of Mutations in rBAT

While rBAT has an alpha-amylase like extracellular domain, the functional role of this domain has not been well established. Overall there are few mutations present (only three) in or near the predicted functional residues (based on possible sugar and calcium binding sites) (Figure 2.3.B). This suggests that these residues may not be functional in rBAT, otherwise mutations would be expected to occur here as was seen for mutations in b(0+)AT (although ConSurf shows that these residues are highly conserved, 11 of the 14 have scores of 8 or 9). It suggests that we do not know what residues are functionally important in rBAT and what function they perform. This makes the structural analysis difficult.

Despite this, a few mutations are located close to “functional” regions of the protein. p.Arg137Gly is one of the residues predicted to bind Calcium, mutation to glycine would lose the positive charge in this region. Similarly, p.Gly140Arg is present within the loop where calcium is modelled to be bound (Figure 2.4.E). This position is completely invariant in homologues (with a maximum ConSurf score of 9) suggesting an important structural/functional role for this residue. Introduction of a positively charged arginine may be expected to interfere with the binding of the positively charged calcium ion (assuming that Calcium does bind here).

p.Thr189Met is located in the alpha helix adjacent to the calcium binding site so it is possible that destabilisation here could affect the calcium binding site (Figure 2.4.E). Three of the mutations (p.Met381Thr, p.Tyr397Cys, p.Gly398Arg) are close to what would be the active site if the protein was an active hydrolase. p.Met381 is highly conserved in orthologues and the mutation to threonine could introduce a polar contact with p.Asp369. Similarly, p.Gly398Arg would introduce a charge and a larger sidechain. For p.Tyr397Cys, the mutation is likely to remove a hydrogen bond (see below).

Overall the structural analysis suggests that, for the majority of the rBAT mutations observed, they may have an effect on the structure or stability of the protein (full structural analysis details in Appendix 1 Table 8). These mutations fall into two main groups. In the first group, a hydrophobic amino acid is replaced by a polar or charged amino acid (examples are p.Tyr151Cys, p.Leu205Ser, p.Leu300Ser, p.Tyr397Cys, p.Tyr461His, p.Met467Thr, p.Ile445Thr, p.Tyr579Asp, p.Phe599Ser) where the hydrophobic side chain is typically buried and packed against other hydrophobic side chains. Of the 49 hydrophobic sidechains that are mutated, 17 are changed to charged amino acids and 15 to polar sidechains (Table 2.3). In the second group, a polar or charged amino acid is typically replaced by a hydrophobic side chain (but in some cases a different polar/charged sidechain) and modelling suggests that these mutations often remove hydrogen bonds or salt bridges that stabilise the protein structure (Figure 2.4.F-I). Of the 45 polar or charged residues that are mutated, 33 are likely to result in loss of hydrogen bonding (Appendix 1 Table 8) and half of them (23) are changed to hydrophobic amino acids (Table 2.3). Examples of these mutations include p.Thr189Met, p.Thr216Met, p.Thr341Ala, p.Arg365Leu, p.Arg452Trp, p.Ser455Leu, p.Ser547Leu (examples shown in Figure 2.4.F-I; Appendix 1 Table 8).

The remaining mutations either change the polarity or charge of the sidechain (Table 2.3) or result in a considerable change in the size of the sidechain. Of the 94 mutations only 21 remain in the same group (i.e. hydrophobic, polar, positive or negative charge; Table 2.3). For the 17 mutations that replace a hydrophobic amino acid with another hydrophobic sidechain, five see a considerable increase in sidechain size (e.g., p.Leu256Phe, p.Gly645Ala) and for a further six the size of the sidechain is reduced (e.g. p.Tyr124Cys, p.Trp255Cys (Figure 2.4.I), p.Tyr480Cys).

Table 2.3: Type of amino acid change for mutations in rBAT.

	Mutation Amino Acid Type				
Original Amino Acid Type	Hydrophobic	Polar	Positive	Negative	Total
Hydrophobic	17	15	13	4	49
Polar	10	2	6	2	20
Positive	10	5	2	0	17
Negative	3	1	4	0	8

The initial sequence analysis suggested clustering of mutations (Figure 2.1). This was also apparent from the structural analysis. There are multiple examples of mutated residues that are close in three dimensions that are not adjacent in sequence. These include: p.Tyr124-p.Tyr151-p.Tyr480, which appear to have Pi interactions between their aromatic sidechains; p.Trp161-p.Asp210, p.Asp179-p.Arg181, p.Arg452-p.Tyr480, p.Arg584-p.Thr417 and p.Tye552His-p.Glu482Lys each of which form a hydrogen bond between them (e.g. Figure 2.4.G); and p.Phe22-p.Glu268-p.Arg270-p.Arg227 and a large group including residues p.Met467-p.Thr471-p.Leu564-p.Leu567-p.Gly568-p.Tyr582. This clustering suggests that these are regions that either have important structural or functional roles, where mutation of any of them results in changes to protein function.

Given the presence of multiple variants that appear to affect protein stability, mCSM (Pires et al. 2014) was used to predict the effects of nsSNVs on protein stability. For mCSM, negative ΔΔG values are destabilising for a protein structure, and positive ΔΔG values are stabilising (Figure 2.2.E-H). rBAT variants known to be associated with cystinuria appear to be distributed more towards negative ΔΔG values than variants present only in ExAC, with median values of –1.122 and –0.668 respectively (p = 7.896e-06, Wilcoxon rank sum test; Figure 2.2.E-F). Compared to cystinuria associated nsSNVs in b(0+)AT, a greater proportion in rBAT are predicted to be highly destabilizing to the protein (median rBAT value of

–1.122, and –0.889 for b(0+)AT) (Figures 2.2.E & 2.2.G), supporting the observation that rBAT variants are more likely to destabilize the protein structure. However, this just falls short of statistical significance (p = 0.05068, Wilcoxon rank sum test).

2.4.5 Automated Prediction of Effects of Mutations in rBAT and b(0+)AT

There are many automated methods available to predict the effect of non-synonymous SNVs. Six of these methods, SIFT, PolyPhen-2, MutationAssessor, FATHMM, Condel, and CADD (note Condel and CADD are consensus methods that combine the output from multiple individual prediction methods to generate an overall prediction) were used to predict the effect of each of the mutations present in rBAT and b(0+)AT (see Methods). This was done to compare the predictions made with clinical data from a cohort of 74 patients in a UK cystinuria clinic (Wong et al. 2015). The predictions made by each of the methods are summarised in Table 2.4 (full predictions in Appendix 1 Tables 9 & 10).

For 20 b(0+)AT (of 58) and 31 (of 94) rBAT mutations all methods make the most deleterious predictions. No mutations in either protein were predicted by all six methods to have the lowest or mildest effect on function. For both proteins, the methods agree for a similar proportion of mutations (32.9% for b(0+)AT, 34.5% for rBAT).

The effects of three mutations in b(0+)AT (p.Gly105Arg, p.Ala182Thr and p.Arg333Trp) have been experimentally characterised. We compared the predictions with the known effects (Table 2.5) and observed good agreement. For b(0+)AT two (p.Gly105Arg and p.Arg333Trp) of the three characterised mutations reduce amino acid transport to 10% of wild type and for both of these mutations all methods predict the greatest effect (Table 2.5). The third mutation (p.Ala182Thr) reduces transport to 60% of wild type and three methods predict that this mutation will have a limited or no effect on the protein, and three predict that it will be damaging.

The divergence between these predictions highlights many important features of pathogenicity prediction tools. Here, these tools are being compared for their predictions of variants that have been experimentally validated. Many of these tools are trained against different datasets, where the reported effects are from clinical associations as opposed to experimental validation. Large-scale benchmarking of variant prediction tools has shown that individual prediction methods have differing performance on clinical association datasets than experimentally validated datasets (Li et al. 2018).

Interestingly, CADD and Condel are both ensemble prediction methods, that integrate the predictions of other tools in their overall prediction, and both correctly predicted all three variants to be pathogenic. Whereas, SIFT, PolyPhen-2, and MutationAssessor performed less well and are all classed as 'function prediction methods' that attempt to predict if a missense variant will change a protein's function. Suggesting that ensemble methods have the better overall performance, in-line with previous findings (Li et al. 2018). However, this delineation in prediction method category and prediction accuracy for these variants breaks down for FATHMM, which is a function prediction method but also correctly predicted that these three variants are pathogenic (Table 2.5).

Overall, the source of diversion for individual variant predictions by each tool can be hard to decipher. These methods are trained on different data sets, use different features of variants in their predictions, have different thresholds for pathogenicity, and calculate an overall pathogenicity prediction using different methods, including different machine learning methods. However, this does highlight the importance of integrating multiple prediction tools to come to a consensus prediction, and also of selecting the best prediction method for a specific question (Li et al. 2018).

Table 2.4: Summary of effects of mutations predicted by the automated methods.

	rBAT	b(0+)AT
SIFT – Tolerated	23	11
SIFT – Damaging	70	48
PolyPhen2 (HumVar) – Benign	12	14
PolyPhen2 (HumVar) – Possibly Damaging	15	16
PolyPhen2 (HumVar) – Probably Damaging	66	29
PolyPhen2 (HumDiv) – Benign	8	15
PolyPhen2 (HumDiv) – Possibly Damaging	11	4
PolyPhen2 (HumDiv) – Probably Damaging	74	40
MutationAssessor – Neutral	2	4
MutationAssessor – Low	17	6
MutationAssessor – Medium	40	29
MutationAssessor – High	34	20
FATHMM – Neutral	0	0
FATHMM – Damaging	93	59
Condel – Neutral	1	7
Condel – Damaging	92	52
CADD- Neutral	5	6
CADD - Deleterious	88	53

Table 2.5: Comparison of predictions with known experimentally characterised effect.

Mutation	Known effect on amino acids transport	SIFT	PolyPhen-2	Mutation Assessor	FATHMM	Condel	CADD
b(0+)AT – G105R	Reduced to 10% of WT	Damaging	Probably Damaging	Medium	Damaging	Damaging	Damaging
b(0+)AT A182T	Reduced to 60% of WT	Tolerated	Benign	Low	Damaging	Damaging	Damaging
b(0+)AT R333W	Reduced to 10% of WT	Damaging	Probably Damaging	High	Damaging	Damaging	Damaging

2.4.6 Comparison of Functional Effect Predictions with Patient Phenotype

We considered how well the predicted effects of the mutations agreed with the observed phenotypes of patients, based on the grouping of the patients by the prediction scores (see Methods). Each mutation was assigned a severity score of either 1 (mild) or 2 (severe) based on the prediction score from the predictive method. Then for each individual an overall severity score was calculated based on the mutations present (see Methods). The patients were grouped according to their overall score and the phenotype data compared between the different groups (see Methods).

As the predictive methods associate a score (or probability) with their predictions we first investigated how altering the threshold between assigning a mutation to the mild or severe group affected the outcome of the comparisons. For each mutation effect prediction method, the number of patients in each severity score group stabilised over a range of prediction score thresholds, e.g. for PolyPhen2 the groupings are stable between thresholds of ≥0.65 and ≥0.80. A single threshold within these ranges was then chosen as the cut-off between mild and severe mutations for that method and used for comparison (Table 2.1 and Appendix 1 Tables 3 & 4).

First b(0+)AT was considered, (Figures 2.5, 2.6.A&B, and Appendix 1 Figures 2 & 3). Each of the prediction methods sorted the patients in to slightly different groupings, but various general trends were observed across the methods. Considering the urinary levels of the amino acids arginine, ornithine, lysine and cystine, for all methods there was a general trend for the levels to be higher in the high severity score groups. For each method the difference seems greatest for arginine, and for SIFT and CADD there was a significant difference between severity score groups 3 and 4 for the levels of arginine (Figures 2.5 & 2.6.A&B). The age of diagnosis and the number of stone episodes and interventions over a three-year period were also considered. Again, the average values for the higher severity score groups typically followed the pattern that may be expected i.e. lower severity score group have a later age of presentation and lower number of stone episodes and interventions (Appendix 1 Figures 2 & 3). Many of these values have large ranges (as demonstrated by the error bars), but some of these comparisons showed statistical significance, e.g. PolyPhen2 predicted groups 3 and 4 show a significant difference for the number of interventions ($p = 0.017$) and age of diagnosis for MutationAssessor and Condel predicted groups 3 and 4 ($p = 0.015$ and $p = 0.014$, respectively).

The equivalent analysis was performed for patients with mutations in rBAT. For rBAT there are only two categories – mild (score 1) and severe (score 2) (details in Methods). For all of the urinary amino acid levels and for all predictive methods the average for severity score group 1 was lower than for severity score group 2 (Figures 2.6.C&D & 2.7). These differences in arginine levels were statistically significant across all methods. The differences in ornithine and lysine levels were statistically significant for PolyPhen2, MutationAssessor, and CADD, and the difference in ornithine was significant for Condel. No method found a statistically significant difference between the groups for cystine levels. However, there are difficulties in the accurate measurement of cystine, so this result is unlikely to be reliable and the levels of the other amino acids are a better indicator of disease severity.

Patients in severity score group 1 for all methods present with disease at a later age than those in severity group 2, though these differences are not statistically significant (Appendix 1 Figures 4 & 5). There is little difference between the average number of stone episodes for the two groups (for all predictive methods) and for SIFT the number of interventions is greater in the lower severity score group (Appendix 1 Figures 4 & 5).

All methods struggle to identify differences in the age of diagnosis, number of stone episodes and number of interventions, but they are better at finding differences in the urinary amino acid levels between the severity score groups. This is perhaps because urinary amino acid levels are a less complex phenotype, which was directly measured. In contrast the other phenotypes are less easily measured or recorded. For example, a patient may present with a large number of stones which all pass spontaneously, whereas another patient may present with fewer more serious stone episodes. The measurements taken may depend on patients accurately recording the number of stones that pass and medical interventions may also affect the number of stone episodes. Therefore, comparisons of urinary amino acid levels may be more informative of disease severity.

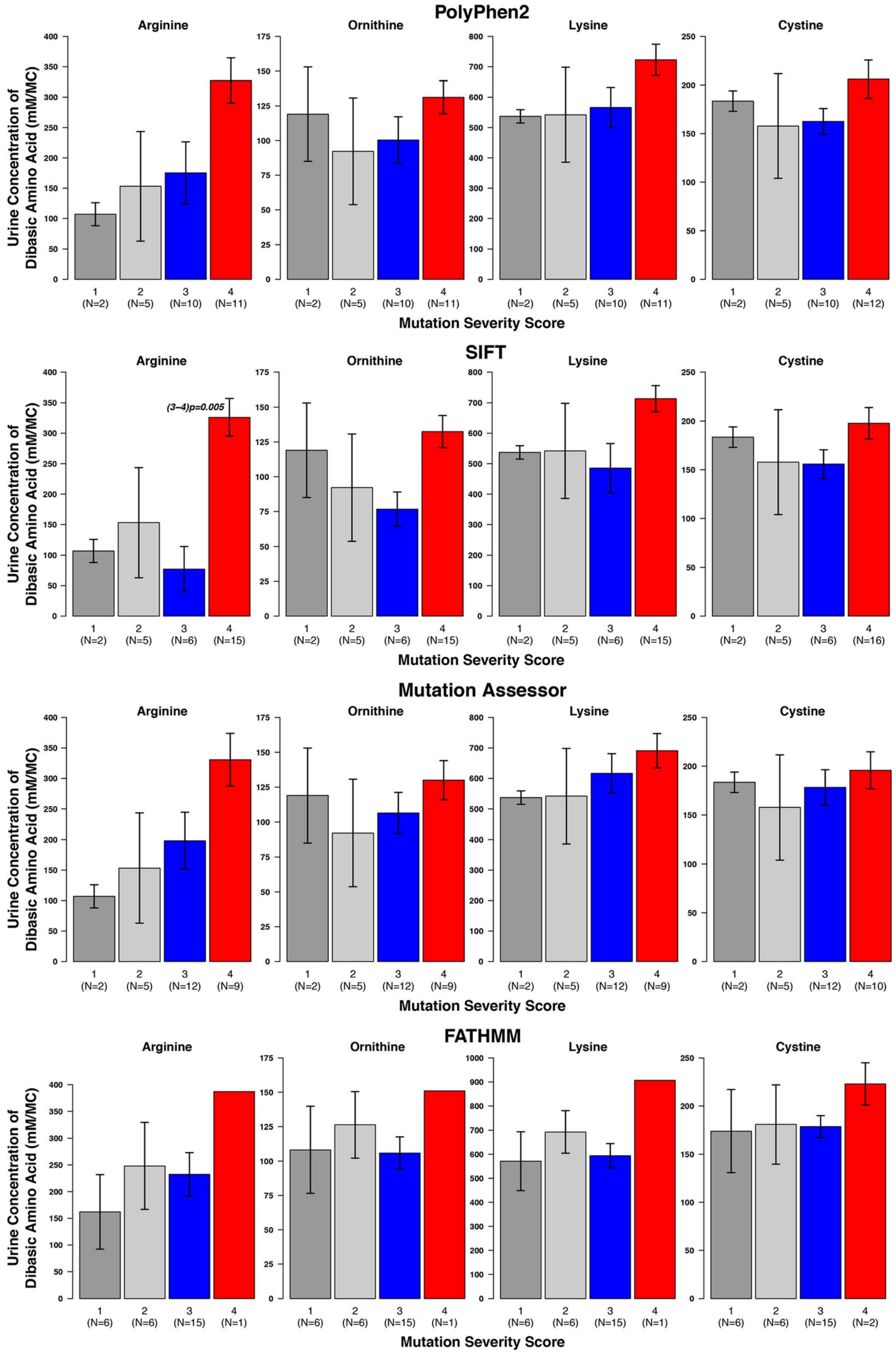


Figure 2.5: Comparison of average urine levels of amino acids between the different severity score groups, for individuals with b(o+)AT mutations. There is one plot per prediction method (PolyPhen2, SIFT, MutationAssessor, and FATHMM). The group numbers are given at the bottom of the plots, with the sample number given in brackets underneath the group name. Where significant differences between groups occur ($p < 0.05$) the *p*-value is displayed on the plot, e.g. (1–2)*p*=0.001 means a significant difference between groups 1 and 2.

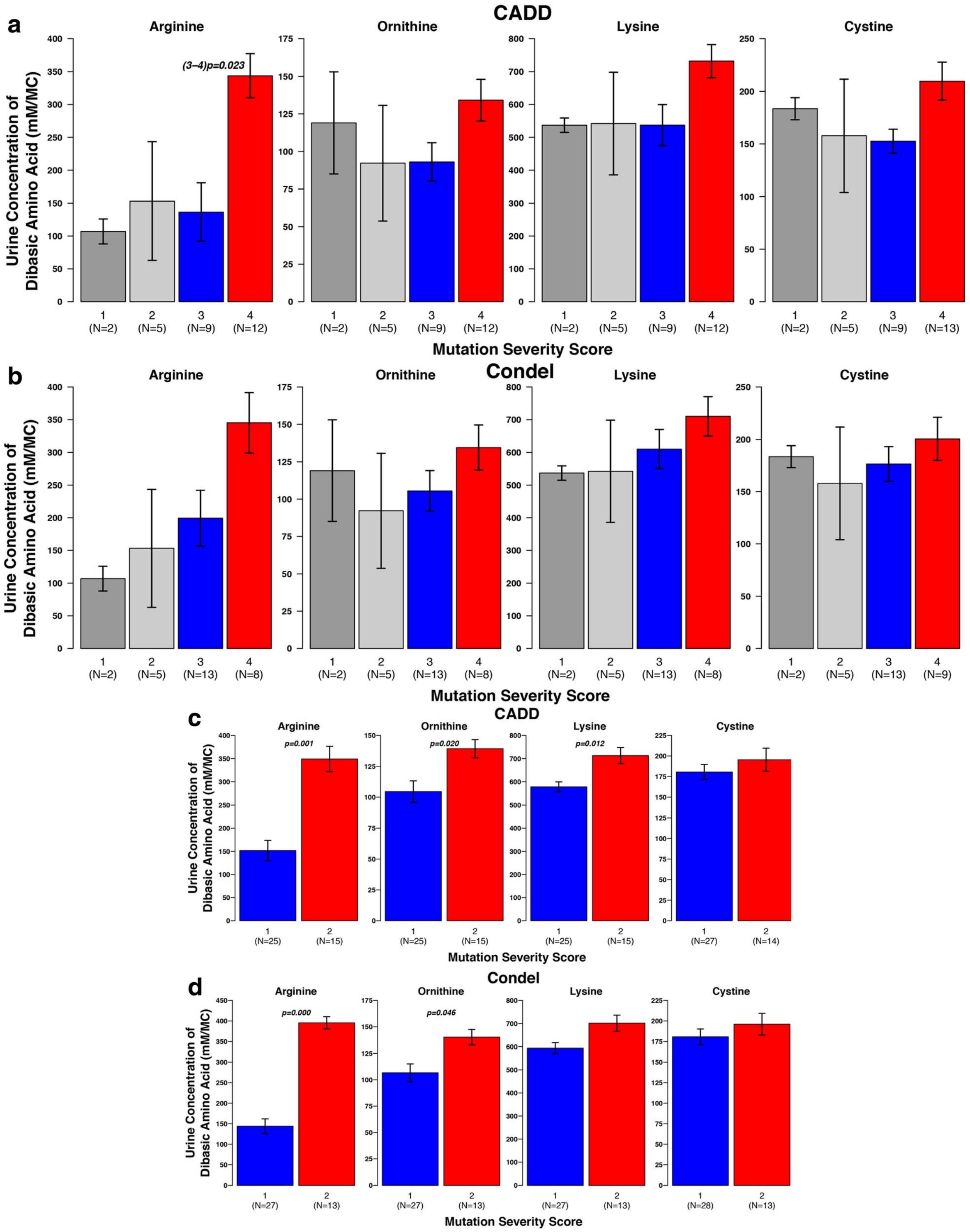


Figure 2.6: Comparison of average urine levels of amino acids between the different severity score groups using Condel and CADD. **A)** and **B)** Individuals with b(o+)AT mutations. **C)** and **D)** Individuals with rBAT mutations. There is one plot per integrated prediction method (CADD and Condel) for each gene. The group numbers are given at the bottom of the plots, with the sample number given in brackets underneath the group name. Where significant differences between groups occur ($p < 0.05$) the p-value is displayed on the plot, e.g. (1–2)p=0.001 means a significant difference between groups 1 and 2.

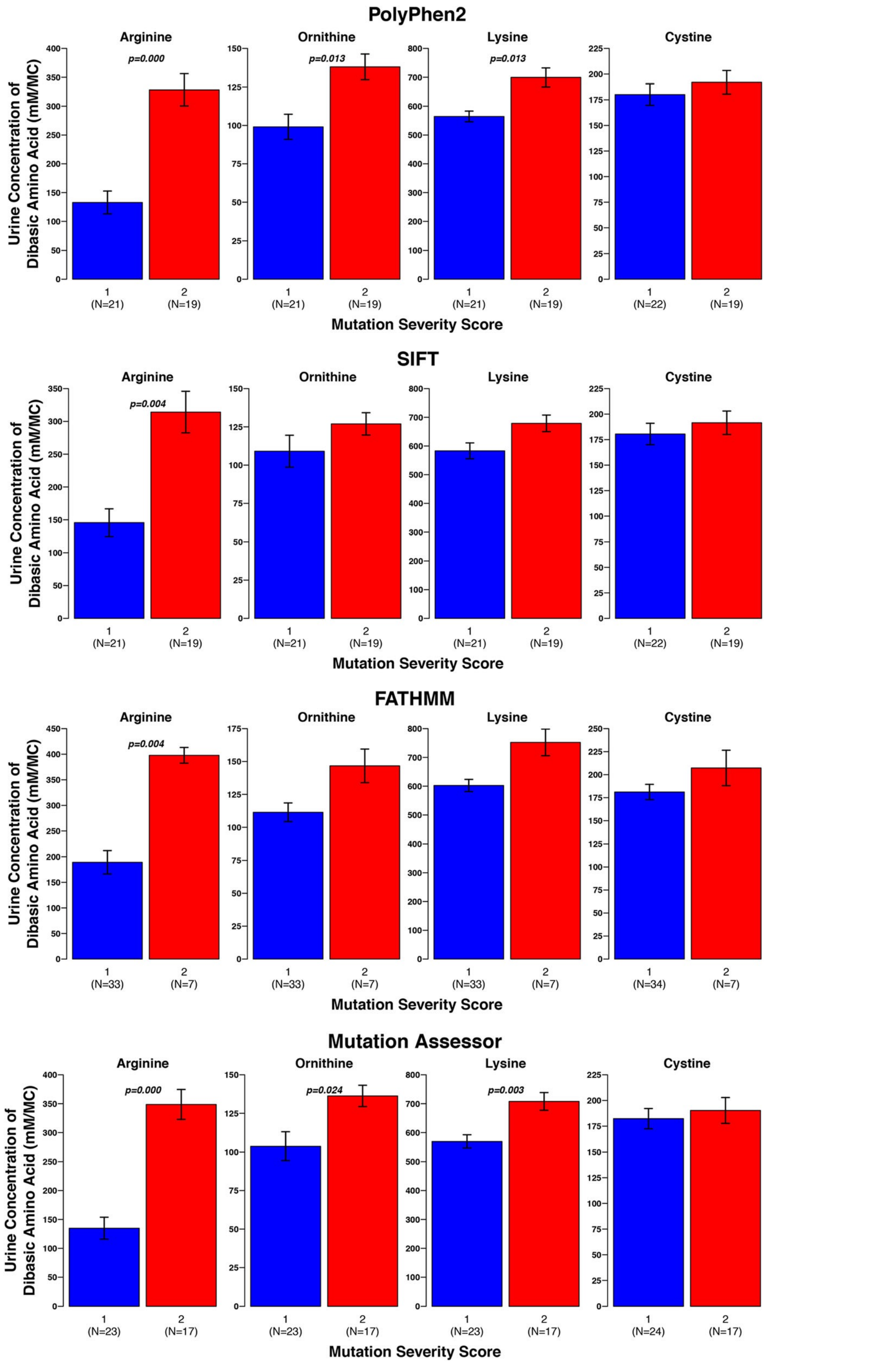


Figure 2.7: Comparison of average urine levels of amino acids between the different severity score groups, for individuals with rBAT mutations. There is one plot per prediction method (PolyPhen2, SIFT, MutationAssessor, and FATHMM). The group numbers are given at the bottom of the plots, with the sample number given in brackets underneath the group name. Where significant differences between groups occur ($p < 0.05$) the p-value is displayed on the plot, e.g. (1–2)p=0.001 means a significant difference between groups 1 and 2.

2.5 Discussion

We have surveyed the mutations present in SLC7A9 and SLC3A1 and their likely effect on the encoded proteins b(0+)AT and rBAT. Across 49 studies, 58 and 94 cystinuria associated point mutations were identified in SLC7A9 and SLC3A1, respectively. Our initial comparison of cystinuria associated variants with variants present only in ExAC, showed that the disease associated variants typically have a lower frequency in the population, they tend to cluster in the protein sequence, largely in different areas of the protein sequence to the ExAC variants which show less clustering. This may suggest that particular regions of the protein sequence cannot be altered without affecting protein function. Additionally, we found that the frequency of ExAC variants was higher for SLC3A1, which may represent the autosomal recessive inheritance of cystinuria when caused by SLC3A1 mutations. Using structural models to investigate mutations in rBAT was more complicated than for b(0+)AT because the function of rBAT is not clearly understood.

Interestingly in rBAT, few mutations directly affected the predicted functional residues that would be associated with the enzyme activity that is typically present in this family of proteins. However, a large number of mutations either remove hydrogen bonds or introduce a charge into a buried or hydrophobic region and could therefore disrupt protein folding or reduce stability. This is consistent with what we know of the rBAT protein; experimental studies suggest that the heavy rBAT subunit is essential for cell surface expression of b(0+)AT and essential for transport of the heterodimer to the plasma membrane (Fernández et al. 2002). The extracellular glycosidase domain may only have a role in cystine transport (Lundgren et al. 1995) and the requirement of chaperone for rBAT to fold correctly. A number of rBAT mutations have been linked with incorrect folding of the protein and/or trafficking to the plasma membrane (Franca et al. 2005).

In contrast, it is known that the light chain b(0+)AT encoded by SLC7A9 forms the exchanger of dibasic amino acids for neutral amino acids (Bartoccioni et al. 2008). In fact, it has been suggested that the light subunit may be fully functional even in the absence of the heavy subunit (Feliubadaló et al. 1999; Chairoungdua et al. 1999; Pfeiffer et al. 1999; Mizoguchi et al. 2001). Given the important functional role of b(0+)AT in amino acid transport, multiple cystinuria associated mutations are identified that affect or are close to predicted functional residues. Additionally, other mutations either seem likely to result in conformational changes or affect protein stability. For example, there are many examples where a buried hydrophobic amino acid is replaced by a charged or polar one but in contrast to rBAT there are few mutations that remove hydrogen bonding, which is likely because b(0+)AT is highly hydrophobic.

Comparison between the different variant effect predictors indicated that they agree for approximately 30–35% of point mutations. The investigation of using these methods to classify patients' disease into mild and severe, has a number of limitations. The methods used have been developed to predict amino acid changes that are likely to cause disease and we see that they do this fairly well for the mutations considered, with most of them predicted to be deleterious. So, we have not used them here for exactly the role they were developed. The scoring system may be overly simple, a patient with two low severity mutations will perhaps fair better than an individual with one severe mutation (or vice versa), but they are treated equally in our scoring system. Additionally, the sample size is relatively small. Finally, the complex inheritance patterns of b(0+)AT makes predictions of mutation effect harder. For rBAT, the pattern is clearer with high severity score groups tending to worse phenotypes (see Figures 2.6.C&D, 2.7, & Appendix 1 Figures 4 & 5).

However, given these limitations the analysis suggests the potential for the use of such methods in this way. Typically, we observed the phenotype differences that would be expected, if those predicted in the lower severity score group actually had a milder form of the disease. For example, the urine levels of nearly all of the amino acids considered (for most of the methods) across both proteins, are lower for the lower severity score group (but not all are statistically significant).

Additionally, for rBAT and b(0+)AT there are general trends for most methods where the age of presentation is higher in the low severity score groups, while the number of stone episodes and number of interventions is greater in the high severity score groups, but this is mostly not statistically significant, which highlights the need to use a larger cohort.

Overall these results are promising. Methods used to predict if mutations are deleterious have been used to categorise mutations and there is some correlation with phenotype. However, it also highlights the limitations of existing methods and improvements are required if they are to be used for even relatively simple precision medicine applications such as the classification of cystinuria disease severity. Additionally, the analyses performed here need to be expanded into a larger cohort of individuals to obtain greater confidence and to identify the most effective way to categorise individuals. With a larger dataset there would be the potential to train a method specifically to classify individuals based on their SLC3A1 or SLC7A9 mutations, which may be more effective than trying to use existing methods that have not been designed specifically to do this. Following this there is the potential to investigate the use of such an approach to provide individual precision treatment in the clinic.

Chapter 3: Identifying Coevolution Within the Human Genome

This paper is entitled: “Identifying Coevolution Within the Human Genome”. This work is currently being prepared for submission.

My contribution to the work was as follows:

1. Helped devise the project with Mark Wass
2. Wrote all of the scripts used in the project and processed all of the data
3. Produced all figures and tables
4. Wrote the manuscript with Mark Wass

3.1 Abstract

Advances in sequencing technologies have enabled large-scale sequencing projects to generate thousands of human genomes, and made it feasible to begin integrating whole genome sequencing in to healthcare in order to deliver precision medicine. One of the key remaining challenges in interpreting a patient’s genome is to identify the small subset of functional or deleterious genetic variants from the large number of variants that naturally occur in healthy humans. Many computational methods have been developed to predict the functional effects of variants, but little progress has been made in predicting the combined effects of multiple variants. In this study we have quantified variant combinations within individual human genomes at the protein level, using 2,504 samples from the 1,000 Genomes Project, and identified both common and rare examples of variant combinations. Some variants within these combinations have potential compensatory effects, suggesting possible coevolutionary relationships between individual residues.

Across the 2,504 samples, 280,329 unique protein-wide non-synonymous variant combinations were identified, with an average of 2,429.71 combinations per human genome (with considerable variation between human populations). A subset of 4,365 combinations occur within close spatial proximity in the structures of the proteins, 42.4% more combinations than expected based on random distribution of the variants in the structures, with an average of 121.73 combinations per human genome. Many of the variant combinations identified as close in space have potentially compensatory effects, with 571 examples of direct amino acid compensation (an amino acid lost in one variant but gained in another variant). Based on the properties of the amino acids in combinations, 474 combinations have potential compensation of amino acid charge, 717 combinations potentially conserve functional groups, and 2,674 are predicted to preserve structural stability. Finally, we identified variant combinations in protein-protein interface sites, both within partners and between partners. A total of 1,246 unique non-synonymous interface variant combinations were found across the genome set, with an average of 25.12 combinations per human genome. Within these, 127 have potential direct amino acid compensation, 96 have potential charge compensation, and 125 have potential functional group compensation.

This work highlights the extent of variant combinations within individual human genomes, with subsets of these representing coevolutionary relationships between protein residues. In the future, variants must not be considered in isolation if true precision medicine is to be achieved. A deleterious variant occurring on its own may have a neutral effect when combined with other variants, and vice versa. Therefore, relationships between variants must be considered in order to interpret their gestalt effect.

3.2 Introduction

Proteins exist on a precipice, delicately balanced between functional and non-functional. Here, the effect of a single amino acid change can be the difference between health and disease. In recent years, thanks to advances in sequencing technologies, there has been a deluge of human genetic variation data (Sherry 2001; Auton et al. 2015; Lek et al. 2016; Pagani et al. 2016; Telenti et al. 2016; Karczewski et al. 2017; Maretty et al. 2017; Landrum et al. 2018). This has provided unprecedented insight in to sequence variation within human populations, but one of the main challenges has been analysing such a large amount of data and distinguishing between variants that are benign and those that cause disease.

Experimental characterisation of the effects of each variant is impractical, and so a large number of computational tools have been developed to tackle this problem (Pappalardo & Wass 2014; Adzhubei et al. 2010; González-Pérez & López-Bigas 2011; Kumar et al. 2009; Kircher et al. 2014; Reva et al. 2011). These tools primarily rely on first principles: using information about protein function, structure, evolutionary conservation, and the physiochemical properties of the wild type and variant amino acids. For simple cases, almost always for protein coding variants, this approach has had some success, with many tools able to distinguish damaging from neutral variants with reasonable accuracy (Daneshjou et al. 2017; Hoskins et al. 2017). However, many commonly used tools have been shown to be less accurate in predicting the effects of de novo variants, frequently predicting neutral variants to be damaging, as these tools are trained on disease and common variants (Miosge et al. 2015).

One thing that most of these tools do not take in to account is the context of other variants. Individual human genomes typically contain between 4.1-5 million variants compared to the reference genome, including 10,000-12,000 that alter the encoded protein sequences (1000 Genomes Project Consortium 2015), in many cases resulting in multiple variants within the same protein. The effect of each variant is context-dependent on the other variants in the protein, in the same way that the effect of mutating one amino acid to another is context-dependent on the sequence location of the amino acid. This phenomenon has long been observed between species, and such context-dependent variants are termed Compensated Pathogenic Deviations (CPDs). These are sequence positions between homologous proteins where the wild type amino acid from one species can cause disease in the other species. This is often due to the fixation of multiple variants in a population, where the individual variants are deleterious but the combination of variants is neutral or advantageous (Kimura 1985).

For example, in human β-haemoglobin the wild type amino acid at position 20 is Val, and the variant Val20Glu is pathogenic (rs33918474). However, in horse β-haemoglobin the wild type amino acid at position 20 is Glu. This discrepancy is attributed to a second difference between the human and horse sequences at position 69, where the wild type amino acids are Glycine and Histidine, respectively (Kondrashov et al. 2002). In the horse β-haemoglobin, Glu20 and His69 can form a hydrogen bond, and in human β-haemoglobin Val20 and Gly69 can form a van der Waals interaction. These pairs of residues form what are known as a Dobzhansky-Muller incompatibility – one amino acid is not present without the other in a species, due to the reduction in fitness caused by having only one (Kondrashov et al. 2002). A generalised example of a Dobzhansky-Muller incompatibility is presented in Figure 3.1.

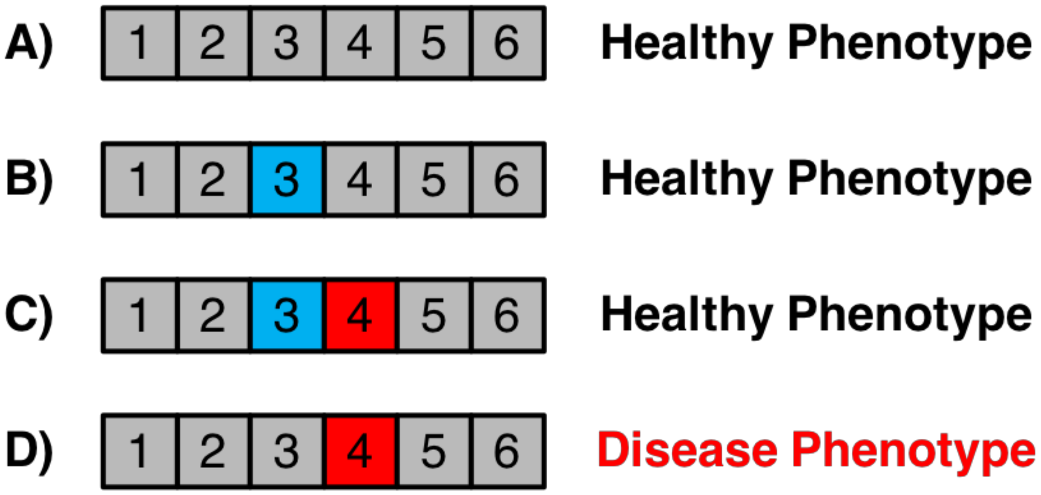


Figure 3.1: (see legend on next page)

Figure 3.1: A generalised example of a Dobzhansky-Muller incompatibility within a protein. **A)** The wild type sequence of a protein, which results in a healthy phenotype **B)** The same protein with a variant at position 3, resulting in a healthy phenotype too. **C)** The same protein with variants at positions 3 and 4, also resulting in a healthy phenotype. **D)** The same protein with a single variant at position 4, which results in a disease phenotype. This disease-phenotype occurs because this variant at position 4 can only be tolerated in conjunction with the second variant at position 3.

The possible mechanisms of such compensations are manifold, but the outcome is the same – a variant in one individual will have different effects compared to the same variant in another individual. Comparison of human disease variants in HumVar and ClinVar with orthologous sequences from 100 vertebrates estimated that 3-12% of human disease variants are found in the sequences of other species and are possible CPDs (Jordan et al. 2015). Experimental analysis of the potential CPDs identified in the genes BBS4, RPGRIP1L, and BTG2 showed that CPDs were able to rescue pathogenic variants and restore wild type function (Jordan et al. 2015). Other examples of compensatory variants with experimental characterisation include mutually compensatory variants in human p53 that preserve protein stability (Mateu & Fersht 1999), and compensatory variants that preserve dimerization stability of the human immunodeficiency virus capsid protein (Del Álamo & Mateu 2005). In both examples there are combinations of variants where the destabilising or damaging effects of individual variants are non-additive, with the full variant combinations better preserving the properties of the wild type protein.

The concept of CPDs has also been applied between individuals of the same species, and is often referred to as “genetic buffering” or “genetic compensation”, where the deleterious effect of one variant is counteracted by the effects of one or more other variant(s). From a large population, Chen *et al.* recently identified individuals carrying variants for severe childhood Mendelian diseases who have reached adulthood with no clinical manifestation of the disease (Chen et al. 2016). Multiple other studies have shown that variants thought to cause specific Mendelian diseases are far more common in human populations than would be expected based on the prevalence of the disease (Minikel et al. 2016; Xue et al. 2012; Piton et al. 2013; Bell et al. 2011). This suggests that some individuals with these disease-causing variants are protected from the diseases by other variants, or that these variants have been previously misclassified as disease causing. Such studies have led to an active search for individuals with disease buffering variants, such as the resilience project (Chen et al. 2016), in the hope of informing future treatment options for specific diseases.

Multiple studies have investigated the prevalence of CPDs between species (Kondrashov et al. 2002; Gao & Zhang 2003; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007; Azevedo et al. 2009; Jordan et al. 2015), and incorporation of CPD information in to algorithms used to predict variant effects has even been suggested as a way of improving the predictive power of such tools (Azevedo et al. 2017). While some work has been done to predict the combined effects of multiple variants (Hopf et al. 2017), no study to our knowledge has looked at the extent of possible compensation events within individual human genomes.

The aim of our study was to quantify the extent of possible epistatically interacting variants within protein coding regions of the human genome. To do this we utilised data from Phase 3 of the 1,000 Genomes Project, generating individual variant sets for each of the 2,504 samples in the study. We then analysed the extent of co-occurring variants within individual proteins, how co-occurring variants are distributed within the 3-dimensional structures of the proteins, and how co-occurring variants affect protein-protein interface sites. The analysis of the combined effects of variant combinations will be essential in the future to better understand complex genotypes, and deliver the promise of precision medicine.

3.3 Methods

3.3.1 Protein Sequence Data

The UniProt human proteome set (UP000005640) was downloaded from the UniProt website, this version was last modified on 22/11/2015 (Bateman et al. 2015). All human ensembl protein sequences were downloaded from the ensembl database, version 84, using the ensembl API (Yates et al. 2016).

3.3.2 Variation Data

The 1,000 Genomes Project phase 3 call set (02/05/2013 release), containing variant calls for 2,504 individuals, was downloaded (Auton et al. 2015). These samples are taken from five different super populations, which are made up of 26 different individual populations, see Appendix 2 Tables 1&2.

The 1,000 Genomes Project phase 3 call set was filtered to contain only single nucleotide variants (SNVs) using VCFtools, version 0.1.13 (Danecek et al. 2011). All variant calls were annotated using ANNOVAR, version 12/11/2014 (Wang et al. 2010). The human genome build version used was GRCh37/hg19, and ANNOVAR was run using the “ensGene” protocol option to give annotations with ensembl identifiers. Where ANNOVAR identified multiple protein isoforms for a single variant only one isoform was used. Canonical isoforms, as identified by UniProt (Bateman et al. 2015), were used where possible. If the canonical isoform was unknown, isoforms present in the Interactome3D data set were chosen first, otherwise the isoform was chosen numerically by its ensembl protein identifier.

ANNOVAR annotates nsSNVs with reference to ensembl protein sequences. Corresponding UniProt and ensembl protein sequences (as identified by UniProt) were aligned using NEEDLE, version 6.6.0 (Rice et al. 2000). All protein sequence positions given in this study are in reference to the UniProt sequence, and variants that could not be mapped to UniProt sequence positions were excluded. ANNOVAR annotates each variant individually and does not consider multinucleotide variants. For each sample in the 1,000 genomes cohort multinucleotide variants were considered. If a sample had two variants in the same codon, both variants were interpreted together as a single change instead of two individual changes.

GnomAD data was used to complement the 1,000 Genomes Project data, by providing the allele frequencies of variants for a much larger population (123,136 individual exomes). However, the GnomAD data is not suitable as a replacement for the 1,000 Genomes Project data, despite many 1,000 Genomes Project samples being included in GnomAD, because the GnomAD data is aggregated. This means that overall occurrences of variants from across the samples are provided, but not the variants for specific individual samples. Therefore, variant combinations for individual samples cannot be calculated. The GnomAD exome VCF dataset was downloaded on 26/04/2017, and the GnomAD allele frequencies of all variants from the 1,000 Genomes Project samples were extracted (Lek et al. 2016; Karczewski et al. 2017).

The Humsavar (Bateman et al. 2015) and ClinVar (Landrum et al. 2018) databases were used to annotate variants with reported clinical significance information. Humsavar (release: July 2018) contains 77,936 variants, and ClinVar (release: August 2018) contains 881,017 variants. For Humsavar, the annotations for each variant are ‘Disease’, ‘Polymorphism’, and ‘Unclassified’. ClinVar provides many more types of annotation than Humsavar, but for simplicity ‘Benign/Likely benign’, ‘Likely benign’ and ‘Benign’ were grouped as ‘Benign’, and ‘Pathogenic’, ‘Likely pathogenic’, and ‘Pathogenic/Likely pathogenic’ were grouped as ‘Pathogenic’. Humsavar annotations are also referred to by these categories to allow for easier comparisons to ClinVar, with Humsavar ‘Disease’ referred to as ‘Pathogenic’, and Humsavar ‘Polymorphism’ referred to as ‘Benign’. For combinations of variants, the clinical annotations were then compared for the constituent variants and a combined annotation is produced for the combination, e.g. a combination with one or more pathogenic variants in a combination with one or more benign variants is annotated as “Pathogenic & Benign”, or a combination with only pathogenic variants is annotated as “Pathogenic”.

3.3.3 Protein Structure Data

All experimental protein structures were downloaded from the Protein Data Bank (PDB), and the Structure Integration with Function, Taxonomy and Sequence (SIFTS) database was used to map the residues in the protein structures to the residues in the protein sequences in the UniProt human proteome set and also to obtain the resolution and protein sequence coverage of each structure (Berman et al. 2000; Velankar et al. 2013). Corresponding UniProt proteins and PDB structures were then aligned using MUSCLE to generate a sequence-structure position map, and calculate the total number of positions in each protein covered by each structure (Edgar 2004).

Protein structures were then selected for each protein, with structures first prioritised by resolution, and if two structures for overlapping sequence regions both had resolutions <3.5Å the structure with the largest sequence coverage was selected. After this structure selection step, all sequence positions covered by an experimental structure have been assigned to the highest quality structure available. For the remaining proteins without experimental structures, and for proteins with partial experimental structures where there are contiguous regions of ≥50 residues without a corresponding experimental structure, structural modelling was performed.

For each protein requiring structural modelling, template structures were identified from a 70% identity representative set of the PDB using hhblits, selecting only high-confidence templates with an hhblits probability ≥80 (Berman et al. 2000; Remmert et al. 2012). Models were selected to cover as many sequence positions as possible in a non-redundant manner, starting from the highest confidence model and selecting additional models that cover ≥50 sequence positions not already covered. Models were then generated for each sequence using the identified template and a protocol based on Phyre2 (Bennett-Lovsey et al. 2008; Kelly et al. 2015), with pulchra used to add and optimise the side chains (Rotkiewicz & Skolnick 2008).

For the set of 20,791 proteins analysed, 16,653 had >0% coverage either by experimental structures, models, or a combination of both. The remaining 4,138 proteins had no experimental structures available and no structural models were identified of adequate quality (Appendix 2 Figure 1). There are 1,609 proteins for which experimental structures cover the entire protein with no sequence gaps ≥50 contiguous residues. The mean sequence coverage for proteins covered entirely by experimental structures was 88.97%, with an average of 1.07 structures per protein (multiple structures correspond to different regions of the protein and are not overlapping).

There are 2,777 proteins with experimental structures available, but with gaps ≥50 contiguous residues. The mean sequence coverage was 38.79%, with an average of 1.19 structures per protein. Of these 2,777 proteins, high-quality templates were identified for 1,653 proteins, and structural modelling was used to increase structure coverage for these proteins. The mean sequence coverage was 56.43%, with an average of 2.91 structures per protein.

For the remaining 16,405 proteins without any experimental structures available, a high-quality template for structural modelling was identified for 12,267 proteins. The mean sequence coverage was 56.33%, with an average of 1.51 structures per protein. Overall, for the UniProt proteome set, combining multiple non-overlapping structures per protein and combining structural modelling with experimental structures greatly increased protein sequence coverage (Appendix 2 Figures 2-5).

3.3.4 Identification of Variant Combinations Within Proteins

For each of the 2,504 samples in the 1,000 Genomes Project data set, combinations of variants were identified within proteins for each individual sample. Each sample can have between zero and two variant combinations per protein, one combination of variants per allele. These variant combinations therefore consist of all of the variants that each sample has per copy of the protein, and they can be distributed anywhere within the 3-dimensional structure of the protein. These are termed Global Combinations (Figure 3.2.A&B).

A second set of variant combinations was then created for variants that are close in 3-dimensional space within protein structures. These are termed Proximal Combinations (Figure 3.2.A-C). All variants in Global Combinations were mapped to structures, and the distances between variants mapped to the same structures were calculated. Variants within 5Å of one another were then grouped in to new combinations. For example, in Figure 3.2.A, Global Combination One contains four variants, of which two are close in space, and these two variants become Proximal Combination One. In Figure 3.2.B, Global Combination Two contains five variants, and two separate Proximal Combinations are derived from this. Proximal Combination Two, contains two variants, and Proximal Combination Three contains three variants. Proximal Combinations Two and Three occur in the same protein, but are distant from one another in space (Figure 3.2.B).

Proximal Combination Three contains three different variants (Figure 3.2.B). The variant in the middle of the three is within 5Å of both of the other variants, but the other two variants are only within 5Å of the middle variant and not each other. This is permitted for Proximal Combinations, but the maximum distance between any two variants in a combination was limited to 15Å to prevent chaining of variants across large distances in the protein structure. For non-synonymous and synonymous variants, the vast majority of Proximal Combinations have a maximum distance between any two variants in the combination of 0-1Å (Appendix 2 Figure 6).

Importantly, the directionality of the side chains of variants in Proximal Combinations is not considered. This can result in scenarios where variants are grouped in to a combination, with potentially interacting effects, when the side chains were not directly interacting in the wild type structure. This reduces the likelihood that the grouped variants are compensating for one another. However, in these scenarios where the side chains are facing in opposite directions there can still be interaction between the variants, as having multiple variants within a localised region of a protein may alter the structure or function of the region differently to having a subset of the variants on their own.

Finally, all unique variant pairs within the Proximal Combinations were identified. For example, a Proximal Combination containing three variants contains three different constituent variant pairs (Figure 3.2.C). PyMOL (https://pymol.org/2/) was used to visualise variant combinations mapped to structures. Where described, variant combinations within proteins are given as a comma-separated list of variants, in ascending order by protein sequence position.

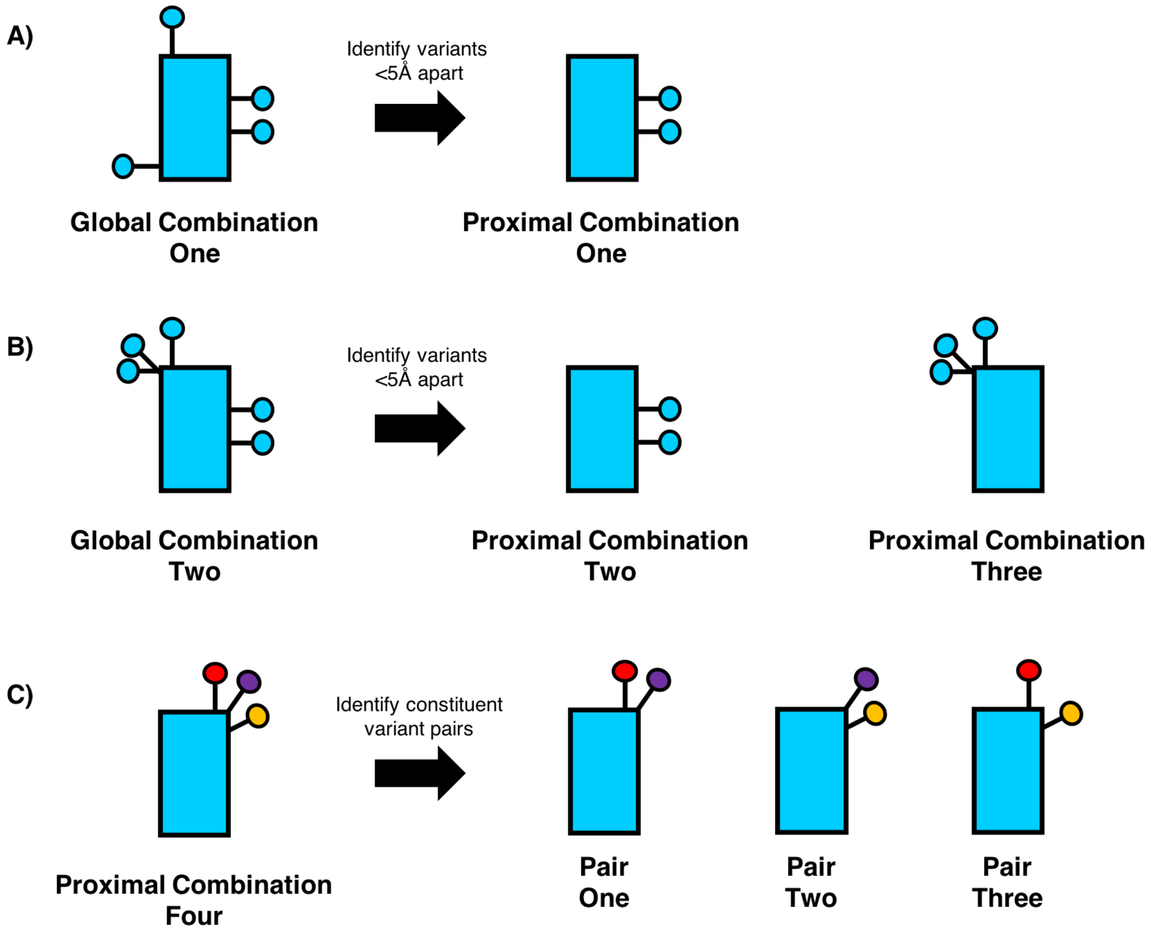

A)
Identify variants <5Å apart
Global Combination One
Proximal Combination One
B)
Identify variants <5Å apart
Global Combination Two
Proximal Combination Two
Proximal Combination Three
C)
Identify constituent variant pairs
Proximal Combination Four
Pair One
Pair Two
Pair Three

Figure 3.2: Identification of variant combinations and variant pairs within protein structures. **A)** A global combination with one constituent proximal combination of variants <5Å apart. **B)** A global combination with two distinct constituent proximal combinations in different areas of the protein. **C)** A proximal combination, which contains three individual constituent variants and three different pairs of constituent variants.

3.3.5 Generation of Random Proximal Combinations

The numbers of Proximal Combinations identified from the sample variation data were compared to Proximal Combinations identified from randomised sequence positions. The positions of variants mapped to structures from Global Combinations were randomly assigned to different positions in the proteins, and Proximal Combinations were then defined for these random combinations. For example, in Figure 3.3.A&B, the variants in Global Combinations are randomly assigned to new positions, but the total number of variants in the protein is kept the same (four variants for Global Combination One and five variants for Global Combination Two). For both non-synonymous and synonymous Global Combinations, 1,000 iterations of this process were performed, creating 1,000 sets of Proximal Combinations for variants assigned to random sequence positions.

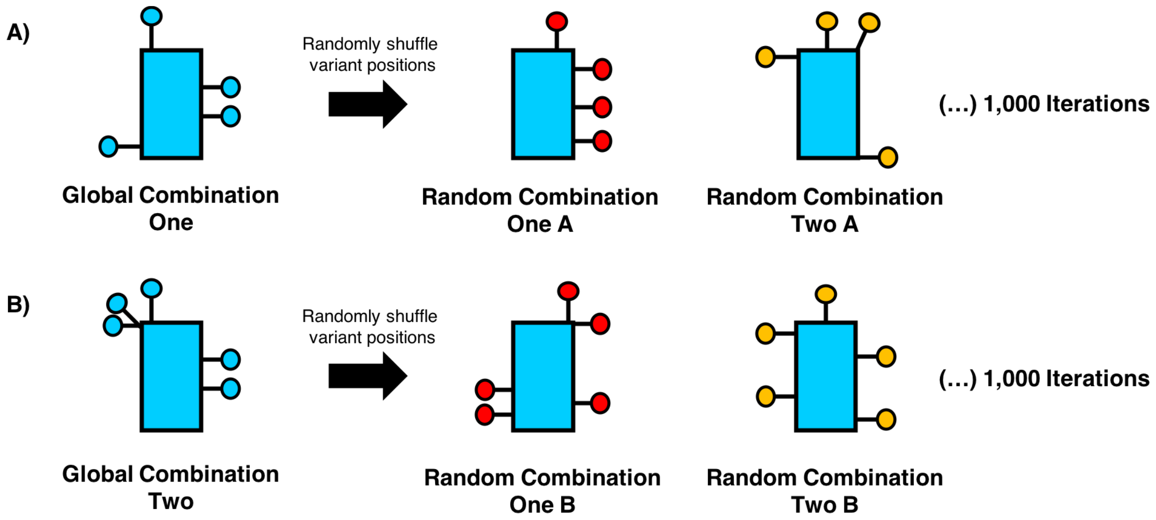


Figure 3.3: Randomisation of variant positions in global combinations. **A)** Randomisation of Global Combination One variant positions, with 2/1000 random iteration outcomes shown as examples. **B)** Randomisation of Global Combination Two variant positions, with 2/1000 random iteration outcomes shown as examples.

3.3.6 Variant Property Compensations

Amino acid properties were assigned using the definitions of Innis et al., and all amino acid properties used are shown in Table 3.1 (Innis et al. 2004). Foldx (Schymkowitz et al. 2005) was used to predict protein structure stability changes for variant combinations. The Foldx RepairPDB function was run for each protein structure with default parameters, and then stability predictions were made using the BuildModel command on the repaired structures with default parameters.

Table 3.1: Properties of amino acids used for finding compensatory amino acid changes

Amino Acid	Charge	Functional Group	Average Atomic Mass
A	Hydrophobic	Hydrophobic	71.08
C	Hydrophobic	Hydrophobic	103.14
D	Negative	Carboxylate	115.09
E	Negative	Carboxylate	129.12
F	Hydrophobic	Phenyl	147.18
G	Hydrophobic	Hydrophobic	57.05
H	Positive	Positive	137.14
I	Hydrophobic	Hydrophobic	113.16
K	Positive	Positive	128.17
L	Hydrophobic	Hydrophobic	113.16
M	Hydrophobic	Hydrophobic	131.19
N	Negative Polar	Amido	114.10
P	Hydrophobic	Hydrophobic	97.12
Q	Negative Polar	Amido	128.13
R	Positive	Positive	156.19
S	Positive Polar	Hydroxyl	87.08
T	Positive Polar	Hydroxyl	101.11
V	Hydrophobic	Hydrophobic	99.13
W	Hydrophobic	Phenyl	186.21
Y	Hydrophobic	Hydroxyl & Phenyl	163.18

3.3.7 Interactome3D Data Processing

The protein-protein interaction data used in this study was taken from the Interactome3D database, version 2017_01 (Mosca et al. 2013). The Interactome3D representative dataset for *Homo sapiens* was used, this contains only the highest-ranking structures and models for each interaction. All experimental structures and global templates were included, while all domain-domain modelled interactions were excluded. The final interaction data set contained 9,642 distinct complexes. Of these, 2,987 are homomeric interactions, and 6,655 are heteromeric interactions. In total, the interaction data set contained 5,500 distinct proteins.

For each protein-protein complex, any residue in one partner of the complex that was within 5Å of a residue in the other partner of the complex was classed as being an interface residue (Lensink et al. 2016). Sequence positions in the structure were mapped to their corresponding positions in the UniProt sequence using MUSCLE (Edgar 2004). This resulted in a list of protein-protein interface residues, mapped to UniProt sequence positions, for each of the 5,500 proteins in the interaction set.

3.3.8 Identification and Classification of Interface Variant Combinations

Using the processed Interactome3D data, non-synonymous and synonymous variants were classified into three different groups: interface variants, non-interface variants, and variants without interaction data. As with combinations of variants within individual proteins, for each of the 2,504 samples in the 1,000 Genomes Project, interface variants were further classified in to one of four variant combination categories: Homomeric Combinations, Heteromeric Combinations, Uni-Partner Combinations, and Singletons (Figure 3.4). These are defined as follows:

1. Homomeric Combinations - interface variants of homomeric complexes, with ≥2 interface variants in each partner of the complex (Figure 3.4.A)
2. Heteromeric Combinations - interface variants of heteromeric complexes, with ≥2 interface variants per partner (Figure 3.4.B)
3. Uni-Partner Combinations – interface variants of heteromeric complexes, with ≥2 interface variants in one partner and no variants in the interface region of the other partner. These can also occur in homomeric complexes, if the complex is asymmetric, i.e. one region of the protein interfaces with a different region of the same protein (Figure 3.4.A&B)
4. Singletons – interface variants of any type of complex, with one interface variant in one partner and no variants in the interface region of the other partner (Figure 3.4.A&B)

Where described, combinations of variants in interfaces use the following notation: <Protein A UniProt ID>:<Comma-Separated List of Variants in Protein

A>_<Protein B UniProt ID>:<Comma-Separated List of Variants in Protein B>, e.g. PA:G21S,R43K_PB:P100H,H403P meaning that, for the interaction PA:PB, G21S and R43K variants both occur in the interface site of PA, and P100H and H403P both occur in the interface site of PB.

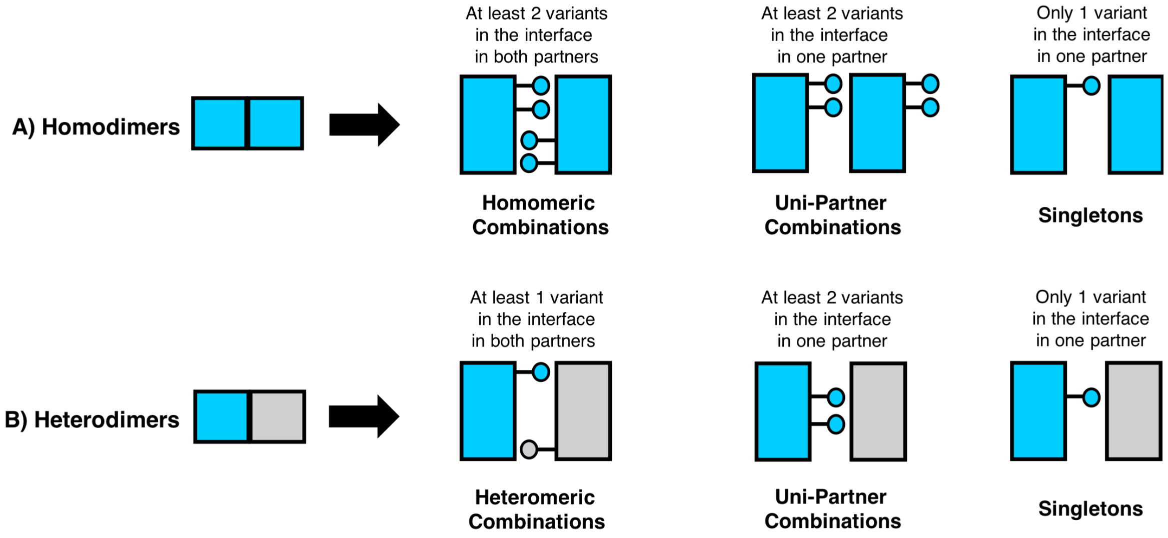

A) Homodimers
At least 2 variants in the interface in both partners
Homomeric Combinations
At least 2 variants in the interface in one partner
Uni-Partner Combinations
Only 1 variant in the interface in one partner
Singletons
B) Heterodimers
At least 1 variant in the interface in both partners
Heteromeric Combinations
At least 2 variants in the interface in one partner
Uni-Partner Combinations
Only 1 variant in the interface in one partner
Singletons

Figure 3.4: Identification of variant combinations within protein-protein interfaces. **A)** Possible interface variant combinations for homodimeric protein-protein interactions – Homomeric Combinations, Uni-Partner Combinations, and Singletons. **B)** Possible interface variant combinations for heterodimeric protein-protein interactions – Heteromeric Combinations, Uni-Partner Combinations, and Singletons.

3.4 Results

3.4.1 Unique Global Combinations

Each individual human genome has a large number of variants compared to the human reference genome, with on average 10,000-12,000 variants that alter protein sequence (1000 Genomes Project Consortium 2015). Individuals also have two distinct copies of each protein and, although there are some highly conserved genes where variants rarely occur, many proteins have multiple variants in them. This means that for 2,504 samples in the 1,000 Genomes Project data set there are a possible 5,008 distinct versions of a protein, as each sample has two alleles and each one can have a different combination of variants. Therefore, for the 20,791 canonical protein isoforms in the UniProt human proteome set there are 104,121,328 possible distinct versions of proteins (20,791 proteins multiplied by 5,008 alleles).

Across all 2,504 samples, 531,560 unique non-synonymous variants and 345,334 unique synonymous variants were identified. A large number of unique variant combinations were identified from these, but as expected there are far fewer unique versions of proteins than the theoretical maximum number, with some proteins having no observed combinations of variants in any individual (Supplementary Table 1). Variant combinations were identified in 10,753 different proteins, with a total of 280,329 non-synonymous Global Combinations, and 342,031 non-synonymous variants occurring on their own across 18,385 different proteins. For synonymous variants, there were 224,541 Global Combinations of variants spread across 11,131 different proteins, and 215,576 variants occurring on their own across 18,258 proteins.

Table 3.2: Numbers of unique combinations per variant combination category, and numbers of single occurrence variants for the total variant sets.

Combination Category	Unique Constituent Variants	Proteins with Combinations	Proteins with Singletons	Unique Combinations	Unique Singletons
Non-Synonymous Global Combinations	531,560	10,753	18,385	280,329	342,031
Non-Synonymous Proximal Combinations	5,596	1,562	N/A	4,365	N/A
Synonymous Global Combinations	345,334	11,131	18,258	224,541	215,576
Synonymous Proximal Combinations	4,615	1,678	N/A	2,558	N/A

3.4.2 Unique Proximal Combinations

When searching for variants that may interact or coevolve, proximity in 3-dimensional protein structure can be used to identify candidate combinations. This has been demonstrated by protein contact and structure prediction methods that use coevolution to identify residues that are close in space (Buslje et al. 2009; Marks et al. 2011; Morcos et al. 2011; Jeong & Kim 2012; Kamisetty et al. 2013; Ekeberg et al. 2013; Schneider & Brock 2014; Seemayer et al. 2014; Kaján et al. 2014; Jones et al. 2015; Adhikari & Cheng 2016; Ovchinnikov et al. 2017). Variants that are distant in space may also coevolve, but it is more difficult to identify such variants.

Many of the variants in the Global Combinations identified will be distant from one another within the 3-dimensional structure of the protein, and so are less likely to be interacting on a functional level, though distant interactions of variants are still possible via allosteric effects, with long-range compensatory effects of variants previously observed within proteins (Del Álamo & Mateu 2005). Here we focus on variants that are close in space within the protein structure and therefore more likely to have coevolutionary relationships. The previously identified Global

Combinations were therefore filtered in to combinations of variants close in space – Proximal Combinations (see Methods).

There are substantially fewer unique Proximal Combinations than unique Global Combinations, with 280,329 unique non-synonymous Global Combinations but only 4,365 unique non-synonymous Proximal Combinations (Table 3.2). A similar trend was observed for synonymous variants, with 224,541 unique Global Combinations and 2,558 unique Proximal Combinations.

The Global Combinations are a result of all of the variants that individuals have within copies of their proteins, many of which are distant from one another in the structure. The Proximal Combination variants occur in close spatial proximity and are confined by the physical limit of the number of residues that can be close in space within a protein structure, as well as any potential selection based on the combined effects of variants. Fewer synonymous combinations are observed for all types of combinations compared to non-synonymous combinations (Table 3.2).

Global Combinations of non-synonymous and synonymous variants are on average larger than the Proximal Combinations, containing on average 1.4 and 2.0 more variants for non-synonymous and synonymous variants, respectively (Table 3.3). The numbers of variants in Global Combinations for non-synonymous and synonymous variants can be very high, with as many as 75 variants in a single combination (Figure 3.5 and Table 3.3). The corresponding non-synonymous and synonymous Proximal Combinations, produced after reducing Global Combinations by proximity in the structure, are on average smaller, with the largest observed combination containing 14 variants (Figure 3.5 and Table 3.3).

Table 3.3: Average sizes of variant combinations for the different combination categories.

Combination Category	Mean Combination Size (variants)	Median Combination Size (variants)	Maximum Combination Size (variants)
Non-Synonymous Global Combinations	4.78	3	75
Non-Synonymous Proximal Combinations	3.38	2	14
Synonymous Global Combinations	4.07	3	68
Synonymous Proximal Combinations	2.07	2	5

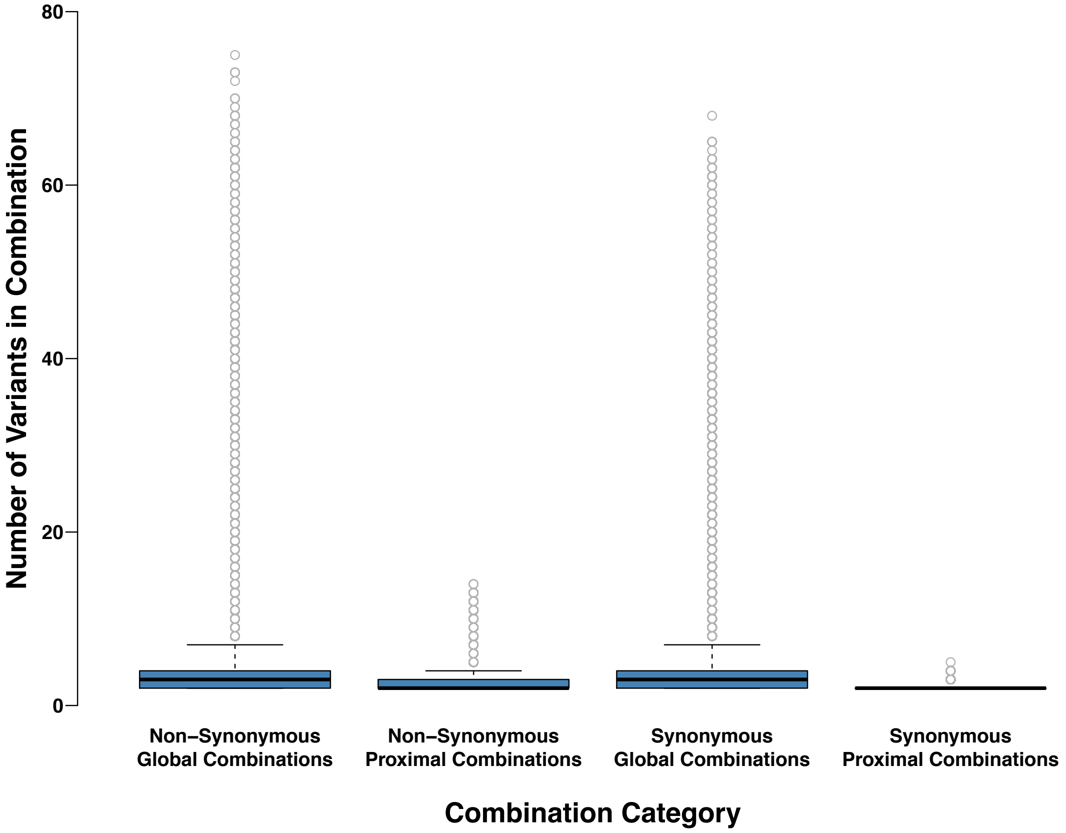


Figure 3.5: Numbers of variants per combination for each combination category.

3.4.3 Observed Proximal Combinations Compared to Random Expectation

If the observed Proximal Combinations are the result of coevolution between residues, one would expect to see more Proximal Combinations than if variants were randomly distributed throughout the protein. To test this, we performed 1,000 random iterations of variant positions in structures (see Methods), and for each iteration regrouped variants in to combinations by spatial proximity. The number of Proximal Combinations identified from the 1,000 Genomes Project sample variation data and the number of Proximal Combinations observed for 1,000 iterations of randomised structure positions were then compared, for non-synonymous and synonymous variants.

For non-synonymous variants, the number of Proximal Combinations observed in the 1,000 Genomes Project sample data is far greater than those obtained for random positioning of variants in the protein (42.4% more observed than the average from random positioning; 4,365 observed combinations, 3,064.61 mean random combinations, and 3,429 maximum random combinations). This may suggest that there is a selective pressure for the variants in the observed Proximal Combinations to be close in space (Figure 3.6.A).

This is not observed for the synonymous variants, where the observed number of Proximal Combinations falls within the distribution of the random iterations (Figure 3.6.B), though the observed number is at the high-end of the distribution (2,558 observed combinations, 2,415.93 mean random combinations, and 2,576 maximum random combinations). This is perhaps expected, as synonymous variants are unable to functionally compensate for each other in the final protein structure. Synonymous variants could however be clustering in space due to sequence proximity, where any selection pressure is likely related to binding motifs in the DNA/RNA or selection of codons for translation speed related to protein folding (Tsai et al. 2008; Zhang et al. 2009; Tuller et al. 2010; Saunders & Deane 2010; Plotkin & Kudla 2011; Gingold & Pilpel 2011; Pechmann & Frydman 2013).

The Proximal Combinations from the random iterations of non-synonymous variants are also smaller on average compared to those from the 1,000 Genomes

Project data, with a mean number of 2.22 variants per combination compared to 3.38 for the sample data (Table 3.4). The maximum sizes of combinations from random iterations were also smaller, with an average value of 7.92, almost half the size of the maximum combination size from the variant data, which was 14 (Table 3.4). For synonymous variants, the mean and maximum sizes of the Proximal Combinations from the sample data are similar to those from the random iterations (Table 3.4).

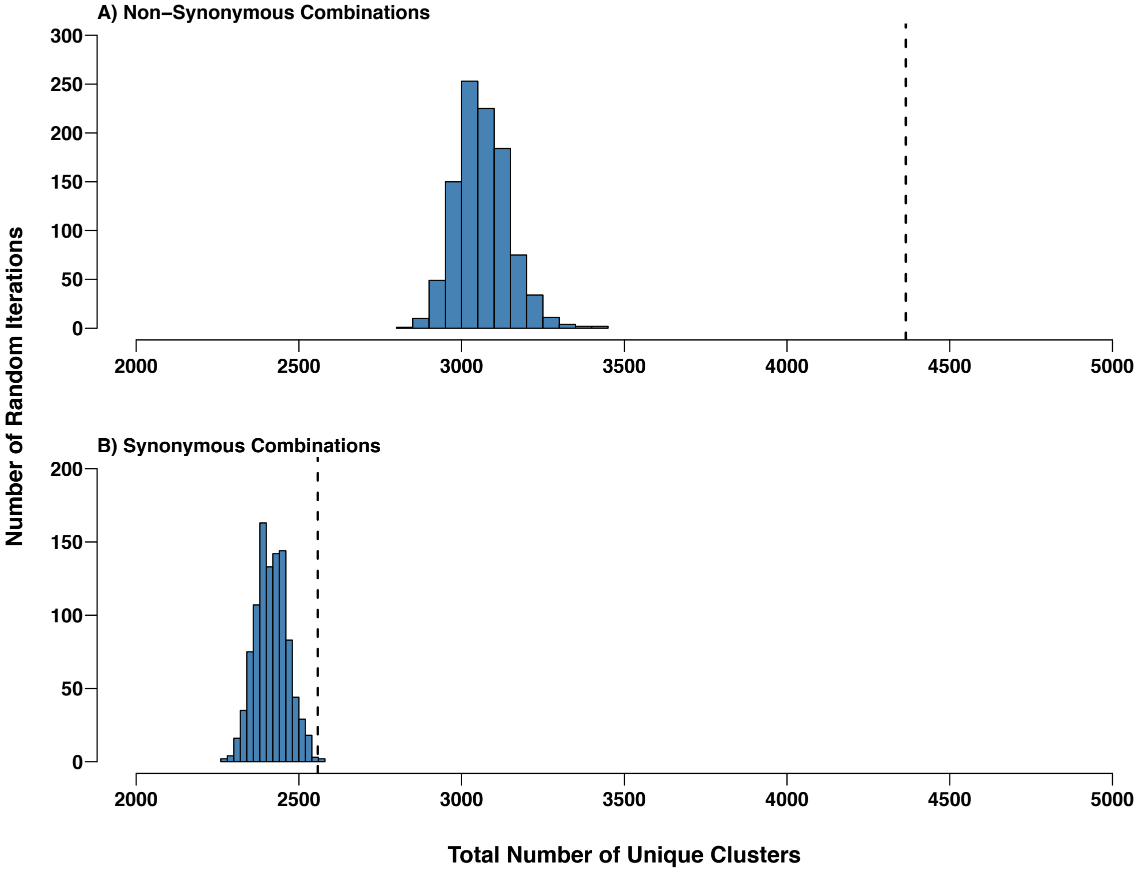


Figure 3.6: Numbers of Proximal Combinations observed from 1,000 random iterations of structure positions vs the observed number of combinations from the 1,000 Genomes Project sample data. **A)** Non-synonymous variants. **B)** Synonymous variants. The blue bars represent the distribution of Proximal Combination numbers from the random iterations. The dashed black line represents the observed number of Proximal Combinations for the 1,000 Genomes Project sample data. Note – y-axes differ in their scales between the subplots.

Table 3.4: Variant combination sizes for Proximal Combinations observed in the 1,000 Genomes Project data and the Proximal Combinations from random iterations.

Combination Category	Mean Combination Size (variants)	Median Combination Size (variants)	Maximum Combination Size (variants)
Non-Synonymous Proximal Combinations	3.38	2	14
Non-Synonymous Random Average Proximal Combinations	2.22	2	7.92
Synonymous Proximal Combinations	2.07	2	5
Synonymous Random Average Proximal Combinations	2.05	2	5.13

3.4.4 Variant Combinations Per Sample

So far, the data presented has considered the number of unique variant combinations observed across all samples in the 1,000 Genomes Project. We next considered the variant combinations of each of the 2,504 samples individually, to analyse the numbers of Global and Proximal Combinations per sample (Figure 3.7 and Table 3.5).

In contrast to the unique set of Global Combinations, each sample has more synonymous Global Combinations than non-synonymous Global Combinations (Figure 3.7, Table 3.2, and Table 3.5). Samples have on average 2,429.71 non-synonymous Global Combinations and 3,005.11 synonymous Global Combinations. There are fewer possible synonymous combinations for a protein, but at the individual sample level more synonymous combinations probably represents the fact that synonymous variants are less likely to have an effect on protein structure and function.

For the Proximal Combinations, each sample has a similar number of non-synonymous and synonymous combinations, with on average 121.73 and 116.89, respectively (Figure 3.7). Therefore, although samples have a large number of proteins with multiple variants in them (Global Combinations), they have relatively few proteins with multiple variants close in space (Proximal Combinations). The highest number of Proximal Combinations observed in a single individual was 177 for non-synonymous variants and 164 for synonymous variants, and the minimum numbers observed were 69 for non-synonymous variants and 84 for synonymous variants. A summary of the per sample variant combination numbers is shown in Table 3.5.

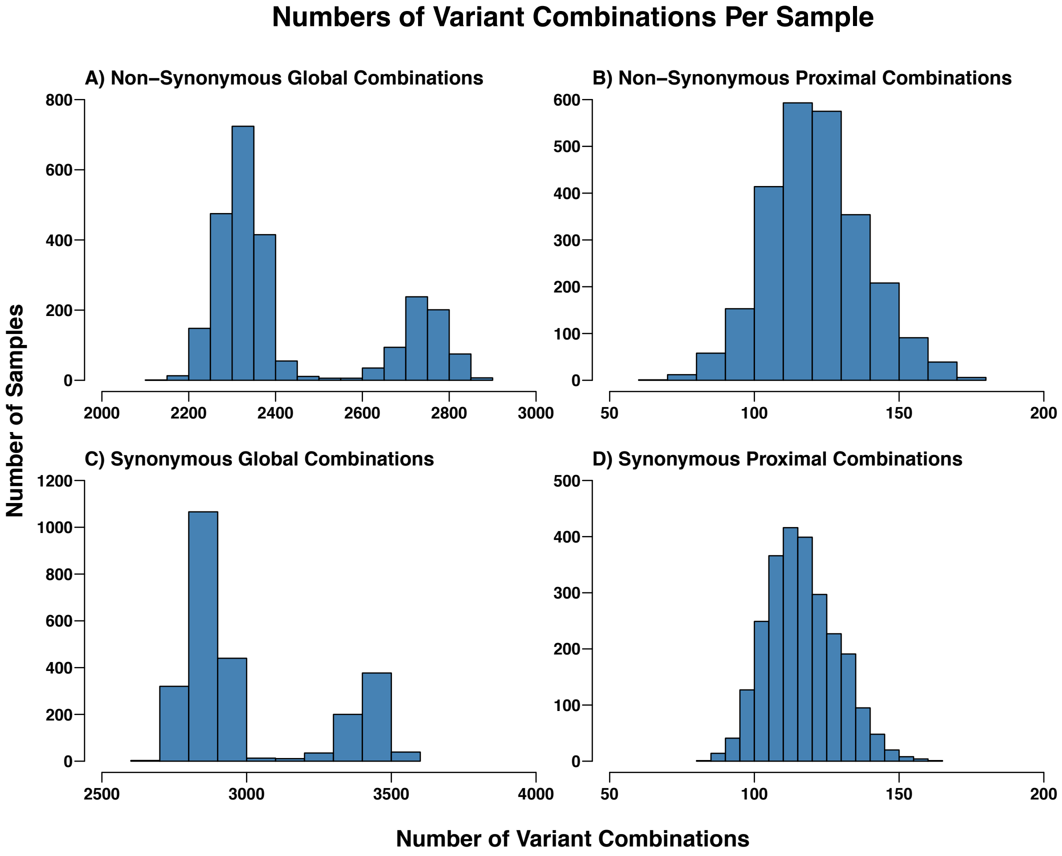


Figure 3.7: Numbers of variant combinations observed per sample for each of the combination categories. **A)** Non-Synonymous Global Combinations. **B)** Non-Synonymous Proximal Combinations. **C)** Synonymous Global Combinations. **D)** Synonymous Proximal Combinations. Note – x- and y-axes differ in their scales between the subplots.

Table 3.5: Numbers of variant combinations observed per sample.

Category	Mean Number of Combinations	Median Number of Combinations	Maximum Number of Combinations	Minimum Number of Combinations
Non-Synonymous Global Combinations Per Sample	2,429.71	2,343.5	2,882	2,115
Non-Synonymous Proximal Combinations Per Sample	121.73	121.0	177	69
Synonymous Global Combinations Per Sample	3,005.11	2,890.0	3,586	2,671
Synonymous Proximal Combinations Per Sample	116.89	116.0	164	84

The non-synonymous and synonymous Global Combinations per sample exhibit a bimodal distribution (Figure 3.7.A&C), which is not observed for the Proximal Combinations (Figure 3.7.B&D). This is caused by population-specific differences between the samples. The 1,000 Genomes Project samples are taken from five different super populations, composed of 26 different individual populations (see Methods). Samples from the African (AFR) super population have on average more non-synonymous and synonymous Global Combinations compared to the other four super populations (Figure 3.8.A&C and Table 3.6), which results in the bimodal distribution of Global Combinations per sample observed (Figure 3.7.A&C). AFR samples also have on average the most Proximal combinations, but the difference between the super populations is much smaller than for the Global Combinations (Figure 3.8.B&D and Table 3.6), resulting in the loss of the bimodal distribution for the Proximal Combinations (Figure 3.7.B&D). This perhaps reflects the fact that the human reference genome is least representative of the AFR super population (1000 Genomes Project Consortium 2015).

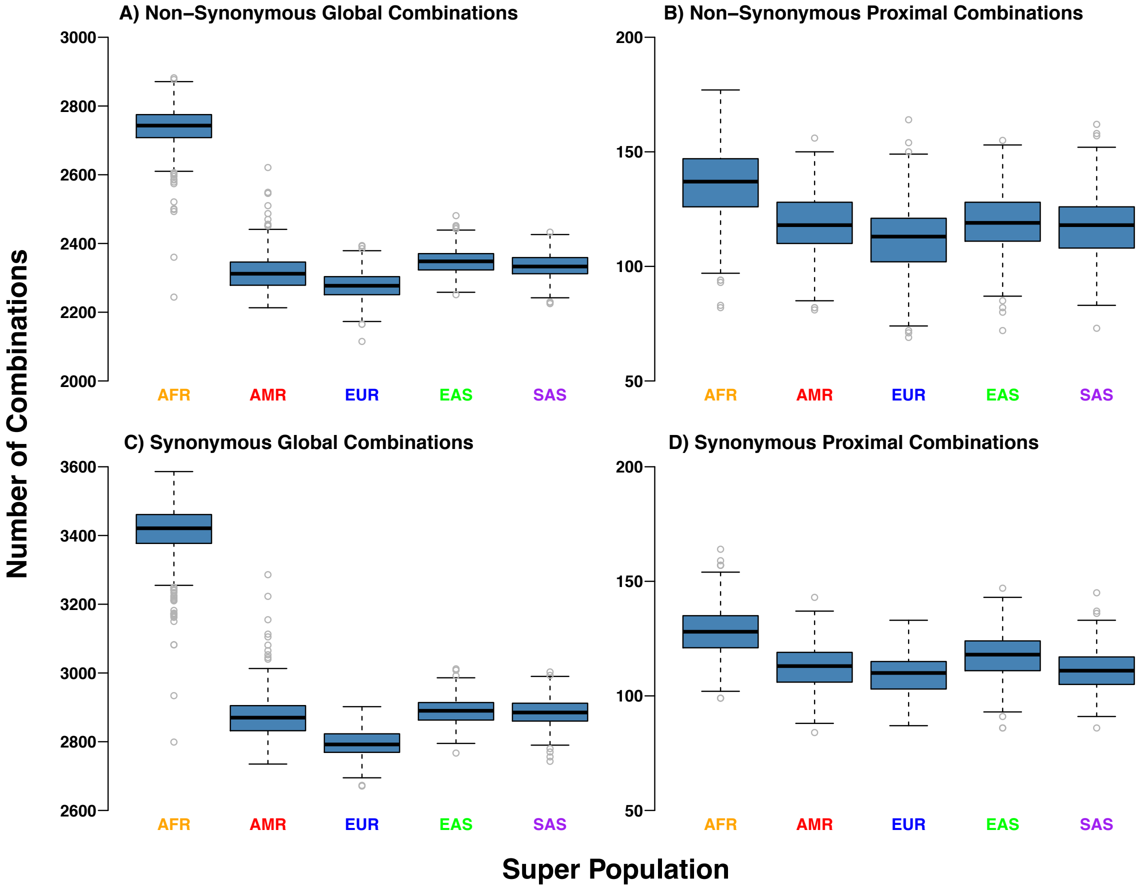


Figure 3.8: Numbers of variant combinations per sample for each of the combination categories, separated in to the five super populations. **A)** Non-Synonymous Global Combinations. **B)** Non-Synonymous Proximal Combinations. **C)** Synonymous Global Combinations. **D)** Synonymous Proximal Combinations. Note – x- and y-axes differ in their scales between the subplots. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

Table 3.6: Mean numbers of each type of variant combination per sample for each of the five super populations.

Super Population	**Non-Synonymous Global Combinations**	**Non-Synonymous Proximal Combinations**	**Synonymous Global Combinations**	**Synonymous Proximal Combinations**
African	2737.57	136.25	3410.56	128.35
American	2318.55	118.25	2875.63	113.06
European	2277.21	111.75	2793.83	109.39
East Asian	2347.46	119.24	2889.51	117.32
South Asian	2334.07	117.41	2885.43	111.36

3.4.5 Variant Combinations Per Protein

We next considered the unique variant combinations for each of the 20,791 proteins in the UniProt human proteome set (see Methods). Each protein has a distinct pattern of Global Combinations, with some proteins having large number of different combinations, and some having none or very few (Figure 3.9). For non-synonymous variants, the average number of Global Combinations per protein was 26.07, and 20.17 for synonymous variants (Table 3.7). The highest number of non-synonymous and the highest number of synonymous Global Combinations for any protein were both observed in the protein 'Protein AHNAK2' (UniProt accession: Q8IVF2), with 3,402 and 3,143, respectively (Table 3.7). This means that there are 3,402 unique non-synonymous and 3,143 unique synonymous combinations of variants for this protein that at least one individual in the 1,000 Genomes Project data set has. Protein AHNAK2 is very large, with the canonical isoform containing 5,795 amino acid residues. There were no Proximal Combinations identified for this protein, however only 93 of the 5,795 residues in the protein were covered by experimental structures, and no high-quality templates were identified for structural modelling. It seems likely though that the large number of unique combinations identified for this protein are at least in part due to its size.

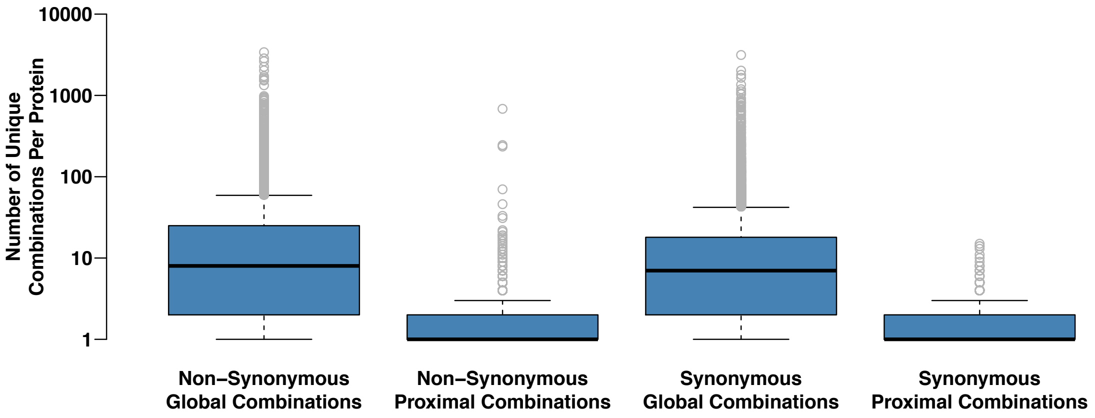


Figure 3.9: Numbers of unique variant combinations observed per protein for each of the combination categories.

Table 3.7: Numbers of unique variant combinations observed per protein.

Combination Category	Mean Number of Combinations	Median Number of Combinations	Maximum Number of Combinations
Non-Synonymous Global Combinations Per Protein	26.07	8	3,402
Non-Synonymous Proximal Combinations Per Protein	2.79	1	684
Synonymous Global Combinations Per Protein	20.17	7	3,143
Synonymous Proximal Combinations Per Protein	1.52	1	15

There are fewer unique non-synonymous and synonymous Proximal Combinations per protein compared to their Global Combination equivalents (Figure 3.9). The non-synonymous and synonymous Global Combinations sets have >20 unique variant combinations on average per protein, whereas the sets of Proximal Combinations on average have only a couple of unique variant combinations per protein (Table 3.7). There are some proteins with a large number of Proximal Combinations, but these are rarer (Figure 3.9 and Table 3.7). For example, HLA class I histocompatibility antigen, B-81 alpha chain (UniProt accession: Q31610) has 684 unique non-synonymous Proximal Combinations (1,525 Global Combinations), which is the most observed for any protein. Due to the role of this protein in the immune system, it may be expected that a large number of unique variant combinations would be observed. Similarly, Immunoglobulin kappa variable 1-5 (UniProt accession: P01602), another immune system protein, contains the most unique synonymous Proximal Combinations of any protein (3,143; Table 3.7).

One conflating factor in comparing the numbers of variant combinations observed between different proteins is the size of the individual proteins. Longer proteins have a higher random probability of being mutated, and therefore would be expected to have more observed variant combinations on average than smaller proteins. Across the dataset, the number of Global Combinations per protein is somewhat correlated with the length of the protein (r=0.54 and r=0.53 for non-synonymous and synonymous Global Combinations, respectively; Appendix 2 Figure 7.A&B). However, there is no correlation between the number of Proximal Combinations per protein and protein length (r=0.01 and r=0.20 for non-synonymous and synonymous Proximal Combinations, respectively; Appendix 2 Figure 7.C&D). The number of Proximal combinations is also not correlated with the numbers of residues covered by structures (this is distinct from protein length, this is the number of residues that could be mapped to a structure), or numbers of residues within 5Å in the protein structure (a proxy for the number of possible proximal combinations, as some structures will have conformations that have the potential for more proximal combinations, with more residues within 5Å; Appendix 2 Figure 7.E-H).

Many of the Proximal Combinations in Q31610 occur in the peptide-binding groove, a region of the protein recently reported to contain multiple pairs of residues that have coevolved in vertebrate evolution (Dib et al. 2018). Table 3.8 shows the ten most common Proximal Combinations in Q31610. Visualisation of these ten most common combinations in the structure of Q31610 shows that they all occur in or near the peptide-binding groove (Figure 3.10). Many other combinations in Q31610 also occur in this peptide-binding groove, with many combinations sharing similar residue positions and also variants. For example, 'S35A,V36M,S48A,Y91F,A93T,Q94N,A95T,D98Y,S121R,Y140F' is the eighth most common Proximal Combination in Q31610 with 176 occurrences (Figure 3.10.H), but an almost identical combination occurs 145 times with Y140S instead of Y140F.

Table 3.8: Top ten Proximal Combinations of non-synonymous variants in Q31610 by number of total occurrences.

Proximal Combination	Total Occurrences
D201E,K202T,E204Q	3,972
V127L,H137Y	416
N104I,R106L,G107R	393
L105A,R106L	341
E69K,N87E	318
E69T,Y91F	317
S101N,N104I,L105A,R106L,G107R	278
S35A,V36M,S48A,Y91F,A93T,Q94N,A95T,D98Y,S121R,Y140F	176
R106L,G107R	157
S35A,V36M,S48A,A93T,Q94N,A95T,D98Y,S101N,L105A,L119W,S121T	154

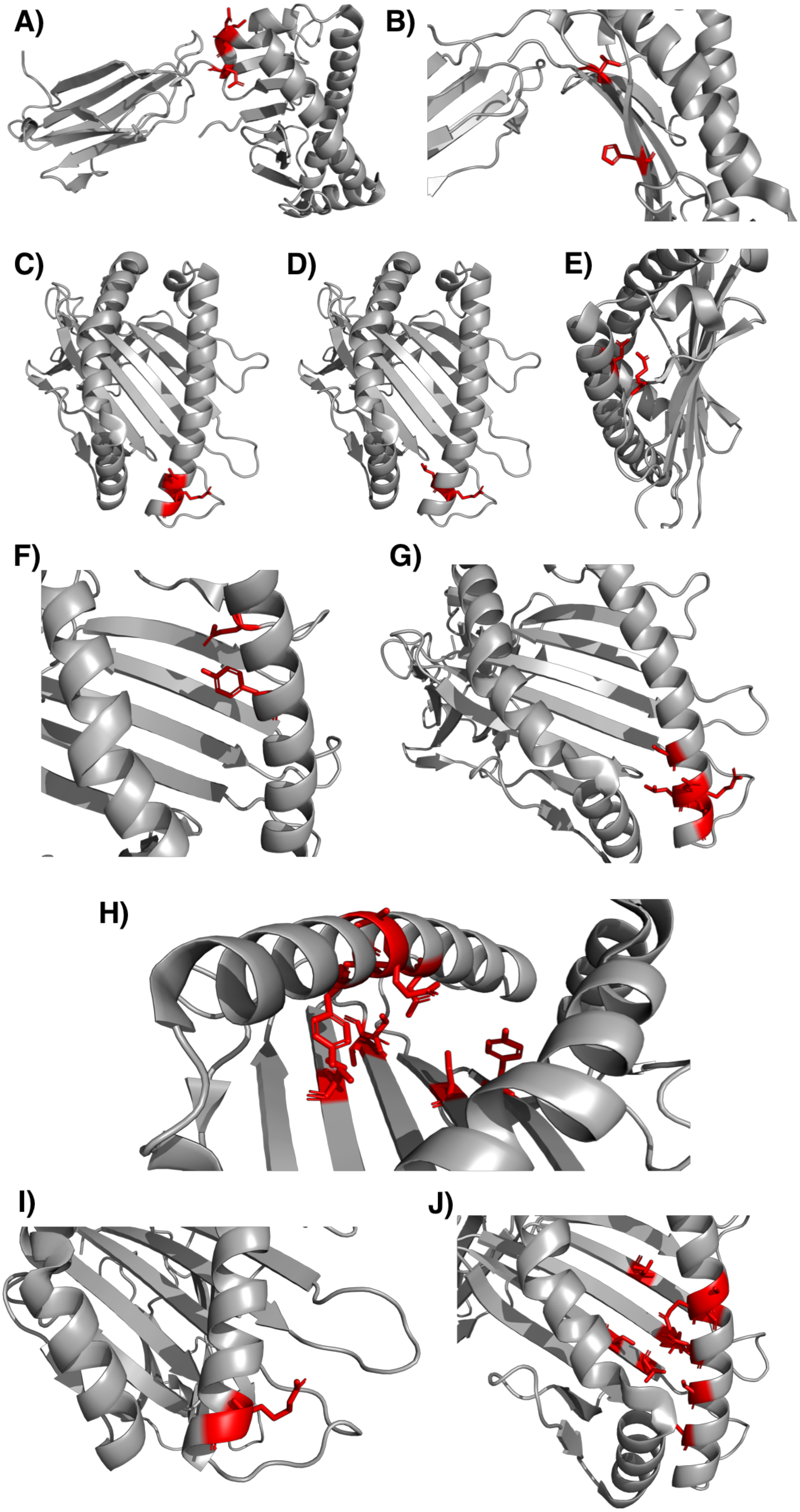


Figure 3.10: (see legend on next page)

Figure 3.10: Visualisation of the top ten Proximal Combinations of non-synonymous variants in Q31610 by number of total occurrences. Proteins are shown in cartoon format and are coloured grey, and variant positions are shown in stick format and are coloured red. **A-J)** Are the ten combinations ordered from most common to least common, e.g. **A)** is the most common combination 'D201E,K202T,E204Q'. See Table 3.8 for the ranked list of Proximal Combinations in Q31610.

Many of the observed variant combinations in proteins are rare, with 173,707 Global Combinations and 2,370 Proximal Combinations of non-synonymous variants only occurring in one heterozygous sample within the 2,504 samples (Figure 3.11 and Table 3.9). Within Q31610, there are 327 Proximal Combinations with only a single occurrence in the 2,504 samples. There are also many combinations that are common, with 1,484 Global Combinations and 64 Proximal Combinations of non-synonymous variants with ≥1,000 occurrences (Figure 3.11 and Table 3.9). Some of these very common combinations reflect positions in the reference genome that do not represent the most common allele observed in human populations.

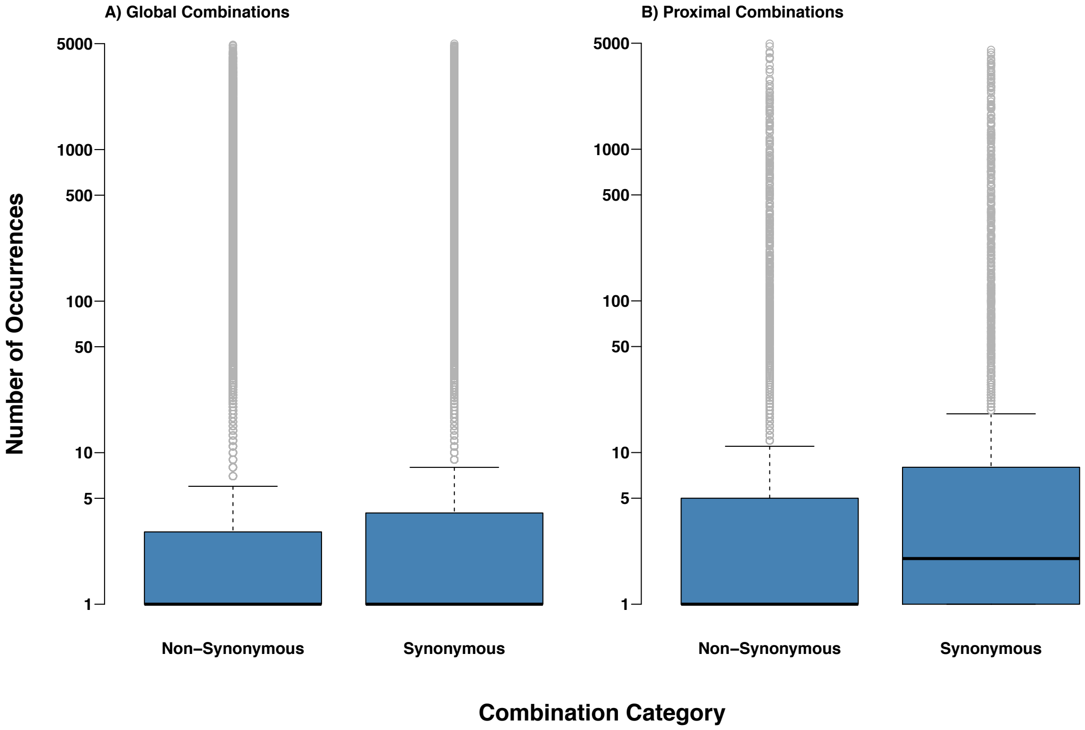


Figure 3.11: Number of occurrences of unique variant combinations within individual proteins, separated by combination category. **A)** Global Combinations. **B)** Proximal Combinations.

Table 3.9: Numbers of rare and common combinations for Global and Proximal variant combinations, and for non-synonymous and synonymous variants.

Variant Type	Global Combinations	Proximal Combinations
Non-Synonymous One Occurrence	173,707	2,370
Synonymous One Occurrence	122,708	1,255
Non-Synonymous ≥25 Occurrences	19,425	554
Synonymous ≥25 Occurrences	22,723	453
Non-Synonymous ≥100 Occurrences	8,016	274
Synonymous ≥100 Occurrences	10,035	271
Non-Synonymous ≥500 Occurrences	2,836	122
Synonymous ≥500 Occurrences	3,703	135
Non-Synonymous ≥1,000 Occurrences	1,484	64
Synonymous ≥1,000 Occurrences	1,952	85

The protein Pancreatic lipase-related protein 3 (UniProt accession: Q17RR3) catalyses the reaction of triacylglycerol with water to produce diacylglycerol and carboxylate. Within this protein, the non-synonymous Proximal Combination ‘V381I,R382G’ occurs 4,763 times (Figure 3.12.A&B), and the combinations ‘A380R,V381I,R382G’ and ‘A380R,R382G’ occur 65 and 22 times, respectively (Figure 3.12.C&D). All three of these combinations contain the variant R382G (4,850 combined occurrences), two contain V381I (4,828 combined occurrences) and two contain A380R (87 combined occurrences). For 4,850 combinations an arginine residue is lost, but for 87 of these (those that contain A380R) the arginine is replaced at a position close in space and the small hydrophobic amino acid alanine is replaced by a small hydrophobic amino acid glycine (Figure 3.12.C&D). Therefore, for a small subset of individuals the arginine lost in the variant R382G may be being compensated for by the variant A280R (see Section 3.4.7).

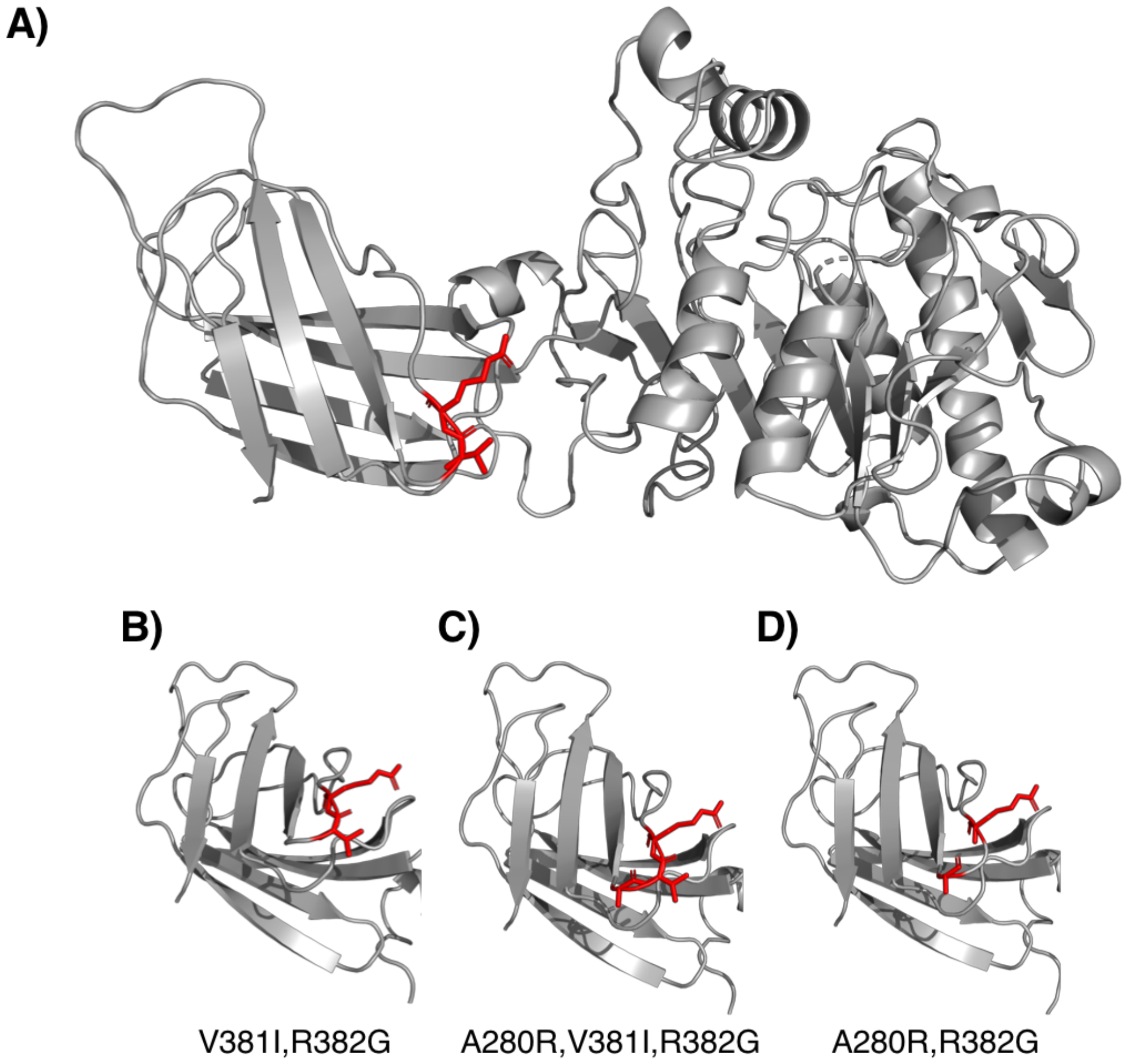


Figure 3.12: Proximal Combinations of non-synonymous variants observed in Pancreatic lipase-related protein 3 (UniProt accession: Q17RR3). Proteins are shown in cartoon format and are coloured grey, and variant positions are shown in stick format and are coloured red. **A)** Proximal combination V38II,R382G shown in the context of the full protein. **B)** Zoomed in view of Proximal combination V38II,R382G. **C)** Zoomed in view of Proximal combination A280R,V38II,R382G. **D)** Zoomed in view of Proximal combination A280R,R382G.

As the combinations in Pancreatic lipase-related protein 3 demonstrate, there can be multiple combinations within a protein, sometimes with overlapping variants. In Figure 3.13 each of the points is a protein, and the colour of the point is determined by the number of unique variant combinations observed for the protein (see Figure 3.13 legend). The x-axis position of the point is determined by the total number of occurrences of any variant combination within the protein, and the y-axis position is determined by the percentage of total combination occurrences accounted for by the most common variant combination in the protein. Therefore, if the most common variant combination accounts for all of the variant combinations in the protein (black points) the point will lie at the very top of the y-axis (100% of combination occurrences are the most common combination). Note, for Proximal Combinations the total occurrences of any combination can exceed 5,008 (2,504 samples multiplied by 2 alleles each), because multiple Proximal Combinations can occur in the same copy of a sample's protein (Figure 3.2.B; see Methods). The distributions of total percentage occurrences for the most common variant combination is shown in Figure 3.14.

For the Proximal Combinations, most proteins have a dominant variant combination, that accounts for the majority of variant combinations observed (Figure 3.13.B&D and Figure 3.14.B&D), even for proteins where there are a large number of total samples with variant combinations in the protein (points on the right of the x-axis in Figure 3.13). Of the 1,562 proteins with non-synonymous Proximal Combinations, 1,098 (70.3%) have a combination that accounts for ≥90% of all combinations, however 544 of these only have one total occurrence (1,352 out of 1,678 for synonymous Proximal Combinations (80.6%), with 754 occurring only once).

For the Global Combinations this trend is less clear, with a lower proportion of proteins having a dominant combination of variants, and most proteins having a large number of different combinations (Figure 3.13.A&C and Figure 3.14.A&C). Of the 10,753 proteins with any non-synonymous Global Combinations, 2,903 (26.9%) have a combination that accounts ≥90% of all combinations, and 1,325 of these have only one total occurrence (3,504 out of 11,131 for synonymous Global Combinations (31.5%), with 1,048 occurring only once).

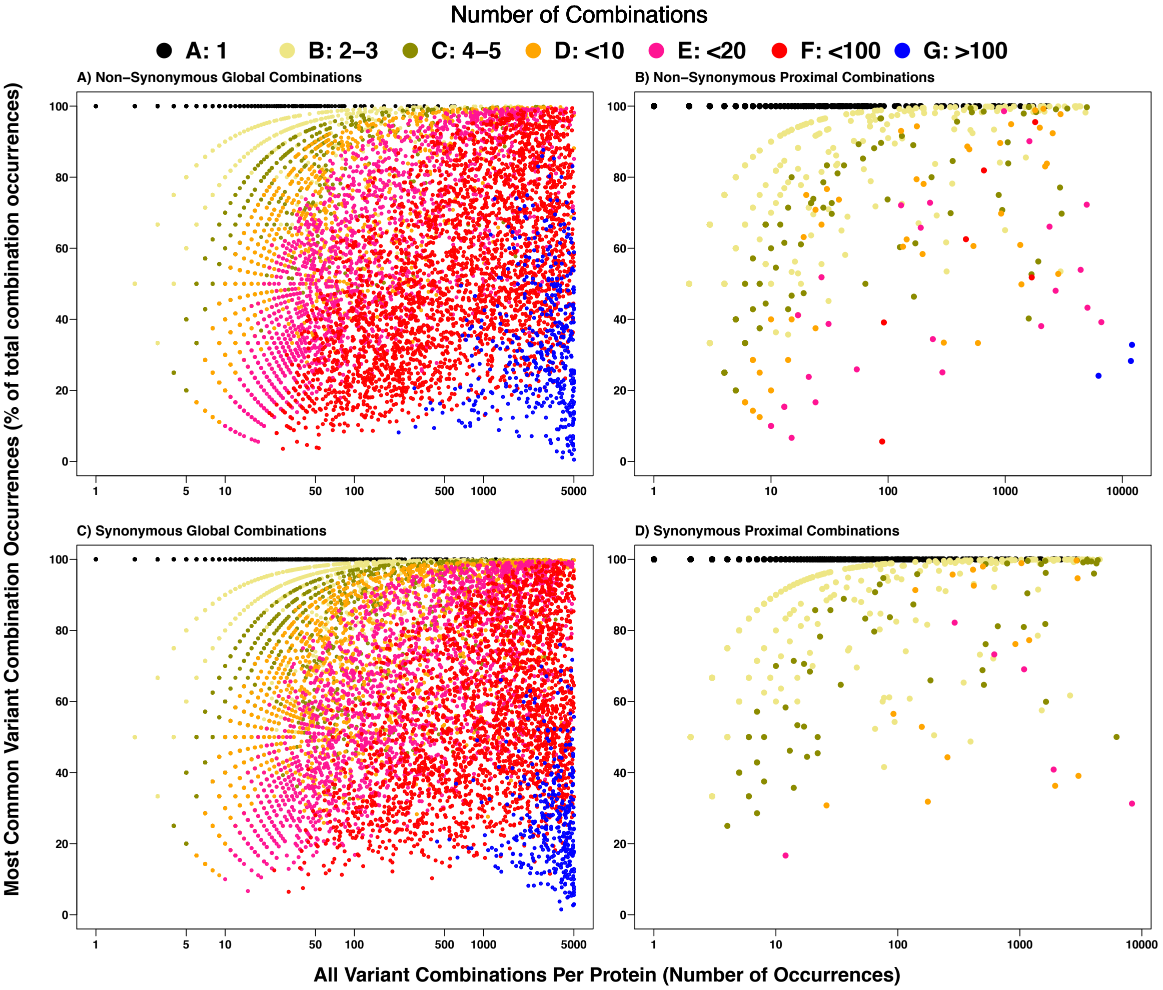


Figure 3.13: Occurrences of the most common variant combinations vs all variant combinations in individual proteins. Each point represents a protein, and the colour of the point is determined by the number of unique variant combinations observed for the protein, see legend. The x-axis position of the point is determined by the number of occurrences of any variant combination within the protein, and the y-axis position is determined by the percentage of total combination occurrences accounted for by the most common variant combination in the protein. If the most common variant combination accounts for all of the variant combinations in the protein (black points) the point will lie at the very top of the y-axis (100% of combination occurrences are the most common combination). **A)** Non-synonymous Global Combinations. **B)** Non-synonymous Proximal Combinations. **C)** Synonymous Global Combinations. **D)** Synonymous Proximal Combinations. Note - x-axes differ in their scales between the subplots.

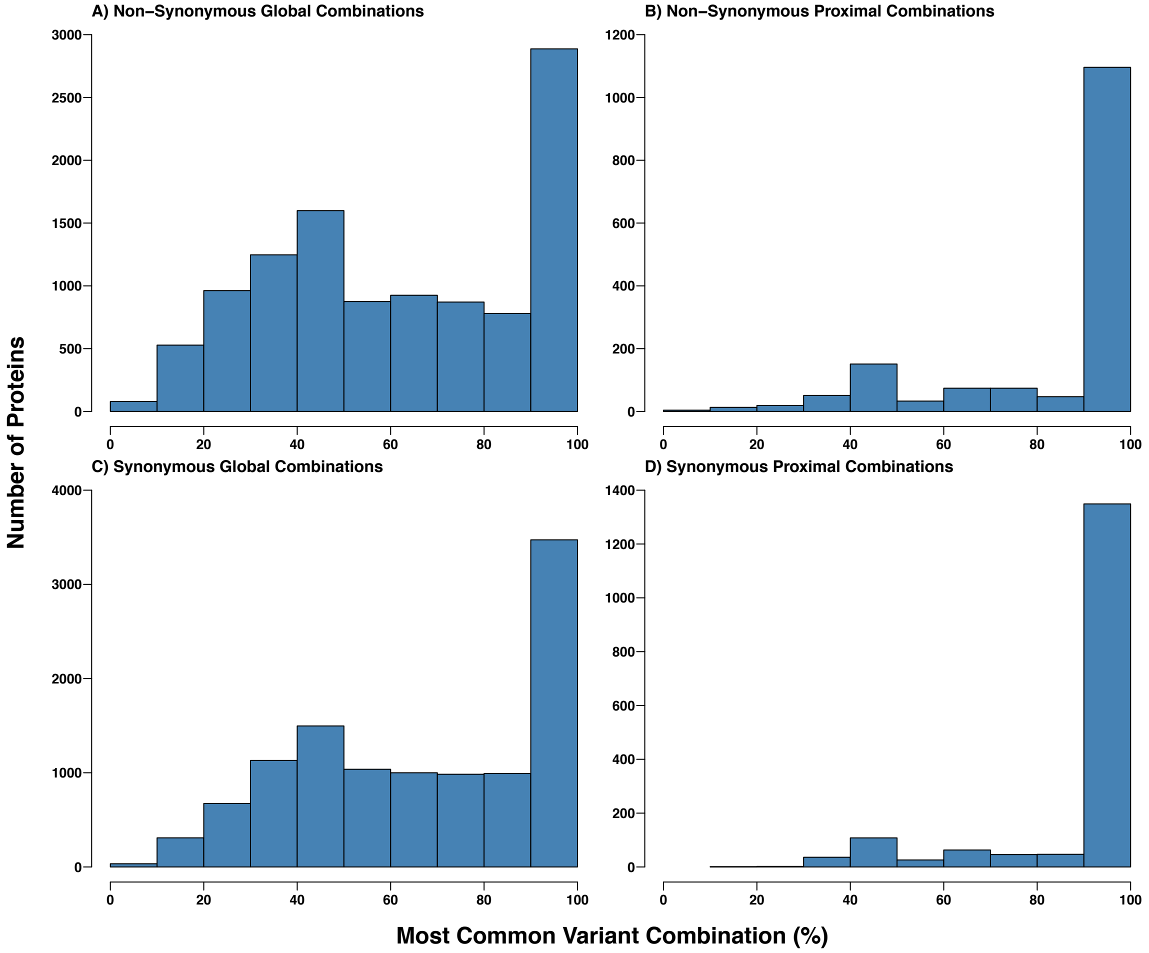


Figure 3.14: Distributions of proportions of variant combination occurrences from the most common variant combination per protein. **A)** Non-Synonymous Global Combinations. **B)** Non-Synonymous Proximal Combinations. **C)** Synonymous Global Combinations. **D)** Synonymous Proximal Combinations. Note – y-axes differ in their scales between the subplots.

In the example of Pancreatic lipase-related protein 3 (Figure 3.12), comparison of the Global and Proximal Combinations of variants highlights the level of noise and complexity when analysing variants that occur together anywhere in the protein compared to those in close spatial proximity. There are 3 different non-synonymous Proximal Combinations in this protein (a dark yellow point in Figure 3.13.B), and the most common of these (‘V381I,R382G’) accounts for 98.21% of all Proximal Combinations in this protein. For Global Combinations in the same protein, there are 42 different combinations of non-synonymous variants (a red point in Figure 3.13.A), and the most common of these (‘V381I,R382G’) accounts for only 64.02% of all Global Combinations in this protein. This is caused by Global Combinations like ‘V381I,R382G,F450Y’ (589 occurrences), where the most common combination ('V381I,R382G') is occurring with another variant (F450Y) which is distant in the protein structure (>5Å away), and therefore less likely to have a coevolutionary relationship with V381I or R382G.

Some of the combinations that are relatively rare at the level of the whole genome set may be fairly common within specific populations. Figures 3.15 & 3.16 show the distributions of non-synonymous Global and Proximal Combinations between the five sample super populations. These plots show the 50 most common combinations for each category. The corresponding heatmaps for synonymous Global and Proximal Combinations are shown in Appendix 2 Figures 8&9, and the distributions within the 26 individual sample populations for all combination types are shown in Appendix 2 Figures 10-13.

For each combination category there are examples of combinations that are spread relatively evenly across the super populations (light blue boxes across the super populations), and others that are concentrated in a subset of super populations (dark blue boxes in specific super populations). For example, the Global Combinations 'Q9BRQ3:Q260R,L263P' and 'O60844:G32S,S162T' occur very rarely in the AFR super population but are common in the other four super populations (Figure 3.15). For Proximal Combinations, 'G3V1Y8:C95R,L96V' is most common in the AFR super population, whereas 'Q8N423:H300Y,C306W' is most common in the EUR super population (Figure 3.16).

The combinations in Figures 3.15 & 3.16 are common combinations, with some having >4,000 total occurrences. Therefore, the population distributions by necessity are evenly distributed among the populations, because almost all individuals in the data set have the combinations. Such combinations reflect positions in the reference genome that are not representative of the most common alleles in human populations.

Differences between super populations can be seen more clearly for combinations that are rarer across the data set (fewer total occurrences). Figure 3.17 shows the 50 most common Proximal Combinations where the total number of combination occurrences is <500. Again, many of these combinations are most common in the

AFR super population, such as ‘A0A0B4J1V5:A91T,K93E’, and ‘Q8WXQ8:L336S,S378G’ - a potential variant compensation of serine (see Section 3.4.7). The combination ‘Q9H339:T78K,R88G’ occurs 486 times in total, with 82.7% of these in the AFR super population, and is another example of a potential variant effect compensation (compensation of a positive charge, with loss of arginine and gain of lysine; see Section 3.4.7). There are also clear examples of combinations that are more common in one or more of the other four super populations, such as A6NGD5:G108S,V109M’, for which 57.6% of occurrences are from the SAS super population (Figure 3.17).

Evenly distributed variant combinations are likely to have first occurred longer ago in human evolution, with population-specific variant combinations likely to have occurred more recently. Similar patterns of super population distributions can be seen for synonymous combinations (Appendix 2 Figures 8&9), as well as the 26 individual populations that make up the five super populations (Appendix 2 Figures 10-13).

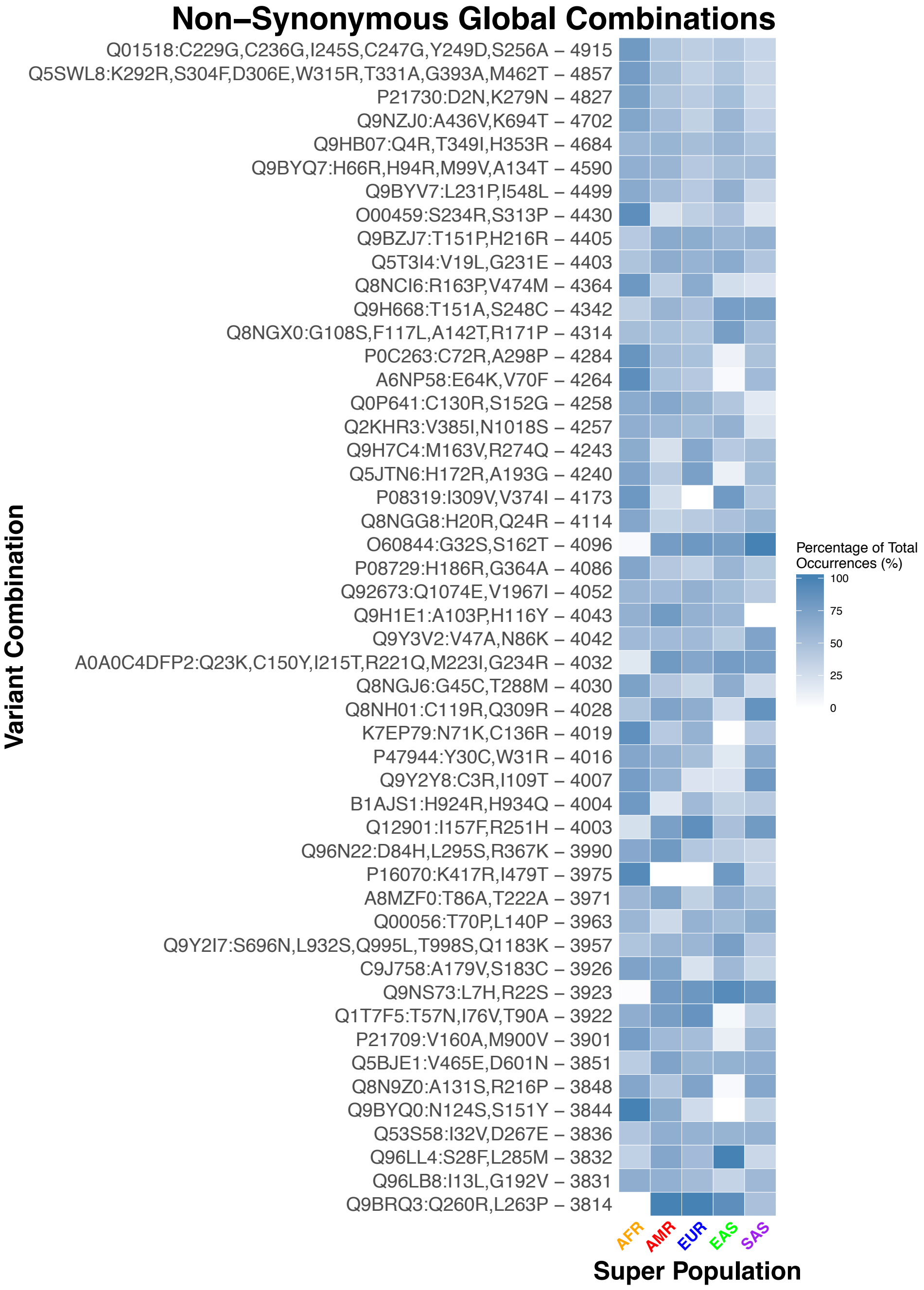


Figure 3.15: Heatmap of occurrences within super populations of the 50 most common non-synonymous Global Combinations. Combinations are given on the y-axis, see Methods for a description of the variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

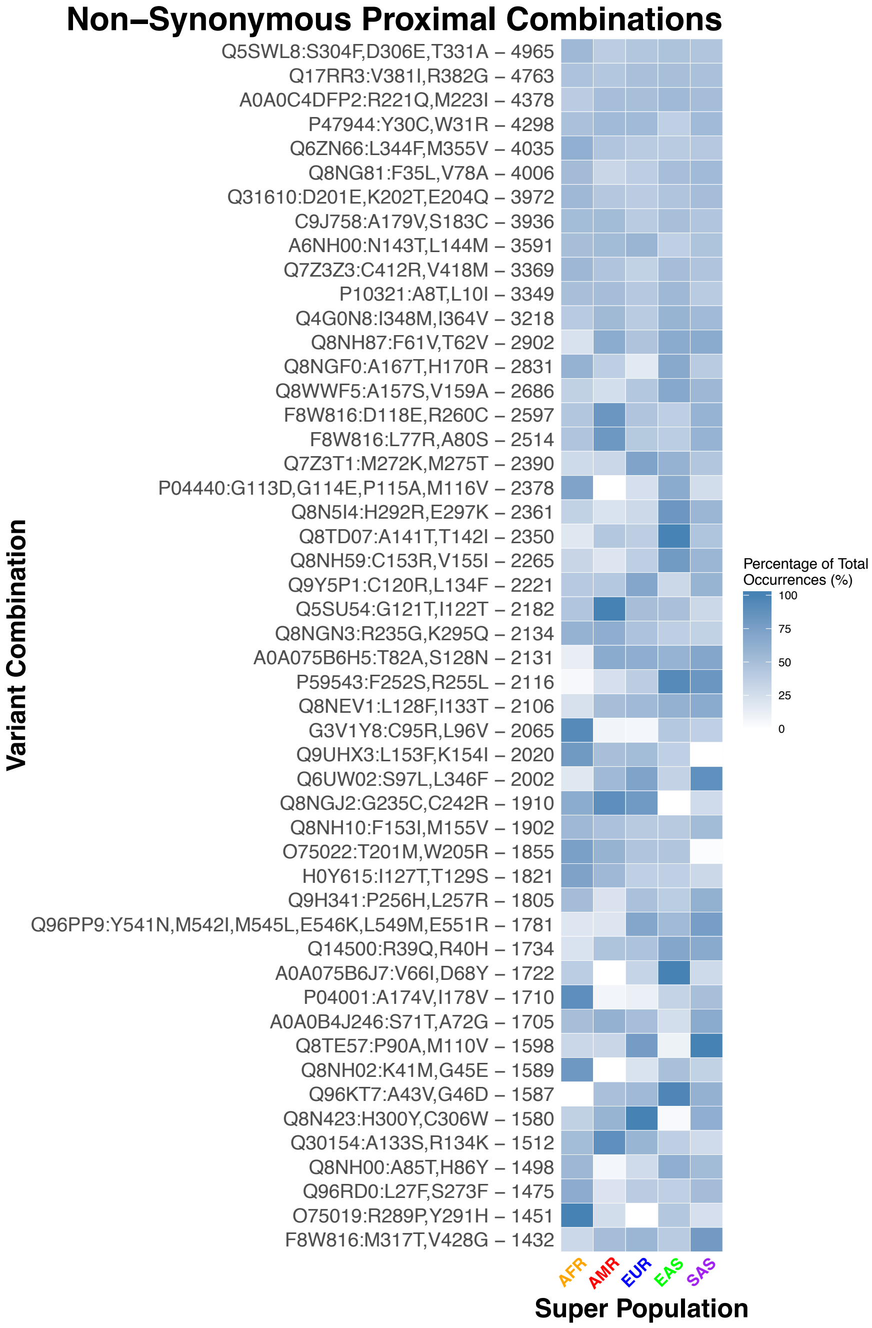


Figure 3.16: Heatmap of occurrences within super populations of the 50 most common non-synonymous Proximal Combinations. Combinations are given on the y-axis, see Methods for a description of the variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

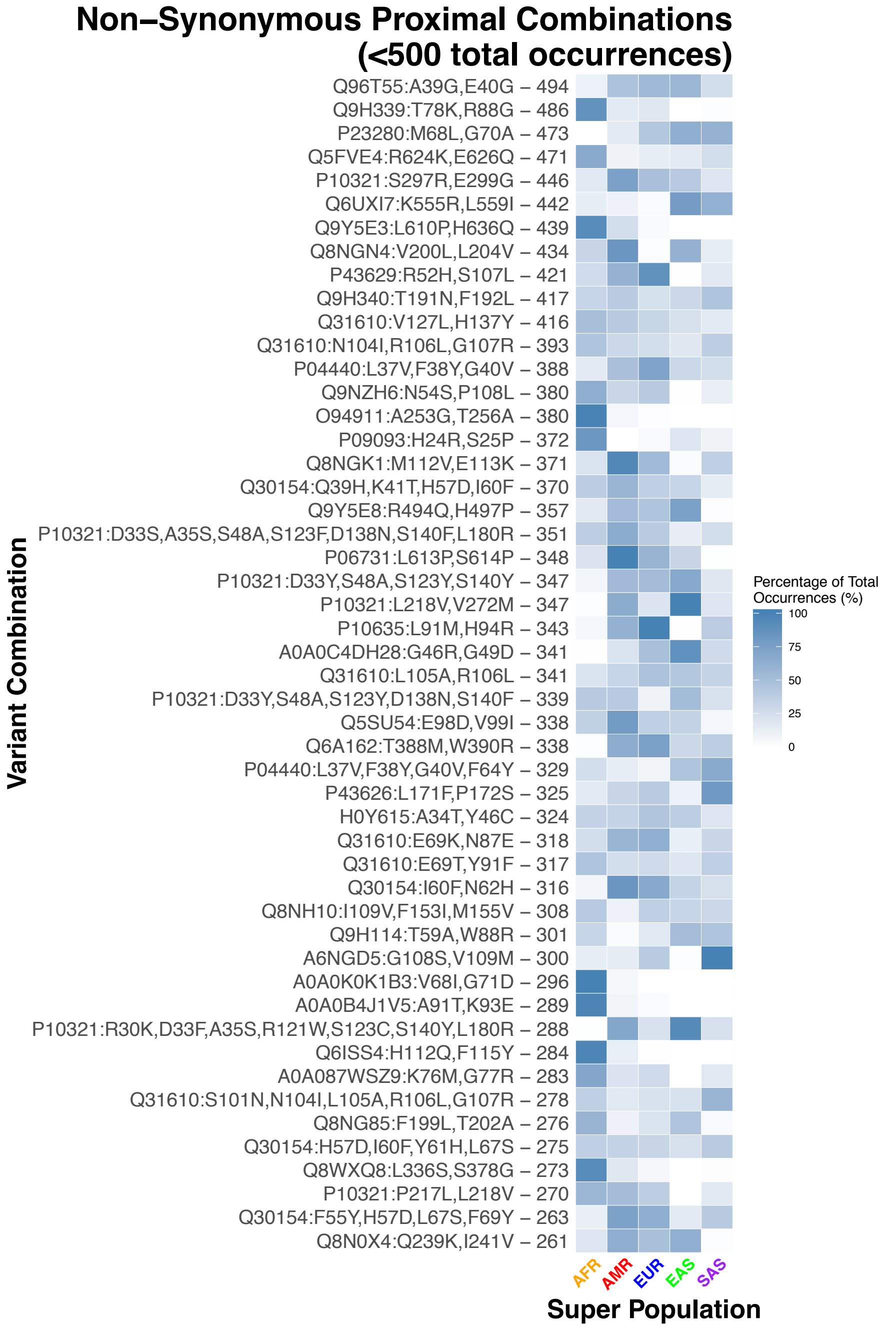


Figure 3.17: Heatmap of occurrences within super populations of the 50 most common non-synonymous Proximal Combinations, for combinations with <500 total occurrences. Combinations are given on the y-axis, see Methods for a description of the variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

While many of the variant combinations in Figures 3.15-3.17 are either evenly spread between the super populations or enriched in the AFR super population, there are individual combinations that are more common within each of the super populations. The distribution of population frequencies for every combination in each combination category is show in Figure 3.18. For each super population there are thousands of combinations where >90% of the total occurrences are from that super population, but the AFR super population clearly has many more of these combinations than the other four super populations (Figure 3.18 & Table 3.10).

Table 3.10: Numbers of unique combinations where ≥90% of total occurrences are from one super population.

Super Population	Non-Synonymous Global Combinations	Non-Synonymous Proximal Combinations	Synonymous Global Combinations	Synonymous Proximal Combinations
African	99,543	1,355	88,495	900
American	22,851	308	15,535	175
European	30,724	505	19,853	230
East Asian	44,469	602	30,100	347
South Asian	42,947	502	30,169	320

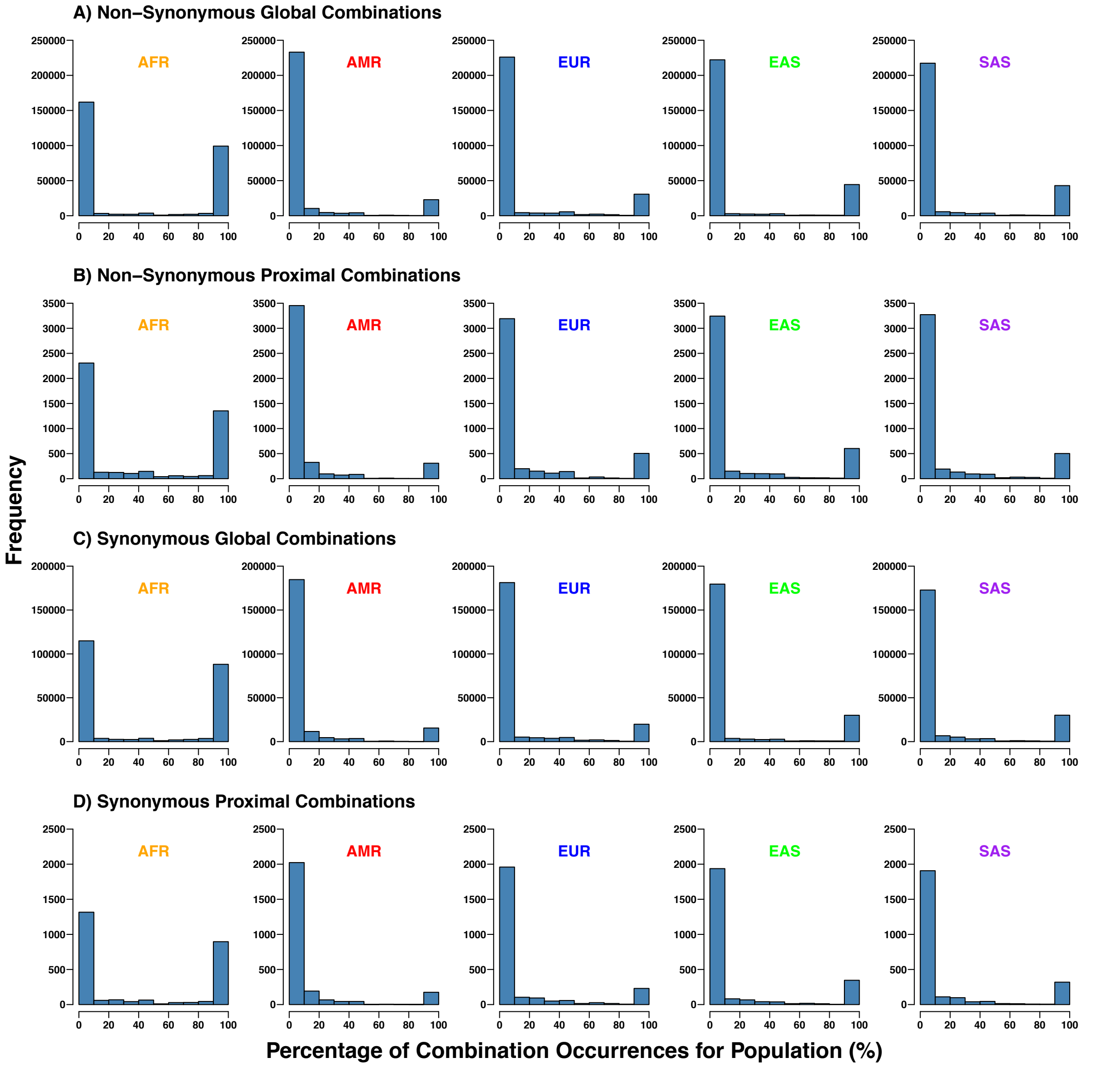


Figure 3.18: Distributions of percentages of total occurrences of each unique variant combination from each super population. A) Non-Synonymous Global Combinations. B) Non-Synonymous Proximal Combinations. C) Synonymous Global Combinations. D) Synonymous Proximal Combinations. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population. Note – y-axes differ between the subplots.

3.4.6 Variant Pairs

The main aim of this study is to identify groups of variants in individual human genomes that have potential coevolutionary relationships with one another. In addition to considering possible variant compensations for the observed variant combinations, we also analysed individual variant pairs within the combinations for possible compensation events (Figure 3.2.C; see Methods). Many of the Proximal

Combinations observed are variant pairs (68.11%; Figure 3.5 and Table 3.3), but variant combinations containing more than two variants can be harder to analyse. For example, a rare variant combination containing three variants may contain a common variant pair with a strong coevolutionary relationship, and one other variant that has a weak coevolutionary relationship with the common variant pair (Figure 3.2.C).

Analysis of the unique constituent variant pairs within combinations can help in such complex cases, such as the proximal combinations in the protein Pancreatic lipase-related protein 3 (Figure 3.12). The combination 'V381I,R382G' occurs frequently (4,763 times) but there is seemingly minimal possible compensation between V381I and R382G. The combinations 'A380R,V381I,R382G' and 'A380R,R382G' occur less frequently (65 and 22 times, respectively) but with an obvious potential compensation of the arginine residue. In this example, the fact that R382G occurs 4,850 times and only 87 of these are with A380R suggests that R382G alone is not severely damaging, but perhaps there is a fitness difference when A380R also occurs.

For a variant pair to be mapped to structure, both variants must not only be mapped to a structure but both to the same structure. For example, for a protein sequence that is represented by two structures, one corresponding to the N-terminus and one to the C-terminus, both variants in the pair must be mapped to one of the structures otherwise the distance between the residues cannot be determined. The majority of the variant pairs mapped to structure are >10Å apart from one another (Table 3.11).

Table 3.11: Numbers of each type of variant pair for non-synonymous and synonymous variants.

Type of Variant Pair	Non-Synonymous	Synonymous
Pairs Mapped to Structures	9,811	9,479
Pairs <5 Angstroms Away	783	502
Pairs <10 Angstroms Away	1,205	888
Pairs >10 Angstroms Away	7,823	8,089

If variants have coevolved, as a way of compensating for negative effects of sub-combinations, they should occur at similar frequencies to one another. However, this may not always be true, in cases where one variant alone is damaging to a protein's function but not lethal to the organism, where one variant can occur alone and is neutral to protein function but can buffer the damaging effects of a second mutation should one occur (Figure 3.1.B), or where individuals have additional variants (not necessarily in the same protein) that compensate for the damaging mutation in a different way.

Figure 3.19 is split in to sub-plots of variant pairs that are <5Å apart, >5Å apart but <10Å apart, and those >10Å apart (see legend). Each point is a pair of variants, with the x-axis position determined by the number of occurrences of either of the variants on their own and the y-axis position determined by the number of occurrences of the two variants together, within the 1,000 Genomes Project samples. Points on the far left of the x-axis are the variant pairs that occur very commonly together and very rarely alone. These pairs of variants are those most likely to have a coevolutionary relationship.

Of the 783 non-synonymous variant pairs <5Å apart, 93 occur 100% of the time together, 206 occur ≥90% of the time together, 239 occur ≥80% of the time together, and 252 occur <10% of the time together (Figure 3.20.A). Of the 502 synonymous variant pairs <5Å apart, 59 occur 100% of the time together, 131 occur ≥90% of the time together, 148 occur ≥80% of the time together, and 198 occur <10% of the time together (Figure 3.20.B). Therefore, there are 100s of variant pairs <5Å apart that always or almost always occur together, and as a fraction of the total number there are more compared to the variant pairs >5Å apart but <10Å apart, and variant pairs >10Å apart (Figure 3.20).

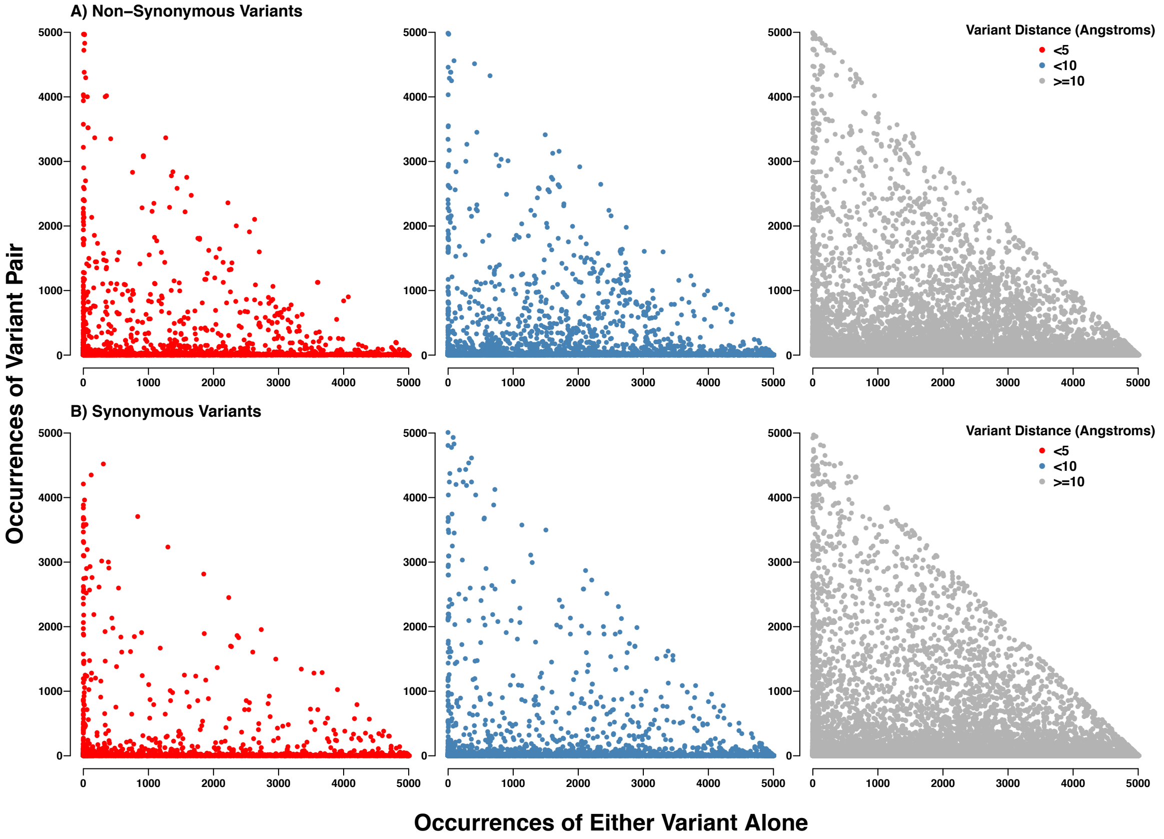


Figure 3.19: Occurrences of variant pairs together vs either variant alone. Variants are coloured by the distance between the two variants, see legend. **A)** Non-Synonymous variant pairs. **B)** Synonymous variant pairs.

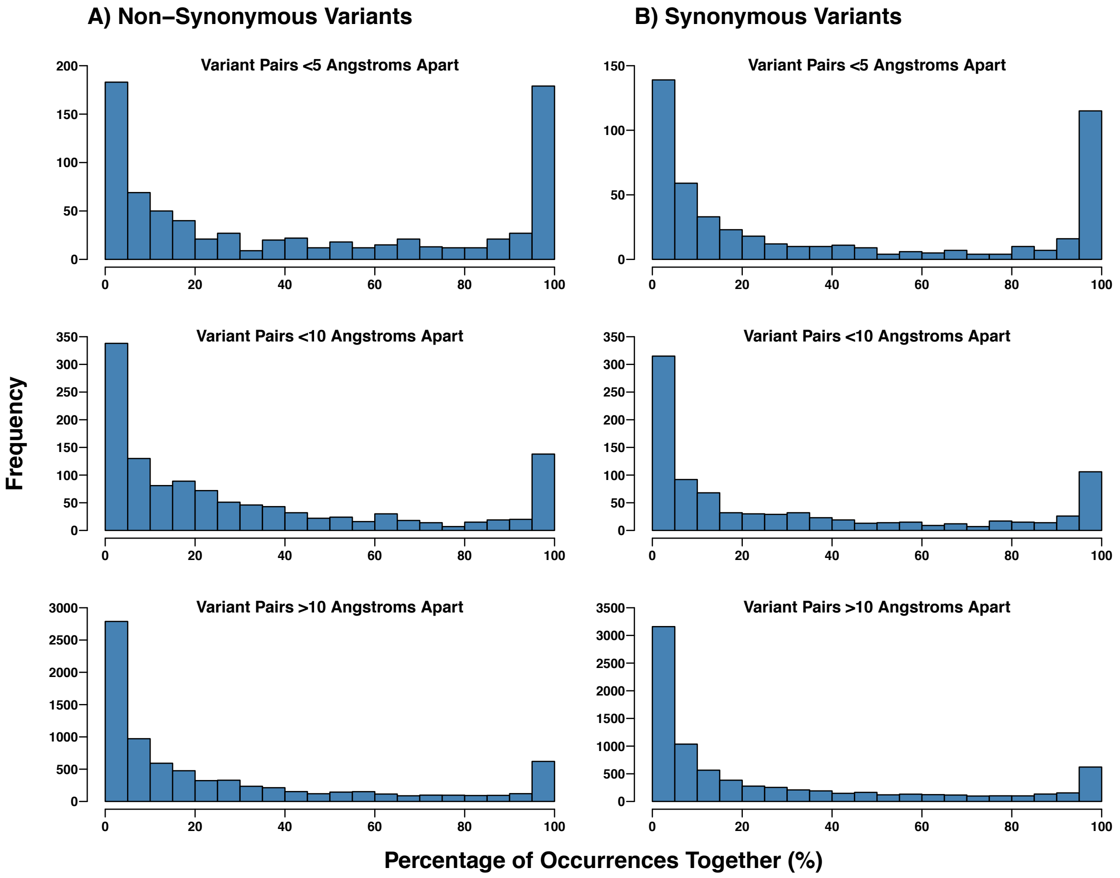


Figure 3.20: Distributions of percentages of occurrences of variant pairs together vs either variant alone, separated by variant pairs <5Å apart, >5Å apart but <10Å apart, and >10Å apart. **A)** Non-Synonymous Variants. **B)** Synonymous variants. Note – y-axes differ in their scales between the subplots.

In addition to comparing how frequently pairs of variants occurred together vs on their own within the 1,000 Genomes Project samples, we also compared the overall occurrences of variants in the observed combinations using the data from GnomAD. The GnomAD data set is comprised of 123,136 exomes, far more individuals than the 1,000 Genomes Project data set (2,504), but the data is aggregated and so it is not possible to identify variant combinations within individual samples. However, this larger dataset is useful for looking at similarities in the overall frequencies of variants.

Using the GnomAD data we considered how frequently the different variants within Proximal Combinations identified from the 1,000 Genomes Project data set occur. For each Proximal Combination, the maximum allele count in GnomAD (i.e. most common variant in a combination) was compared to the mean allele count for the combination (Figure 3.21.A). The closer the allele count for the different variants in a combination is, the smaller the difference between the maximum count and the mean count. The same analysis was performed comparing the median and minimum allele counts with the maximum counts (Figure 3.21.B&C), and the distributions of mean counts, median counts, and minimum counts as a percentage of maximum allele counts is shown in Figure 3.22.

When comparing the mean allele count to the maximum allele count, there is a clear set of combinations that have very similar mean and maximum allele counts (points closer to the x=y line in Figure 3.21.A) showing that there are variant combinations in which the frequencies of all the variants are similar. There are 149 Proximal Combinations of non-synonymous variants where the mean allele count is identical to the maximum allele count, 225 where the mean count is >99% of the maximum count, 352 where the mean count is >90% of the maximum count, and 476 where the mean count is >80% of the maximum count (Table 3.12). This appears to relate more to combinations that are pairs (blue points in Figure 3.21.A), although there are some larger combinations that are also close to having equal mean and maximum allele counts. The graph also shows an off-diagonal line of points (gradient ~0.5), containing combinations that are primarily pairs, this represents combinations where one variant is common but the other variant is rare, so the mean allele count is approximately half of the maximum allele count (Figure 3.21.A & Figure 3.22.A).

There are a large number of points in the bottom left corner of Figure 3.21.A where the maximum allele count for the combination is below 50,000 (1,551 combinations). To highlight these combinations, separate subplots comparing mean and maximum allele counts for combinations were produced, filtering for combinations with maximum allele counts ≤50,000, ≤10,000, ≤5,000, and ≤1,000 (Figure 3.21.E-H). For these comparatively rarer combinations, there are some combinations of 3-4 variants with similar mean and maximum allele counts - red points close to the x=y line (six combinations with mean count ≥90% of maximum count and 22 combinations ≥80%). These were not visible in Figure 3.21.A, as the scale has to cover the full range of allele counts.

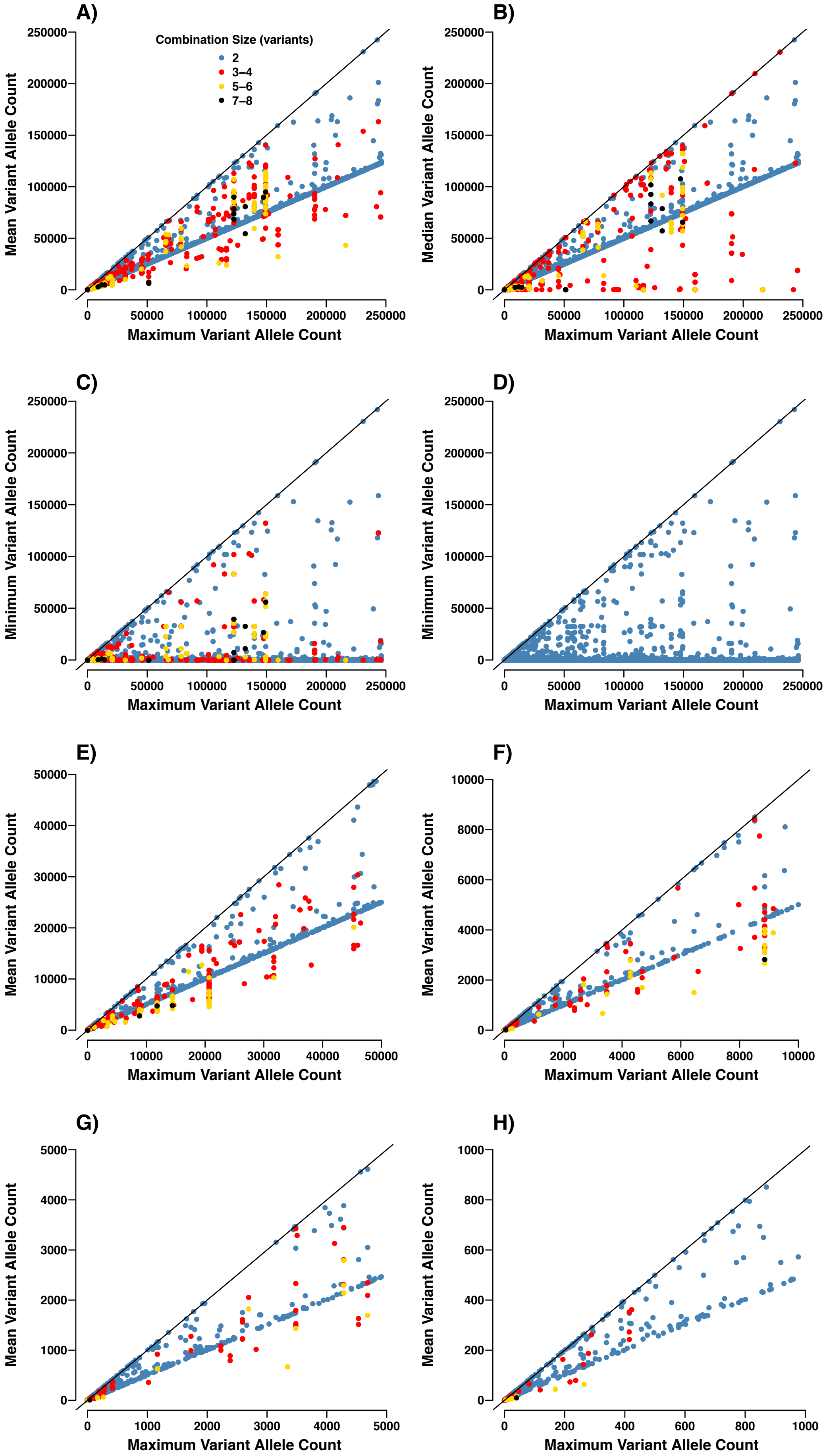


Figure 3.21: (see legend on next page)

Figure 3.21: Allele counts of variants in GnomAD for the observed Proximal Combinations. Each point corresponds to a variant combination, and the points are coloured by the number of variants in the combination, see legend. **A)** Comparison of the maximum allele count with the mean allele count for each combination. **B)** Comparison of the maximum allele count with the median allele count for each combination. **C)** Comparison of the maximum allele count with the minimum allele count for each combination. **D)** Comparison of the maximum allele count with the minimum allele count for the constituent variant pairs within Proximal Combinations. **E)** Comparison of the maximum allele count with the mean allele count for each combination, where maximum count <50,000. **F)** Comparison of the maximum allele count with the mean allele count for each combination, where maximum count <10,000. **G)** Comparison of the maximum allele count with the mean allele count for each combination, where maximum count <5,000. **H)** Comparison of the maximum allele count with the mean allele count for each combination, where maximum count <1,000. Note – x- and y-axes differ in their scales between the subplots.

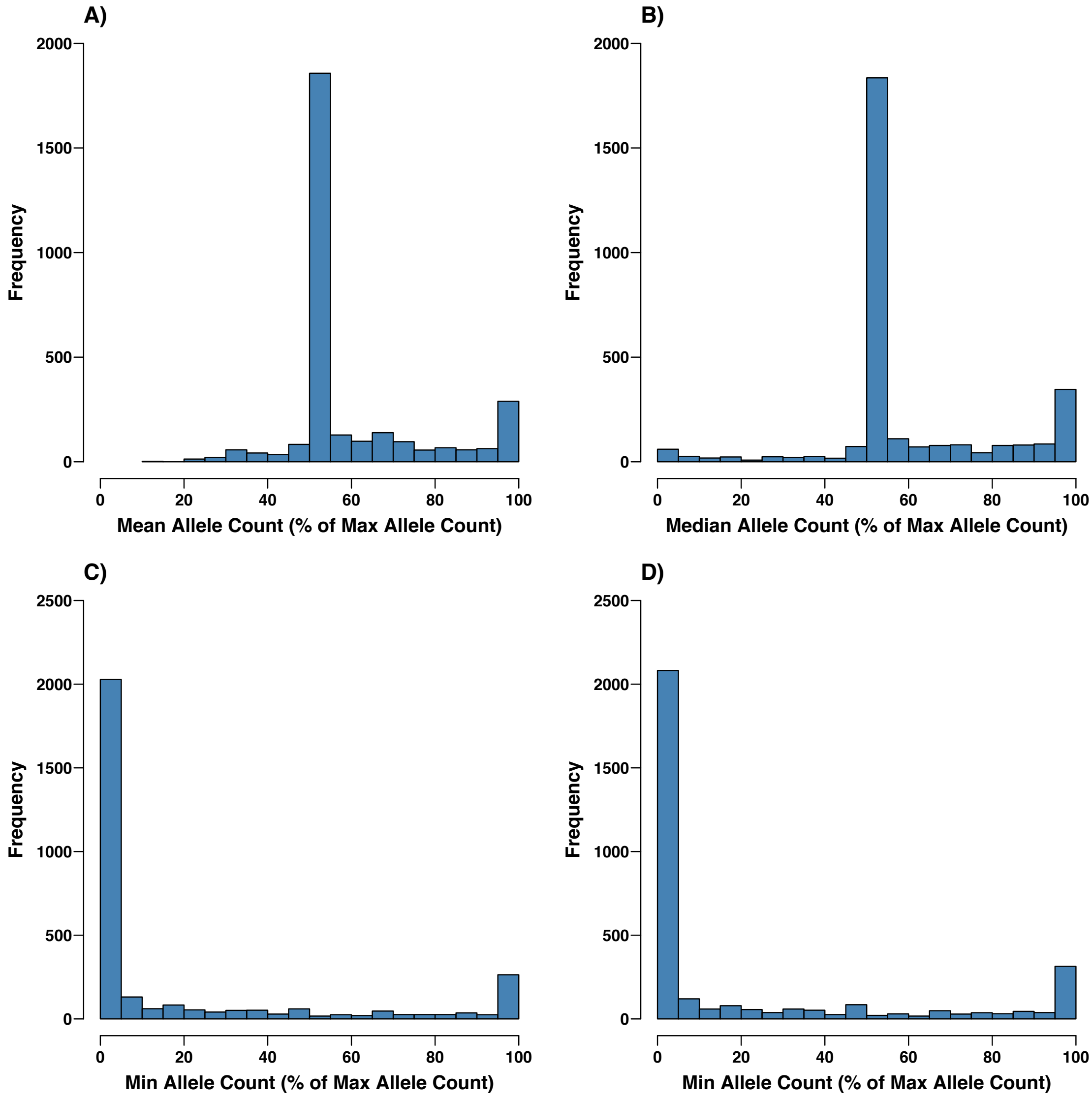


Figure 3.22: (see legend on next page)

Figure 3.22: Distributions of mean, median, and minimum allele counts within Proximal Combinations and their constituent variant pairs as a percentage of the maximum allele count within the combination. **A)** Mean allele count as a percentage of the maximum allele count within Proximal Combinations. **B)** Median allele count as a percentage of the maximum allele count within Proximal Combinations. **C)** Minimum allele count as a percentage of the maximum allele count within Proximal Combinations. **D)** Minimum allele count as a percentage of the maximum allele count within variant pairs. Note – y-axes differ in their scales between the subplots.

Table 3.12: Mean, median, and minimum allele counts per non-synonymous Proximal Combination as a percentage of the maximum allele count, and minimum allele count as a percentage of the maximum allele count for constituent variant pairs of non-synonymous Proximal Combinations

Threshold (% of Maximum Allele Count)	Mean Allele Count for Proximal Combinations	Median Allele Count for Proximal Combinations	Minimum Allele Count for Proximal Combinations	Minimum Allele Count for Variant Pairs
100	149	156	149	178
>99	225	260	207	243
>90	352	431	289	352
>80	476	589	351	428

Comparison of the maximum and median allele counts (Figure 3.21.B and Figure 3.22.B) identified a set of combinations (mainly containing three or four variants; red points in Figure 3.21.B), where the median allele count is much closer to the maximum allele count than the mean allele count is (Figure 3.21.A&B Figure 3.22.A&B). This suggests that there are some combinations observed where a subset of the variants have similar frequencies but the remaining variants in the combination occur rarely (e.g. a combination containing three variants could have two with similar frequencies and a third variant that only occurs in a single individual).

The three overlapping Proximal Combinations in Pancreatic lipase-related protein 3 are an example of a common combination where two variants occur commonly together (V381I and R382G), and sometimes those same two variants occur with a third rarer variant (A280R; Figure 3.12). V381I and R382G are both very common in GnomAD (allele counts of 241,737 and 242,927 respectively), but

A380R is very rare, with no occurrences in GnomAD. Similarly, in the Olfactory receptor (UniProt accession: A6NLW9) the combination ‘A6NLW9:P60L,F66S ,R67M’ occurs once, but the sub-combination ‘A6NLW9:F66S,R67M’ occurs 1,079 times. In accordance with this, in GnomAD, F66S and R67M have similar allele counts (allele counts of 50,481 and 50,734 respectively), but P60L is much rarer (allele count of 2).

There are 156 Proximal Combinations of non-synonymous variants where the median allele count is identical to the maximum allele count, 260 where the median count is >99% of the maximum count, 431 where the median count is >90% of the maximum count, and 589 where the median count is >80% of the maximum count (Table 3.12). This is more combinations than when comparing the mean allele count to the maximum allele count for each percentage threshold, and may indicate that some subsets of variants occur frequently together, while the remaining variants in the combination do not (Table 3.12).

This is further supported by comparison of the minimum and maximum allele counts (Figure 3.21.C and Figure 3.22.C), which shows that for combinations >2 variants in size the minimum allele count is almost always much lower than the maximum allele count, with very few exceptions. This indicates that for most of these combinations containing >2 variants, that at most two of them have similar frequencies, while the remaining variants have much lower frequencies. There are 149 Proximal Combinations of non-synonymous variants where the minimum allele count is identical to the maximum allele count, rising to 207 where the minimum count is >99% of the maximum count (Table 3.12).

Analysis of the constituent variant pairs of Proximal Combinations (i.e. all pairs from Proximal Combinations of any size) highlights the large number of variant pairs with very similar allele counts (Figure 3.21.D and Figure 3.22.D). There are 178 constituent pairs of variants where the median allele count is identical to the maximum allele count, 243 where the median count is >99% of the maximum count, 352 where the median count is >90% of the maximum count, and 428 where the median count is >80% of the maximum count (Table 3.12).

For the GnomAD data it is impossible to know whether the variants occur in the same individuals and in the same copies of the gene, due to the aggregated nature of the GnomAD data set, but in combination with the 1,000 Genomes Project data it can provide additional clues for variants with a likely coevolutionary relationship. With 121,136 samples in the GnomAD dataset, there are a possible 242,272 alleles for an autosomal genome position. If the maximum allele count of a variant in a combination is above 121,136 (half of the alleles) then it must occur with one or more of the other variants in at least one individual human.

There are 902 combinations where the maximum allele count is >121,136 (122 combinations for mean counts >121,136, 150 for median counts, and 23 for minimum counts). For the protein Metallothionein-4 (UniProt accession: P47944), the combination ‘P47944:Y30C,W31R’ occurs 4,398 times (Figure 3.23.A), and the allele counts in GnomAD are 230,457 and 230,953 respectively. In the protein T cell receptor alpha variable 30 (UniProt accession: A0A087WSZ9), the combination ‘A0A087WSZ9:K76M,G77R’ (a potential positive charge compensation, with loss of lysine and gain of arginine – see Section 3.4.7) occurs 283 times (Figure 3.23.B), with allele counts in GnomAD of 7,479 and 7,477 respectively.

For 153 of the observed Proximal Combinations the allele counts of the constituent variants are identical in GnomAD (with 30 of these having allele counts ≥25), and it is more likely for these variants to have a strong coevolutionary relationship. In the protein Advanced glycosylation end product-specific receptor (UniProt accession: Q15109), the combination ‘Q15109:W271R,C301S’ occurs only once (with no individuals having either variant on its own; Figure 3.23.C), but both variants have allele counts of 500 in GnomAD. In the protein Heat shock 70 kDa protein 6 (UniProt accession: P17066), the combination ‘P17066:N153S,D154N’ (a potential direct amino acid compensation of asparagine – see Section 3.4.7) occurs 63 times (Figure 3.23.D), and both variants have allele counts of 561 in GnomAD.

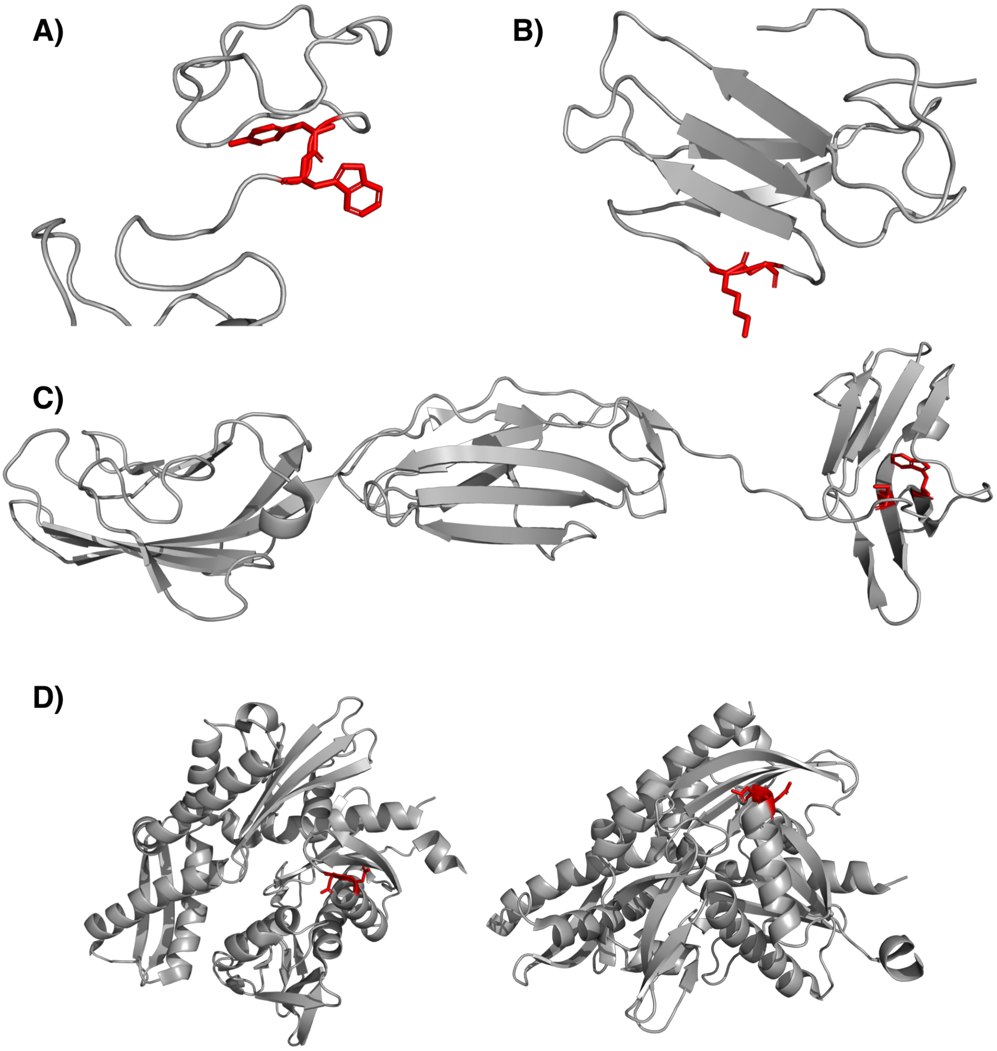


Figure 3.23: Four examples of Proximal Combinations with similar allele counts in GnomAD. Proteins are shown in cartoon format and are coloured grey, and variant positions are shown in stick format and are coloured red. **A)** The combination ‘P47944:Y30C,W31R’. **B)** The combination ‘A0A087WSZ9:K76M,G77R’. **C)** The combination ‘Q15109:W271R,C301S’. **D)** The combination ‘P17066:N153S,D154N’ (from two different angles).

3.4.7 Proximal Combination Effect Compensations

Coevolution occurs when a subsequent variant arises that compensates for a previous variant. We have identified a set of variant combinations that co-occur in close proximity within protein structures. To further identify if these variants have coevolved, we considered if the variants observed could compensate for each other. We analysed various properties of the wild type and variant amino acids within these Proximal Combinations.

We first considered potential direct amino acid compensations, which is the simplest and most complete type of compensation event. For example, in a pair of variants this would occur when one position changes from glycine to serine and the second position changes from serine to glycine. In these exchanges, the net change of amino acids is zero, the positions of the amino acids have changed within the structure, although the variants were already in close proximity to each other.

Not all direct compensation events involve every amino acid in the combination being balanced, like in the serine>glycine & glycine>serine example. A serine>glycine mutation could occur in close proximity to a valine>serine mutation, and in this case the serine would be compensated for but not the loss of glycine. This direct amino acid compensation could occur in 100 samples, and the same variants could occur with an additional leucine>isoleucine variant in another 100 samples. This serine compensation would be counted as two different combinations with compensations (each occurring in 100 samples), but one unique compensation event that occurs in 200 samples.

For the set of 4,365 Non-Synonymous Proximal Combinations, 571 unique direct amino acid compensations were observed (Figure 3.24 and Table 3.13). Some individual amino acids are compensated for more often than others, with 85 unique arginine compensations (14.8% of total), the most for any amino acid (Figure 3.25). There are a large number of compensations of other amino acids, particularly serine (64 - 11.2%), valine (59 - 10.3%), and alanine (50 - 8.8%; Figure 3.25). The other amino acids have between 10 (cysteine) and 39 (threonine) unique compensations, except for tryptophan, which has no compensation events (Figure 3.25).

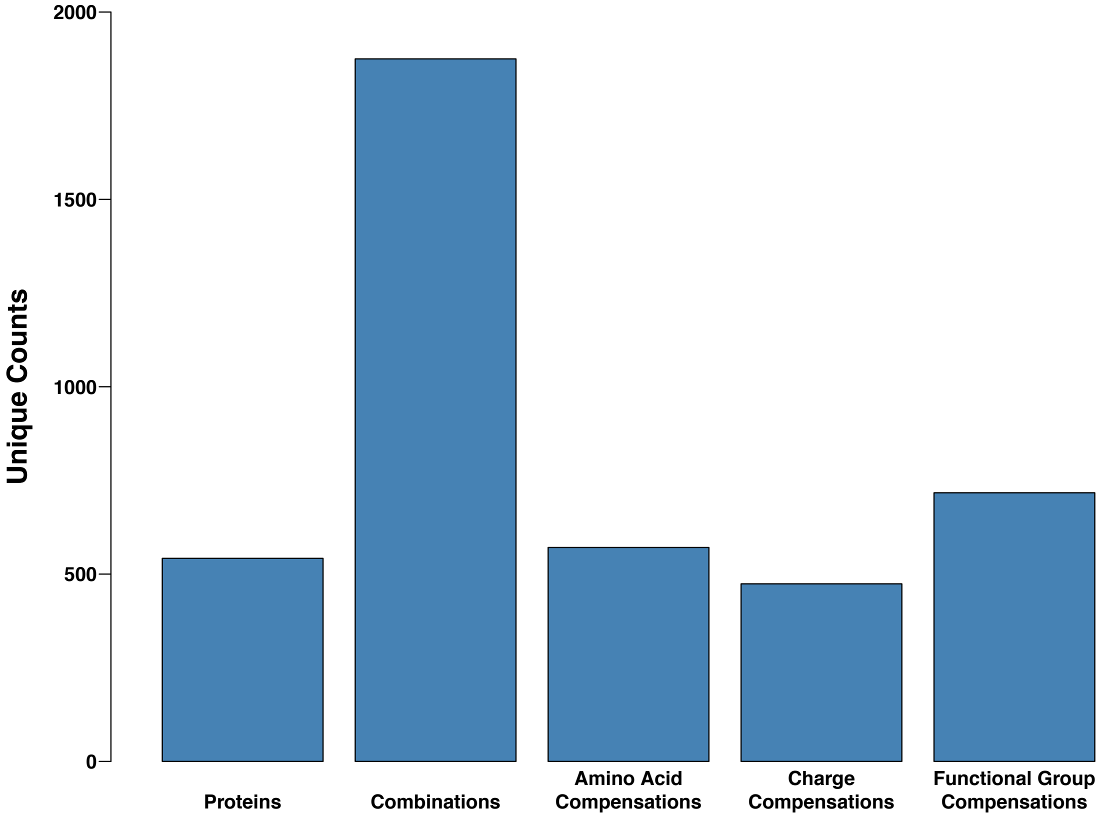


Figure 3.24: Numbers of potential property compensations within Non-Synonymous Proximal Combinations.

Table 3.13: Numbers of potential compensation types for Non-Synonymous Proximal Combinations.

Compensation Category	**Unique Compensations**
Proteins with Compensation Events	542
Combinations with Compensation Events	1,875
Amino Acid Compensations	571
Charge Compensations	474
Functional Group Compensations	717

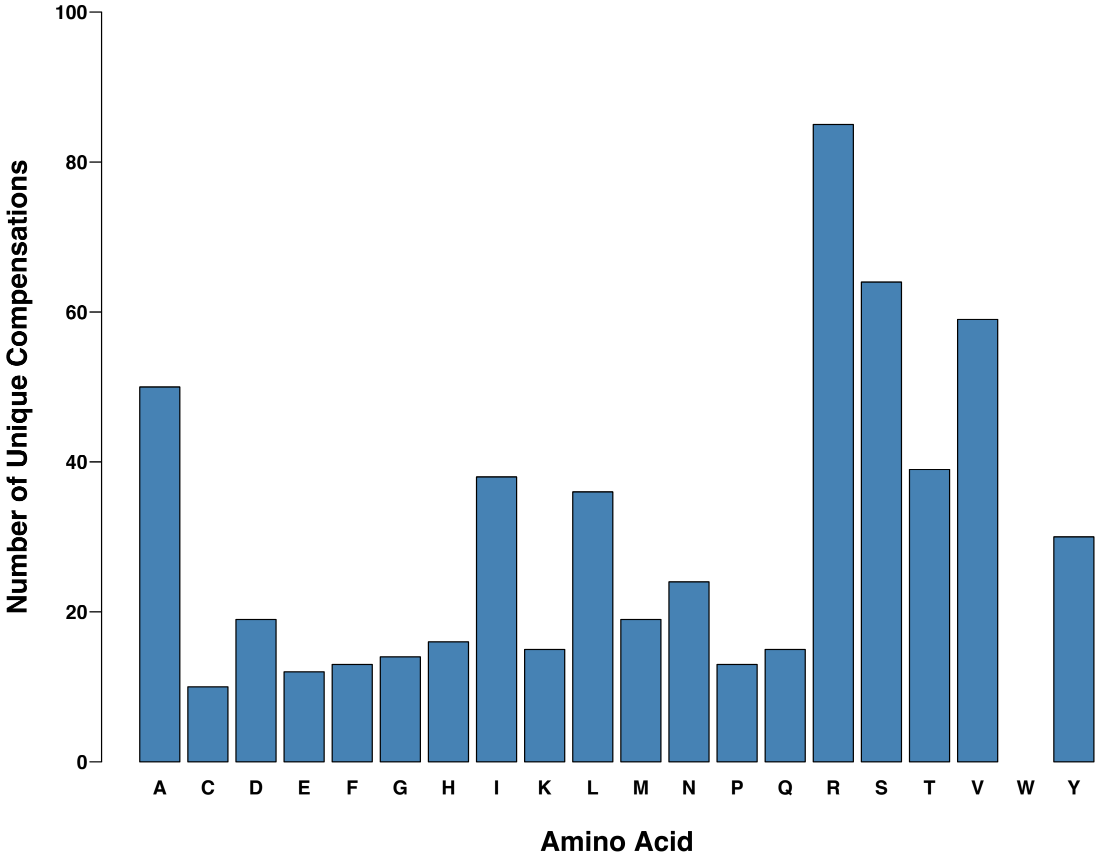


Figure 3.25: Numbers of unique occurrences of potential direct amino acid compensations within Non-Synonymous Proximal Combinations.

The lack of observed direct tryptophan compensations is likely a combination of the fact that tryptophan occurs at a low frequency in proteins compared to other amino acids, with only 1.29% of total residues in the current release of UniProt (Bateman et al. 2015), and the fact that there is only a single codon that encodes tryptophan (TGG), which limits the number of ways that a mutation to tryptophan can occur. In the current release of UniProt cysteine residues are rarer than tryptophan residues across the complete database, accounting for only 1.20% of total residues (Bateman et al. 2015), and this is reflected in the fact that cysteine compensations are the second rarest type of direct compensation observed. However, cysteine has two codons that encode it (TGC and TGT), which makes it more likely that a random mutation results in a cysteine residue and therefore direct compensation events are more likely than for tryptophan residues.

Across the dataset, amino acid composition was only weakly correlated with the number of compensations observed for each amino acid (r=0.46), but a relatively high correlation was observed between the number of compensations observed for each amino acid and the number of codons that each amino acid is encoded by (r=0.77; Appendix 2 Figure 7.I&K). This high correlation with the number of codons for a given amino acid is logical. If a random mutation occurs, an amino acid with more codons that encode it is more likely to be introduced than an amino acid with fewer codons that encode it. Therefore, direct compensation is more likely for amino acids with more encoding codons.

The previously discussed non-synonymous Proximal Combinations in the protein Q17RR3 (‘V381I,R382G’, ‘A380R,V381I,R382G’, and ‘A380R,R382G; Figure 3.12; see Section 3.4.5) represent a complex case of a potential direct amino acid compensation. The 87 individuals with the A380R variant have a potential compensation of the arginine residue lost by the variant R382G. However, the majority of the individuals (4,763) with the R382G variant have no potential arginine compensation.

A less complex example of a direct amino acid compensation occurs in the protein Ret finger protein-like 4A (UniProt accession: A6NLU0). The function of this protein is not well characterised, but it has been experimentally observed, and is known to contain a degenerate zinc finger RING-type domain (Bateman et al. 2015) (often associated with proteins involved in the ubiquitination pathway). In this protein, there are 1,186 occurrences of the variant combination ‘A6NLU0:S276A,A277S’ (Figure 3.26.A), where the serine and alanine losses are both directly compensated for by serine and alanine gains. There are only two occurrences of S276A without A277S, and one occurrence of A277S without S276A. Additionally, this combination is skewed between the super populations, with 43.08% from AFR, 10.29% from AMR, 9.27% from EUR, 12.98% from EAS, and 24.37% from SAS.

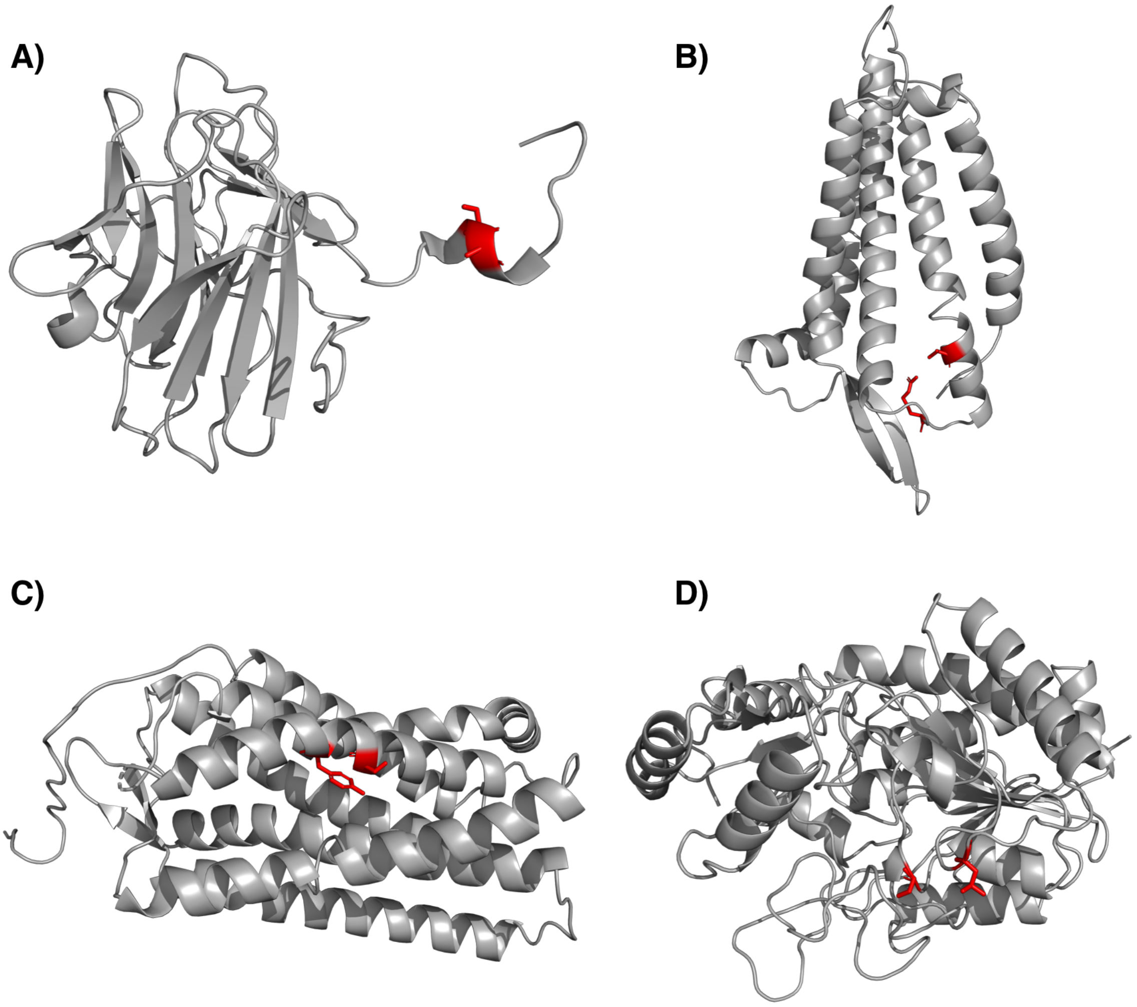


Figure 3.26: Four examples of potential variant combination property compensations. Proteins are shown in cartoon format and are coloured grey, and variant positions are shown in stick format and are coloured red. **A)** The potential direct amino acid compensation in the combination ‘A6NLU0:S276A,A277S’. **B)** The potential positive charge compensation in the combination ‘Q9H339:T78K,R88G’. **C)** The potential functional group compensation in the combination ‘Q8NGA0:S249F,Y252C’. **D)** The potential direct amino acid compensation combined with a potential charge/functional group compensation in the combination ‘Q8WXQ8:L336S,S378G’.

Where direct amino acid exchanges were not observed, we considered the charges and the functional groups of the wild type and variant amino acids to identify indirect compensations via preservation of amino acid properties (see Methods). For example, a charge compensation could be the maintenance of a positive charge or of a hydrophobic environment, and a functional group compensation could be the maintenance of a hydroxyl group or a phenyl group. There may be other ways that variants can compensate for each other but these are not easily measurable. Charge and functional group compensations were not counted if they were caused by a direct amino acid compensation, e.g. a direct compensation of an arginine residue is also a positive charge compensation, but would be counted only as a direct amino acid compensation.

Of the 4,365 unique Non-Synonymous Proximal Combinations, 474 contain potential charge compensations, and 717 contain potential functional group compensations (Figure 3.24 and Table 3.13). Functional group compensations are therefore the most common type of potential compensation observed, followed by direct amino acid compensations, and charge compensations are the least common (Figure 3.24 and Table 3.13).

For the five possible charge compensations (positive, negative, positive polar, negative polar, and hydrophobic), hydrophobic compensations are by far the most common type, with 262 unique hydrophobic compensations (55.27% of all charge compensations; Figure 3.27). This may be expected due to hydrophobic amino acids being the most common type of amino acid charge (Table 3.1). Each of the other four possible charge compensations were observed at least once: 72 positive, 19 negative, 93 positive polar, 28 negative polar (Figure 3.27).

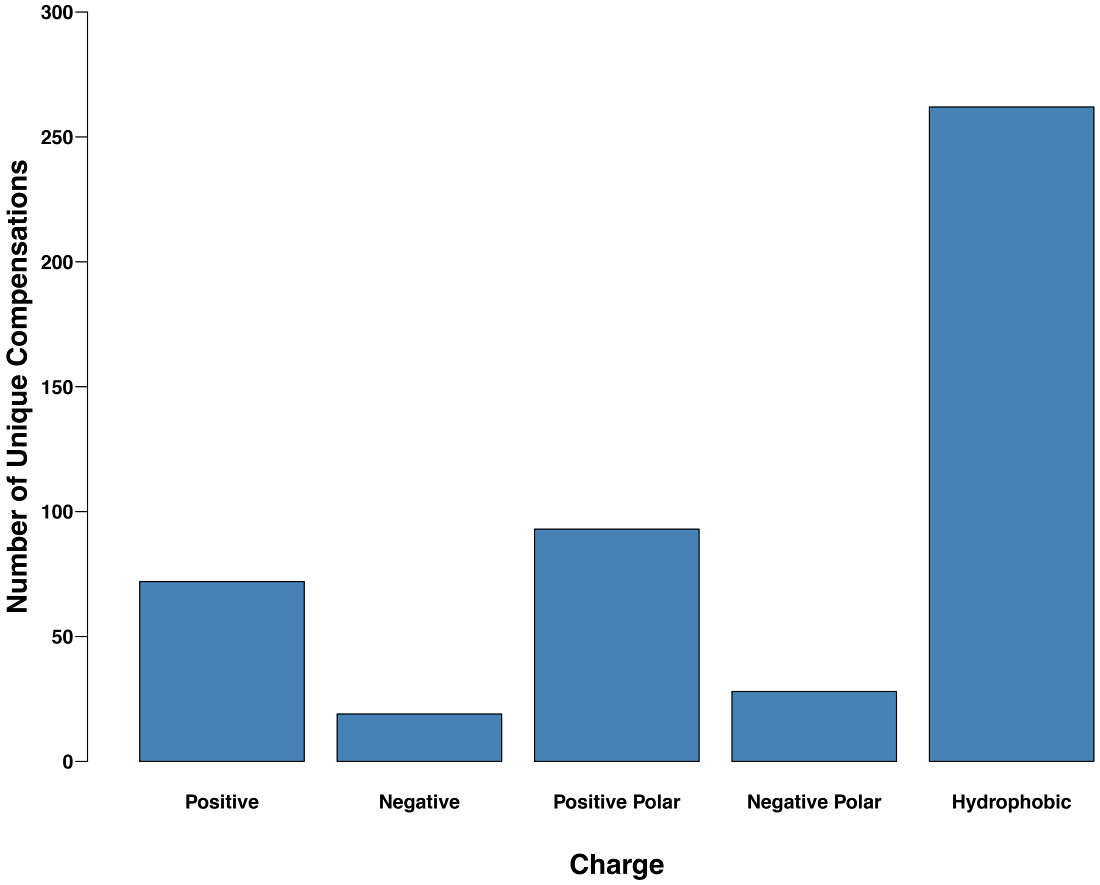


Figure 3.27: Numbers of unique occurrences of potential charge compensations within Non-Synonymous Proximal Combinations.

The previously discussed combination ‘Q9H339:T78K,R88G’ (see Section 3.4.5), of note because 82.7% of the 486 total occurrences are within the AFR super population, is one example of a potential charge compensation event. A positive charge is lost in the R88G variant, but a positive charge is gained in the T78K variant, which maintains a positively charged residue in this region of the protein structure (Figure 3.26.B). Q9H339 corresponds to the protein Olfactory receptor 51B5, which is involved in the sensory perception of smell. Both T78K and R88G occur in an extracellular domain of the protein (Bateman et al. 2015), therefore these residues may be involved with ligand binding necessary for olfactory perception, and loss of the positively charged arginine residue without compensation by the gain of the positively charged lysine residue could result in altered ligand binding. Across the data set, there is only one occurrence of T78K without R88G, and only 42 occurrences of R88G without T78K.

For the six possible functional group compensations (positive, carboxylate, amido, hydroxyl, phenyl, and hydrophobic), hydroxyl (250 unique; 34.86% of total) and hydrophobic (244 unique; 34.03% of total) stand out as by far the most common two types (Figure 3.28). Each of the other four possible functional group compensations were observed at least once: 104 phenyl, 72 positive, 28 amido, and 19 carboxylate (Figure 3.28).

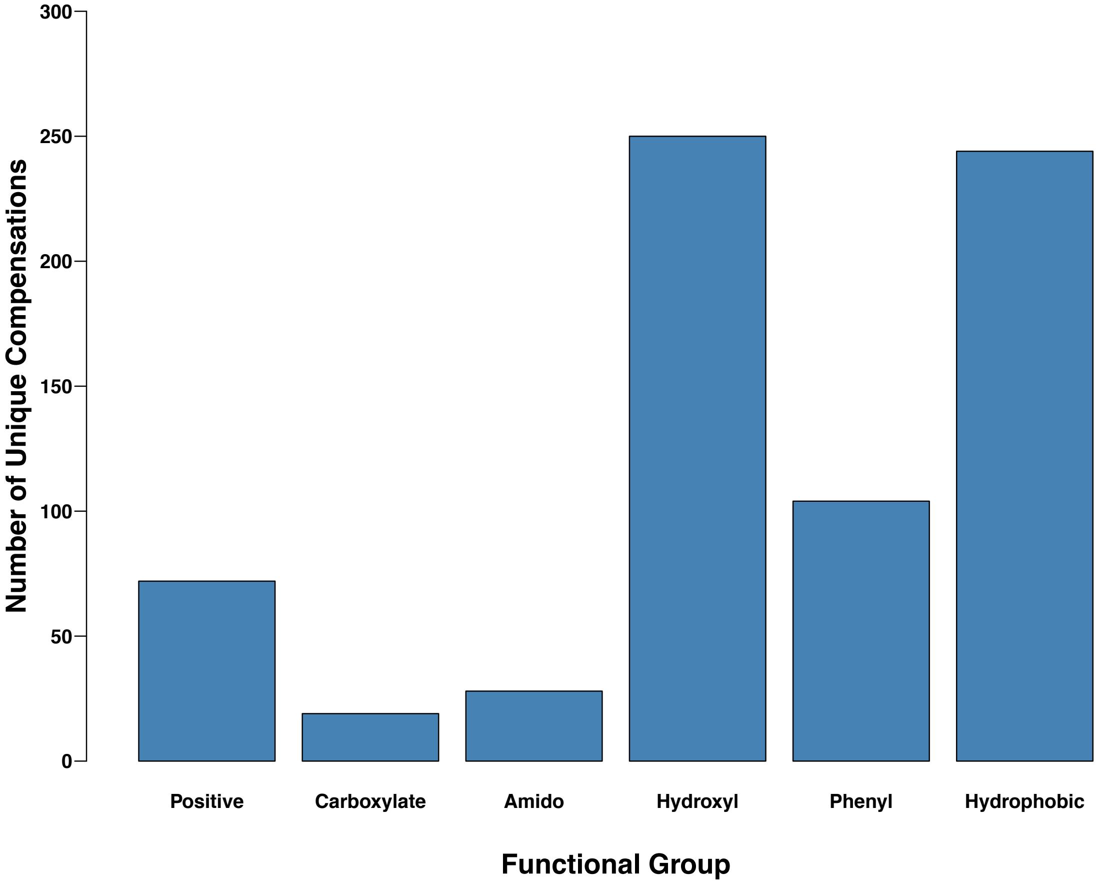


Figure 3.28: Numbers of unique occurrences of potential functional group compensations within Non-Synonymous Proximal Combinations.

In the protein Olfactory receptor 7G1 (UniProt accession: Q8NGA0), the combination ‘Q8NGA0:S249F,Y252C’ occurs 139 times (Figure 3.26.C), and is one example of a potential functional group compensation – with the phenyl group lost with the variant Y252C potentially compensated for by the gain of the phenyl group with S249F. Interestingly, in this example two hydroxyl groups are lost within close spatial proximity without replacement. This combination is most common in the AFR and SAS super populations, and is very rare in the AMR and EUR super populations (44.60% from AFR, 1.44% from AMR, 0.00% from EUR, 7.19% from EAS, and 46.76% from SAS).

Variant combinations can involve potential combinations of different compensation types. The previously discussed combination ‘Q8WXQ8:L336S,S378G’ in the protein Carboxypeptidase A5 (UniProt accession: Q8WXQ8; see Section 3.4.5) is an example of a potential direct amino acid compensation and a charge/functional group compensation (Figure 3.26.D). In this example, one serine residue is lost and another is gained, and one small hydrophobic amino acid is lost and another is gained (leucine is lost and glycine is gained). This combination occurs 273 times, and predominantly occurs in the AFR super population (Figure 3.17; 88.64% in AFR, 7.33% in AMR, 3.30% in EUR, 0.00% in EAS, and 0.73% in SAS).

Overall, a potential functional compensation involving direct compensation of one or more amino acids, compensation of one or more charge types, or compensation of one or more functional groups was identified for 1,875 Proximal Combinations across 542 proteins. This corresponds to potential compensatory effects of variants in 42.96% of the 4,365 total Proximal Combinations identified (Figure 3.24 and Table 3.13).

3.4.8 Variant Combination Stability Effects

If variants compensate for each other, this may be to maintain protein stability, i.e. observed combinations of non-synonymous variants frequently occur together because they better preserve the wild type stability of the protein structure, compared to sub-combinations of constituent variants. Foldx (Schymkowitz et al. 2005) was used to analyse the effects of variant combinations on protein stability. Starting with the wild type protein, each individual variant was introduced separately and the stability effect of each was estimated. This process of introducing variants was then expanded to all possible sub-combinations of the observed variant combination, and finally the stability effect of the observed combination was predicted (see Methods).

For the analysis of protein stability, Proximal Combinations containing only two variants were separated from larger combinations. This was done because there are no sub-combinations for combinations containing only two variants. Figure 3.29 shows the distributions of the stability change predictions for combinations containing ≥2 variants, Figure 3.29.A shows single variants on their own, Figure 3.29.B shows all sub-combinations, and Figure 3.29.C shows the observed variant combinations.

Single variants are most closely clustered around zero (Figure 3.29.A), with a median value of +0.48 kcal/mol, although there are clearly some single variants with large effects on protein stability. Sub-combinations have a median value of +4.42 kcal/mol (Figure 3.29.B), and full combinations have a median value of +3.97 kcal/mol (Figure 3.29.C). The distribution of single variants is significantly lower than the distribution of full combinations (*p*=2.20e-16, Wilcoxon rank sum test), and from the sub-combinations (*p*=2.20e-16). The distributions of full combinations of variants is significantly lower than the sub-combinations (*p*=6.75e-05).

Figure 3.30 shows a comparison of the stability effect distributions of single variants and full combinations (both variants) in Proximal Combinations containing only two variants. Sub-combinations (single variants) in these Proximal Combinations have a median value of +0.18 kcal/mol (Figure 3.30.A), with the full combinations having a median value of +0.49 kcal/mol (Figure 3.30.B). The distribution of full combinations is significantly higher than that of the sub-combinations for these Proximal Combinations containing two variants (*p*=4.36e-13).

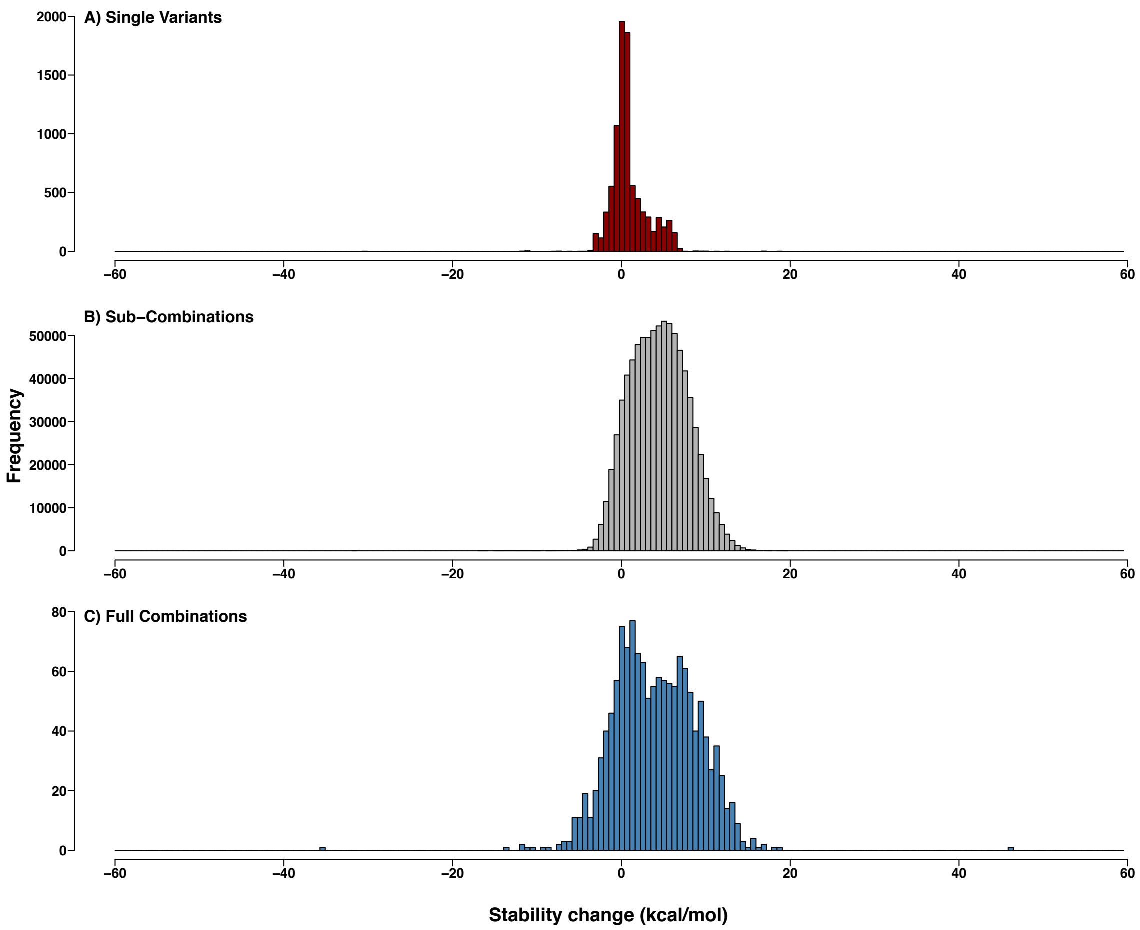


Figure 3.29: Foldx protein stability change prediction distributions for Non-Synonymous Proximal Combinations containing three or more variants. **A)** Single variants within the observed variant combinations. **B)** All sub-combinations for the observed variant combinations. **C)** Full observed variant combinations. Note – the y-axes differ in their scales between the subplots.

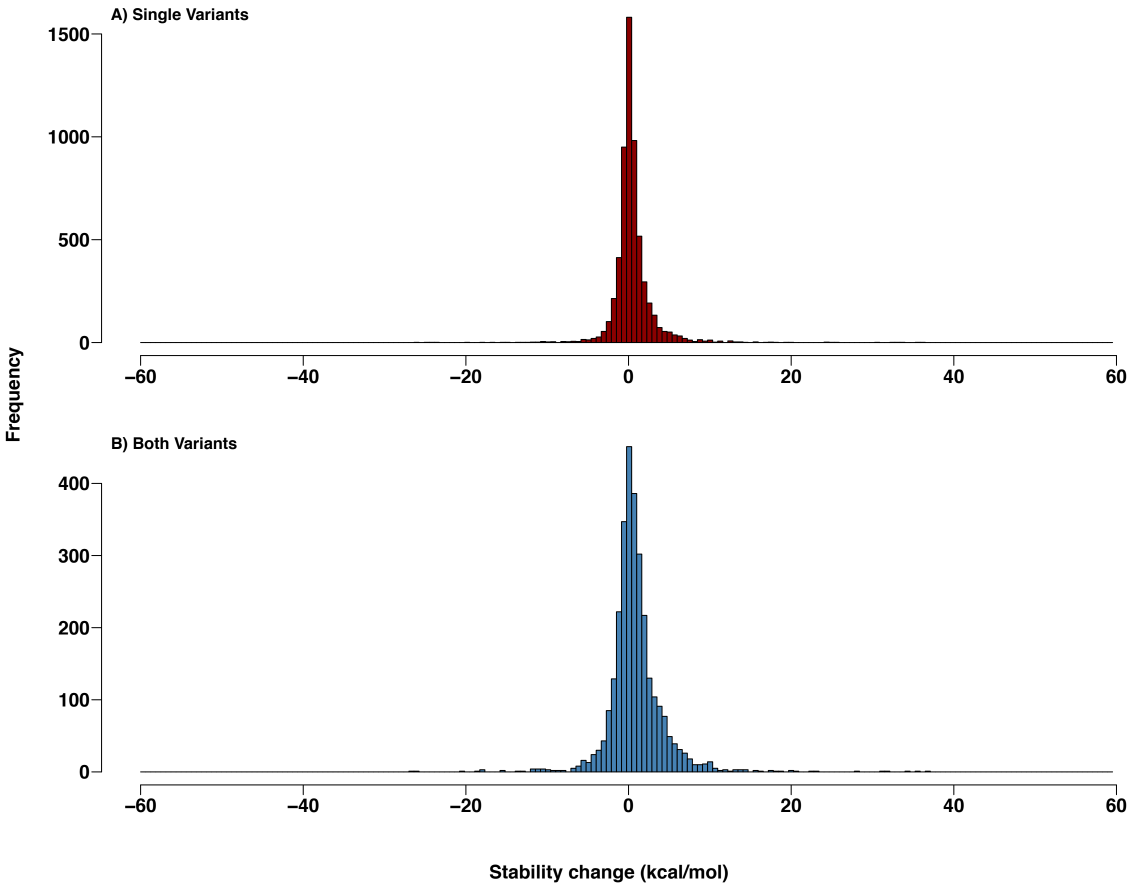


Figure 3.30: Foldx protein stability change prediction distributions for Non-Synonymous Proximal Combinations containing two variants. **A)** Single variants within the observed combination. **B)** Both variants together in the observed combination. Note – the y-axes differ in their scales between the subplots

In these distributions it is clear that many single variants, sub-combinations and full combinations are predicted to have highly destabilising or stabilising effects that could alter protein structure and function (Figure 3.29 and Figure 3.30). To understand if the observed Proximal Combinations better preserve the stability of the wild type protein than sub-combinations of constituent variants, it is essential to compare combinations and sub-combinations directly for each protein. The stability change for the full combination of variants was compared to the highest and lowest stability changes of sub-combinations of variants (Table 3.14). For each Proximal Combination, the full combination stability change was lower than the highest stability change for any sub-combination for 3,223 combinations (74.1%), and higher for 1,125 combinations (25.9%; 1,258 (90.5%) and 132 (9.5%)

respectively for Proximal Combinations containing >2 variants). Full combination stability changes were lower than the lowest stability change for any sub-combination for 649 combinations (14.9%), and higher for 3,699 combinations (85.1%%; 57 (4.1%) and 1,333 (95.9%) respectively for Proximal Combinations containing >2 variants).

Table 3.14: Comparisons of the predicted stability change of full combinations vs the predicted stability changes of the highest and lowest changes from sub-combinations, for all Non-Synonymous Proximal Combinations and also Non-Synonymous Proximal Combinations >2 variants in size (not pairs)

Category	Full Combination Change < Highest Single Change	Full Combination Change > Highest Single Change	Full Combination Change < Lowest Single Change	Full Combination Change > Lowest Single Change
All Proximal Combinations	3,223	1,125	649	3,699
Proximal Combinations with >2 Variants	1,258	132	57	1,333

Comparison of the absolute stability changes of variant combinations, distance from 0 kcal/mol predicted stability change regardless of if the values are positive or negative, showed that 2,674 full combinations of variants (61.5%) are predicted to be closer to the wild type stability of the protein than one or more sub-combination of variants, and 1,674 (38.5%) are further from the wild type stability than the most severe sub-combination of variants (1,208 (86.9%) and 182 (13.1%) respectively for Proximal Combinations containing >2 variants; Figure 3.31 and Table 3.15). Therefore, for the majority of Proximal Combinations there are sub-combinations of variants that are predicted to cause a larger change in structural stability (stabilising or destabilising), indicating possible selection of combinations of variants that preserve the wild type protein stability (Figure 3.31.A). This is more pronounced for the Proximal Combinations containing >2 variants, where 86.91% of combinations are closer to the wild type protein stability than one or more of the sub-combinations of variants (Figure 3.31.B).

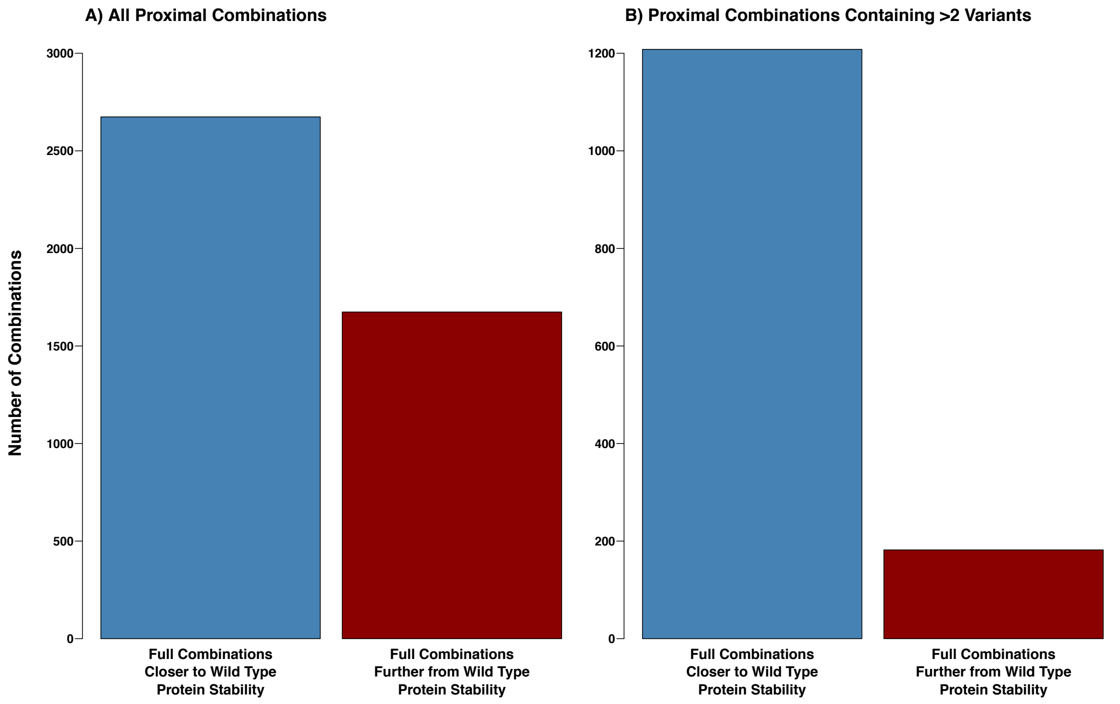


Figure 3.31: Comparison of absolute stability change prediction values for full combinations and the sub-combinations with the largest predicted change in stability. **A)** All Non-Synonymous Proximal Combinations. **B)** Non-Synonymous Proximal Combinations containing >2 variants. Note – y-axes differ in their scales between the subplots.

Table 3.15: Comparisons of the stability change predictions of full combinations vs the stability change predictions of the highest and lowest changes from sub-combinations, in absolute values (distance from 0 kcal/mol predicted change regardless of positive or negative values (destabilising or stabilising))

Category	Full Combination Change Closer to Zero than Highest Single Change	Full Combination Change Further from Zero than Highest Single Change
All Proximal Combinations	2,674	1,674
Proximal Combinations with >2 Variants	1,208	182

As previously discussed, there are a large number of combinations observed in the protein Q31610 (see Section 3.4.5; Figure 3.10 and Table 3.8). The combination ‘Q31610:S35A,V36M,S48A,A93T,Q94N,A95T,D98Y,S101N ,L105A,L119W,S121T,Y140F’ was observed 21 times (Figure 3.32.A). This combination of 12 variants is predicted to have a stability change of +4.53 kcal/mol, and is therefore predicted to be destabilising. However, one of the sub-combinations within this combination (‘Q31610:S11A,A69T,Q70N,A71T,D74Y ,S77N,L95W') is predicted to have a far more destabilising effect, with a stability change of +19.21 kcal/mol. This suggests a potential stability compensation for the full variant combination observed.

One of the largest predicted stability differences between a full combination and a sub-combination was observed in the protein Tryptase gamma (UniProt accession: Q9NRR2), which is a serine protease. There was a single occurrence of the combination ‘Q9NRR2:G224W,T239I’ (Figure 3.32.B), which is predicted to be quite highly destabilising (+10.13 kcal/mol). However, G224W on its own is predicted to be far more destabilising (+30.52 kcal/mol). On its own, T239I is predicted to be mildly stabilising (-2.51 kcal/mol), and is predicted to buffer the destabilising effect of G224W. Interestingly, T239I is a commonly occurring variant (4,395 total occurrences in 1,000 Genomes Project samples, and an allele count of 222,599 in GnomAD), but G224W is very rare, only occurring once and in conjunction with T239I, where perhaps its effects are mitigated (zero occurrences in GnomAD).

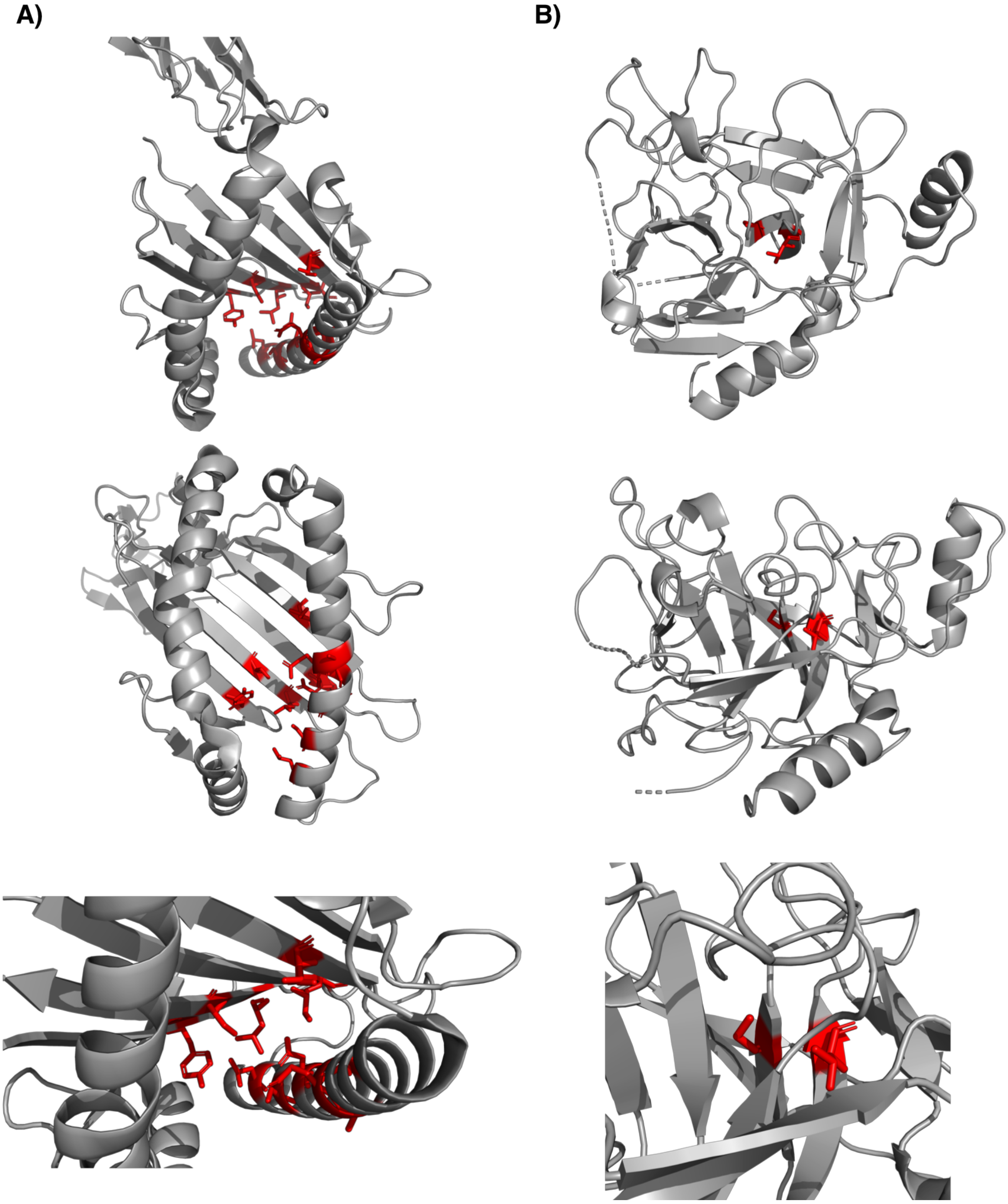


Figure 3.32: Examples of Proximal Combinations with large differences between the predicted stability change of the full combination and one or more sub-combinations. Proteins are shown in cartoon format and are coloured grey, and variant positions are shown in stick format and are coloured red. Each combination is presented from two different sides and then zoomed in. **A)** The variant combination 'Q31610:S35A,V36M,S48A,A93T,Q94N,A95T,D98Y,S101N,L105A,L119W ,S121T,Y140F'. **B)** The variant combination 'Q9NRR2:G224W,T239I'.

The destabilising/stabilising effects of variants can be in part due to differences in the masses of the amino acids, for example replacement of a glycine residue (average atomic mass 57.05) with a tryptophan residue (average atomic mass 186.21) is an increase in average atomic mass of +129.16. Therefore, a G>W variant increases the mass at given structure position by more than the average mass of a single amino acid (118.89). Whereas, other variants are far more conservative in terms of change in atomic mass, e.g. an N>D variant increases the atomic mass by +0.99.

To investigate potential mass compensations of non-synonymous Proximal Combinations, the total atomic masses of wild type amino acids and variant type amino acids in each combination were compared (see Methods). Importantly, the potential mass differences for combinations of variants is much higher than that of single variants. The distribution of total atomic mass changes is shown in Figure 3.33.A, with a mean change of +14.84, median change of +9.09, minimum change of -228.29, and a maximum change of +345.3998. The majority of combinations (84.8%) have a total atomic mass change between -100 and +100, less than the difference caused by a single G>W variant (3,700 combinations; 2,548 combinations between -50 and +50 (58.4%)).

Additionally, comparison of the total atomic mass change of each combination with the number of occurrences of the combination showed that the majority of combinations with more extreme atomic mass differences are rare (Figure 3.33.B). Together these data suggest that the majority of the observed non-synonymous Proximal Combinations are fairly conservative in terms of mass change, and that there may be coevolutionary relationships between residues that conserve total mass within localised areas of protein structures.

Importantly, the mass of the amino acids is only a proxy for the volume of the amino acid, which is a key factor in potential instability introduced by missense mutations. Each of the amino acid features within combinations compared here (charge, mass, and functional group) are inadequate on their own to describe the properties of a variant combination. These properties must be considered in a combined manner, and in some cases, there will be factors unique to individual proteins (such as ligand-binding sites) that will complicate the interpretation of the amino acid property changes within variant combinations.

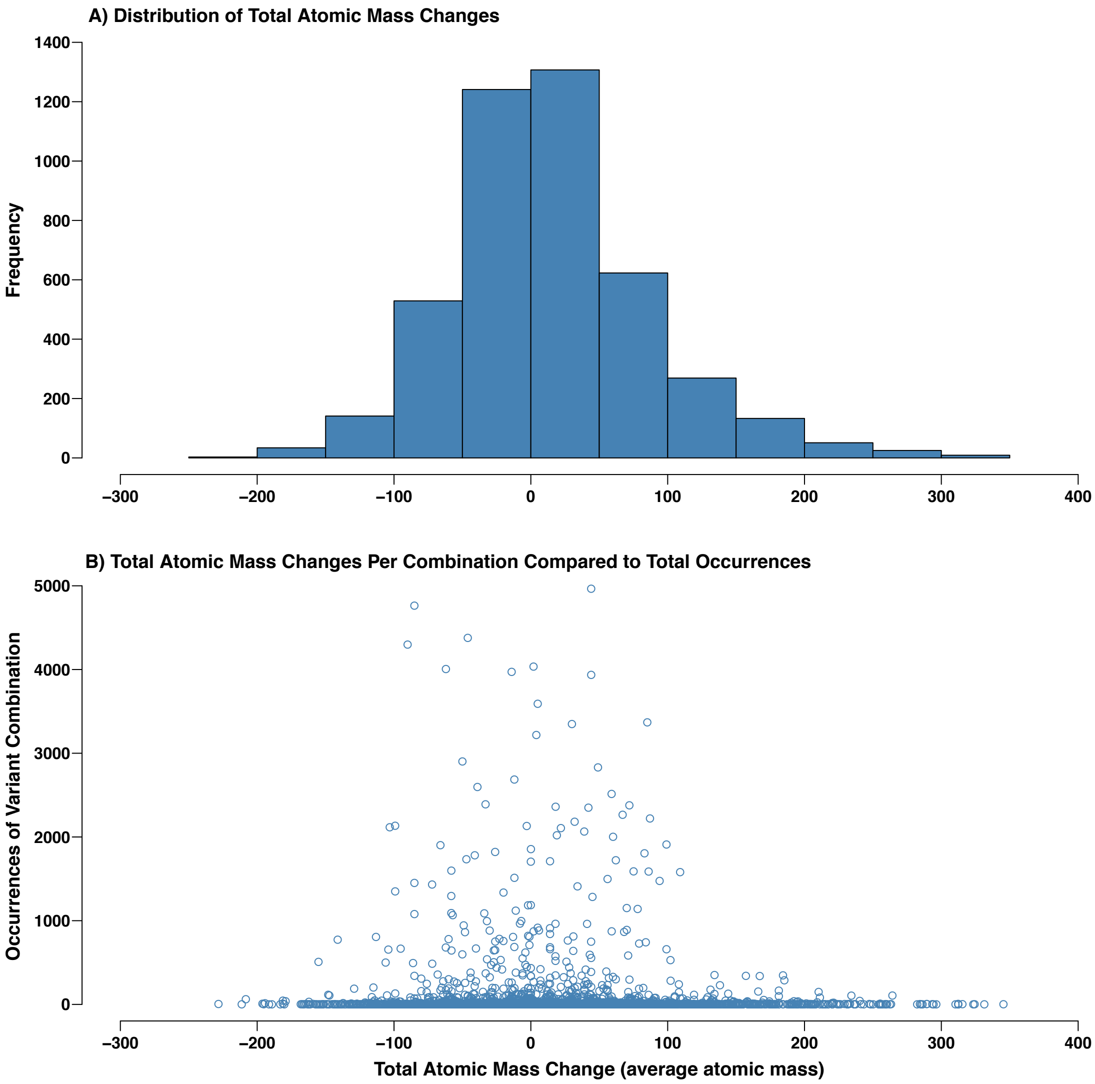


Figure 3.33: Total mass changes for Non-Synonymous Proximal Combinations. **A)** The distribution of total mass changes. **B)** Comparison of the total mass change of each combination to the total number of occurrences of the combination.

3.4.9 Clinical Significance of Proximal Combination Variants

To identify potential compensation/buffering effects of known deleterious variants by other co-occurring known benign variants, clinical annotations from Humsavar and ClinVar were used to annotate non-synonymous Proximal Combinations (see Methods). The vast majority of these combinations are composed solely of variants annotated as benign and/or unclassified (Table 3.16), highlighting the need for better clinical annotation of variants.

Table 3.16: Combined clinical annotations of all Non-Synonymous Proximal Combinations.

Combination Classification	Humsavar	ClinVar
Pathogenic	1	0
Pathogenic & Benign	2	0
Pathogenic & Unclassified	3	2
Benign	582	1
Benign & Unclassified	1,935	37
Unclassified	1,842	4,325

Only eight of the combinations contain variants annotated as pathogenic by Humsavar or ClinVar, and the majority of these are rare combinations (Table 3.17). The combination ‘P02649I:C130R,R132C’ (Figure 3.34.A) was identified in the protein Apolipoprotein E. This combination contains the unclassified variant R132C and the pathogenic variant C130R, which is associated with Alzheimer disease 2 (AD2) and Hyperlipoproteinemia 3 (HLPP3) (Bateman et al. 2015). Within this combination R132C could potentially be compensating for the effect of C130R, with this combination being a potential direct amino acid compensation of both arginine and cysteine. This combination is only observed once in the data set. Intriguingly, in GnomAD there are zero occurrences of the pathogenic C130R, but 22,731 occurrences of the R132C variant. Individuals with this R132C variant may be able to buffer some of the negative effects of the C130R variant if it were to arise. However, it should be noted that in this example the side chains of the two residues are not pointing in the same direction, suggesting that any potential buffering effects are indirect. This is due to the fact that the directionality of the side chains is not considered when determining proximal combinations (see Methods).

The most common of these eight combinations that contain variants annotated as pathogenic is ‘Q7RTS7:E271D,F274S’ (Figure 3.34.B), which occurs 87 times (F274S occurs once without E271D, and E271D occurs 3,955 times without

F274S). This combination occurs in the protein Keratin, type II cytoskeletal 74, and occurs almost exclusively in the SAS super population (0.00% in AFR, 0.00% in AMR, 0.00% in EUR, 1.15% in EAS, and 98.85% in SAS). The variant E271D is classified as benign, and the variant F274S is classified as pathogenic, with an association to Ectodermal dysplasia 7, hair/nail type (ECTD7) (Bateman et al. 2015). As with 'P02649I:C130R,R132C', the pathogenic variant (F274S) is rarer than the benign variant (E271D; allele counts of 2,572 and 180,372 respectively in GnomAD). Therefore, E271D could be buffering the effects of this variant in some of these individuals, but potential compensatory effects are less obvious than for the combination 'P02649I:C130R,R132C'.

Table 3.17: Combined clinical annotations of Non-Synonymous Proximal Combinations for combinations containing one or more variant annotated as pathogenic.

Protein	Combination	Total Occurrences	Combined Annotation	Annotation Source
Q7RTS7	E271D,F274S	87	Pathogenic & Benign	Humsavar
P20933	R161Q,C163S	2	Pathogenic	Humsavar
P20933	R161Q,C163S	2	Pathogenic & Unclassified	ClinVar
P23352	V534I,E539K	2	Pathogenic & Benign	Humsavar
O00187	E124K,P126L	1	Pathogenic & Unclassified	Humsavar
P02649	C130R,R132C	1	Pathogenic & Unclassified	Humsavar
Q96EU7	D131E,N202K	1	Pathogenic & Unclassified	ClinVar
Q9BQ52	S217L,L224F	1	Pathogenic & Unclassified	Humsavar

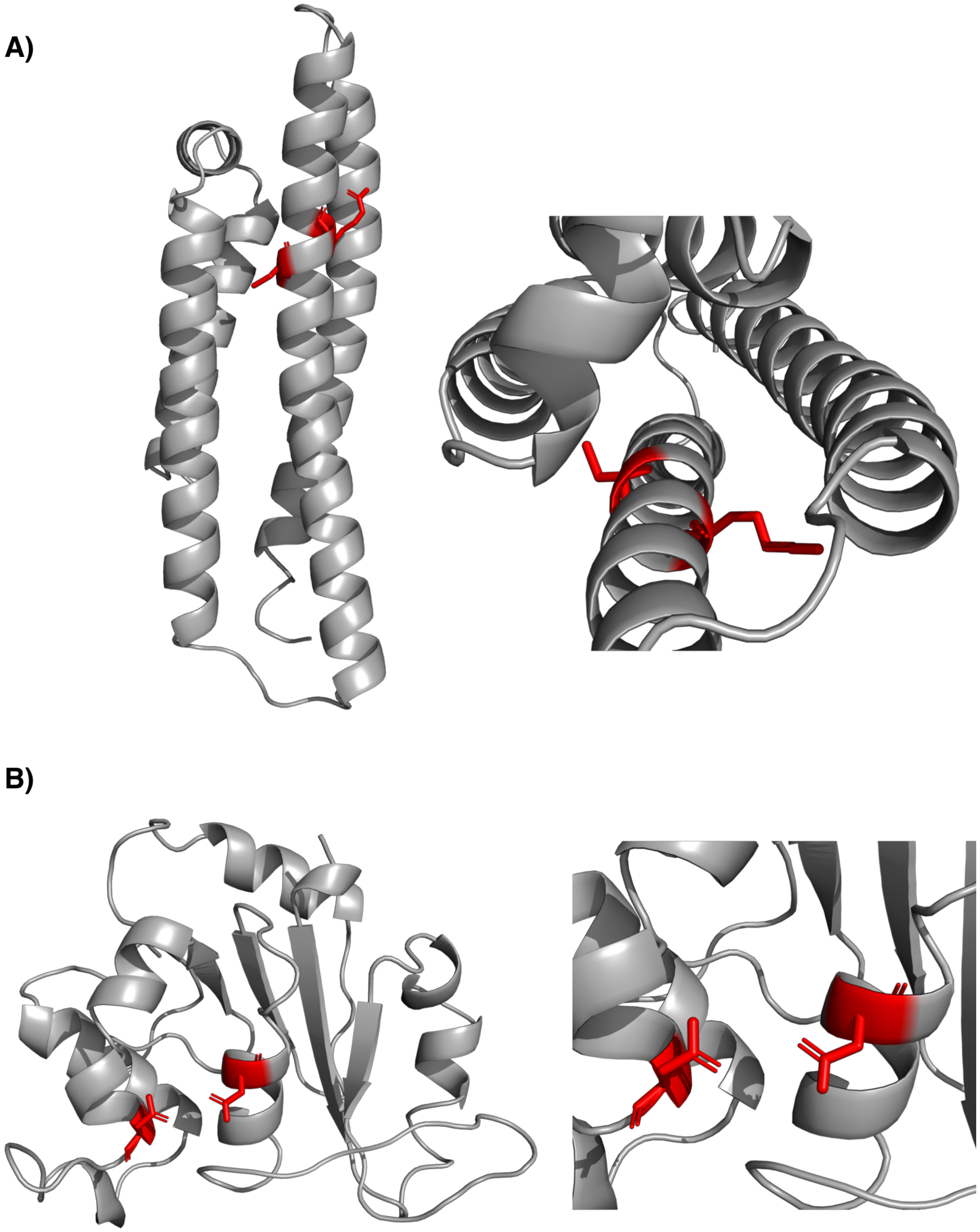


Figure 3.34: Two examples of Non-Synonymous Proximal Combinations containing variants classified as pathogenic by Humsavar or ClinVar. Proteins are shown in cartoon format and are coloured grey, and variant positions are shown in stick format and are coloured red. Each example is shown with the whole structure in view and then zoomed in on the variant combination. **A)** The combination ‘P02649I:C130R,R132C’. **B)** The combination ‘Q7RTS7:E271D,F274S’.

3.4.10 Variant Combinations in Protein-Protein Interfaces

Protein-Protein interactions are essential for cellular function and are a key component of cellular complexity. The human interactome is estimated to contain between ~650,000 (Stumpf et al. 2008) and >1,000,000 (Ranea et al. 2010) unique protein-protein interactions, and as of August 2018 there are 334,684 non-redundant human protein-protein interactions listed in BioGRID build 3.4.163 (Chatr-Aryamontri et al. 2015). Human disease variants have previously been shown to be enriched in protein-protein interface sites (David et al. 2012; Wang et al. 2012), and we are interested in combinations of variants that occur in interfaces that could have compensatory effects. To investigate this, we expanded our analysis of variant combinations within protein structures to combinations of variants in protein-protein interface sites.

We used interaction data for 9,642 protein-protein interactions with residue-level structural characterisation from Interactome3D (Mosca et al. 2013), to study combinations of variants that occur within interface sites (see Methods). Across homomeric and heteromeric protein-protein complexes, we considered three different types of protein-protein interface variant combinations: Homomeric, Heteromeric, and Uni-Partner (Figure 3.4; see Methods). These three combination types cover all possible compensation scenarios between interface variants. Homomeric and Heteromeric combinations involve variants in the interface sites of both partner proteins, with the difference between the two combination types being whether the complex is a homomer or a heteromer (Figure 3.4). Uni-Partner combinations involve multiple variants within the interface site of one of the two partners in an interaction, and can occur in both homomeric and heteromeric complexes (Figure 3.4). We also consider the scenario of a single variant within an interface site, where variant compensation is not possible, and group these as Singletons (Figure 3.4).

3.4.11 Unique Interface Variant Combinations

In total, 6,824 unique non-synonymous and 6,385 synonymous variants were found to be in the 9,642 unique protein-protein interfaces in the Interactsome3D representative set (see Methods for further Interactome3D details). For non-synonymous variants, 592 form at least 1 variant combination and 947 synonymous variants form combinations (Table 3.18). Non-synonymous combinations are spread across 233 proteins and 272 complexes, and across 443 proteins and 449 complexes for synonymous combinations (Table 3.18).

Singleton interface variants are far more common than interface variant combinations, likely due to the probability of one variant occurring being higher than the probability of two variants occurring. There are 6,582 non-synonymous and 6,054 synonymous singleton variants, with 11,285 unique non-synonymous singletons and 10,846 unique synonymous singletons (Table 3.19). There are more unique non-synonymous and synonymous singletons than variants in singletons, because singleton variants can occur in multiple different protein-protein interfaces. Non-synonymous singletons are spread across 2,627 proteins and 4,296 complexes, and across 2,652 proteins and 4,851 complexes for synonymous variants (Table 3.19).

Across all samples, for non-synonymous variants there are 147 Homomeric, 364 Heteromeric, and 735 Uni-Partner combinations, and for synonymous variants there are 108 Homomeric, 672 Heteromeric, and 737 Uni-Partner combinations (Table 3.20). Homomeric combinations being the rarest type of combination for non-synonymous and synonymous variants reflects that they only occur in homomeric protein-protein interactions and only 2,987 of the 9,642 complexes in the Interactome3D representative set are homomeric (see Methods).

Table 3.18: Numbers of constituent variants in combinations, numbers of combinations, numbers of proteins with combinations, and numbers of complexes with combinations, for non-synonymous and synonymous interface variant combinations.

Variant Type	Variants in Combinations	Combinations	Proteins with Combinations	Complexes with Combinations
Non-Synonymous	592	1,246	233	272
Synonymous	947	1,517	443	449

Table 3.19: Numbers of constituent singleton interface variants, numbers of singletons, numbers of proteins with singletons, and numbers of complexes with singletons, for non-synonymous and synonymous interface variant combinations.

Variant Type	Variants in Singletons	Singletons	Proteins with Singletons	Complexes with Singletons
Non-Synonymous	6,582	11,285	2,627	4,296
Synonymous	6,054	10,846	2,652	4,851

Table 3.20: Numbers of unique interface variant combinations types, for non-synonymous and synonymous interface variant combinations.

Variant Type	Heteromeric Combinations	Homomeric Combinations	Uni-Partner Combinations
Non-Synonymous	364	147	735
Synonymous	672	108	737

As with combinations of variants within structures (Figure 3.5), the majority of the interface variant combinations contain only a few variants, with 525 non-synonymous and 890 synonymous combinations containing only two variants (Figure 3.35). However, some interface variant combinations are much larger, containing as many as 12 variants for non-synonymous combinations, and eight variants for synonymous combinations (Figure 3.35).

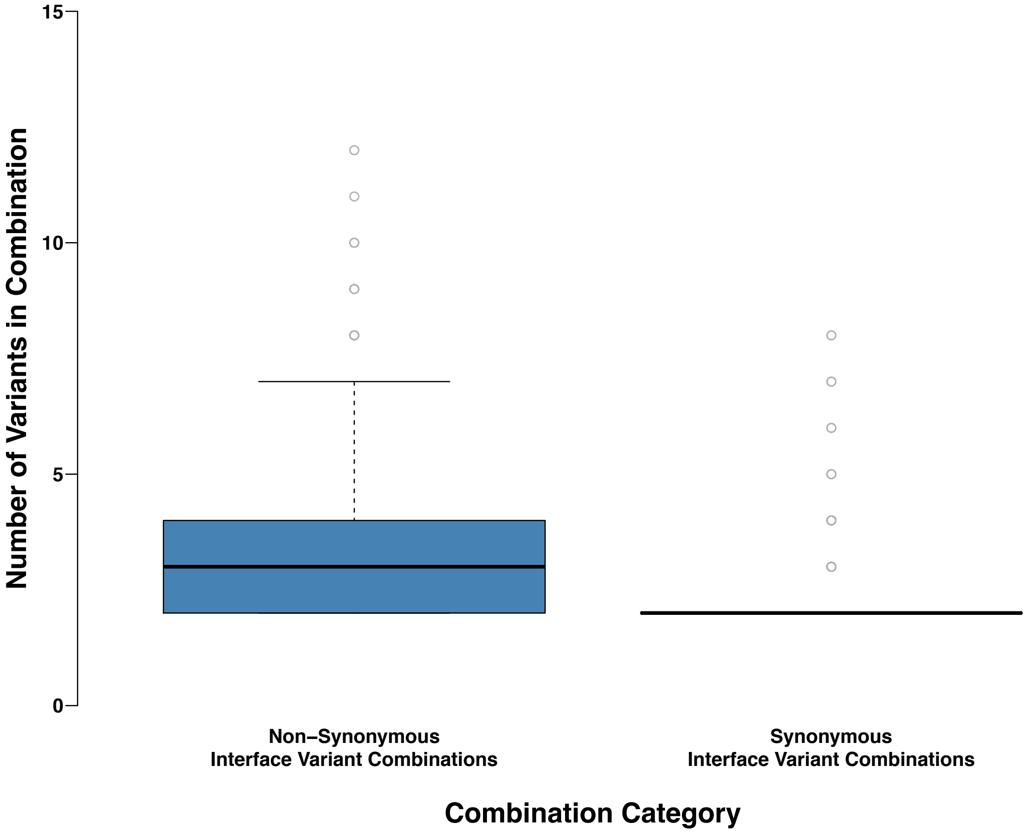


Figure 3.35: Combination sizes for non-synonymous and synonymous interface variant combinations.

One of these larger variant combinations occurs in the heteromeric protein-protein complex P04440:P20036. One individual has the Heteromeric variant combination ‘P04440:L37V,F38Y,G40V,A65V,A85E,E86D,G114E,P115A_P20036:E59D,T103I,L104A,P127A’, which contains 12 variants (Figure 3.36; see Methods for a description of the interface variant combination notation format). The complex P04440:P20036 is a heteromeric interaction between two immune system proteins: HLA class II histocompatibility antigen DP beta 1 chain and HLA class II histocompatibility antigen DP alpha 1 chain. From the full combination of interface variants (Figure 3.36.A), there is a clear cluster of variants in one area of the interaction containing the variants L37V,F38Y,G40V,A65V,A85E,E86D in P04440, and T103I,L104A in P20036, with the remaining variants from the full combination of interface variants in different areas of the interface (Figure 3.36.B).

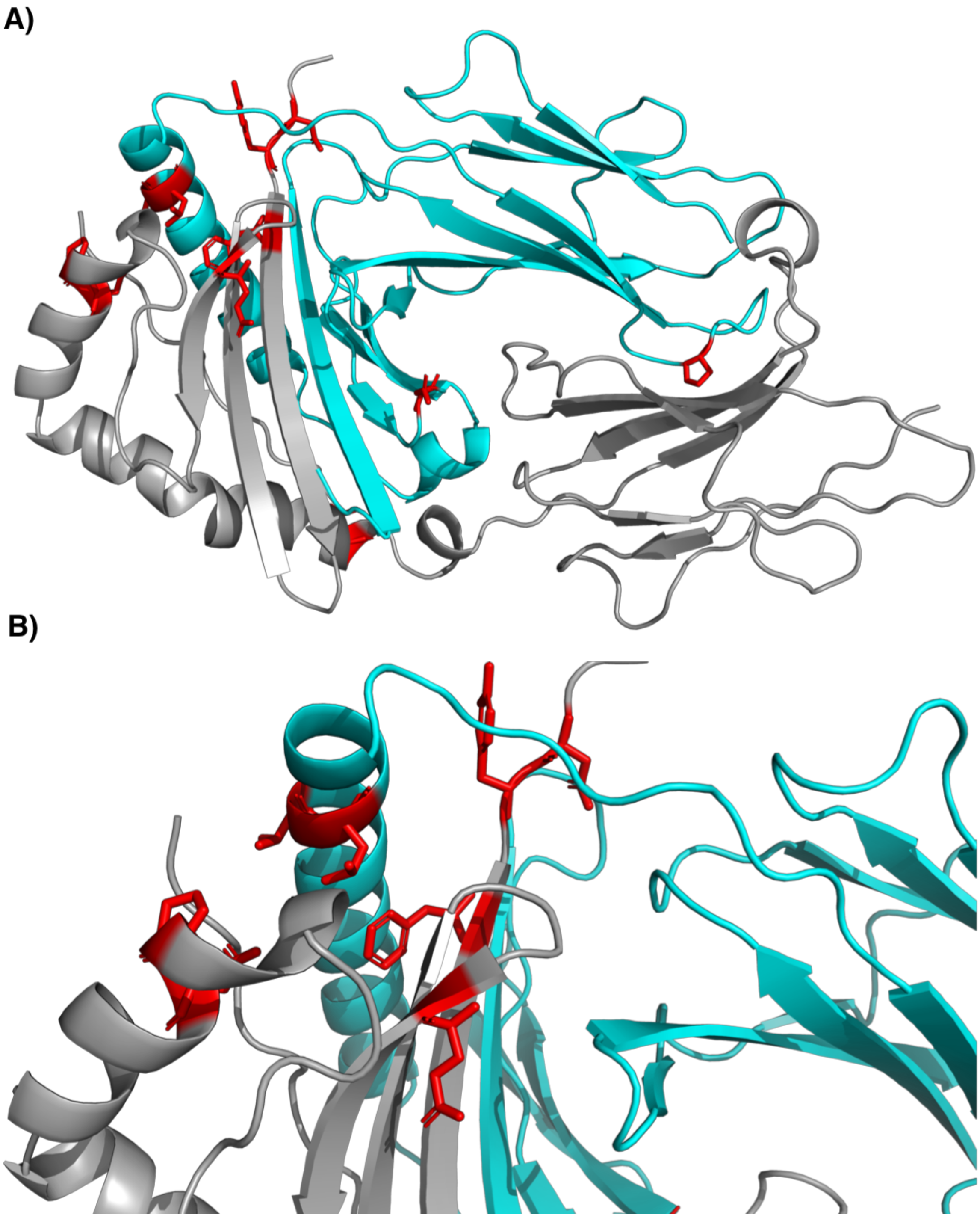


Figure 3.36: Visualisation of the Homomeric interface variant combination ‘P04440:L37V,F38Y,G40V,A65V,A85E,E86D,G114E,P115A_P20036:E59D,T103I,L104A,P127A’, which occurs in the interface P04440:P20036. In all subplots, chain A is coloured grey, chain B is coloured cyan, and positions of variants are shown in stick format and coloured red. **A)** Cartoon representation of the entire interface. **B)** Zoomed in view of the area of the interface containing the most variants.

In the homomeric complex P78324:P78324, there are 209 occurrences of the interface variant combination ‘P78324:T52I,R54H,A57V,D95E,L96S,N100K _P78324:T52I,R54H,A57V,D95E,L96S,N100K’, which is the same six variants (T52I, R54H, A57V, D95E, L96S, N100K) occurring in both partners as this is a homomeric complex (Figure 3.37). This is a complex of two Tyrosine-protein phosphatase non-receptor type substrate 1 proteins, also known as signal-regulator protein (SIRP-α). SIRP-α is a cognate receptor for CD47, and the interaction of the two is involved in regulation of Interleukin-12 levels, possibly as a homeostatic mechanism to prevent escalation of the inflammatory immune response (Latour et al. 2001).

The variants in this combination have a complex set of potentially compensatory effects on one another. The variants R54H and N100K are potentially compensatory, with a large positively charged amino acid lost in R54H and a large positively charged amino acid gained in N100K (arginine loss compensated by lysine gain). The variant T52I results in loss of a hydroxyl group and gain of a hydrophobic residue and the variant L96S potentially compensates with loss of a hydrophobic residue and gain of a hydroxyl group (leucine loss compensated for by isoleucine gain, and threonine loss compensated for by serine gain). Finally, the variants D95E and A57V are both conservative changes, with a negative charge at position 95 maintained in D95E and a small hydrophobic amino acid maintained at position 57 in A57V. Therefore, despite the occurrence of six different variants (some of which are radical changes), the overall properties of the positions in the interface are largely maintained (loss of N and gain of H in N100K R54H are the only changes without potential compensation).

However, as with many interface variant combinations, this variant combination is more complex than it first appears. Interface variant combinations are a result of combined variants in different partners of an interaction, and there may be a mixture of different interface variant combinations for the same complex within one individual. For example, given a homomeric complex of the protein ‘PA’ (PA:PA), if an individual has a wild type form of PA (PA-WT) and a variant form (PA-VAR), then there are two possible interface variant combinations: PA-WT:PA-VAR and PA-VAR:PA-VAR.

For the complex P78324:P78324 there were 66 different interface variant combinations observed (though 45 of these occur 10 or fewer times). The ten most common interface variant combinations for P78324:P78324 are shown in Table 3.21. These ten combinations highlight the complexity of interface variant combinations, with different compensatory effects observed within them, e.g. the most common combination involves the potential hydroxyl compensation of T52I and L96S, but here R54H is occurring without N100K to compensate for the large positively charged amino acid lost. Similarly, the second most common combination results in a net gain of a hydroxyl group in the interface, as L96S is occurring without T52I. This complexity is a major challenge in the analysis of interface variant combinations.

Table 3.21: The ten most common interface variant combinations in the interface P78324:P78324.

Variant Combination	Total Occurrences	Combination Category
P78324:T52I,R54H,A57V,D95E,L96S	1,274	Uni-Partner
P78324:D95E,L96S	1,114	Uni-Partner
P78324:T52I,R54H,A57V,D95E,L96S,N100K	818	Uni-Partner
P78324:T52I,R54H,A57V,D95E,L96S_P78324:T52I,R54H,A57V,D95E,L96S	725	Homomeric
P78324:D95E,L96S_P78324:D95E,L96S	557	Homomeric
P78324:A57V,D95E,L96S	416	Uni-Partner
P78324:T52I,R54H,A57V,D95E,L96S,N100K_P78324:T52I,R54H,A57V,D95E,L96S,N100K	411	Homomeric
P78324:T52I,R54H,A57V,D95E,L96S_P78324:T52I,R54H,A57V,D95E,L96S,N100K	377	Homomeric
P78324:T52I,R54H,A57V,D95E,L96S,N100K_P78324:T52I,R54H,A57V,D95E,L96S	377	Homomeric
P78324:A57V,D95E,L96S_P78324:A57V,D95E,L96S	208	Homomeric

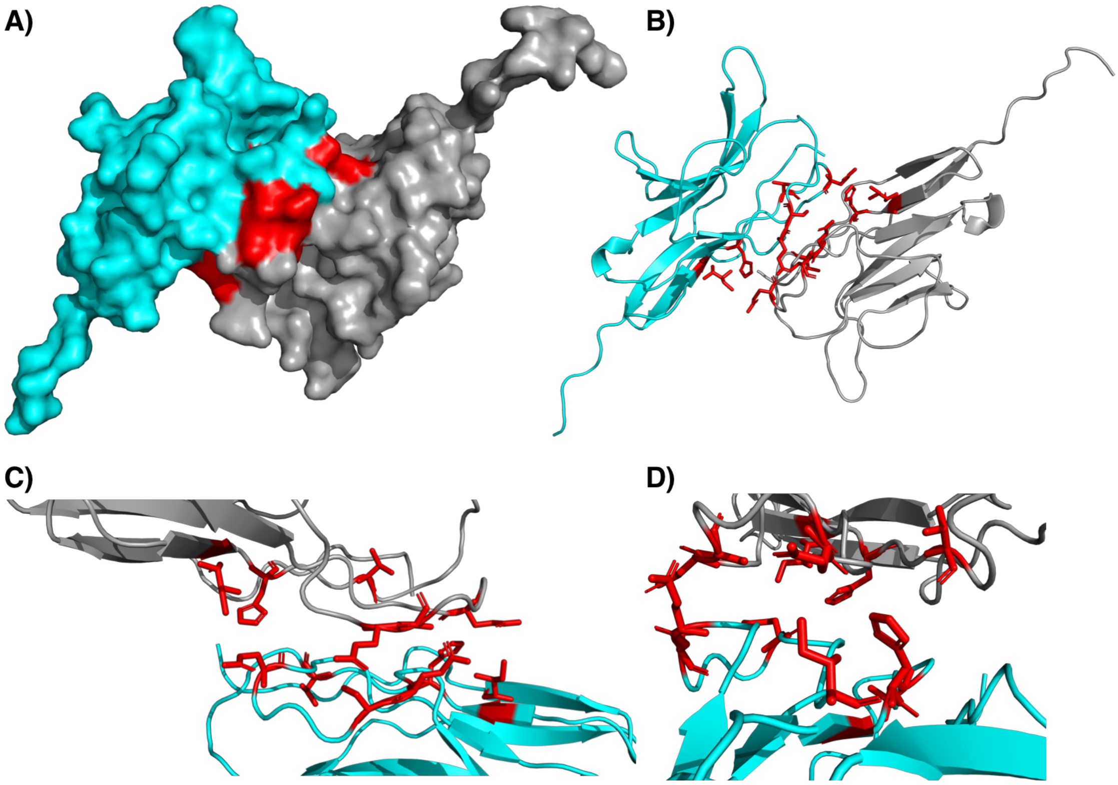


Figure 3.37: Visualisation of the Homomeric interface variant combination 'P78324:T52I,R54H,A57V,D95E,L96S,N100K_P78324:T52I,R54H,A57V,D95E,L96S,N100K', which occurs in the interface P78324:P78324. In all subplots, chain A is coloured grey, chain B is coloured cyan, and positions of variants are shown in stick format and coloured red. **A)** Surface representation. **B)** Cartoon representation. **C)** Zoomed in view A. **D)** Zoomed in view B.

3.4.12 Interface Variant Combinations Per Sample

The numbers of interface variant combinations for each sample is shown in Figure 3.38, split in to the different categories of interface variant combinations. For each interface variant combination category, individuals have more synonymous combinations than non-synonymous combinations, presumably due to the limited functional effects of synonymous variants on interfaces (Figure 3.38). This differs from the combinations of variants observed to be close in space within protein structures, where individuals tended to have more non-synonymous Proximal Combinations (but more synonymous Global Combinations; Figure 3.7 and Table 3.5). For both non-synonymous and synonymous categories, Singletons are the most common type per individual, which reflects the fact that only a single mutation is required. For the three non-synonymous categories involving multiple variants, Uni-Partner combinations are the most common, with Heteromeric and

Homomeric combinations occurring at similar levels. Table 3.22 summarises the per sample interface variant combinations.

For non-synonymous variants, each sample has on average 25.12 total interface variant combinations comprising 4.38 Homomeric Combinations, 2.77 Heteromeric Combinations, and 17.97 Uni-Partner Combinations, and for synonymous variants 34.04 total, 3.29 Homomeric, 10.46 Heteromeric, and 20.29 Uni-Partner (Table 3.22). Each sample has more interface Singletons than combination types with multiple variants, with on average 110.29 non-synonymous Singletons and 417.19 synonymous Singletons.

As with combinations of variants within structures (Figure 3.8 and Table 3.6), samples from the AFR super population have on average more interface variant combinations than samples from the other four super populations (Figure 3.39 and Tables 3.23 & 3.24). This trend is observed for each type of non-synonymous interface variant combination, except for Homomeric combinations (where samples have similar numbers between the five super populations), but the differences between super populations are smaller for synonymous combinations (Figure 3.38 and Tables 3.23 & 3.24).

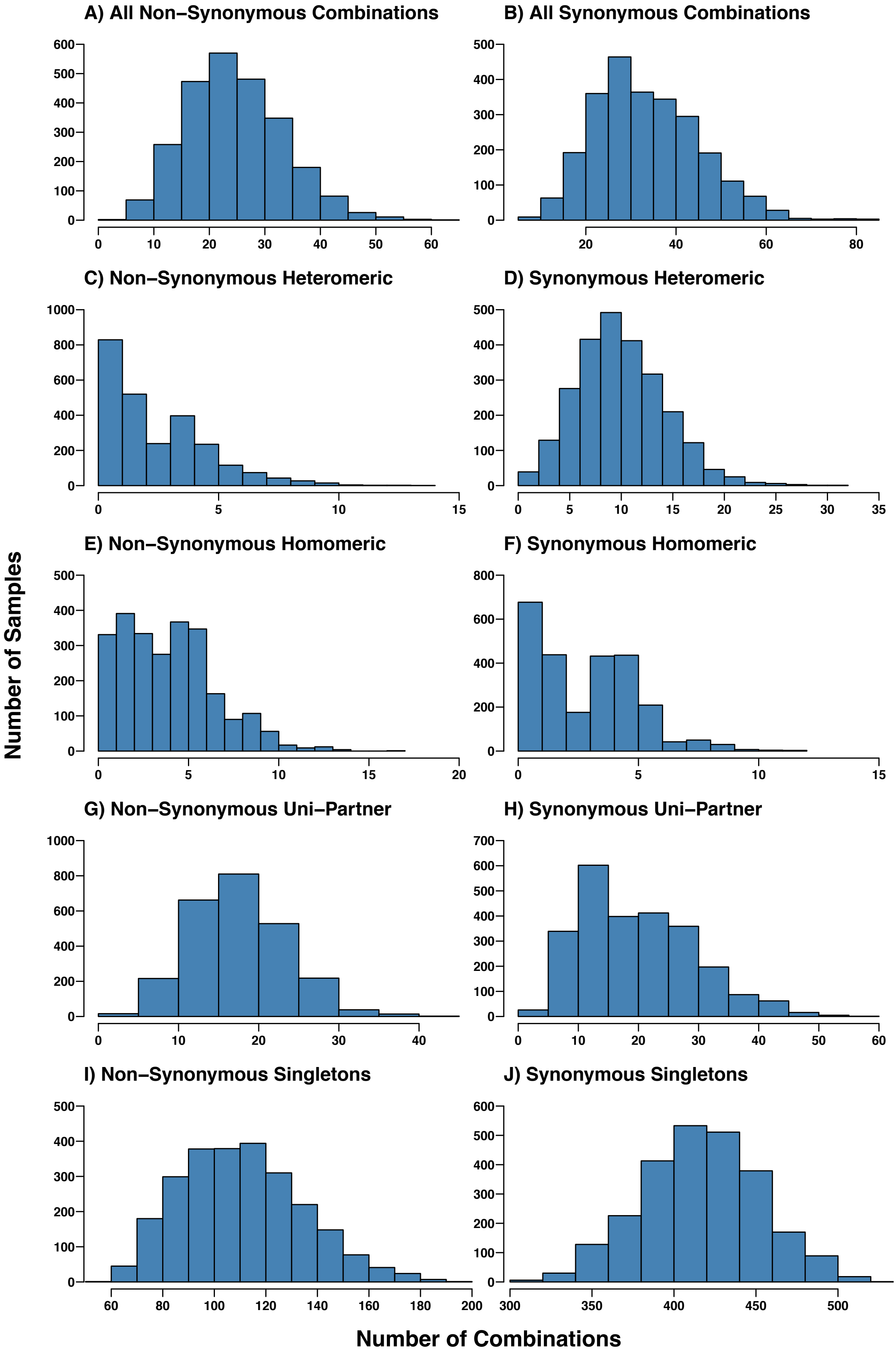


Figure 3.38: (see legend on next page)

Figure 3.38: Number of interface variant combinations observed per sample for each combination category. **A)** All Non-Synonymous Combinations. **B)** All Synonymous Combinations. **C)** Non-Synonymous Heteromeric Combinations. **D)** Synonymous Heteromeric Combinations. **E)** Non-Synonymous Homomeric Combinations **F)** Synonymous Homomeric Combinations. **G)** Non-Synonymous Uni-Partner Combinations. **H)** Synonymous Uni-Partner Combinations. **I)** Non-Synonymous Singletons. **J)** Synonymous Singletons.

Table 3.22: Numbers of interface variant combinations observed per sample.

Combination Category	Mean Number of Combinations	Median Number of Combinations	Minimum Number of Combinations	Maximum Number of Combinations
Non-Synonymous Combinations	25.12	24	4	64
Synonymous Combinations	34.04	33	7	81
Non-Synonymous Heteromeric	2.77	2	0	14
Synonymous Heteromeric	10.46	10	0	31
Non-Synonymous Homomeric	4.38	4	0	17
Synonymous Homomeric	3.29	3	0	12
Non-Synonymous Uni-Partner	17.97	18	3	43
Synonymous Uni-Partner	20.29	19	2	56
Non-Synonymous Singletons	110.29	110	59	192
Synonymous Singletons	417.19	417	307	522

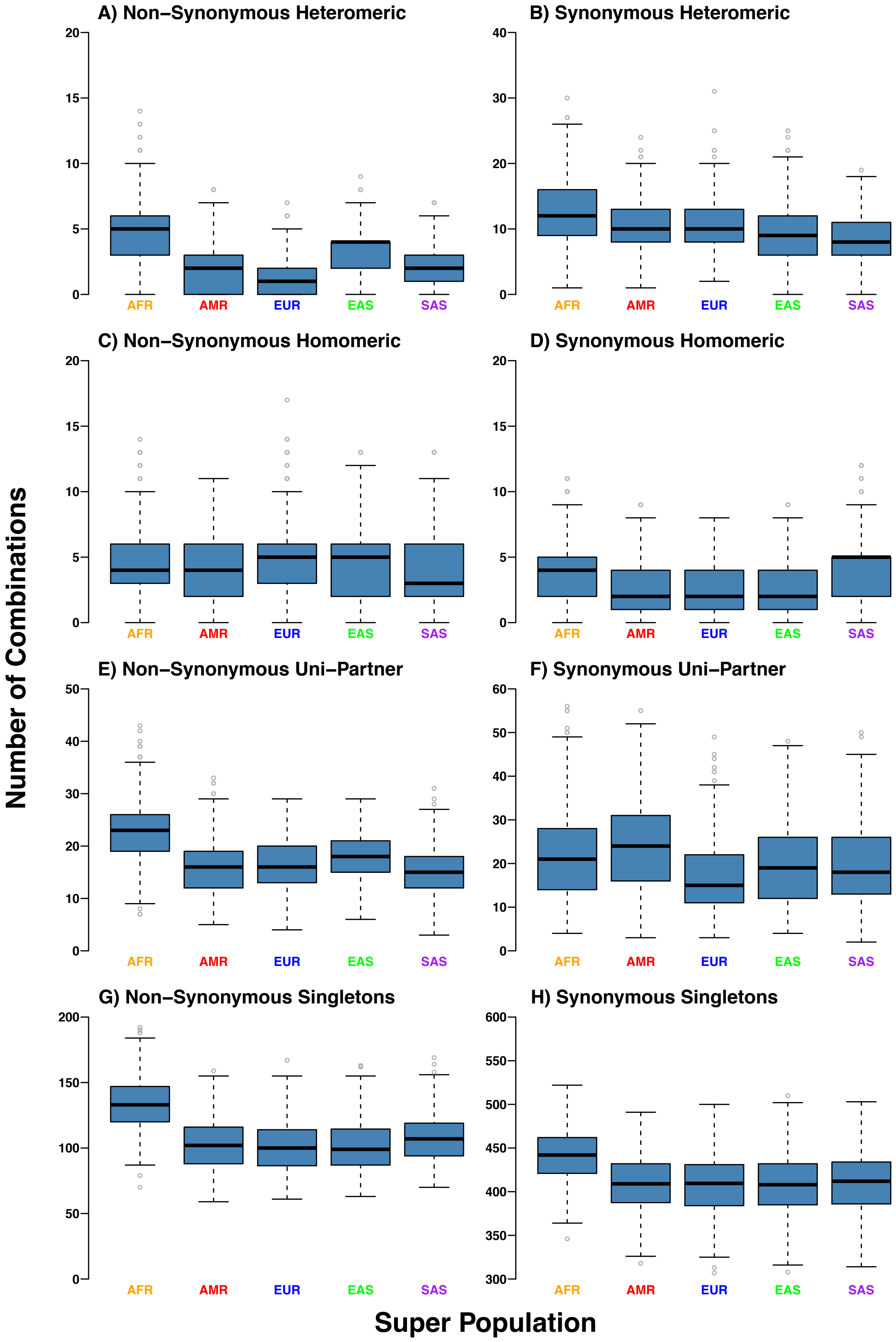


Figure 3.39: (see legend on next page)

Figure 3.39: Numbers of interface variant combinations per sample for each of the combination categories, separated in to the five super populations. **A)** Non-Synonymous Heteromeric Combinations. **B)** Synonymous Heteromeric Combinations. **C)** Non-Synonymous Homomeric Combinations. **D)** Synonymous Homomeric Combinations. **E)** Non-Synonymous Uni-Partner Combinations. **F)** Synonymous Uni-Partner Combinations. **G)** Non-Synonymous Singletons. **H)** Synonymous Singletons. Note – x- and y-axes differ in their scales between the subplots. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

Table 3.23: Mean numbers of each type of non-synonymous interface variant combination per sample for each of the five super populations.

Super Population	Non-Synonymous Heteromeric Combinations	Non-Synonymous Homomeric Combinations	Non-Synonymous Uni-Partner Combinations	Non-Synonymous Singletons
African	4.91	4.35	22.50	134.02
American	1.85	4.18	15.95	102.68
European	1.18	4.97	15.98	101.05
East Asian	3.07	4.50	18.09	101.43
South Asian	1.85	3.85	15.19	107.40

Table 3.24: Mean numbers of each type of synonymous interface variant combination per sample for each of the five super populations.

Super Population	Synonymous Heteromeric Combinations	Synonymous Homomeric Combinations	Synonymous Uni-Partner Combinations	Synonymous Singletons
African	12.64	3.73	21.72	441.44
American	10.49	2.84	24.16	408.44
European	9.27	2.58	19.60	408.34
East Asian	10.63	2.74	16.97	407.68
South Asian	8.53	4.29	19.77	409.58

3.4.13 Interface Variant Combinations Per Protein Complex

As with the Proximal Combinations within structures, where individual proteins had distinct patterns, different protein complexes were observed to have different patterns of interface variant combinations. Figure 3.40 shows the numbers of unique interface variant combinations for each of the categories. For the non-synonymous variant combinations, Singletons are the largest category (11,285), Uni-Partner combinations are the second largest category (735), followed by Heteromeric combinations (364), and then Homomeric combinations (147).

However, many of these combinations are rare (Figure 3.41 and Table 3.25), with some having fewer than 25 total occurrences (11,632 non-synonymous and 10,860 synonymous). Far fewer combinations can be considered common, with only 149 non-synonymous and 502 synonymous combinations occurring ≥500 times (Table 3.25). Removal of combinations with fewer than 25 total occurrences retains the same pattern (Singletons > Uni-Partner > Heteromeric > Homomeric), but the overall numbers of each fall (Figure 3.40 and Table 3.25).
These differences are also likely in part due to the composition of the Interactome3D dataset, which contains more heteromeric complexes than homomeric complexes (6,655 and 2,987 complexes, respectively).

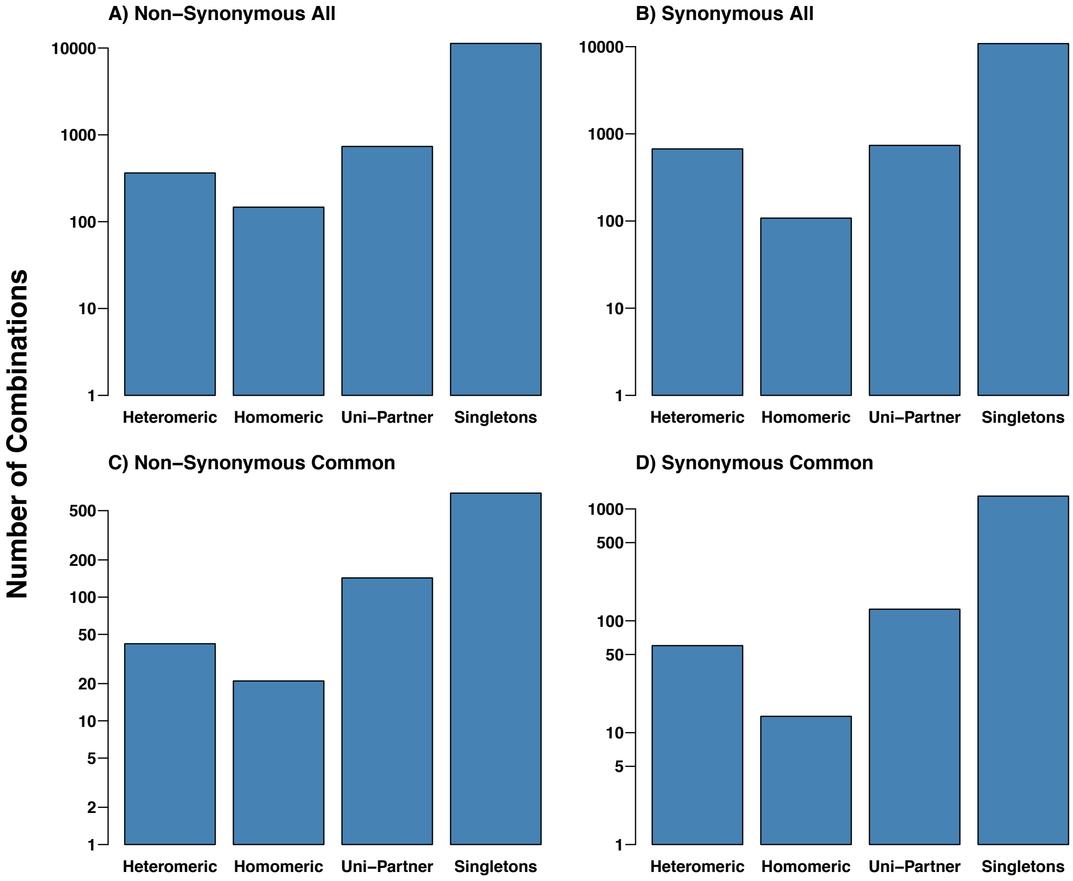


Figure 3.40: Numbers of unique interface variant combinations observed for each combination category. **A)** All non-synonymous combinations. **B)** All synonymous combinations. **C)** Common non-synonymous combinations (≥25 occurrences). **D)** Common synonymous combinations (≥25 occurrences). Note – y-axes differ in their scales between the subplots.

Table 3.25: Numbers of rare and common combinations for each interface variant combination category and for non-synonymous and synonymous interface variants

Variant Type	Heteromeric Combinations	Homomeric Combinations	Singletons	Uni-Partner Combinations
Non-Synonymous One Occurrence	145	77	7,257	240
Synonymous One Occurrence	274	57	5,885	283
Non-Synonymous ≥25 Occurrences	42	21	693	143
Synonymous ≥25 Occurrences	60	14	1,302	127
Non-Synonymous ≥100 Occurrences	15	16	304	73
Synonymous ≥100 Occurrences	31	8	764	64
Non-Synonymous ≥500 Occurrences	3	6	118	22
Synonymous ≥500 Occurrences	12	3	459	28
Non-Synonymous ≥1,000 Occurrences	0	2	68	10
Synonymous ≥1,000 Occurrences	7	2	318	19

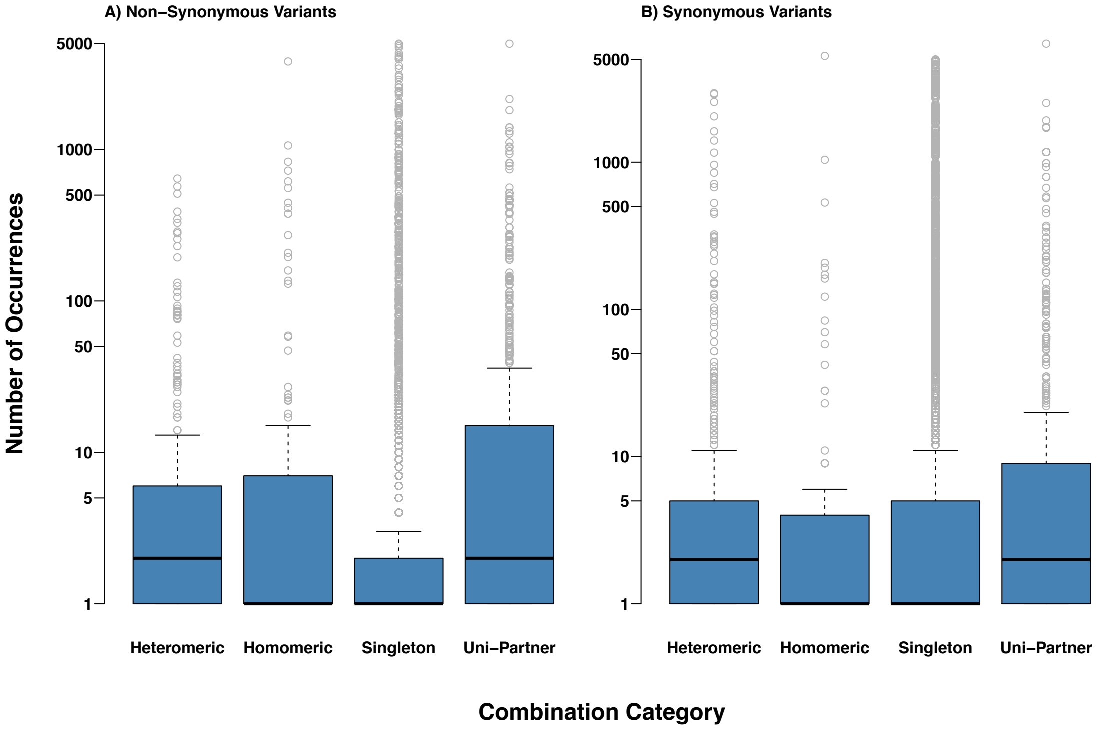


Figure 3.41: Number of occurrences of unique interface variant combinations, separated by combination category. **A)** Non-synonymous variants. **B)** Synonymous variants.

As with combinations within structures, many of the interface variant combinations are rare, but for each complex there is usually a dominant variant combination, which accounts for the majority of the total combinations in the interface (Figure 3.42). The format of Figure 3.42 is identical to Figure 3.13 (which describes combinations of variants within structures). Each of the points is a protein-protein complex, and the colour of the point is determined by the number of unique variant combinations observed for the complex (see Figure 3.42 legend). The x-axis position of the point is determined by the number of occurrences of any variant combination within the complex, and the y-axis position is determined by the percentage of total combination occurrences accounted for by the most common variant combination in the complex. If the most common variant combination accounts for all of the variant combinations in the complex (black points), the point will lie at the very top of the y-axis (100% of combination occurrences are the most common combination). The distributions of total percentage occurrences for the most common variant combination is shown in Figure 3.43, for each of the interface variant combination categories.

The patters of common interface variant combinations show similarities across Heteromeric, Homomeric, and Uni-Partner Combinations. For each category, in the majority of complexes the most common variant combination accounts for >90% of all combinations in the complex (Figure 3.42 & Figure 3.43). For non-synonymous variants, this is true for 96 Heteromeric Combinations (64.0%), 28 Homomeric Combinations (60.0%), and 124 Uni-Partner Combinations (72.9%; 186 (62.2%), 28 (60.9%), and 211 (74.0%) respectively for synonymous variants).

For Singletons (Figure 3.42.G&H), a large proportion of the complexes have a Singleton that occurs far more frequently than any other in the complex, with 2,185 non-synonymous Singletons (50.86% of total complexes) that account for >90% of all Singletons in individual complexes, and 2,868 for synonymous Singletons (59.12% of total complexes; Figure 3.43.G&H).

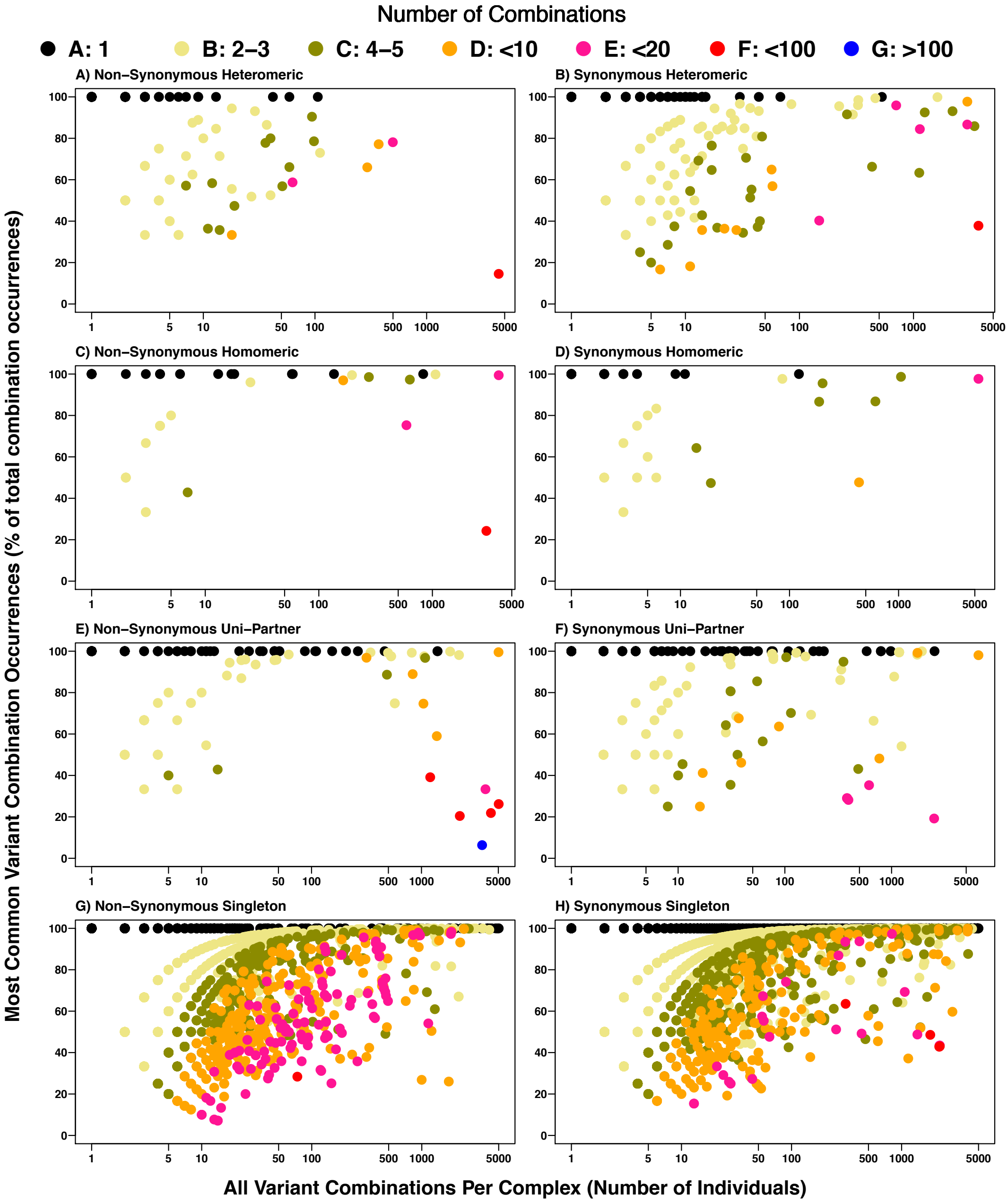


Figure 3.42: Occurrences of the most common interface variant combinations vs all interface variant combinations for each complex. Each point represents a protein-protein complex, and the colour of the point is determined by the number of unique variant combinations observed for the complex, see legend. The x-axis position of the point is determined by the number of occurrences of any variant combination within the complex, and the y-axis position is determined by the percentage of total combination occurrences accounted for by the most common variant combination in the complex. If the most common variant combination accounts for all of the variant combinations in the complex (black points) the point will lie at the very top of the y-axis (100% of combination occurrences are the most common combination). **A)** Non-Synonymous Heteromeric combinations. **B)** Synonymous Heteromeric combinations. **C)** Non-Synonymous Homomeric combinations. **D)** Synonymous Homomeric combinations. **E)** Non-Synonymous Uni-Partner combinations. **F)** Synonymous Uni-Partner combinations. **G)** Non-Synonymous Singletons. **H)** Synonymous Singletons. Note – x- and y-axes differ in their scales between the subplots.

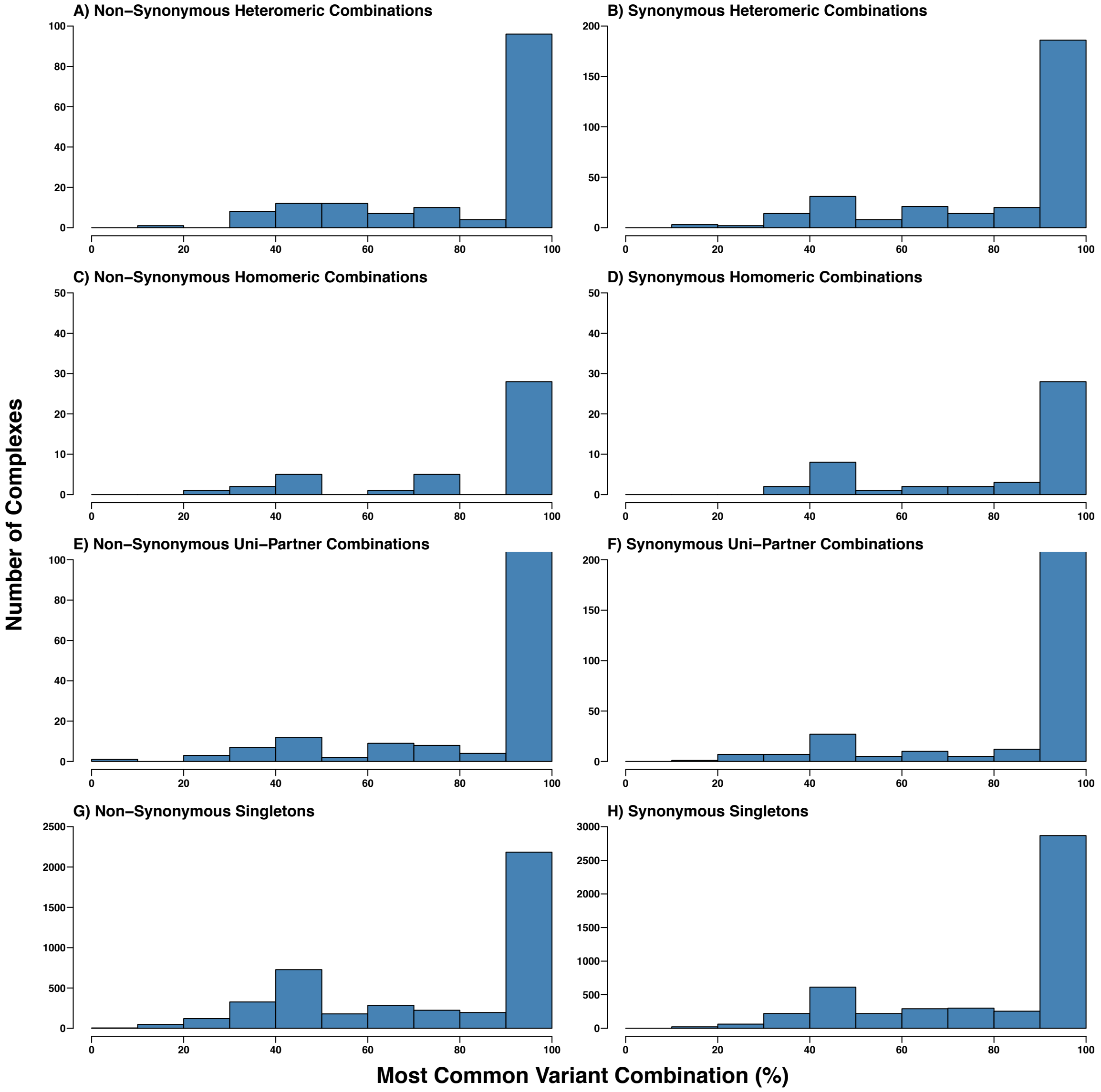


Figure 3.43: Distributions of proportions of interface variant combination occurrences from the most common variant combination per complex. **A)** Non-Synonymous Heteromeric Combinations. **B)** Synonymous Heteromeric Combinations. **C)** Non-Synonymous Homomeric Combinations. **D)** Synonymous Homomeric Combinations. **E)** Non-Synonymous Uni-Partner Combinations. **F)** Synonymous Uni-Partner Combinations. **G)** Non-Synonymous Singletons. **H)** Synonymous Singletons. Note – y-axes differ in their scales between the subplots.

The protein Immunoglobulin heavy variable 3-23 (UniProt accession: P01764) plays a role in antigen recognition, and forms a homomeric complex. In this complex (P01764:P01764), the variant combination ‘P01764:S73G,G75S’ occurs 154 times (Figure 3.44). This is a clear example where the combination of variants could be compensatory. In both chains of the complex, serine is lost at position 73 and a glycine is gained, and then at position 75 a glycine is lost and a serine is gained. The wild type and variant type amino acids are balanced, with no net change in the number of serine or glycine residues. Neither of these variants were observed on their own in any individual out of the 2,504 samples in the 1,000 Genomes Project - they only occur together. One possible advantage of the two variants occurring together, and neither alone, is maintaining the hydroxyl group of serine. In the wild type structure S73 acts as a hydrogen bond donor, forming a hydrogen bond with the backbone of G75 (Figure 3.44), which is likely stabilising the tight turn in the structure.

This combination (‘P01764:S73G,G75S’) in P01764:P01764 is classed as a Uni-Partner combination, despite being a homomer, as the structure is asymmetrical (see Methods). Positions 73 and 75 in chain A of the structure are ≤5Å from residues in chain B, but residues 73 and 75 in chain B are too far away. Therefore, the combination is classed as Uni-Partner. Clearly in this example, the distinction between Uni-Partner combinations and Homomeric combinations is arbitrary, and the compensatory effect is likely between the variants within each chain. However, this compensatory effect within the chains could help to maintain the necessary fold for the dimer to form.

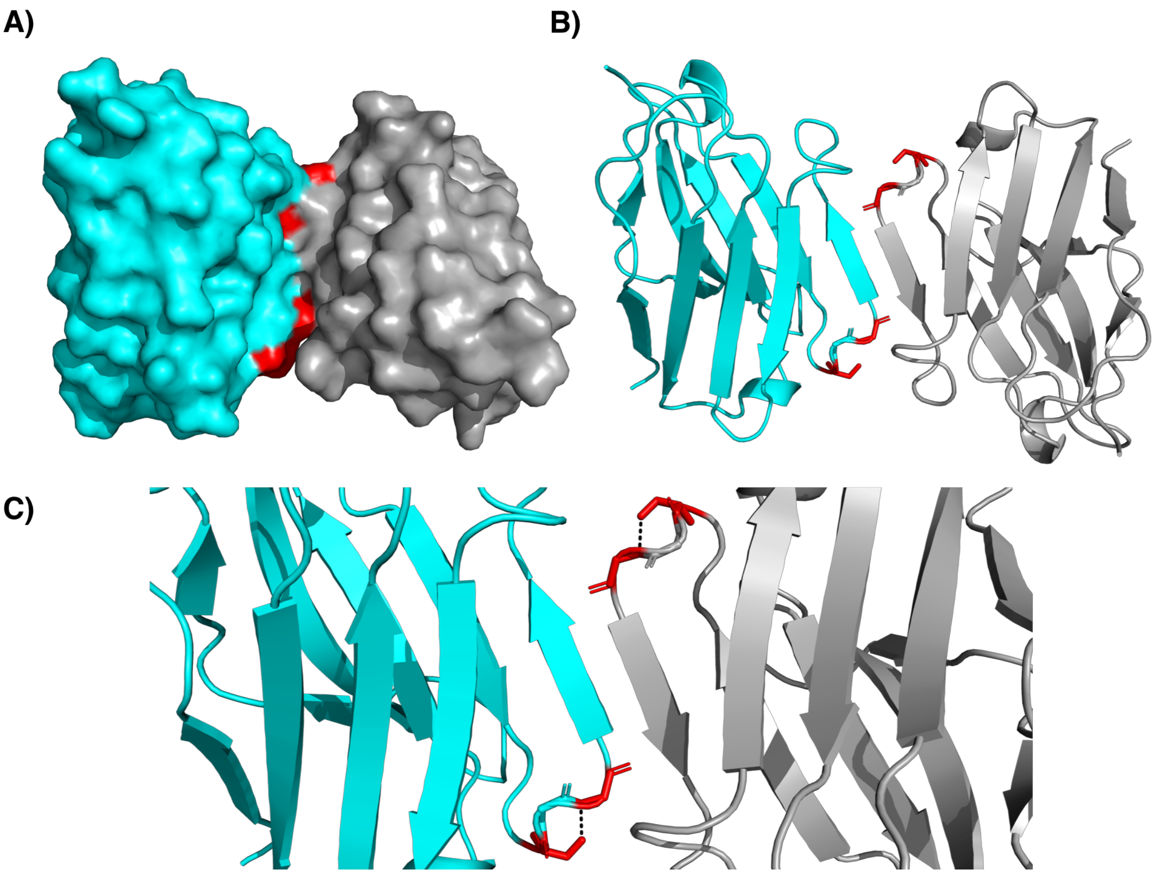


Figure 3.44: Visualisation of the interface variant combination 'P01764:S73G,G75S', which occurs in the interface P01764:P01764. In all subplots, chain A is coloured grey, chain B is coloured cyan, and positions of variants are shown in stick format and coloured red. **A)** Surface representation. **B)** Cartoon representation. **C)** Zoomed in view showing the hydrogen bond between the hydroxyl group of S73 and the backbone of G75, with the hydrogen bond represented by a dashed black line.

As with combinations of variants within individual proteins (Figures 3.15-3.17 and Appendix 2 Figures 8-13), some of the interface variant combinations that are relatively rare at the level of the whole genome set may be fairly common within specific populations. These evenly distributed interface variant combinations are likely to have first occurred longer ago in human evolution, with population-specific interface variant combinations likely to have occurred more recently.

Figures 3.45-3.47 show the distributions of interface variant combinations between the five super populations represented in the 1,000 Genomes Project data set. These plots show the 50 most common combinations for each category. For each interface combination category there are examples of combinations that are spread relatively evenly across the super populations (light blue boxes across the super populations), and others that are concentrated in a subset of super populations (dark blue boxes in specific populations). Similar patterns can be seen for the 26 individual populations that make up the 5 super populations, as well as for synonymous variant combinations, see Appendix 2 Figures 14-22).

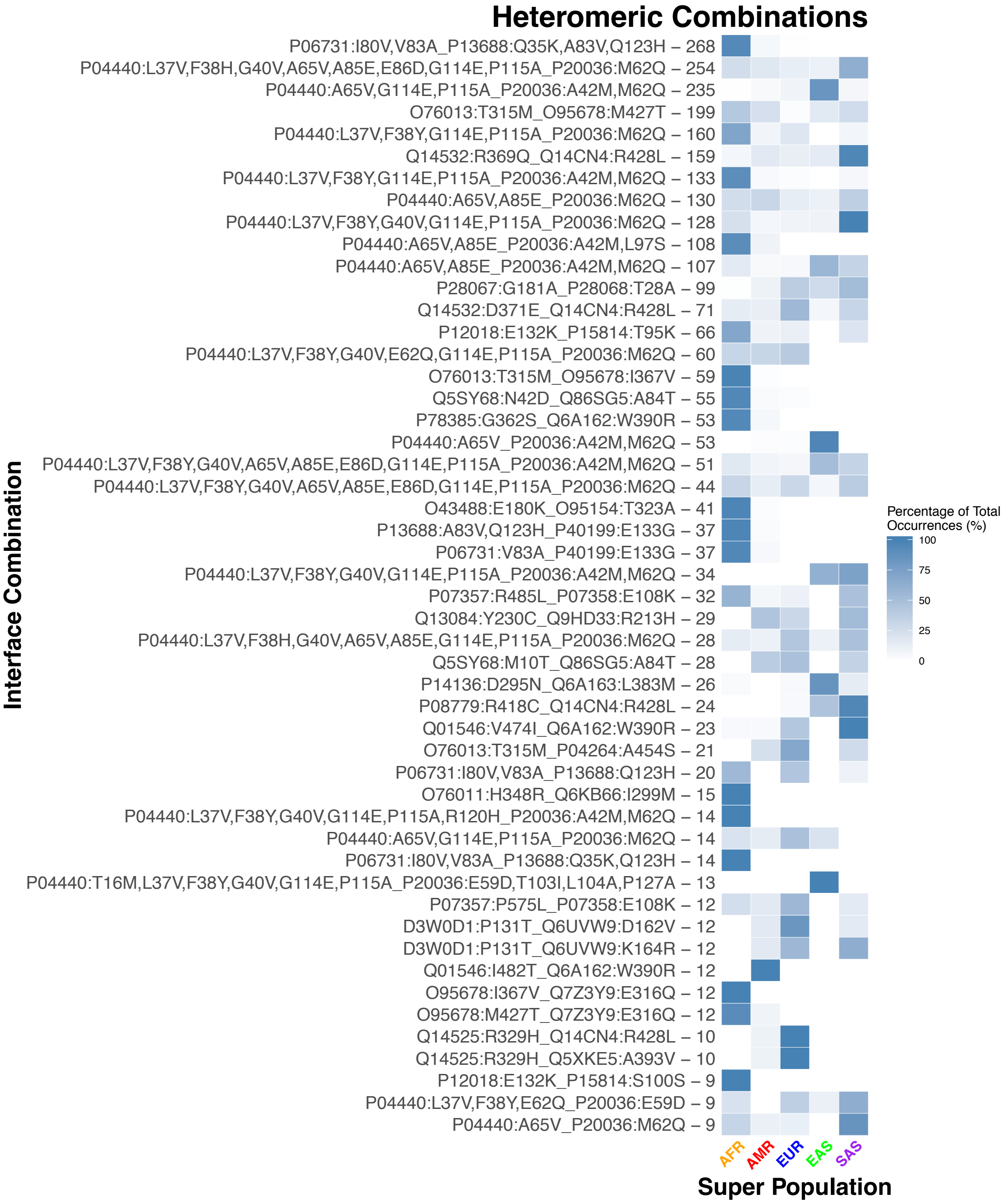


Figure 3.45: Heatmap of occurrences within super populations of the 50 most common Heteromeric non-synonymous interface variant combinations. Combinations are given on the y-axis, see Methods for a description of the interface variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

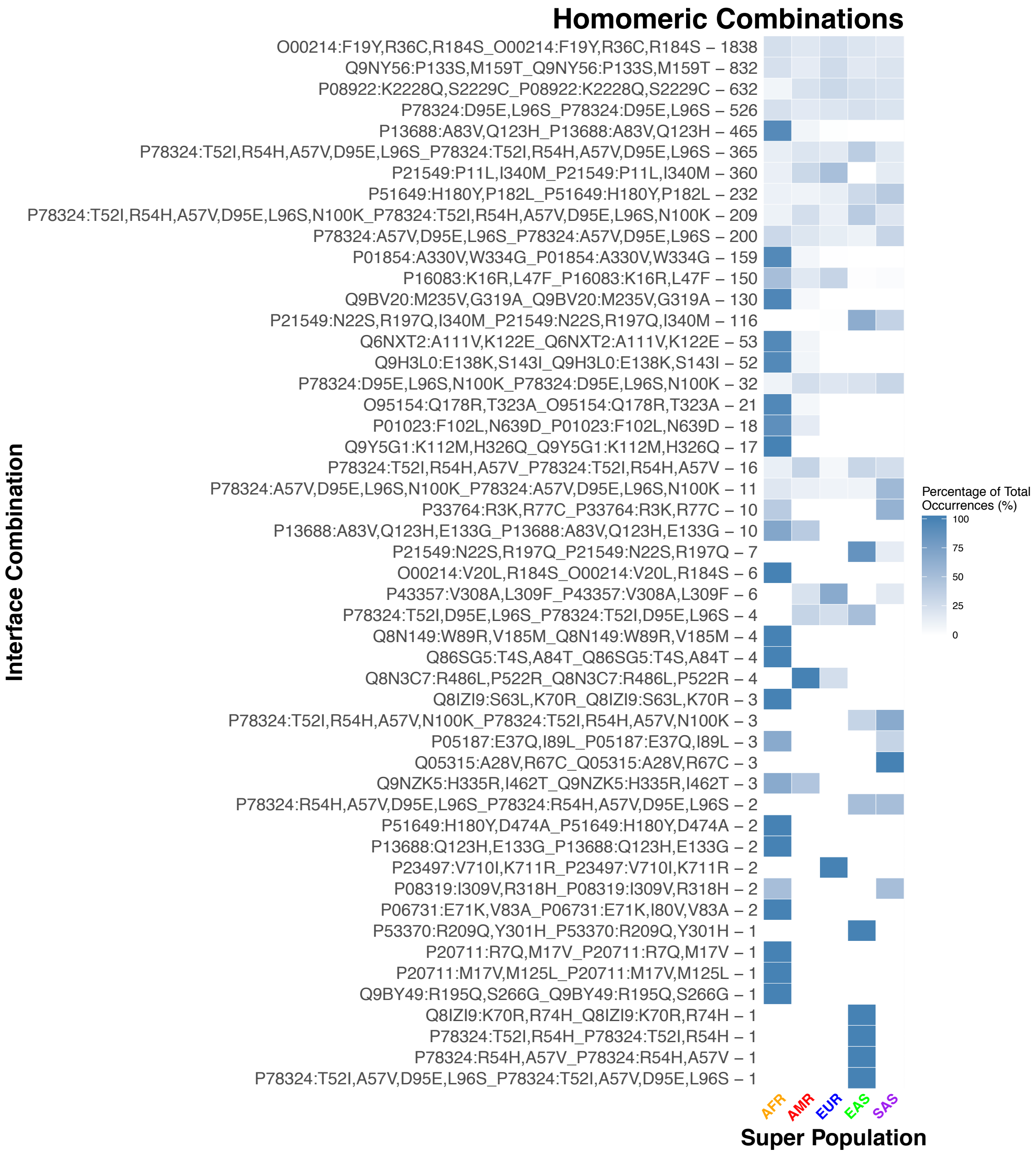


Figure 3.46: Heatmap of occurrences within super populations of the 50 most common Homomeric non-synonymous interface variant combinations. Combinations are given on the y-axis, see Methods for a description of the interface variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

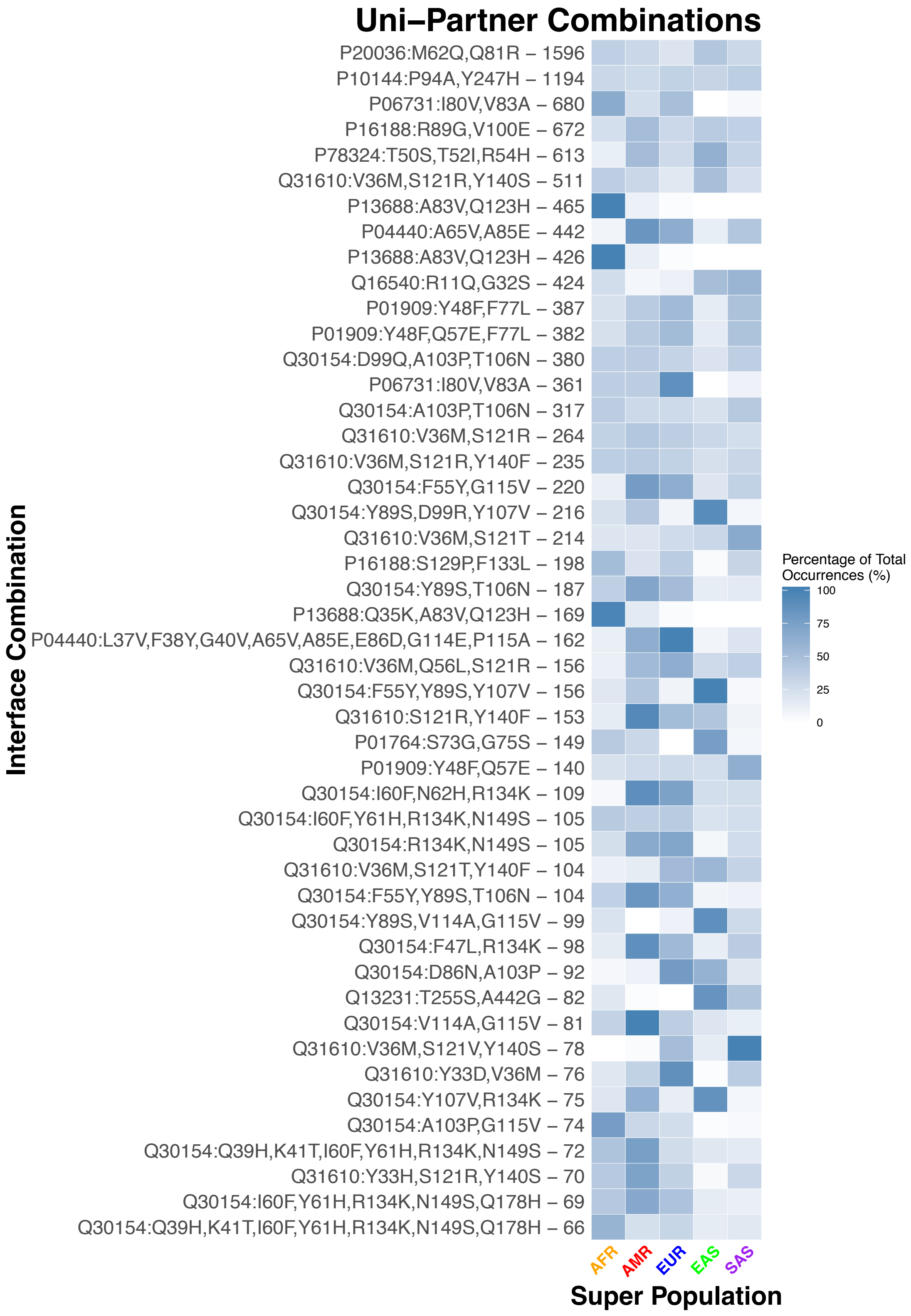


Figure 3.47: Heatmap of occurrences within super populations of the 50 most common Uni-Partner non-synonymous interface variant combinations. Combinations are given on the y-axis, see Methods for a description of the interface variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

The distributions between super populations for each interface variant combination vary, with some combinations being common or rare in certain populations, and other combinations occurring at roughly the same frequency between populations (Figures 3.45-3.47). In Figure 3.48, the distributions of these population frequencies are shown for non-synonymous Heteromeric, Homomeric, and Uni-Partner combinations (Figure 3.48; Appendix 2 Figure 23 for synonymous variants). There is a clear difference for the AFR super population compared to the other four super populations, with many more combinations that occur almost exclusively in the AFR super population for all three types of interface variant combinations (Figure 3.48 and Table 3.26). There are 267 combinations when >90% of the occurrences are in the AFR super population (72 for AMR, 66 for EUR, 108 for EAS, 96 for SAS). Many of these combinations are rare, and when filtering for combinations with ≥25 total occurrences this number falls to 22 for the AFR super population (0 for AMR, 0 for EUR, 3 for EAS, 0 for SAS). As with variant combinations within structures (Figure 3.18), this perhaps reflects the fact that the human reference genome is least representative of the AFR super population (1000 Genomes Project Consortium 2015).

The previously discussed combination ‘P01764:S73G,G75S’ in P01764:P01764 (Figure 3.44), which has a potential compensatory effect of maintaining the hydroxyl group of serine, also has an interesting distribution of occurrences between the super populations. This combination occurs in 149 individuals, 38.26% from AFR, 10.74% from AMR, 46.98% from EAS, 4.03% from SAS, and 0.00% from EUR (Figure 3.47). Similarly, the combination ‘P78324:T52I,R54H ,A57V,D95E,L96S,N100K_P78324:T52I,R54H,A57V,D95E,L96S,N100K’ (Figure 3.37) occurs 209 times, with the most occurrences in the EAS super population: 10.04% AFR, 18.66% AMR, 11.96% EUR, 39.71% EAS, 19.62% SAS (Figure 3.46).

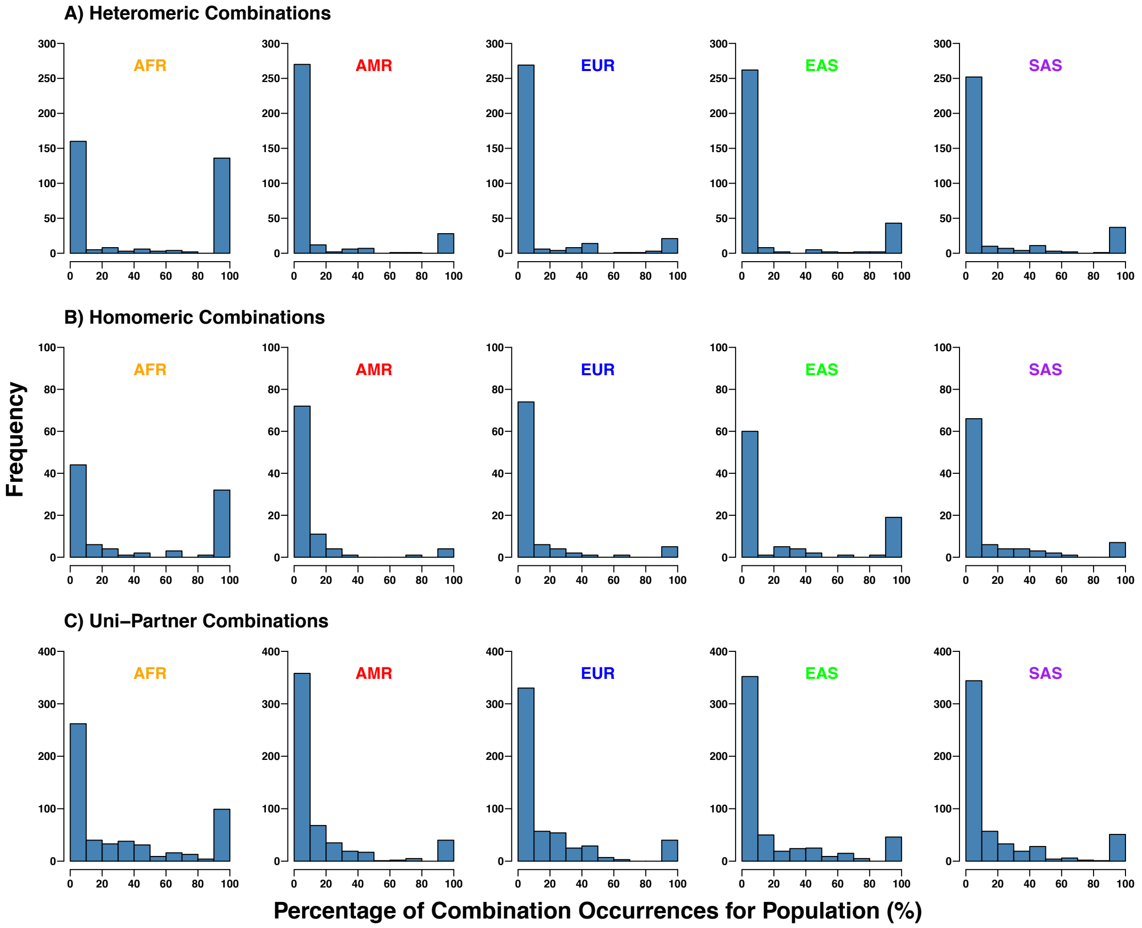


Figure 3.48: Distributions of percentages of total occurrences of each unique non-synonymous interface variant combination from each super population. **A)** Heteromeric Combinations. **B)** Homomeric Combinations. **C)** Uni-Partner Combinations. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

Table 3.26: Numbers of unique non-synonymous interface variant combinations where ≥90% of total occurrences are from one super population.

Super Population	Heteromeric Combinations	Homomeric Combinations	Uni-Partner Combinations
African	136	32	99
American	28	4	40
European	23	5	40
East Asian	43	19	44
South Asian	37	7	50

3.4.14 Interface Variant Combination Effect Compensations

As with combinations of variants within proteins, interface variant combinations were analysed to identify combinations with potential compensation events (see Methods and Section 3.4.7). The total numbers of each type of potential compensation event (Figure 3.49 and Table 3.27) show similar patterns to compensations within structures (Figure 3.24 and Table 3.13). As with Proximal Combinations, the number of direct amino acid compensations was only weakly correlated with the amino acid composition of the interfaces analysed (r=0.51), but a higher correlation was seen with the number of codons that encode each amino acid (r=0.62; Appendix 2 Figure 7.J&L).

Patterns within each type of compensation were also similar, with arginine amino acid compensations being the most common type of amino acid compensation (18.9%; Appendix 2 Figure 24), hydrophobic (55.2%) and positive (26.0%) types being the most common charge compensations (Appendix 2 Figure 25), and hydrophobic (41.6%) and hydroxyl (24.0%) types being the most common types of functional group compensations (Appendix 2 Figure 26; e.g. 'P01764:S73G,G75S' shown in Figure 3.44). Additionally, as with combinations within structures (Figure 3.33) the total mass changes of interface variant combinations are on average fairly conservative (mean -11.0, median -12.1, min -244.4, max +224.176, 86.6% between -100 and +100 total mass change; Appendix 2 Figure 27).

Table 3.27: Numbers of potential compensation types for interface variant combinations.

Compensation Category	Unique Compensations
Interfaces with Compensation Events	101
Combinations with Compensation Events	595
Amino Acid Compensations	127
Charge Compensations	96
Functional Group Compensations	125

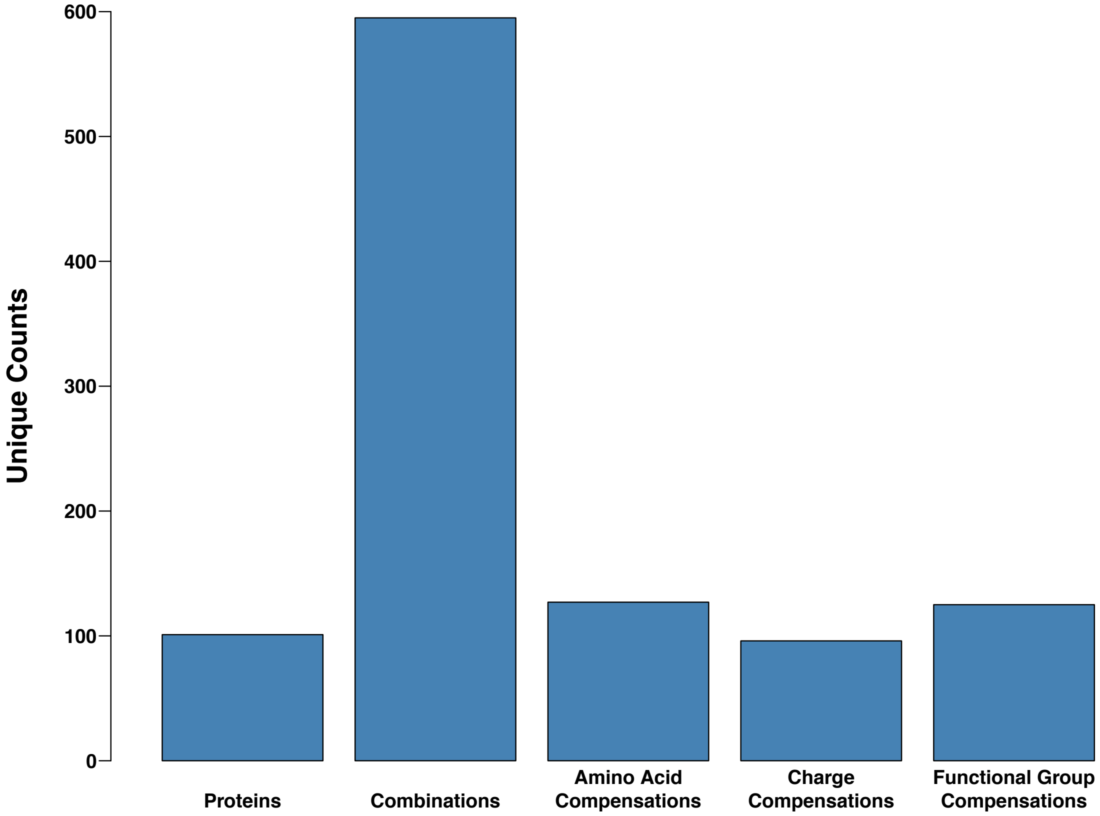


Figure 3.49: Numbers of potential property compensations within interface variant combinations.

3.4.15 Clinical Significance of Variants in Interface Combinations

As with variant combinations within structures (see Section 3.4.9), to identify potential compensation/buffering effects of deleterious variants, non-synonymous interface variant combinations were annotated using Humsavar and ClinVar (see Methods). However, none of the observed interface variant combinations contain variants annotated as pathogenic in Humsavar or ClinVar (Tables 3.28 & 3.29 respectively). As with variant combinations within structures (Table 3.16), the majority of interface variant combinations involve Unclassified variants, again highlighting the need for better clinical annotation of genetic variants.

Table 3.28: Combined clinical annotations of all non-synonymous interface variant combinations with Humsavar.

Combination Classification	Heteromeric Combinations	Homomeric Combinations	Uni-Partner Combinations
Benign	116	58	270
Benign & Unclassified	166	78	324
Unclassified	82	11	141

Table 3.29: Combined clinical annotations of all non-synonymous interface variant combinations with ClinVar.

Combination Classification	Heteromeric Combinations	Homomeric Combinations	Uni-Partner Combinations
Benign	270	58	116
Benign & Unclassified	324	78	166
Unclassified	141	11	82

3.5 Discussion

In this study we have quantified variant combinations in individual human genomes at three different levels: variant combinations across the lengths of protein sequences (280,329 non-synonymous), variant combinations in close proximity within the 3-deminsonal space of protein structures (4,365 non-synonymous), and variant combinations within protein-protein interface sites (1,246 non-synonymous). These combinations show considerable variability in their distributions between populations, with some combinations distributed evenly across populations and others heavily skewed towards individual populations. At the protein level most proteins have one predominant combination of variants, especially for variant combinations in close spatial proximity.

Within the 4,365 non-synonymous Proximal Combinations (those close in 3-dimensional space), there are 571 examples of potential direct amino acid compensations, 474 examples of potential charge compensation, and 2,674 predicted to preserve structural stability. Given these potential compensatory effects, and the fact that there are 42.4% more Proximal Combinations observed than expected based on random distribution of the variants in the structures, many of these combinations represent likely coevolutionary relationships between residues in human proteins. Potential compensation events were also observed within the 1,246 non-synonymous interface variant combinations (those in protein-protein interface sites), with 127 potential direct amino acid compensations, 96 potential charge compensations, and 125 potential functional group compensations.

Positively charged amino acids play important roles in protein structure and function, such as in the formation of salt bridges. Potential compensation of positively charged side chains was a common phenomenon observed in Proximal Combinations, with 14.8% of direct amino acid compensations being arginine compensations (18.9% for interface variant combinations), and 15.2% of all charge compensations being positive charge compensations (26% for interface variant combinations). For functional groups, potential compensation of hydroxyl groups was most common (34.9% of total functional group compensations in Proximal Combinations and 24.0% for interface variant combinations). Hydroxyl groups are uniquely functional in proteins, having the ability to form stabilising hydrogen bonds and also undergo dynamic phosphorylation, a common regulatory mechanism of protein function.

The majority of Proximal Combinations (61.5%) are predicted to better conserve protein stability compared to sub-combinations, with an even more pronounced trend for combinations containing >2 variants (86.9%). Many of the observed combinations are fairly conservative in terms of total amino acid mass change, an obvious potential mechanism for maintaining structural stability, in addition to the observed amino acid property compensations. For example, disease variants have previously been shown to be enriched in the hydrophobic cores of proteins (David et al. 2012), frequently causing structural instability. Correspondingly, 55.3% of the potential charge compensations observed in Proximal Combinations were hydrophobic compensations.

The variant combinations that we have identified in this study are likely a fraction of the true number that occur within human genomes. Firstly, we have only considered protein coding changes (and only single amino acid variants), and there are undoubtedly combinations involving non-coding variants with compensatory effects. Secondly, for simplicity we only considered the canonical isoform of each protein. There will be variant combinations that only occur within these non-canonical isoforms, although recent work has suggested that for most proteins there is only one functionally relevant isoform at the protein level (Tress et al. 2017). Finally, Proximal Combinations can only be identified for proteins with an experimental or modelled structure. For the 20,791 proteins in the human proteome set used in this study, 4,138 have 0% sequence coverage by structures, with the remaining proteins having an average coverage of 59.5% (see Methods). Similarly, a structure for each protein-protein complex is required in order to determine interface variant combinations. The release of Interactome3D (Mosca et al. 2013) used in this study contains 9,642 unique protein-protein complexes (see Methods), a small fraction of the ~650,000 (Stumpf et al. 2008) to >1,000,000 (Ranea et al. 2010) protein-protein interactions predicted to constitute the human interactome.

Nevertheless, this study is an important step in the interpretation of the combined effects of variants, highlighting the scale of variant combinations in human genomes and the importance of the development of appropriate methods for their interpretation. It will also be essential to generate additional public data sets like the 1,000 Genomes Project, where the individual combinations of variants within genomes can be analysed, in contrast to aggregated data sets, like GnomAD, for which this is not possible. Combinations of variants in individual human genomes will determine resilience or susceptibility to a host of selection pressures, from inherited disease, to drug response, to pathogenic infection. The interpretation of gestalt effects of variant combinations will be a key challenge in the future of precision medicine.

Chapter 4: Investigating the Molecular Determinants of Ebolavirus Pathogenicity

This paper is entitled: "Investigating the Molecular Determinants of Ebolavirus Pathogenicity". This work is currently being prepared for submission.

My contribution to the work was as follows:

1. Wrote all of the scripts used in the project, including processing of the ebolavirus genomes, and the identification and subsequent processing of the specificity determining positions
2. Produced all of the figures and tables
3. Helped analyse the data with all other authors
4. Helped write the manuscript with all other authors

4.1 Abstract

Ebolaviruses have gained significant attention in recent years as one of the deadliest human pathogenic viruses. However, only four of the five known *Ebolavirus* species are pathogenic to humans. In this study, a set of 1,408 ebolavirus genomes has been used to compare the protein sequences of the human pathogenic and non-pathogenic species, in order to identify the molecular determinants of human pathogenicity. We identified 166 specificity determining positions (SDPs) across the seven ebolavirus proteins that differentiate the human pathogenic and non-pathogenic species, and using structural analysis identified a subset of nine positions that are likely to have a functional impact on the proteins. The results are highly similar to those previously generated using a much smaller set of 196 ebolavirus genomes, with a 73% overlap of protein positions between the analyses. The set of likely functional SDPs was strikingly similar between the analyses, with nine SDPs identified using 1,408 genomes, which included all eight of the likely functional SDPs previously identified using a set of 196 genomes. The one new functional SDP detected using the larger genome set is located in the VP24-Karyohperin α5 interface, where three previously identified functional SDPs are located. This updated analysis supports previous studies that have suggested the ebolavirus protein VP24 plays an important role in human pathogenicity. The high similarity between the SDPs identified, and crucially the agreement of proposed functional SDPs, suggests that our approach is robust enough to be applied more broadly for investigating viral phenotypes. Rapid identification of functional differences between viruses will be essential in the future in order to maximise the benefits of the rise of in-field sequencing during virus outbreaks. Our approach has the potential to be used as a computational tool for quickly identifying functional sequence differences between groups of viruses.

4.2 Introduction

The last few decades have seen the regular spill-over of zoonotic viruses, including SARS, MERS, Zika virus, and ebolaviruses. At the same time, our ability to sequence viruses during outbreaks has rapidly advanced. This is best demonstrated by the recent West Africa *Ebola virus* outbreak, which infected more than 28,000 individuals, killing over 11,000 and during which in-field sequencing was performed extensively, including with nanopore based sequencers (CDC 2018; Hoenen et al. 2016; Quick et al. 2016). There are now more than 2,000 *Ebola virus* genomes available from this outbreak, which enables a more detailed analysis of virus evolution than was previously possible. Our interest is in using these genome sequences to investigate molecular determinants of ebolavirus pathogenicity.

Four of the five identified *Ebolavirus* species are known to cause Ebolavirus Disease (EVD), which is a haemorrhagic fever characterised by fatigue, nausea, severe hypotension, organ damage, internal and external bleeding, and a high mortality rate (Feldman & Geisbert 2011). The mortality rate of EVD varies between outbreaks but has been observed as high as >90% within an outbreak. Endemic to Africa, these four species – *Ebola virus*, *Sudan virus*, *Bundibugyo virus*, and *Tai Forest virus* – are classed as the pathogenic *Ebolavirus* species. However, the fifth species in the *Ebolavirus* genus, *Reston virus*, is not considered to be pathogenic in humans. Though data is limited due to the small number of reported outbreaks, *Reston virus* is thought to be able to infect humans but does not cause EVD (Miranda et al. 1999; World Health Organization 2009). Importantly, *Reston virus* has been shown to be pathogenic in non-human primates (Miranda & Miranda 2011), suggesting that the differences in pathogenicity between ebolavirus species is relatively subtle.

Previous studies have investigated the biological properties of *Reston virus* that result in this lack of pathogenicity. Recent work proposed 180 Specificity Determining Positions (SDPs) across the seven ebolavirus proteins - positions that are differentially conserved between *Reston virus* and the other four human pathogenic ebolavirus species (Pappalardo et al. 2016). These SDPs may make *Reston virus* replication, host cell entry, or host immune system suppression impossible, or at least severely impaired compared to the pathogenic ebolavirus species.

It is important to establish exactly which SDPs have a direct effect on pathogenicity. The previous study was performed during the West Africa Ebola virus outbreak using 196 genome sequences, which was a large sample set at the time. Since then many more ebolavirus genome sequences have become available. In this study, we expand the original analysis to a much larger sample set (1,408 genomes), in order to refine this set of SDPs, and assess how many sequences are required to confidently perform such analyses. Increasing the sample size will give more accurate results, highlighting previously unseen variation and conservation of sequence positions. Finally, we discuss the application of our approach as a more general tool for the investigation of the molecular determinants of phenotypic differences between viruses, and the necessity of such an approach in analysing the large and rapidly growing amount of virus sequence data.

4.3 Methods

4.3.1 Ebolavirus Nomenclature

The virus nomenclature in this report follows the recommendations set by Kuhn et al., *Filoviridae* is the family, in the order *Mononegavirales*. Both of these terms are always italicised when referenced. The genus is known as *Ebolavirus*, and is only italicised when referring to the genus, but not when referring to physical viruses, virus properties, or constituent virus parts such as proteins or genomes. Ebolavirus Disease (EVD) also remains unitalicised. The five individual species are subsequently referred to as *Bundibugyo ebolavirus* (type virus: *Bundibugyo virus*, BDBV), *Reston ebolavirus* (type virus: *Reston virus*, RESTV), *Sudan ebolavirus* (type virus: *Sudan virus*, SUDV), *Tai Forest ebolavirus* (type virus: *Tai Forest virus*, TAFV) and *Zaire ebolavirus* (type virus: *Ebola virus*, EBOV) (Kuhn et al. 2014).

4.3.2 Collection of *Ebolavirus* Genomes

All ebolavirus genome sequences were obtained from the National Center for Biotechnology Information (NCBI) (Brister et al. 2015), the Virus Pathogen Resource (ViPR) (Pickett et al. 2012), as well as taken from a repository obtained from (Urbanowicz et al. 2016), available here: https://github.com/ebov/space-time. Duplicate sequences present in >1 of the databases were filtered out during initial sample collection, with the order of source preference being NCBI > ViPR > Urbanowicz et al. Table 4.1 summarises the sources used for the set of *Ebolavirus* genomes.

Table 4.1: Summary of source databases used to obtain Ebolavirus genomes for analysis.

Species	NCBI	ViPR	Urbanowicz	Total
Ebola virus	1,469	43	505	2,017
Sudan virus	14	5	0	19
Bundibugyo virus	7	2	0	9
Tai Forest virus	1	3	0	4
Reston virus	18	9	0	27
Total	1,509	62	505	2,076

4.3.3 Genome Processing and Filtering

For each sample genome, open reading frames (ORFs) were identified using the EMBOSS getorf tool (Rice et al. 2000), and the resulting ORFs were matched to the UniProt *Ebola virus* reference protein sequences using BLAST (Camacho et al. 2009; Bateman et al. 2015). The top ORF hit for each *Ebola virus* protein was then used as the protein sequence for that sample, for all proteins except GP. The ebolavirus GP protein is the result of mRNA editing, due to a slippery 7A-motif that is translated as eight A nucleotides, with the regular ORF containing an early stop codon (Volchkov et al. 1995). The GP ORF hits were further processed by editing the identified ORF to swap the 7A-motif for 8 A nucleotides. ORFs were then re-identified for the edited sequence and BLAST was used to search against the *Ebola virus* reference proteins.

After these steps, ebolavirus samples that did not have a BLAST hit with >90% coverage compared to the *Ebola virus* reference protein, for each of the seven proteins, was removed. Samples with poor metadata, such as unknown host or data were also removed (partial dates were allowed, e.g. if only the year of collection was known). This was to ensure that only high-quality samples were analysed, as incomplete data could affect subsequent analyses. Table 4.2 summarises the samples that were removed in this step, and a full list of the samples that were retained can be found in Appendix 3 Table 1.

Table 4.2: Summary of the sequences removed from the initial set of ebolavirus genome sequences.

Species	Starting	Removed	Final
Ebola virus	2,017	661	1,356
Sudan virus	19	5	14
Bundibugyo virus	9	1	8
Tai Forest virus	4	1	3
Reston virus	27	0	27
Total	2,076	668	1,408

4.3.4 Genome Sequence Alignment and Identification of Specificity Determining Positions

Clustal Omega was used to generate sequence alignments for each of the ebolavirus proteins (Fabian Sievers et al. 2011), and the individual sequence identities were obtained from the Clustal Omega output. Jensen-Shannon divergence scores were then calculated for each protein (Capra & Singh 2007). S3det was used in supervised mode to find specificity determining positions (SDPs), with sequences assigned to two groups prior to running S3Det (Rausell et al. 2010). Group 1 contained all of the human pathogenic sequences (*Ebola virus*, *Sudan virus*, *Bundibugyo virus*, and *Tai Forest virus*) and group 2 contained all of the human non-pathogenic sequences (*Reston virus*). All SDPs are referred to by the amino acid in the *Ebola virus* protein sequence, the position in the *Ebola virus* reference protein sequence, and the corresponding amino acid in the *Reston virus* protein sequence, e.g. G20A meaning at position 20 *Ebola virus* has a glycine residue and *Reston virus* has an alanine residue*.*

4.3.5 Structural Analysis of SDPs

All available ebolavirus protein structures were downloaded from the Protein Databank (PDB) (Berman et al. 2000), and SDPs were mapped to the highest quality structure available, based on structure resolution and coverage. Multimeric protein structures were used to analyse the effects of the SDPs on partner interactions. Where structures were unavailable from the PDB, proteins were modelled using Phyre2 with default settings (Kelly et al. 2015). This was the case for the protein L, which was modelled on the PDB structure 5A22, chain A, with 90% coverage (alignment: residues 8-2010) and 100% confidence. Table 4.3 summarises the structures used for analysis. PyMOL (https://pymol.org/2/) was used to visualise the identified SDPs in the protein structures and generate images.

Table 4.3: Summary of the structures used for SDP investigation.

Protein	Species	PDB Structure ID	Oligomeric Form	Residue Coverage
VP24	EBOV	4M0Q	Homodimer	11 - 237
	EBOV	4U2X	Heterodimer (with KPNA5)	16 - 231
VP30	EBOV	2I8B	Homodimer	142 - 272
VP35	EBOV	4IBC	Homodimer	215 - 340
	EBOV	3L26	Homodimer (bound to RNA)	215 - 340
VP40	EBOV	4LDB	Homodimer	44 - 326
	EBOV	4LDD	Homo 6-mer	44 - 326
	EBOV	4LDM	Homo 8-mer	44 - 188
NP	EBOV	4QB0	Monomer	641 - 739
	EBOV	4YPI	Heterodimer	38 - 385
GP	EBOV	5JQ3	Hetero 6-mer	32 – 501 & 502 – 632
L	EBOV	N/A (Phyre2 model)	Monomer	8 - 2010
Nucleocapsid (NP with VP24)	EBOV	6EHM	Hetero 4-mer	NP: 1 – 739 VP24: 1 - 251

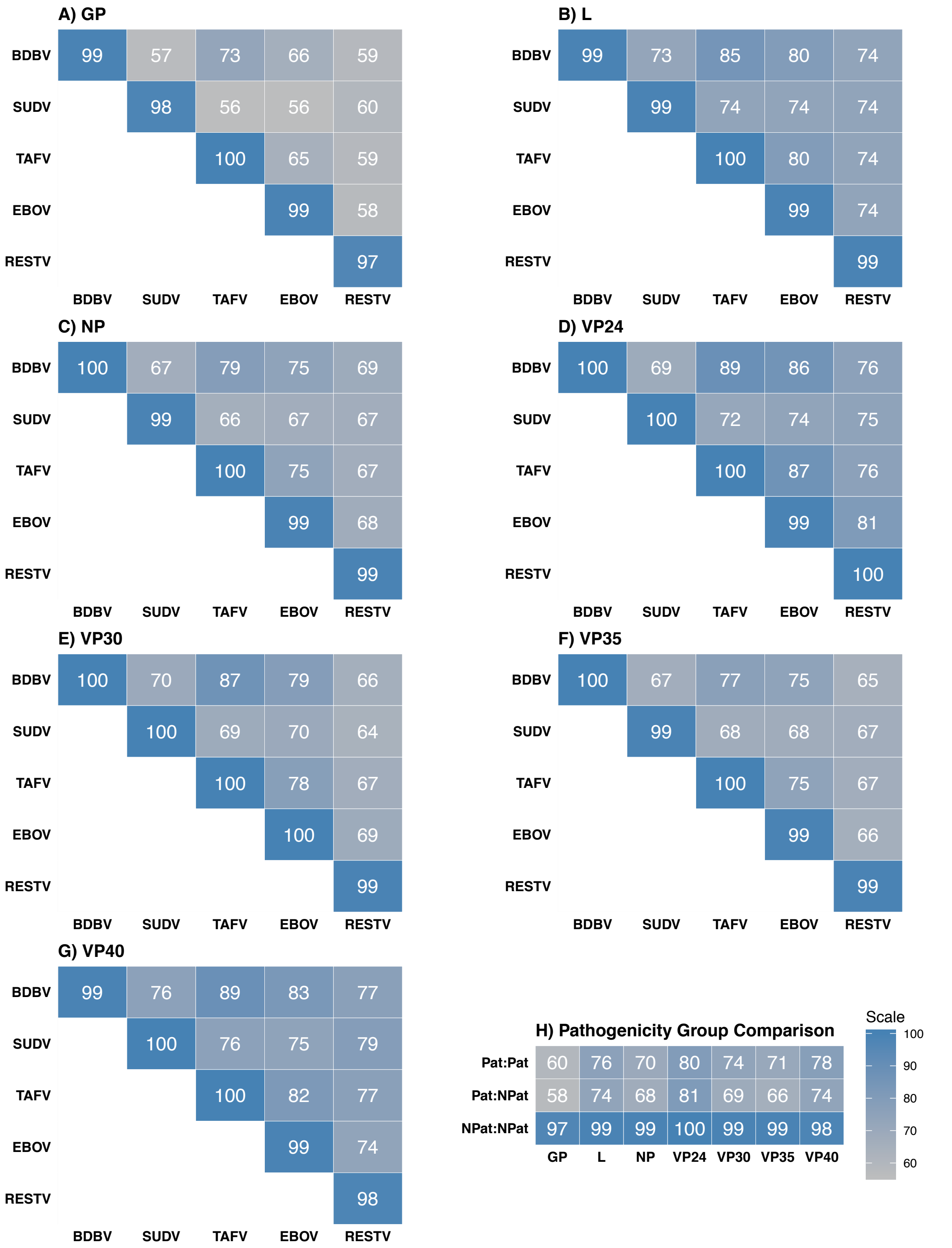
For the subset of SDPs mapped to structures, multiple computational tools were used to predict the functional effects of each SDP. mCSM was used to predict the effect on protein stability (Pires et al. 2014), where the change in stability (ΔΔG) is measured in kcal/mol, with negative values being destabilising and positive values being stabilising. Relative solvent accessibility of SDP residues was also calculated using mCSM. BLOSUM62 scores were assigned to each SDP, with the score calculated for the change between the *Ebola virus* sequence and the *Reston virus* sequence wherever there was variation amongst the pathogenic species.

4.4 Results

4.4.1 Ebolavirus Protein Conservation

From an initial set of 2,076 ebolavirus sequences, obtained from multiple sources, 1,408 sequences were retained for analysis after filtering (see Methods). This represents 7.5 times more sequences than the previous SDP-based analysis of ebolavirus pathogenicity (Pappalardo et al. 2016). This data set contains 27 *Reston virus* sequences (increased from 17 previously), and for the human pathogenic species: 1,356 *Ebola virus*, 14 *Sudan virus*, 8 *Bundibugyo virus* and 3 *Tai Forest virus* genomes.

Very high levels of conservation were observed within each species, with greater variation between species (Figure 4.1). The greatest variation between species is observed in GP, with conservation typically between 50-60%, see Figure 4.1.A. This high degree of variation is likely due to GP being the only ebolavirus surface protein, which is under selective pressure from the host species (Liu et al. 2015). Comparison of *Reston virus* proteins to the other four human pathogenic species showed that there is greater divergence in GP, NP, VP30 and VP35, with conservation between 58%-69%, whereas VP24, L and VP40 have a higher level of conservation (74-81%; Figure 4.1.H).



A) GP

	BDBV	SUDV	TAFV	EBOV	RESTV
BDBV	99	57	73	66	59
SUDV		98	56	56	60
TAFV			100	65	59
EBOV				99	58
RESTV					97

B) L

	BDBV	SUDV	TAFV	EBOV	RESTV
BDBV	99	73	85	80	74
SUDV		99	74	74	74
TAFV			100	80	74
EBOV				99	74
RESTV					99

C) NP

	BDBV	SUDV	TAFV	EBOV	RESTV
BDBV	100	67	79	75	69
SUDV		99	66	67	67
TAFV			100	75	67
EBOV				99	68
RESTV					99

D) VP24

	BDBV	SUDV	TAFV	EBOV	RESTV
BDBV	100	69	89	86	76
SUDV		100	72	74	75
TAFV			100	87	76
EBOV				99	81
RESTV					100

E) VP30

	BDBV	SUDV	TAFV	EBOV	RESTV
BDBV	100	70	87	79	66
SUDV		100	69	70	64
TAFV			100	78	67
EBOV				100	69
RESTV					99

F) VP35

	BDBV	SUDV	TAFV	EBOV	RESTV
BDBV	100	67	77	75	65
SUDV		99	68	68	67
TAFV			100	75	67
EBOV				99	66
RESTV					99

G) VP40

	BDBV	SUDV	TAFV	EBOV	RESTV
BDBV	99	76	89	83	77
SUDV		100	76	75	79
TAFV			100	82	77
EBOV				99	74
RESTV					98

H) Pathogenicity Group Comparison

	GP	L	NP	VP24	VP30	VP35	VP40
Pat:Pat	60	76	70	80	74	71	78
Pat:NPat	58	74	68	81	69	66	74
NPat:NPat	97	99	99	100	99	99	98

Figure 4.1: Intra- and inter-species protein sequence conservation for each of the 7 ebolavirus proteins, and conservation between pathogenic and non-pathogenic groups. **A)** GP gene. **B)** L gene. **C)** NP gene. **D)** VP24 gene. **E)** VP30 gene. **F)** VP35 gene. **G)** VP40 gene. **H)** Comparison of pathogenicity groups for all 7 proteins.

4.4.2 Ebolavirus Specificity Determining Positions

SDPs were identified using S3Det (see methods) (Rausell et al. 2010), which returned a total of 166 SDPs spread across all seven ebolavirus proteins, see

Table 4.4. Overall, 3.43% of residues can be considered SDPs, with the highest percentage identified in VP30 (6.94%) and the lowest in L (2.39%). Previously, 180 SDPs were identified (reported as 189 but SDPs in sGP and GP were identical as they share a common N-terminus), so the increased number of sequences has reduced the total number of SDPs. This is due to positions in the protein alignments that now have greater variation and are not sufficiently conserved to be classed as SDPs. The increased number of sequences used has also resulted in positions that with a smaller set were too variable to be classed as SDPs, but with many more sequences the level of variation is reduced and the positions are classed as SDPs. Figure 4.2 provides a visual summary of the overlap between the original set of SDPs and the new set of SDPs. Overall, 34 SDPs were lost, while 20 new SDPs were identified, with a 73% overlap of SDPs.

Very few SDPs that were previously identified were lost for the structural proteins (VP24, VP30, VP35 and VP40), with only a single SDP lost in VP30. New SDPs were identified for each of these proteins ranging from two for VP24 to seven for VP40 (Figure 4.2 and Table 4.4). In contrast, for NP, GP and L, many SDPs were lost (Figure 4.2 and Table 4.4), ranging from five for NP to 17 in L. At the same time, very few new SDPs were identified in these proteins, only four between all three proteins. Overall, in comparison with the smaller study we see a range between a net gain of seven SDPs (for VP40) to a net loss of 17 (for L). The full SDP list, including lost, retained and gained SDPs, can be found in Appendix 3 Tables 2-8.

Table 4.4: Summary of the numbers of SDPs lost, retained, and gained in the updated SDP set.

Protein	SDPs in Original Set	SDPs Lost	SDPs Retained	SDPs Gained	SDPs in Updated Set
NP	29	5	24	0	**24**
VP35	19	0	19	3	**22**
VP40	9	0	9	7	**16**
GP	30	11	19	1	**20**
VP30	17	1	16	4	**20**
VP24	9	0	9	2	**11**
L	67	17	50	3	**53**

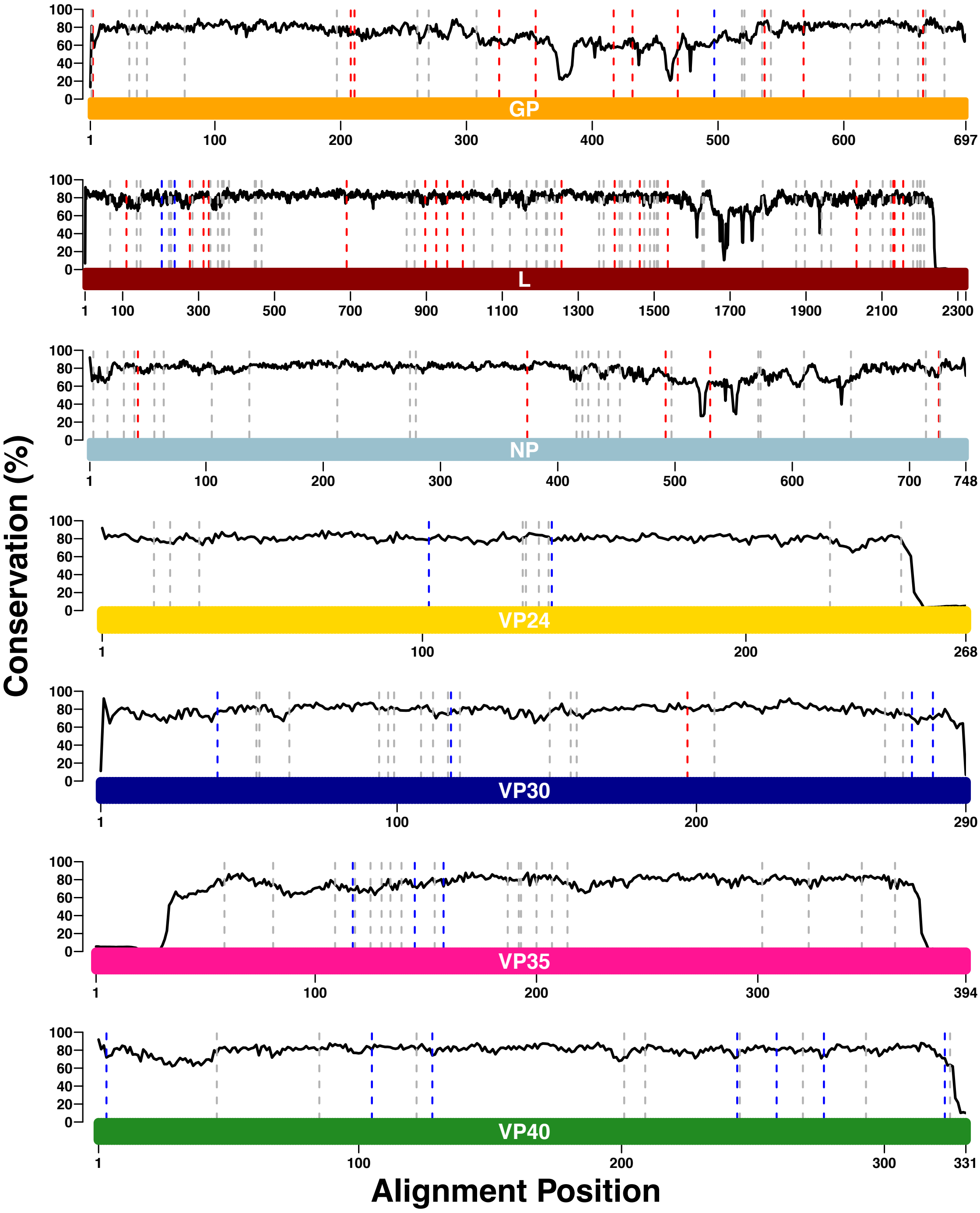


Figure 4.2: Comparison of SDPs previously identified with 196 ebolavirus genomes to the set of SDPs identified with 1,408 genomes. The coloured bars represent the lengths of the protein sequence alignments, and each bar is labelled with the name of the protein that it represents. The solid black line represents the Jensen-Shannon conservation score. Dashed red lines represent previously identified SDPs that were lost in the updated analysis, dashed grey lines represent SDPs that were retained, and dashed blue lines represent new SDPs that have been identified. Note – x-axes differ in their scales between subplots.

4.4.3 Structural Analysis of SDPs

The previous study mapped 47 of 180 SDPs onto ebolavirus protein structures or models (Pappalardo et al. 2016). In the current study, 88 of 166 SDPs were mapped onto proteins structures. Therefore, ~53% of the SDPs were mapped onto protein structures, with the increase partly due to the ability to model the structure of the RNA dependent RNA polymerase (L; see Methods). Figure 4.3 shows some basic functional characterisation of the SDPs, divided in to the lost, retained, and gained categories of SDPs. Figure 4.3.A shows that very few of the SDPs are predicted to have a stabilising effect on the protein structure (positive $\Delta\Delta G$ values), with none having a large stabilising effect, and the majority of the SDPs have a destabilising effect (negative $\Delta\Delta G$ values). Interestingly, some of the SDPs lost with the larger genome set are predicted to be fairly destabilising (7 of the lost SDPs have a predicted stability change < -1.0 kcal/mol) and could have falsely been interpreted as being potentially functional. Figure 4.3.B shows that the majority of the SDPs are relatively solvent inaccessible, though multiple SDPs are located on the surface of the proteins in each SDP category. The majority of the SDPs are fairly conservative changes, see Figure 4.3.C, and some of the most radical changes are within the lost SDPs set. The SDPs being mostly conservative changes is consistent with the majority of the SDP set having a small impact on protein function and stability, with relatively few SDPs having a major impact.

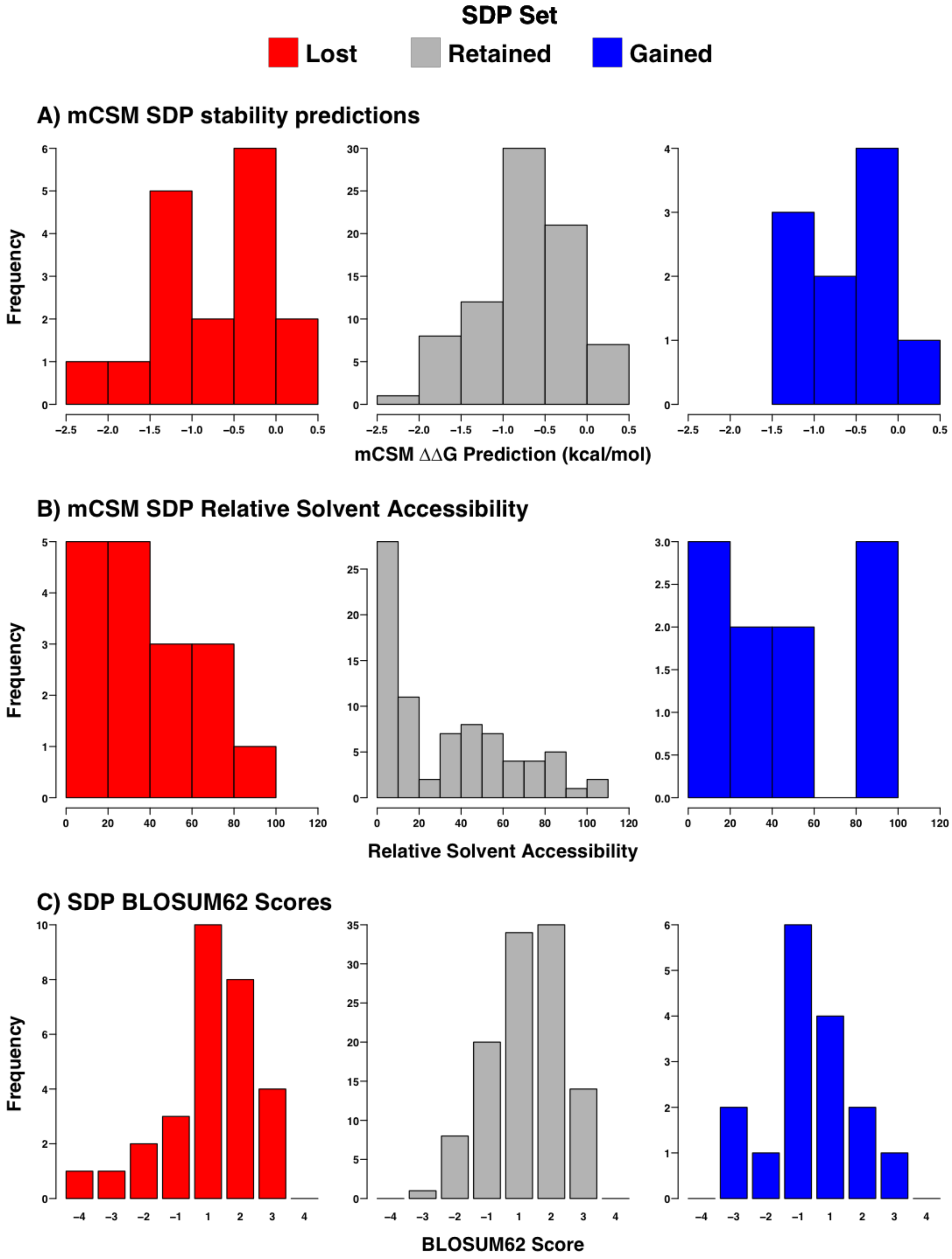


Figure 4.3: Functional characterisation of SDPs. **A)** mCSM stability change predictions for SDPs mapped to structures. **B)** Relative solvent accessibility of SDPs mapped to structures. **C)** BLOSUM62 scores for all SDPs. SDPs only found with 196 genomes are shown in red (lost SDPs), SDPs found in both sets are shown in grey (retained SDPs), and SDPs only found when using the 1,408 genomes set are shown in blue (gained SDPs). Note – y-axes differ in their scales between subplots.

4.4.4 SDPs Affecting Protein Function

Of the 47 SDPs previously mapped onto protein structures, eight (four in VP24, two in VP40 and one in each of VP30 and VP35) of them were proposed to be likely to alter protein structure or function, and a further five (three in GP and two in NP) were proposed to have a possible effect but support/evidence was limited (Pappalardo et al. 2016). Using the larger genome set, 12 of these 13 SDPs were retained, with only one of the lower confidence SDPs lost from NP (A705R). Of the 20 new SDPs, 10 were mapped onto protein structures and we propose that one additional SDP is likely to have an effect on protein structure and function, see Table 4.5. A summary of the SDPs with proposed functional impacts is shown in Table 4.6, including SDPs only found using 196 genomes and SDPs only found using 1,408 genomes. Overall, the set of likely functional SDPs is highly similar between both analyses, with only one low-confidence SDP lost and one high-confidence SDP gained.

Table 4.5: Summary of SDPs per ebolavirus protein, and the predicted functional impacts.

Protein	Length	SDPs	%Residues SDPs	SDPs Modelled	Probable Integrity	Probable Interface	Possible Integrity	Possible Interface
NP	739	24	3.25	10	0	0	1	0
VP35	340	22	3.47	4	0	1	0	0
VP40	326	16	4.91	13	1	1	0	0
GP	676	20	2.96	10	0	0	0	3
VP30	288	20	6.94	5	0	1	0	0
VP24	251	11	4.38	10	1	4	0	0
L	2,212	53	2.39	36	0	0	0	0

Table 4.6: Summary of the SDPs with proposed functional impacts identified using the 196 and the 1,408 genome sets.

Protein	SDP	Functional Effect	Confidence	Status
NP	R105K	Stability: Loss of hydrogen bonding	Possible	Retained
NP	A705R	Stability: Introduction of salt bridge with E694	Possible	Lost
VP35	E269D	Interface: Dimeric VP35 interface	Probable	Retained
VP40	P85T	Interface: Octameric VP40 interface	Probable	Retained
VP40	Q245P	Stability: Breaks an alpha helix	Probable	Retained
GP	I260L	Interface: Within GP glycan cap	Possible	Retained
GP	T269S	Interface: Within GP glycan cap	Possible	Retained
GP	S307H	Interface: Within GP glycan cap	Possible	Retained
VP30	R262A	Interface: Dimer interface, loss of hydrogen bond	Probable	Retained
VP24	T131S	Interface: With KPNA5	Probable	Retained
VP24	M136L	Interface: With KPNA5	Probable	Retained
VP24	Q139R	Interface: With KPNA5	Probable	Retained
VP24	R140S	Interface: With KPNA5	Probable	Gained
VP24	T226A	Stability: Loss of hydrogen bond	Probable	Retained

4.4.5 SDPs in the VP24-Karyopherin-α5 Binding Site

One of the key determinants of a virus' ability to successfully infect a host is its ability to downregulate the host immune system, with the human interferon (IFN) response being a key defence mechanism against viral infection (Haller et al. 2006). Ebolavirus VP24 is able to downregulate the host IFN response, in order to prevent attenuation of the virus' replication cycle, and functions by binding to several human proteins, including STAT1 and its binding partners karyopherin alpha proteins (KPNA; KPNA1, KPNA5, and KPNA6). The interaction of STAT1 and KPNA proteins is a critical step in the upregulation of IFN signalling, and in binding to these proteins VP24 is able to prevent IFN upregulation (Reid et al. 2006; Reid et al. 2007; Zhang et al. 2012; Xu et al. 2014).

Using the larger genome set, 11 SDPs were identified in VP24, including two that were not previously identified. In the original study, three of the VP24 SDPs (T131S, M136L, Q139R) were present in the binding site with human karyopherin α5 (KPNA5). These SDPs were proposed as likely to alter the interaction of *Reston virus* VP24 with KPNA5 (and possibly other human karyopherin α proteins), and were predicted to reduce the ability of *Reston virus* to inhibit the human IFN response (Pappalardo et al. 2016).

Of the two newly identified SDPs (I102L and R140S), R140S is also in the KPNA5 binding site (Figure 4.4) and is adjacent to the Q139R SDP. Structural analysis shows R140 can form hydrogen bonds with E476 (backbone) and Y477 (sidechain) in KPNA5, and also with the sidechain of VP24 E113. S140, the *Reston virus* amino acid, would still have the potential to form hydrogen bonds but not as extensively as R140. Interestingly, residue 140 also differs between *Ebola virus* and *Bundibugyo virus* (R140H) and *Tai Forest virus* (R140Q), and has been implicated in a reduced efficiency of Bundibugyo virus VP24 in downregulating IFN signalling (Schwarz et al. 2017).

In the original SDP study, these differences in the *Reston virus* VP24 were predicted to make it less able to downregulate IFN signalling, contributing to its lack of human pathogenicity (Pappalardo et al. 2016). Recent experimental work has confirmed that *Reston virus* VP24 is less able to downregulate IFN signalling. In a study using reporter assays to analyse the effects of different ebolavirus VP24 proteins on IFN signalling, *Reston virus* VP24 was shown to downregulate IFN signalling by ~30%. In contrast, the VP24 proteins of the four pathogenic *Ebolavirus* species were able to downregulate IFN signalling by ~80-90% (Guito et al. 2017). Mutations in VP24, including in the VP24-KPNA5 interface site, have also been shown to be essential in adaptation of the virus to a new host (Volchkov et al. 2000; Ebihara et al. 2006; Reid et al. 2006; Reid et al. 2007; Zhang et al. 2012; Dowall et al. 2014; Xu et al. 2014; de La Vega et al. 2015; Pappalardo et al. 2017). Our updated SDP results combined with this experimental work place greater confidence in the role of VP24 in ebolavirus pathogenicity.

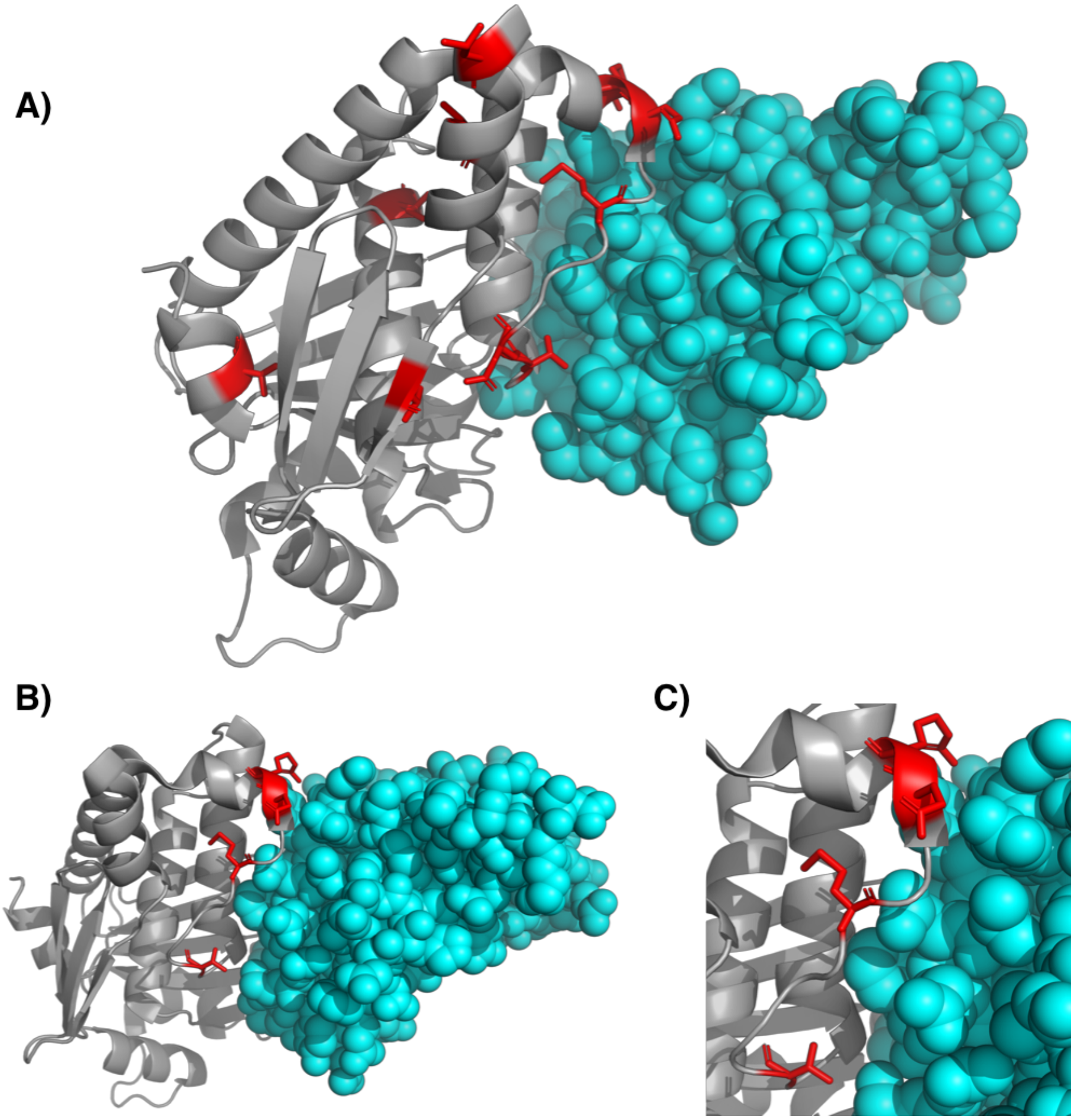


Figure 4.4: SDPs identified in VP24. *Ebola virus* VP24 is shown in cartoon format and coloured grey, KPNA5 is shown in sphere format and coloured cyan, and SDPs are shown in stick format and coloured red. **A)** All SDPs identified in VP24. **B)** VP24 SDPs in the interface with KPNA5. **C)** Zoomed in view of the VP24 SDPs in the KPNA5 interface.

4.4.6 New SDPs in VP40 and L

The most dramatic increase in SDP number was seen for VP40, rising from nine to 16 with the increase in genome number. However, while we were able to map five of these new SDPs to structure, we could find no possible functional impact for these SDPs (Figure 4.5.A). The most confident functional SDPs in VP40 are still P85T, which is in the interface of the octameric form of VP40 and is likely to affect this interaction, and Q245P, which breaks an alpha helix and is likely to affect the protein's integrity (Table 4.6).

Similarly, for the protein L we were able to map 36 SDPs to structure that we previously could not (Figure 4.5.B). However, the specific functional regions of L are still poorly understood, and therefore functional interpretation of these SDPs is not possible. Given the known role of L as an RNA polymerase, it is more difficult to see how changes in this protein could be related to differences in pathogenicity. However, L may have additional roles that we currently have no knowledge of, and so it is not impossible for these SDPs to have an effect on pathogenicity. No other new SDPs were mapped to protein structures.

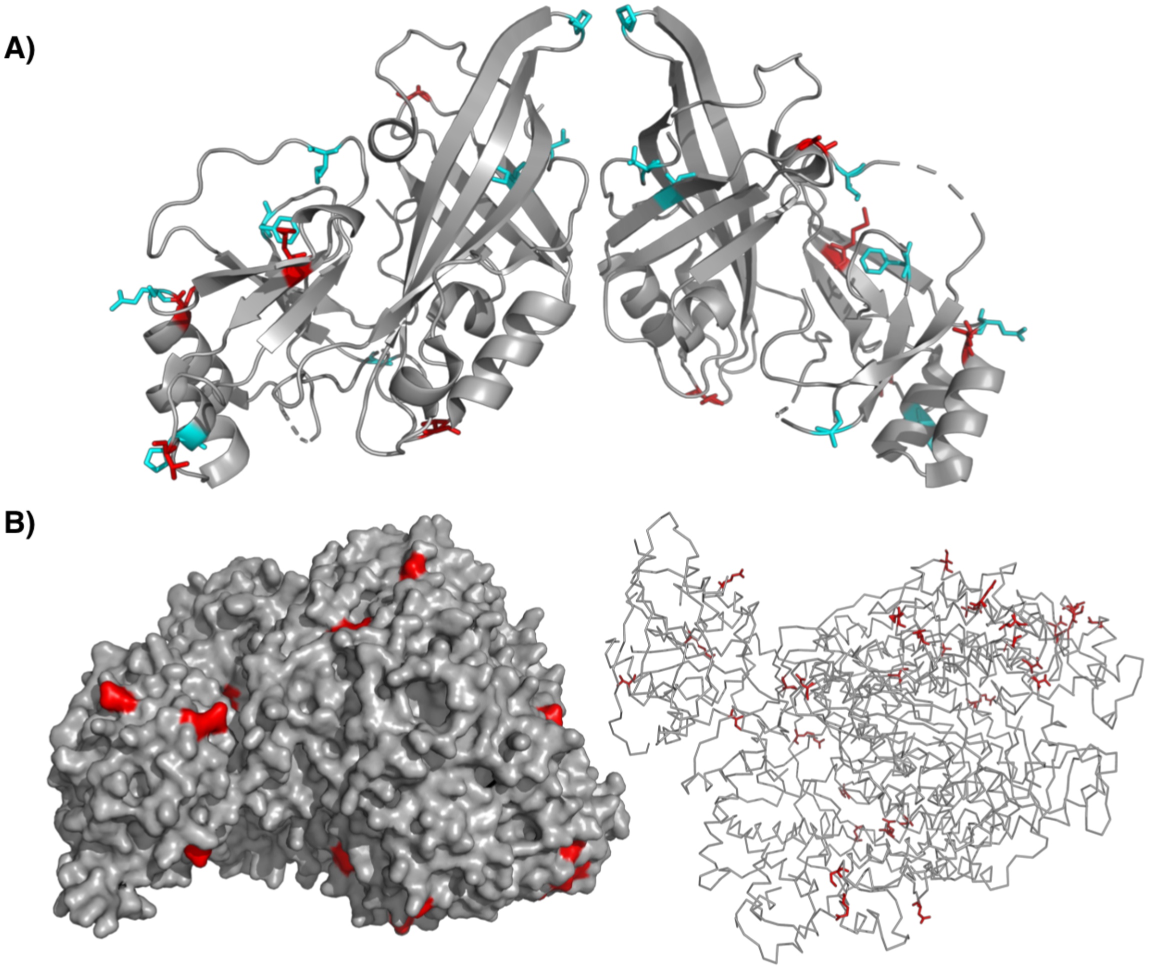


Figure 4.5: SDPs mapped to VP40 and L. **A)** SDPs identified in VP40 – VP40 is shown in cartoon format and coloured grey, SDPs are shown in stick format with retained SDPs coloured cyan and gained SDPs coloured red. **B)** SDPs mapped to the Phyre2 structure of L, shown as a surface representation and as a ribbon representation – L is shown in grey and SDPs are shown in red.

4.5 Discussion

Ebolaviruses represent a severe threat to public health. This has been highlighted by the three ebolavirus outbreaks in the past 5 years, including the 2013-2016 West African epidemic, which was the largest outbreak in history (CDC 2018). As such, identification of the key molecular determinants of ebolavirus pathogenicity is essential for effectively tackling the future threat of ebolaviruses. We previously published an approach, combining computational methods with detailed structural analysis, as a tool for understanding ebolavirus pathogenicity (Pappalardo et al. 2016). In this work we have presented an important update of the previous study, utilising the large number of ebolavirus genomes that have since become available. We have refined the previous set of 180 SDPs to a set of 166 SDPs that are associated with human pathogenicity. We were able to map 88 of these SDPs to structures, and we again performed detailed structural analyses to determine which SDPs could have a functional role in pathogenicity. We identified nine SDPs that are likely to have an impact on protein function, including all eight of the SDPs we previous identified to have a likely functional impact. The additional likely functional SDP was found in the VP24-KPNA5 interface where three other likely functional SDPs are located.

We cannot rule out a role in pathogenicity for other positions in the proteins, especially as not all of the refined set of 166 SDPs could be mapped to structures, and insight for SDPs mapped to L was limited by a lack of knowledge about the functionally important regions of the L protein. There may also be changes at the non-coding level that affect pathogenicity, such as changes to recently identified ebolavirus miRNAs (Duy et al. 2018). However, this small set of functional SDPs that we have identified again suggests that a small number of changes at key positions in the *Reston virus* genome could be enough to make it pathogenic to humans (Pappalardo et al. 2016; Pappalardo et al. 2017). *Reston virus* is known to circulate within pig breeding populations (Barrette et al. 2009; Marsh et al. 2011; Miranda & Miranda 2011; Fischer et al. 2017), and so a human pathogenic *Reston virus* would represent a severe threat to public health if it emerged in pigs, where it could spill-over in to humans.

The 73% overlap of SDPs observed at the sequence level when using 196 and 1,408 genomes, combined with the highly similar set of likely functional SDPs, suggests that our approach is fairly robust. Firstly, these highly similar results place greater confidence in the roles of the nine likely functional SDPs in the pathogenicity of ebolaviruses. This confidence is further increased by the recent experimental evidence for the functional differences between *Reston virus* VP24 and the VP24s of the human pathogenic ebolaviruses. Secondly, the robustness of our approach, even when used on a relatively small number of sequences, highlights its potential more generally in identifying functional differences between groups of viruses. Our approach can now more confidently be applied to many other groups of viruses, and become a useful tool for rapid computational characterisation of viruses. This type of analysis will be crucial in the future for extracting meaningful insights from the large amount of data generated during virus outbreak from in-field sequencing, and unpicking the molecular determinants of virus phenotypes.

Chapter 5: Transition-to-Transversion Bias in the Evolution of *Ebola virus*

This paper is entitled: “Transition-to-Transversion Bias in the Evolution of *Ebola virus*”. This work is currently being prepared for submission.

My contribution to the work was as follows:

1. Helped devise the research with Mark Wass, Martin Michaelis, and Tim Fenton
2. Wrote all of the scripts used in the project and processed all of the data
3. Produced all figures and tables
4. Wrote the manuscript with help from Mark Wass

5.1 Abstract

Transition-to-transversion (Ti:Tv) bias is a commonly observed mutation pattern across diverse species. We have utilised 1,307 *Ebola virus* genome sequences from the 2013-2016 West Africa outbreak, as well as 48 genome sequences from previous outbreaks, to study Ti:Tv bias in *Ebola virus* evolution. The West Africa outbreak was the largest in history, and provides the opportunity to study the mutational effects on the virus of long-term exposure to human hosts.

Compared to the 1976 reference genome, transition mutations were found to be far more common than transversion mutations, with a mean Ti:Tv of 5.40 across the genome, 6.70 in coding regions, and 4.10 in non-coding regions. This broad trend was observed across outbreaks. Ti:Tv bias was variable by genome region, for example non-coding region 4 (region 5,457-6,038) had a mean Ti:Tv of 15.30 and a maximum Ti:Tv of 25.00. Samples from a 2014 outbreak in the Democratic Republic of the Congo exhibited a distinct Ti:Tv bias, with a mean Ti:Tv bias of 9.50 across the genome. For the gene NP, these samples had a mean Ti:Tv of 29.40, and a maximum Ti:Tv of 31.00, compared to 6.00 mean and 7.14 maximum for samples in the West Africa outbreak. Within the West Africa outbreak the mean Ti:Tv bias was 6.66, with a maximum Ti:Tv bias of 18.00.

Analysis of the protein coding mutations, showed that the majority of transition mutations result in synonymous outcomes, but non-synonymous outcomes from transition mutations still outnumber synonymous and non-synonymous outcomes from transversions. There was also no clear difference observed in the severity of non-synonymous changes for transition and transversion mutations, based on BLOSUM62 scores, suggesting a mutational cause, not selective. No clear trend was observed for 5' or 3' context preferences of transition mutations, suggesting that the mutational causes of the Ti:Tv bias are context-independent. Our data show that there has been a large Ti:Tv bias in the evolution of *Ebola virus* between and within outbreaks, and it is likely that this is mostly a mutational rather than selective process, with possible enzymatic contribution from host ADAR and APOBEC editing.

5.2 Introduction

A single nucleotide in a DNA or RNA sequence can undergo three possible substitutions, which allows for a total of 12 types of nucleotide substitutions. Of these, four are classified as ‘Transitions’ (A>G, G>A, C>U, & U>C), which means that both the original and substituted nucleotides are either both purines or both pyrimidines. The remaining eight types of substitutions are transversions, where a purine is substituted for a pyrimidine, or *vice versa*.

For each nucleotide’s 3 possible substitutions, 2 will be transversions and 1 will be a transition. Therefore, by pure chance, the ratio of transitions to transversions (Ti:Tv) would be 0.5. However, this is rarely observed in nature. It is well established that there is a general evolutionary bias for nucleotide substitutions to be transitions as opposed to transversions (Vogel & Kopun 1977; Wakeley 1996). The degree of bias differs by organism and also by genome region (Gojobori et al. 1982; Kumar 1996; Rosenberg et al. 2003; Keller et al. 2007; Denver et al. 2009; Lynch 2010; Zhu et al. 2014). Ti:Tv bias is generally observed to be 2-4 (Kumar 1996; Rosenberg et al. 2003; Stoltzfus & Norris 2016), but has been observed to be ~15 for the control region of human mitochondrial DNA (Tamura & Nei 1993; Purvis & Bromham 1997), and ~19 across the mammalian phylogeny of cytochrome b (Meyer et al. 1999). However, transition bias is not universal, with one study of Grasshopper genomes demonstrating a complete lack of transition bias (Keller et al. 2007).

The cause of this general bias remains unclear. There are two main hypotheses for the bias, which are not mutually exclusive. The first is the mutational cause hypothesis – there is a molecular process that directly leads to more transitions than transversions (Vogel & Kopun 1977). The second hypothesis is the selective hypothesis – there is an evolutionary advantage for transitions over transversions, and so transversions are selected against. The mutational cause hypothesis is supported by the observation that, in addition to protein coding regions of genomes, there is also an observed transition bias within introns, non-coding regions, and pseudogenes (Gojobori et al. 1982; Zhang & Gerstein 2003; Jiang & Zhao 2006). In addition to observations of biased accumulation of transitions over transversion in evolution, some studies have shown that the rate of transition mutations is also higher than that of transversions, and is often similar to the evolutionary transition bias for the same organism and genome region (Denver et al. 2004; Pauly et al. 2017).

The second hypothesis, the selective hypothesis, is based on the idea that transitions are generally less damaging than transversions. Due to the degeneracy in the genetic code, nucleotide substitutions do not always cause changes at the amino acid level. The composition of the genetic code means that transitions are less likely to change encoded amino acids and also amino acid changes caused by transitions are more conservative than those caused by transversions (Miyata et al. 1979; Wakeley 1996; Rosenberg et al. 2003; Keller et al. 2007; Stoltzfus & Norris 2016). This allows transitions to occur more frequently than transversions without changes to the fitness of the organism. However, estimation of the severity of amino acid substitutions based on the properties of the amino acids can be a very crude method, and the actual severity of an amino acid substitution also depends on the context within the protein and system. Additionally, the argument that transitions are less damaging and so occur more frequently can be circular – a high transition mutation rate will inevitably lead to more conservative amino acid changes if transitions are more likely to cause conservative changes (Dagan et al. 2002; Yampolsky & Stoltzfus 2005; Lyons & Lauring 2017).

Two recent studies have analysed direct fitness effects of transition and transversion mutations to test the hypothesis that transition mutations are more conservative. The first study combined 8 different data sets of non-synonymous mutations, 7 of which are virus data sets and 1 of which is a beta lactamase gene (Stoltzfus & Norris 2016), and concluded that selection at the protein level plays at best a minor role in the observed Ti:Tv bias. The second study analysed the fitness effects of transitions and transversions in multiple influenza virus and human immunodeficiency virus data sets (Lyons & Lauring 2017), and concluded that selection is a major contributor to the observed Ti:Tv bias. These two studies used different data sets and statistical frameworks and came to contradictory conclusions about the cause of Ti:Tv bias. It seems likely that in reality Ti:Tv bias is a complex interplay of both mutational cause and selection, and that the extent of the bias depends heavily on the organism, genome region, and additional selection and mutation pressures.

The 2013-2016 West Africa *Ebola virus* outbreak is the largest outbreak in humans ever observed. Due to their nature, viruses are subject to the mutational biases of their host cell machinery. Meaning that if a virus uses host polymerases, and host polymerases have a Ti:Tv bias, viruses replicating inside host cells will have a Ti:Tv bias with a mutational cause. Invading viruses will also be targeted by host defence mechanisms. As *Ebola virus* does not naturally circulate within humans, one key question is how did long-term exposure to human hosts during the West Africa outbreak affect the evolution of the virus, and what unique mutational pressures was the virus under while circulating in human populations. Analysis of genome sequences from the West Africa outbreak suggested signs of editing by the human adenosine deaminase enzyme ADAR (Park et al. 2015; Whitmer et al. 2018), similar to previous reports of ADAR editing in viruses (Wong et al. 1989; Murphy et al. 1991; Cattaneo 1994; Rueda et al. 1994; Zahn et al. 2007; Carpenter et al. 2009; Gelinas et al. 2011; Tong et al. 2015). ADAR enzymes deaminate adenine (A) bases in to inosine (I) bases, and inosine is read as a guanosine (G) base, resulting in A>G mutations. The APOBEC family of enzymes are another set of host enzymes associated with mutation signatures in viruses, which are cytidine deaminases that convert cytidine (C) bases in to thymidine (T) bases resulting in C>T mutations (Bishop et al. 2004). Some APOBEC proteins are known to be part of the innate immune system, for example APOBEC3G is upregulated by interferon signalling and is able to restrict virus replication (Wang et al. 2008; Milewska et al. 2018), with some viruses even encoding gene products to inhibit cytidine deaminases, such as Viral infectivity factor (Vif) encoded by Human Immunodeficiency Virus (HIV) (Rose et al. 2004).

In this study we have utilised the substantial increase in the number of sequenced *Ebola virus* genomes, which resulted from the 2013-2016 West Africa outbreak (Carroll et al. 2015; Arias et al. 2016; Baize et al. 2014; Gire et al. 2014; Hoenen et al. 2016; Park et al. 2015; Quick et al. 2016; Simon-Loriere et al. 2015; Tong et al. 2015; Whitmer et al. 2018; Urbanowicz et al. 2016), to study *Ebola virus* mutation patterns. Prior to this outbreak we found 53 sequences for *Ebola virus* genomes, within public databases. From the recent outbreak alone, we found genome sequence data for 1,307 unique samples. This provides the opportunity to study *Ebola virus* mutation patterns in more depth than ever before, both within the recent outbreak and over time compared to the 1976 Mayinga *Ebola virus* reference genome, including searching for signatures of mutational pressure by host ADAR and APOBEC enzymes.

5.3 Methods

5.3.1 *Ebola virus* Genome Sequences

Ebola virus sequence data was downloaded on 04/09/2017 from three separate databases: NCBI (Brister et al. 2015), ViPR (Pickett et al. 2012), and a database of sequences compiled from (Urbanowicz et al. 2016), which is available here: https://github.com/ebov/space-time. For NCBI and ViPR, the options to download only full-length sequences were used. From this set of sequences, a unique sample set was produced using unique sample GenBank identifiers.

5.3.2 Genome Sequence Processing

All open reading frames (ORFs) were identified for each *Ebola virus* sample using the EMBOSS getorf program (Rice et al. 2000). A BLAST search was performed for each sample's ORFs against the *Ebola virus* reference set of proteins, which were downloaded from UniProt (Bateman et al. 2015). For six of the seven proteins encoded by the *Ebola virus* genome NP, VP35, VP40, VP30, VP24, and L, the top ORF hit was selected as the corresponding protein for the sample.

The final protein, GP, ORF hits were then further processed, as full-length GP is the result of mRNA editing and is not directly encoded in the ebolavirus genome. Within the GP ORF there is a slippery sequence that leads to ribosome slipping and insertion of an adenine base, which causes a frame-shift that will avoid the first stop codon in the original ORF (Volchkov et al. 1995). This adenine insertion is within a stretch of seven consecutive adenine residues and results in an 8-adenine motif. For each of the original GP ORF hits from BLAST the original genomes were searched from the ORF start-point for this 7-adenine motif and were edited to an 8-adenine motif. New ORFs were then identified for each of the edited sequences, and the ORFs were searched against the *Ebola virus* reference protein set using BLAST. For each sample, the top GP hit for the edited ORFs was then selected as the ORF for full-length GP.

5.3.3 Sample Filtering

To remove low-quality or partial-genome sequences from the data set, any sample that did not have a BLAST hit for each of the seven *Ebola virus* reference proteins, which covered ≥90% of the reference protein length, was removed from the set. Additionally, samples were removed if the host was not known to be *Homo sapiens* or the sample date was unknown (partial dates, e.g. only the year, were allowed). The resulting filtered dataset contained 1,355 samples and is summarised in Appendix 4 Table 1.

5.3.4 Mutation Calling

All genome sequences were aligned using Clustal Omega (Fabian Sievers et al. 2011). Differences compared to the 1976 reference genome (GenBank identifier AF086833) were then called as mutations for each sample, except in cases where the difference was a gap. Mutations were also removed if either the 5' or the 3' nucleotide was a gap, as it was essential to know the context of the mutations for this work. This led to the removal of fewer than 10 mutations on average per sample (see Appendix 4 Figure 1). For the analysis of mutations within the 2013-2016 West Africa outbreak the oldest sample from the outbreak was used as the reference (GenBank identifier KX000399). Ti:Tv values were calculated for each sample as: Ti:Tv = Total Number of Transitions / Total Number of Transversions. If a sample had no transitions or no transversions for a region then the Ti:Tv value is assigned NA. For non-synonymous mutations BLOSUM scores were calculated using the BLOSUM62 matrix.

5.3.5 Expected Mutation Numbers Per Region

For each genome region, consisting of the seven genes (NP, VP35, VP40, GP, VP30, VP24, and L) and eight non-coding regions (non-coding regions 1-8), the length and base composition was determined. Then, for each genome region in each individual sample, the expected number of each mutation type was calculated. The expected number of mutations was calculated per sample using the total number mutations across the genome and the base composition of each region compared to the whole genome. For example, if a region contains 10% of the total C bases in the genome, then 10% of all C>A, C>U, and C>G mutations would be expected in that region, if the mutations occur randomly throughout the genome.

5.4 Results

5.4.1 Mutations Compared to the 1976 Reference Genome

To understand general mutation patterns in *Ebola virus* over time we first compared each sample to the 1976 Mayinga reference genome, and analysed the 12 types of single nucleotide mutations. The number of mutations for each sample compared to the 1976 reference *Ebola virus* genome is shown in Figure 5.1, separated in to the different mutation types. There is a clear trend of four mutation types being far more prevalent than the other eight types: A>G, G>A, C>U, and U>C. This corresponds to there being one dominant mutation per starting nucleotide, i.e. for starting C bases, C>U mutations are far more common than C>A and C>G mutations (Figure 5.1). In each case the most prevalent mutation is a transition, and for each starting base the number of transitions is significantly higher than the number of transversions ($p < 2.2e\text{-}16$, Wilcoxon rank sum test). These 4 transition mutations are by far the most prevalent for the set of unique mutations from across the samples (Appendix 4 Figure 2).

Low-end outliers in Figure 5.1 are caused by older samples in the set, including samples from the same year as the reference genome. As would be expected, these have far fewer mutations compared to the reference genome than more recent samples, as far less time had elapsed between when they were sequenced and when the reference genome was sequenced. Therefore, these older samples are more similar to the virus reservoir from which the reference genome emerged.

This trend at the whole genome level is also mostly observed for each of the individual coding and non-coding regions of the genome (Appendix 4 Figures 3 and 4). While the general trend is observed for each region, there are clearly differences between regions. The four VP genes (VP24, VP30, VP35, and VP40) are highly conserved and have very few mutations generally. These genes still show a bias towards transitions, but for example VP24 has more G>A, C>U, and U>C transition mutations than transversions, but a low number of A>G transitions. VP30 has a similar pattern, with a low number of C>U mutations. For the genes NP, GP and L the pattern of transition bias is clear for all four starting nucleotide bases.

Using the base composition of the different regions of the genome and the number of mutations of each type across the genome the expected number of mutations for each region based on a random distribution of variants was calculated (see methods). Appendix 4 Figures 5-16 show the observed vs expected numbers of each mutation type for each region of the genome. Transition types G>A, C>U, and U>C generally have fewer mutations than expected in the L gene, whereas A>G has more than expected. For the genes NP and GP, there are more A>G and U>C mutations than expected by chance, but fewer G>A mutations than expected, and more C>U mutations than expected in GP but fewer than expected in NP. All transition types occur more frequently than expected in non-coding region 8.

This mutation pattern translates to an average Ti:Tv of 5.4 across the whole genome (Figure 5.2). The Ti:Tv is slightly higher in coding regions compared to non-coding regions, with mean values of 6.7 and 4.1 respectively. Table 5.1 summarises the Ti:Tv rates for the different regions of the genome. Non-coding region number 4 (reference genome region 5,457-6,038) is a notable exception to the other non-coding regions. It has a mean Ti:Tv of 15.3, which is the highest for any genome region, and the second highest maximum Ti:Tv value of 25. Similarly, VP30 stands out from all of the other coding regions. It has a mean Ti:Tv of 2.135, which is much closer to non-coding regions 2, 6, and 7 than any of the other coding regions.

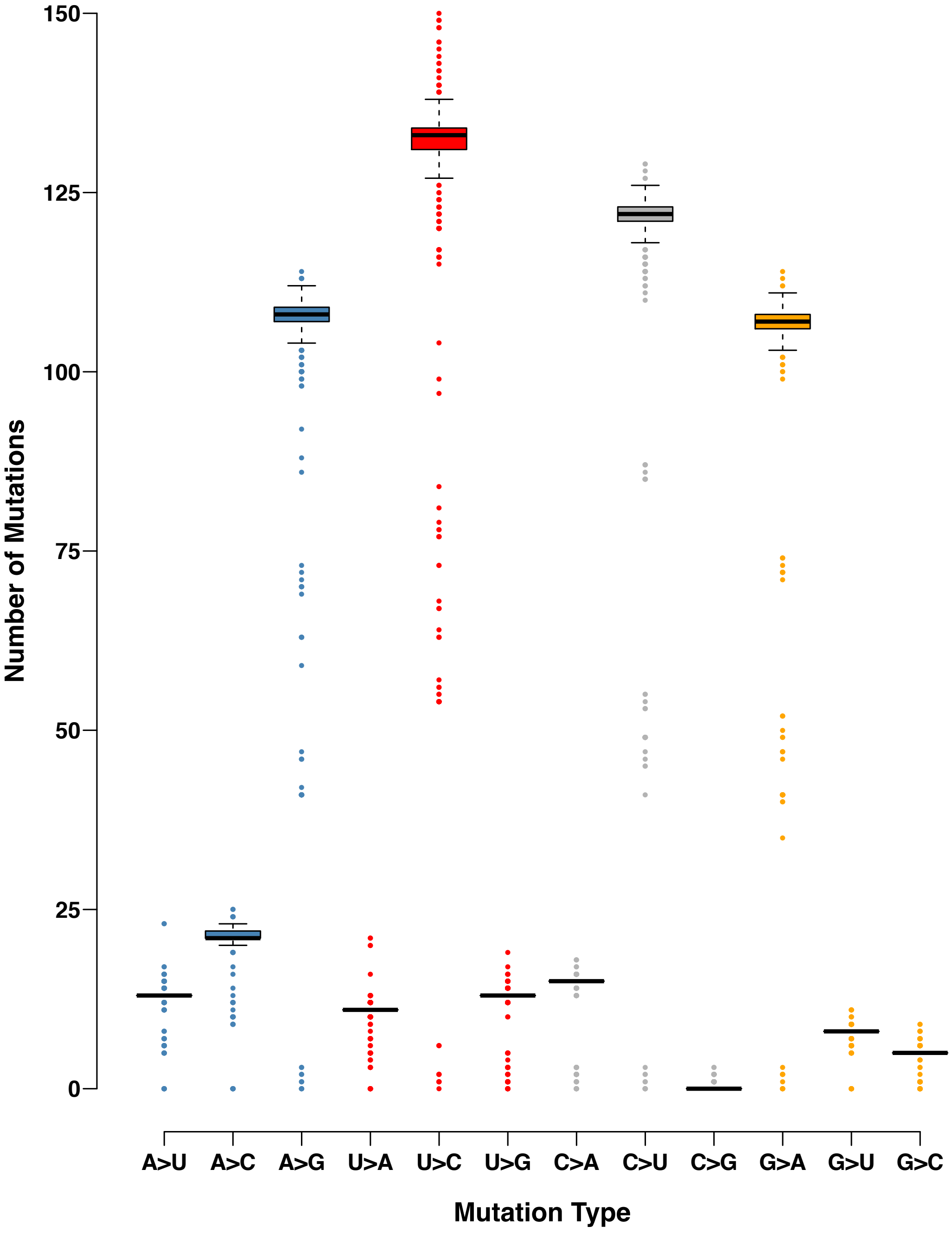


Figure 5.1: Observed mutation types for each sample compared to the 1976 reference genome. Colours indicate the starting base of the mutation – blue A, red U, grey C, and orange G.

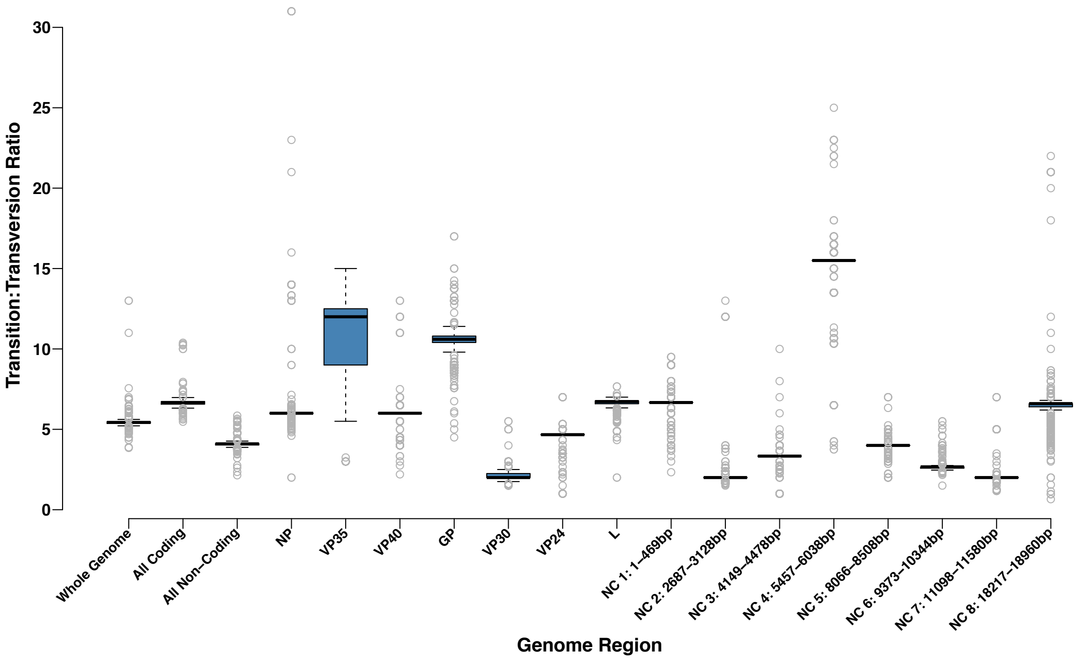


Figure 5.2: Ti:Tv values observed for each sample's mutations compared to the 1976 reference genome.

Table 5.1: Summary statistics of Ti:Tv values for all *Ebola virus* samples.

Genome Region	Mean Ti:Tv	Median Ti:Tv	Minimum Ti:Tv	Maximum Ti:Tv
Whole Genome	5.45	5.43	3.85	13.00
Coding Regions	6.65	6.67	5.47	10.39
Non-Coding Regions	4.09	4.09	2.14	5.85
NP Gene	6.25	6.00	2.00	31.00
VP35 Gene	10.37	12.00	3.00	15.00
VP40 Gene	5.96	6.00	2.20	13.00
GP Gene	10.38	10.60	4.50	17.00
VP30 Gene	2.14	2.00	1.50	5.50
VP24 Gene	4.55	4.67	1.00	7.00
L Gene	6.64	6.73	2.00	7.67
Non-Coding Region 1	6.67	6.67	2.33	9.50
Non-Coding Region 2	2.11	2.00	1.50	13.00
Non-Coding Region 3	3.31	3.33	1.00	10.00
Non-Coding Region 4	15.30	15.50	3.75	25.00
Non-Coding Region 5	3.96	4.00	2.00	7.00
Non-Coding Region 6	2.69	2.67	1.50	5.50
Non-Coding Region 7	2.07	2.00	1.17	7.00
Non-Coding Region 8	6.49	6.60	0.66	22.00

5.4.2 Mutation Patterns Over Time

The mutation patterns seen in Figure 5.1 and Figure 5.2 are heavily skewed to the 2013-2016 West Africa outbreak, due to the large number of these samples in the data set. Figure 5.3 shows that, for the whole genome, the mutation pattern is fairly consistent over time and between outbreaks, with transitions being more common than transversions.

However, there are differences over time between the regions of the genome, and also differences between the types of mutations within the same regions. Patterns over time for the individual coding regions of the genome are shown in Appendix 4 Figures 17-23. The gene GP has many A>G, U>C, and C>U mutations, but comparatively few G>A mutations. Additionally, the A>G, U>C, and C>U mutations have increased at different points in time. For example, samples from the West Africa outbreak have twice as many C>U mutations in GP compared to the previous outbreak, but A>G and U>C mutations are at equivalent levels to the previous outbreak. For the gene L, A>G, U>C, and C>U mutations seemingly steadily increase over time, but G>A mutations show a large increase for the West Africa outbreak compared to the previous outbreak.

Differences between mutation types and genome regions can also be seen for the non-coding regions (Appendix 4 Figures 24-31). Non-coding region 4 is again striking, with transition type mutations increasing steadily over time, and also for a number of U>C high-end outliers within the West Africa outbreak.

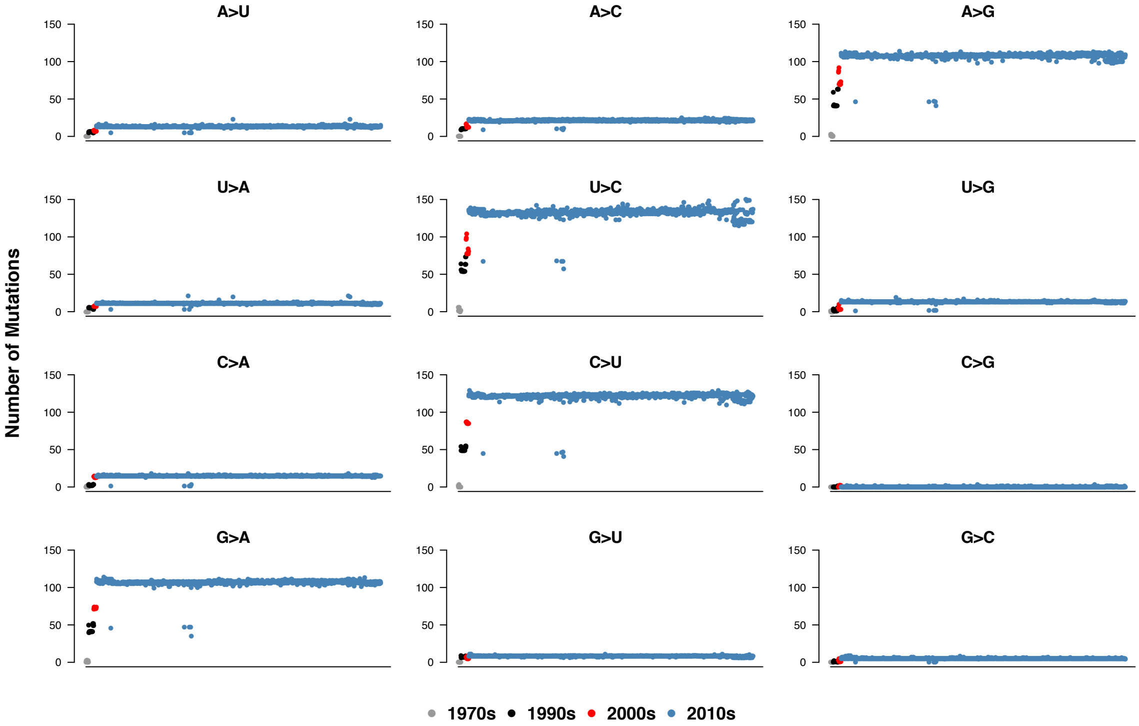


Figure 5.3: Observed mutation types over time for each sample compared to the reference 1976 genome. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

5.4.3 Democratic Republic of the Congo Outbreak

Within outbreaks mutation numbers are fairly constant (Figure 5.3). For the 2010s samples there are 5 low-end outliers that appear to break this trend. However, these samples are actually from a 2014 outbreak in the Democratic Republic of the Congo (DRC), which was a separate outbreak from the main West Africa outbreak and resulted from a separate crossover event (The World Health Organisation 2014; Maganga et al. 2014). The Ti:Tv rates for these 5 samples are striking compared to the West Africa outbreak samples (Table 5.2). For coding regions the mean Ti:Tv is 9.5 for the DRC samples, compared to 6.6 for the West Africa samples. Specifically for the gene NP, DRC samples have a mean Ti:Tv of 29.4, compared to only 6.0 for the West Africa samples. For the gene VP35 this difference is reversed, with West Africa samples having a Ti:Tv of 12.4, compared to only 6 for the DRC samples. Similarly, for non-coding region 8, the Ti:Tv is far higher for West Africa samples than DRC samples, with means of 6.41 and 2.45, respectively, and maximum Ti:Tv values of 18 and 3.67, respectively.

Table 5.2: Comparison of Ti:Tv statistics for 2013-2016 West Africa outbreak samples and 2014 DRC samples.

Genome Region	Mean Ti:Tv West African	Mean Ti:Tv DRC	Median Ti:Tv West African	Median Ti:Tv DRC	Min Ti:Tv West African	Min Ti:Tv DRC	Max Ti:Tv West African	Max Ti:Tv DRC
Whole Genome	5.41	6.46	5.43	6.87	3.85	4.70	5.68	7.56
Coding Regions	6.64	9.49	6.67	10.23	5.47	6.28	7.14	10.39
Non-Coding Regions	4.06	4.09	4.10	4.24	2.14	3.21	4.44	5.07
NP Gene	5.97	29.40	6.00	31.00	4.60	23.00	7.14	31.00
VP35 Gene	12.41	6.00	12.00	6.00	3.25	6.00	15.00	6.00
VP40 Gene	5.96	9.35	6.00	11.00	2.20	2.75	13.00	11.00
GP Gene	10.35	NA	10.60	NA	4.50	NA	14.25	NA
VP30 Gene	2.10	4.80	2.00	5.00	1.60	4.00	3.00	5.00
VP24 Gene	4.64	1.10	4.67	1.00	2.00	1.00	7.00	1.50
L Gene	6.68	5.67	6.73	6.00	4.87	4.33	7.10	6.00
Non-Coding Region 1	6.70	8.00	6.67	8.00	3.00	8.00	9.50	8.00
Non-Coding Region 2	2.04	NA	2.00	NA	1.50	NA	3.80	NA
Non-Coding Region 3	3.37	1.00	3.33	1.00	2.00	1.00	10.00	1.00
Non-Coding Region 4	15.47	16.00	15.50	16.00	10.33	16.00	25.00	16.00
Non-Coding Region 5	3.98	5.00	4.00	5.00	2.22	5.00	6.33	5.00
Non-Coding Region 6	2.65	3.18	2.67	3.60	2.17	1.50	3.33	3.60
Non-Coding Region 7	2.02	6.60	2.00	7.00	1.20	5.00	3.50	7.00
Non-Coding Region 8	6.41	2.45	6.60	2.00	0.66	1.56	18.00	3.67

5.4.4 Mutation Contexts

We have shown that transition mutations are by far the most common types of mutations in *Ebola virus* genomes. This is true within individual genomes and also for the unique set of mutations across samples, suggesting that regardless of possible positive selection of transitions over transversions there must be a process by which transitions happen more frequently than transversions.

The cause of this mutational bias, for example targeting by host enzymes or from physiological biases in polymerase proteins, may be sequence context specific. The identification of such context preferences could provide clues as to the mutational cause. As discussed in the introduction, ADAR enzymes cause A>G mutations and APOBEC enzymes cause C>T mutations, but due to the life cycle of *Ebola virus* the combination of the two enzymes could result in A>G, G>A, C>T, and T>C mutations. Not only do these enzymes have a specific mutation that they cause, they also have specific preferences for sequence context. In human ADAR enzymes, hADAR1 and hADAR2 both have a 5' preference of U > A > C > G, and have 3' preferences of G > C ~ A > U and G > C > U ~ A, respectively, for dsRNA (Eggington et al. 2011). APOBEC enzymes are commonly reported to have a 5' preference for T bases (Pham et al. 2003; Bishop et al. 2004; Beale et al. 2004; Henry et al. 2009; Thielen et al. 2010; Roberts et al. 2012; Alexandrov et al. 2013). However, humans have 11 different genes in the APOBEC family of genes: APOBEC1, APOBEC2, APOBEC3A, APOBEC3B, APOBEC3C, APOBEC3D/E, APOBEC3F, APOBEC3G, APOBEC3H, APOBEC4, and activation-induced cytidine deaminase (AID). These genes have been shown to have different 5' and 3' preferences in DNA, of variable window sizes. For example, APOBEC3G has a stronger preference for a 5' C base (Bishop et al. 2004; Henry et al. 2009). APOBEC3G is thought to play a key role in innate viral immunity, although its antiviral effect is thought to be due to interaction with viral transcripts rather than its enzymatic activity (Fehrholz et al. 2012; Milewska et al. 2018).

As described, ADAR enzymes cause A>G mutations and APOBEC enzymes cause C>T mutations. However, *Ebola virus* is a negative single-stranded RNA virus, which creates a positive strand copy of itself as part of its life cycle (Messaoudi et al. 2015). *Ebola virus* sequence data is stored as the positive strand copy, but signatures of ADAR and APOBEC mutations will depend on whether the positive or the negative strand of the virus was mutated by the enzymes (Figure 5.4). The actions of ADAR and APOBEC enzymes on both the positive and negative strand copies of *Ebola virus* during its lifecycle would result in increased levels of A>G, G>A, C>T, and T>C mutations – the four transition mutation types. This means that ADAR and APOBEC enzymes are a potential mutational cause for transitions and could have a large effect on observed transition biases.

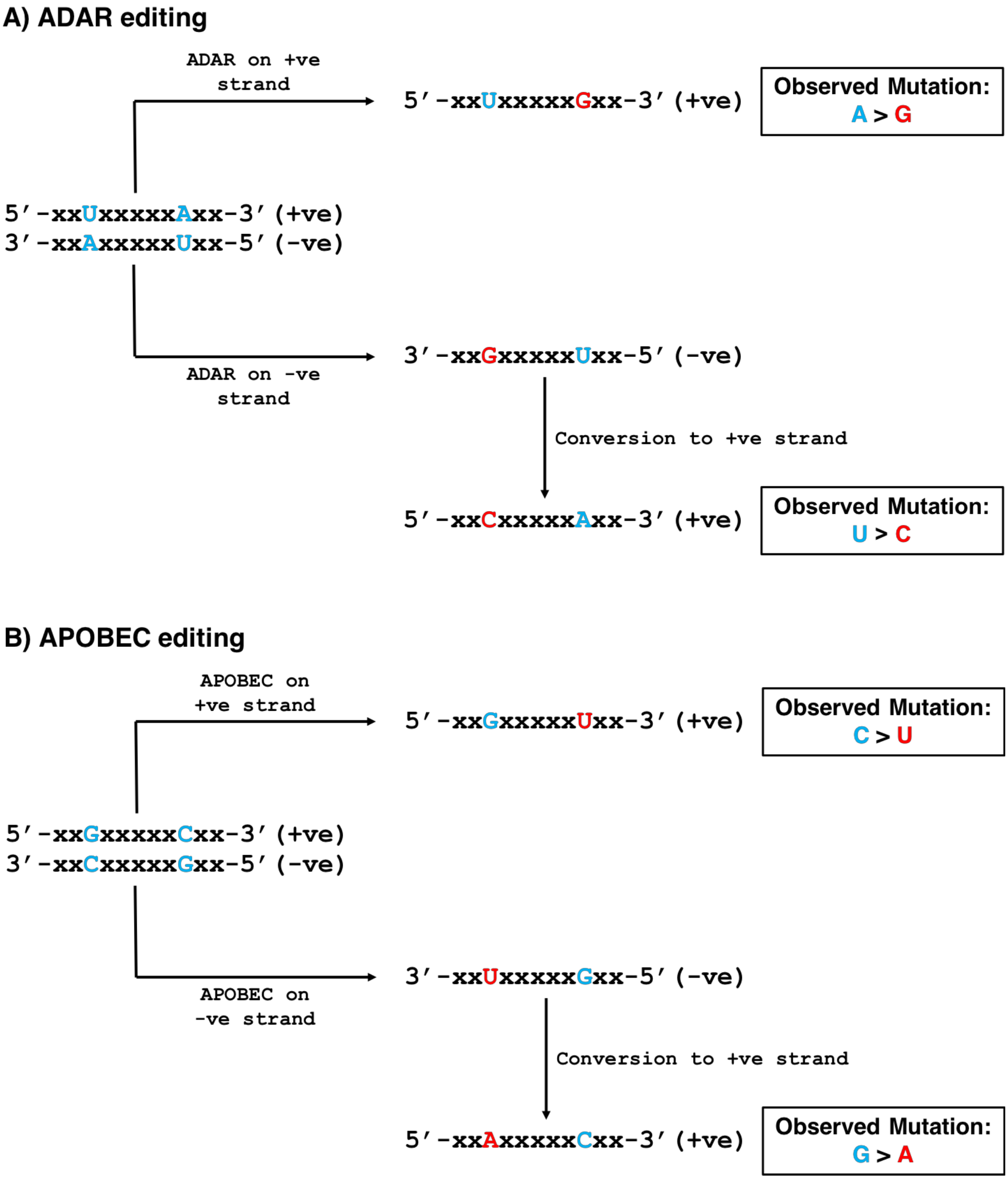


Figure 5.4: Overview of the possible actions of ADAR and APOBEC enzymes acting on the positive and negative strands of *Ebola virus* and the expected outcomes observed in the sequenced genome.

To investigate whether there was a context bias for any of the four transition mutations the 5’ and 3’ nucleotide contexts of all mutations were analysed (Figure 5.5). Here, the 5’ and 3’ contexts are shown for the mutated positions in the positive strand of the genome. Therefore, for G>A and U>C mutations, which are potentially caused by the actions of APOBEC and ADAR on the negative strand of the genome, the 5’ and 3’ contexts would be reversed in the negative strand and the starting bases would be their complementary bases. For example, 5’-A(G>A) context on the positive strand is (C>U)U-3’ on the negative strand. So, for the commonly reported APOBEC preference for a 5’-T base at C>U editing sites, the contexts would be (G>A)A for negative strand editing and U(C>U) for positive strand editing (Figure 5.5.D&G respectively).

Limited 5’ or 3’ context bias was observed for the four transition type mutations (Figure 5.5). Figure 5.5.C shows a possible 5’ A base preference for G>A mutations and Figure 5.5.H shows a possible 3’ A preference for C>U mutations. While statistically significant ($p < 2.2e\text{-}16$, Wilcoxon rank sum test), the differences in mutation numbers between all of the contexts are fairly small. With a larger number of total mutations these contexts may be clearer. There is no clear bias observed for 5’ and 3’ context combinations (Appendix 4 Figures 32-35). This result indicates that the Ti:Tv bias observed is not sequence context specific.

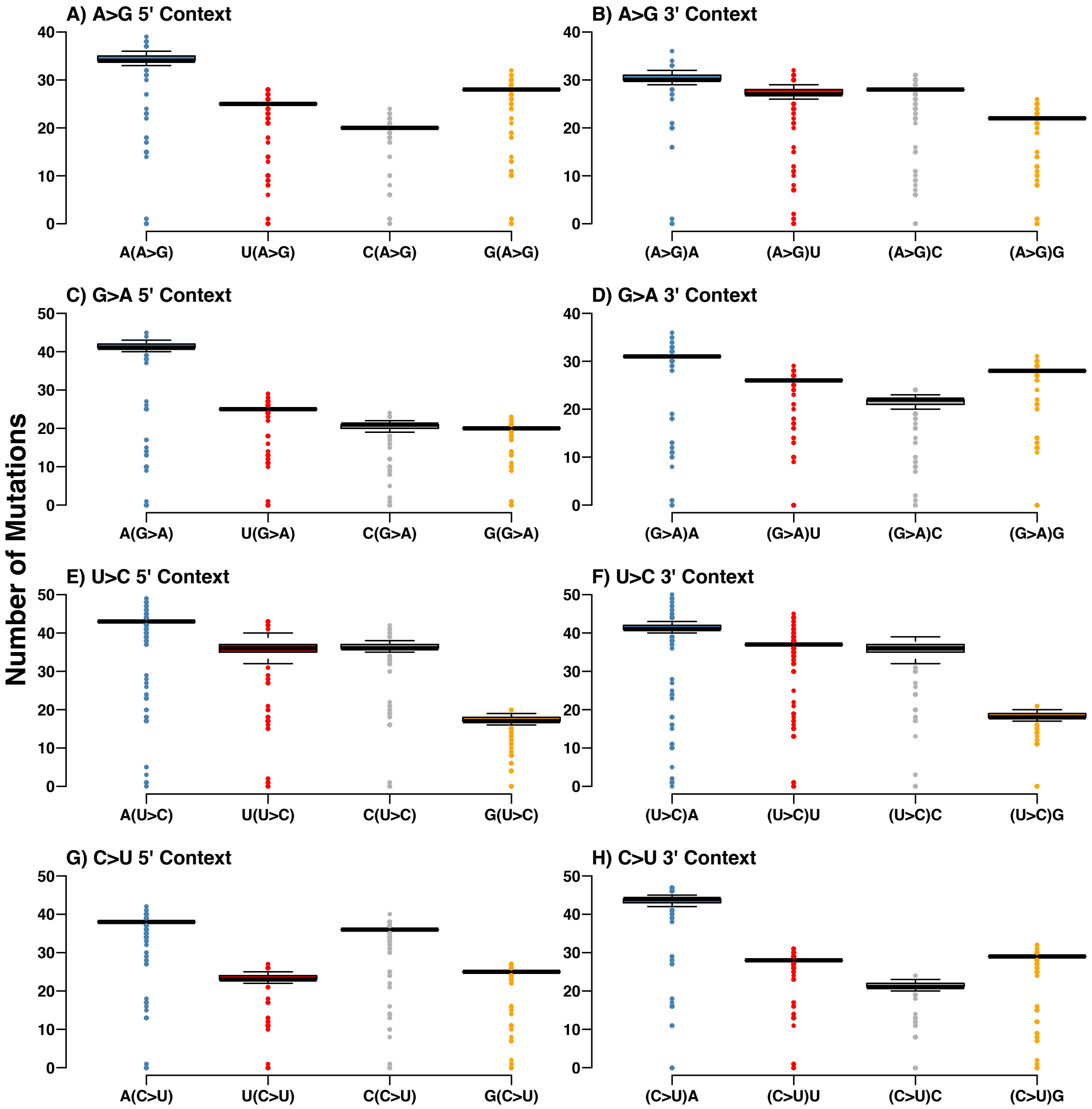


Figure 5.5: Mutation contexts of transition type mutations for all *Ebola virus* samples compared to the 1976 reference genome. **A)** 5’ A>G mutation contexts. **B)** 3’ A>G mutation contexts. **C)** 5’ G>A mutation contexts. **D)** 3’ G>A mutation contexts. **E)** 5’ U>C mutation contexts. **F)** 3’ U>C mutation contexts. **G)** 5’ C>U mutation contexts. **H)** 3’ C>U mutation contexts. Note, y-axes differ in their scales between the subplots.

5.4.5 Protein Coding Effects of Mutation Types

The purifying selection hypothesis posits that Ti:Tv biases are the result of transition mutations being generally less damaging than transversion mutations. The rationale being that the genetic code makes transitions more likely to be synonymous mutations, and more likely that non-synonymous transitions are more conservative changes than non-synonymous transversions (Miyata et al. 1979; Wakeley 1996; Rosenberg et al. 2003; Keller et al. 2007; Stoltzfus & Norris 2016).

Analysis of the mutations affecting protein coding regions of individual *Ebola virus* genomes showed that the majority of transition mutations result in a synonymous protein outcome (Figure 5.6). However, non-synonymous outcomes from transitions are still frequent, with non-synonymous outcomes from G>A, A>C, and U>C mutations being more common than synonymous outcomes for all transversions, and non-synonymous C>U mutations occurring at a similar frequency to the most common synonymous transversions. This trend is more pronounced for the set of unique mutations from all samples, with more non-synonymous outcomes for each type of transition than synonymous outcomes for any transversion (Appendix 4 Figure 36).

For each starting nucleotide, there is no clear trend of the transition mutation being more conservative than the two transversion mutations based on BLOSUM62 scores (Figure 5.7). On average, each sample's U>C mutations are slightly more conservative than U>A or U>G mutations, but A>G and G>A mutations are on average in-between their alternative transversions, and C>U mutations are less conservative than C>A or C>G mutations. For the unique set of mutations there is no clear difference in BLOSUM62 scores between the transitions and transversions (Appendix 4 Figure 37).

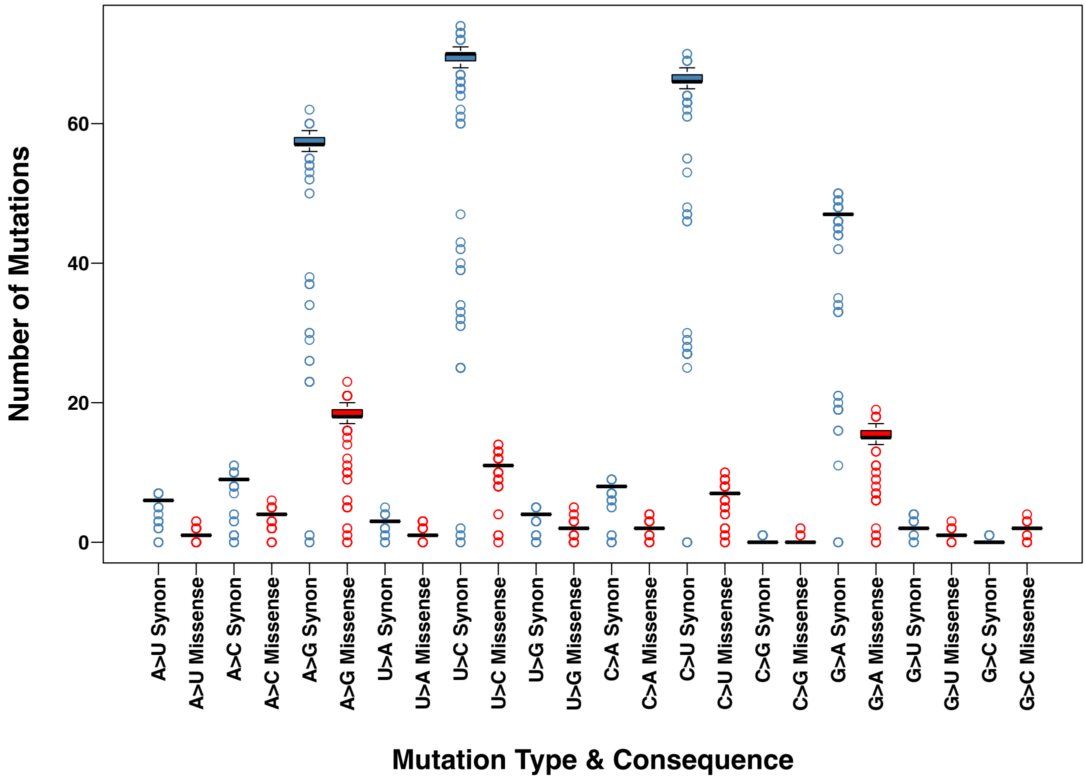


Figure 5.6: Protein coding consequences of each sample's mutations compared to the 1976 reference genome, separated by mutation type.

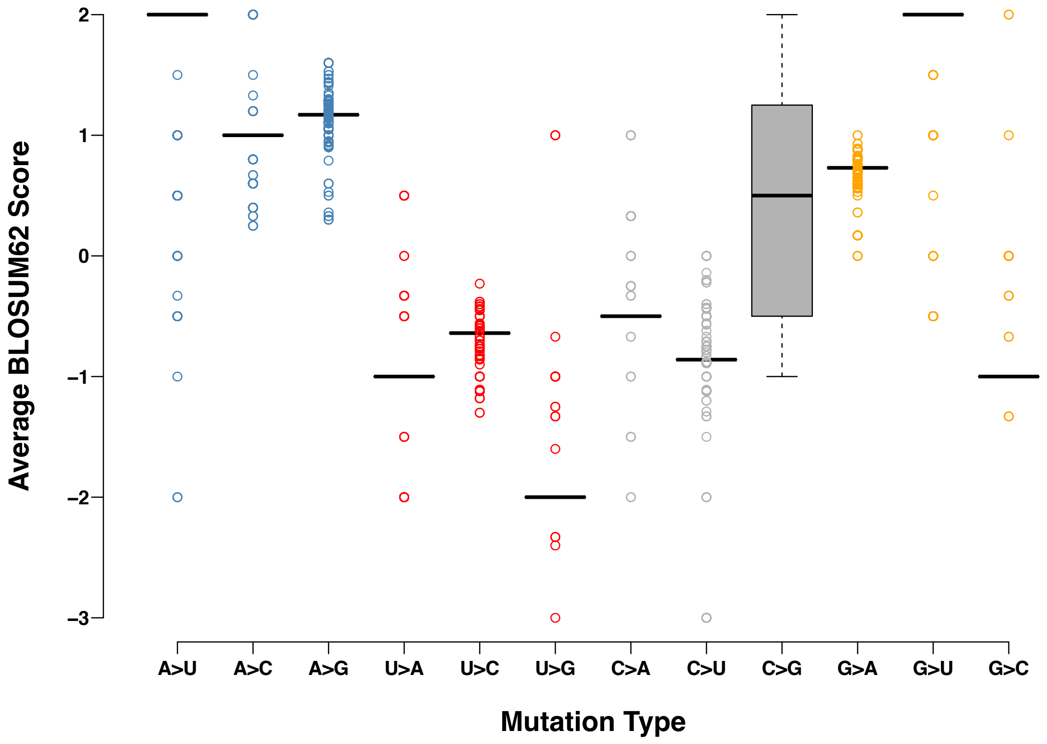


Figure 5.7: Average BLOSUM62 matrix scores for each sample's mutations compared to the 1976 reference genome, separated by mutation type.

5.4.6 Mutation Patterns Within the 2013-2016 West Africa Outbreak

The 2013-2016 West Africa outbreak was the largest *Ebola virus* outbreak to date. Modern sequencing technologies also meant that this was the first outbreak where it was possible to sequence large numbers of patient samples and study the evolution of the virus in real-time over the course of the outbreak (Quick et al. 2016; Hoenen et al. 2016). As discussed, *Ebola virus* ordinarily circulates within non-human hosts, and this outbreak represents the opportunity to study how the virus is affected by long-term exposure to human hosts. Our data presented in Figure 5.3 shows that transitions and transversions compared to the 1976 reference genome did not greatly increase over the 2013-2016 West Africa outbreak. However, this is compared to mutations on a time-scale of multiple decades compared to the four years of the outbreak and it is clear that there is some variation within the outbreak.

To analyse this variation further mutations were called for each sample in the outbreak compared to the earliest outbreak sample available (see Methods). Any sample could have been used as the reference here, and the choice of which early outbreak sample was fairly arbitrary. The reference sample was chosen only because it is likely to be one of the most similar to the first sample to spill-over in to humans, and mutations compared to this sample are most likely to represent the mutations that occurred over course of the outbreak. The aim here is only to determine which types of nucleotide changes are present between samples.

Within the outbreak, transition mutations are again far more common than transversion mutations for each starting nucleotide, i.e. for starting C bases, C>U mutations are far more common than C>A and C>G mutations (Figure 5.8). These differences are statistically significant for each starting base ($p < 2.2\text{e-}16$ for C, T, and G starting bases, and $p < 8.5\text{e-}253$ for starting A bases, Wilcoxon rank sum test).

Notably, as with mutations compared to the 1976 reference genome, there are many high-end outliers, especially for U>C mutations. The average Ti:Tv across the genome was 6.7, higher than the 5.4 observed when comparing to the 1976 reference genome (Figure 5.9 and Table 5.3). The mean Ti:Tv was much higher in the coding regions compared to the non-coding regions (6.5 and 2.8, respectively), and the maximum Ti:Tv observed for any sample across the whole genome was 18.

Over the course of the outbreak the four transition types show a general increase over time, with a number of late-outbreak samples being high-end outliers (Figure 5.10). High-end outliers are not seen for the transversion mutations to the same extent, suggesting that this is not general hypermutation, but is skewed towards transitions. As with the mutations compared to the 1976 reference genome, there is no clear pattern of 5' or 3' context preference (Figure 5.11), or 5' and 3' context combination preferences (Appendix 4 Figures 38-41). Context patterns are harder to analyse for within outbreak mutations, due to the much smaller number of mutations overall.

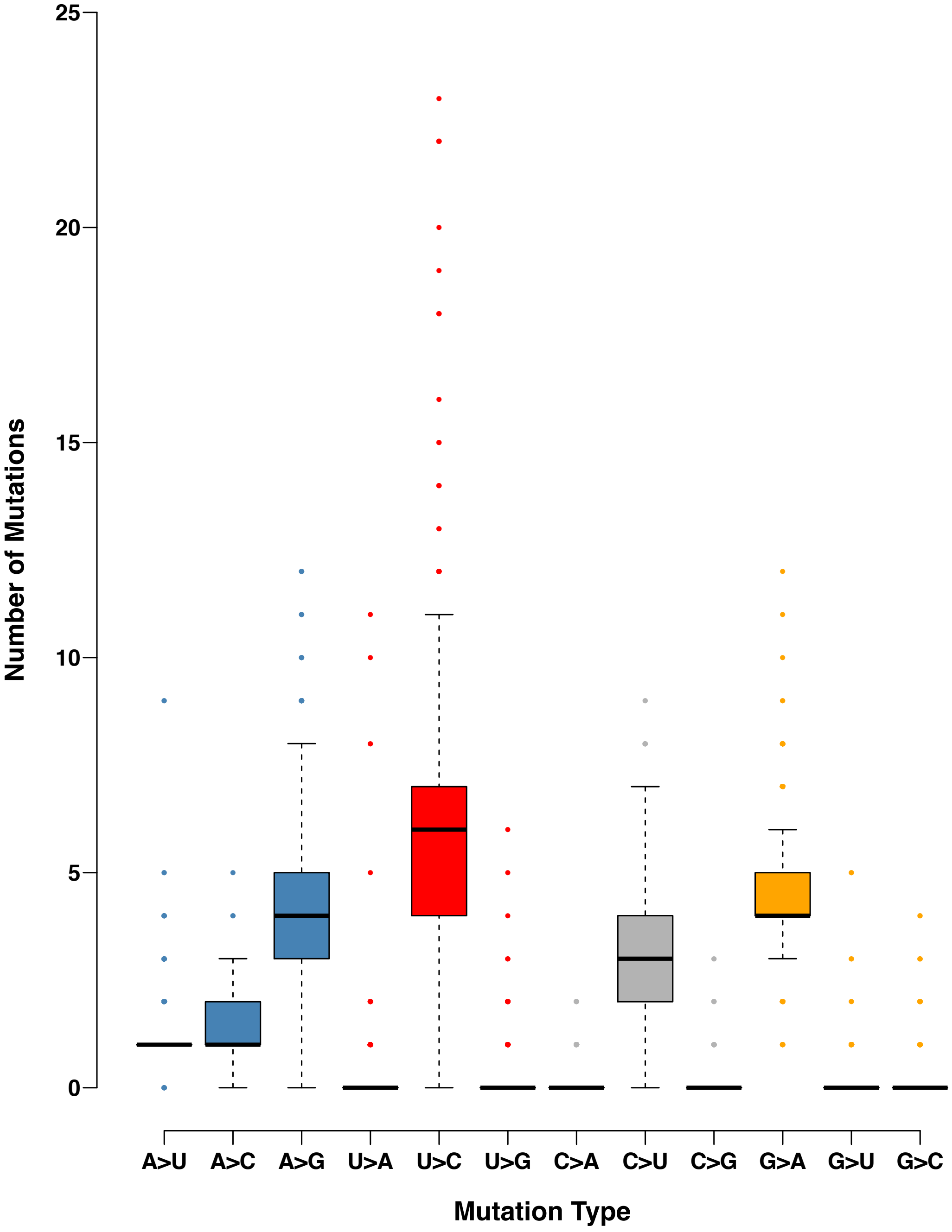


Figure 5.8: Observed mutation types for each sample compared to the early West Africa outbreak reference. Colours indicate the starting base of the mutation – blue A, red U, grey C, and orange G.

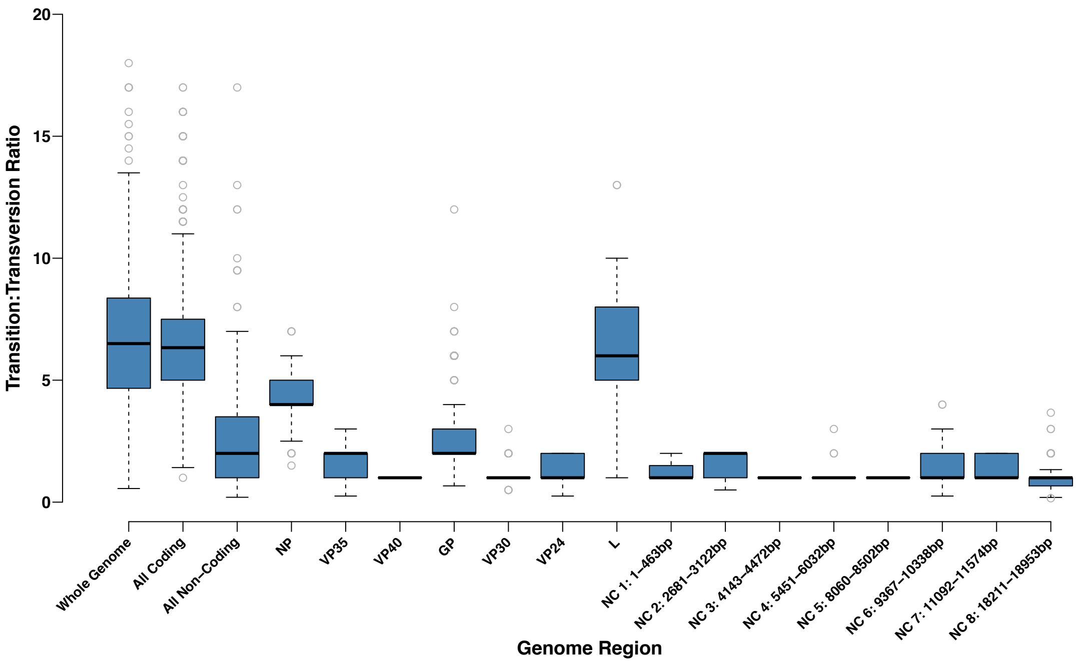


Figure 5.9: Ti:Tv values observed for each sample's mutations compared to the early West Africa outbreak reference.

Table 5.3: Summary statistics of Ti:Tv values for mutations within the 2013-2016 West Africa outbreak.

Genome Region	Mean Ti:Tv	Median Ti:Tv	Minimum Ti:Tv	Maximum Ti:Tv
Whole Genome	6.66	6.50	0.56	18.00
Coding Regions	6.49	6.33	1.00	17.00
Non-Coding Regions	2.75	2.00	0.20	17.00
NP Gene	4.21	4.00	1.50	7.00
VP35 Gene	1.52	2.00	0.25	3.00
VP40 Gene	1.00	1.00	1.00	1.00
GP Gene	2.21	2.00	0.67	12.00
VP30 Gene	1.04	1.00	0.50	3.00
VP24 Gene	1.30	1.00	0.25	2.00
L Gene	6.21	6.00	1.00	13.00
Non-Coding Region 1	1.29	1.00	1.00	2.00
Non-Coding Region 2	1.50	2.00	0.50	2.00
Non-Coding Region 3	1.00	1.00	1.00	1.00
Non-Coding Region 4	1.27	1.00	1.00	3.00
Non-Coding Region 5	1.00	1.00	1.00	1.00
Non-Coding Region 6	1.52	1.00	0.25	4.00
Non-Coding Region 7	1.35	1.00	1.00	2.00
Non-Coding Region 8	1.05	1.00	0.16	3.67

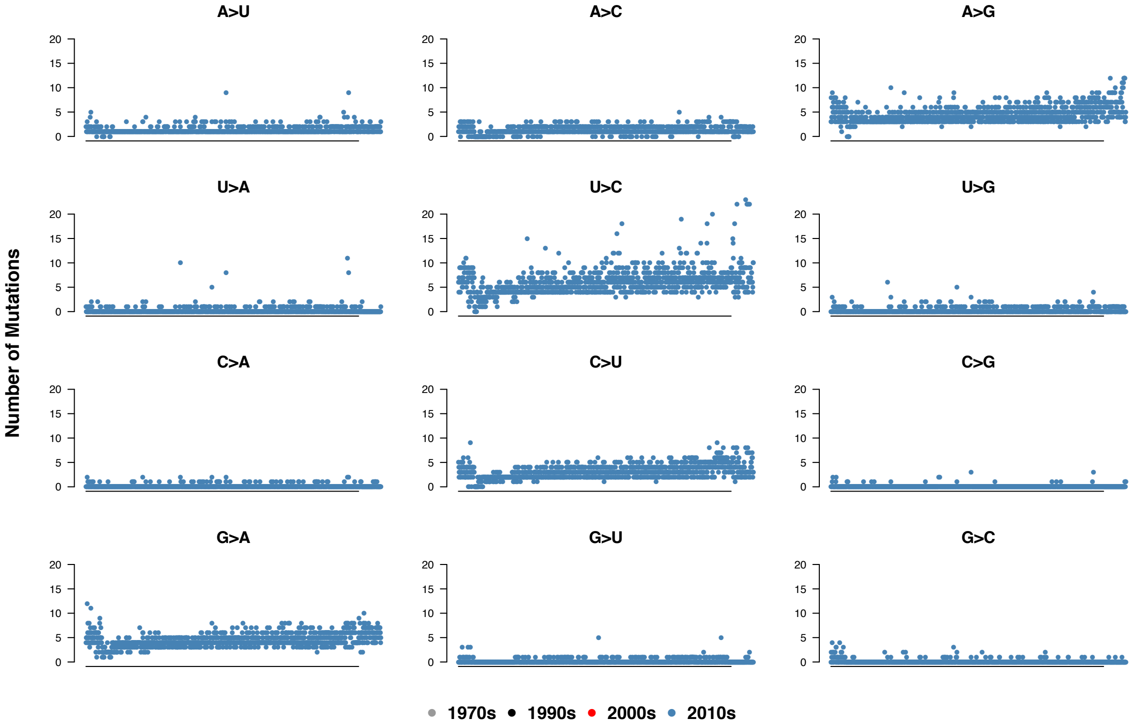


Figure 5.10: Observed mutation types over time for each sample compared to the early West Africa outbreak reference. Samples are ordered left to right from oldest to newest. The colour of the dot indicates the decade that the sample is from.

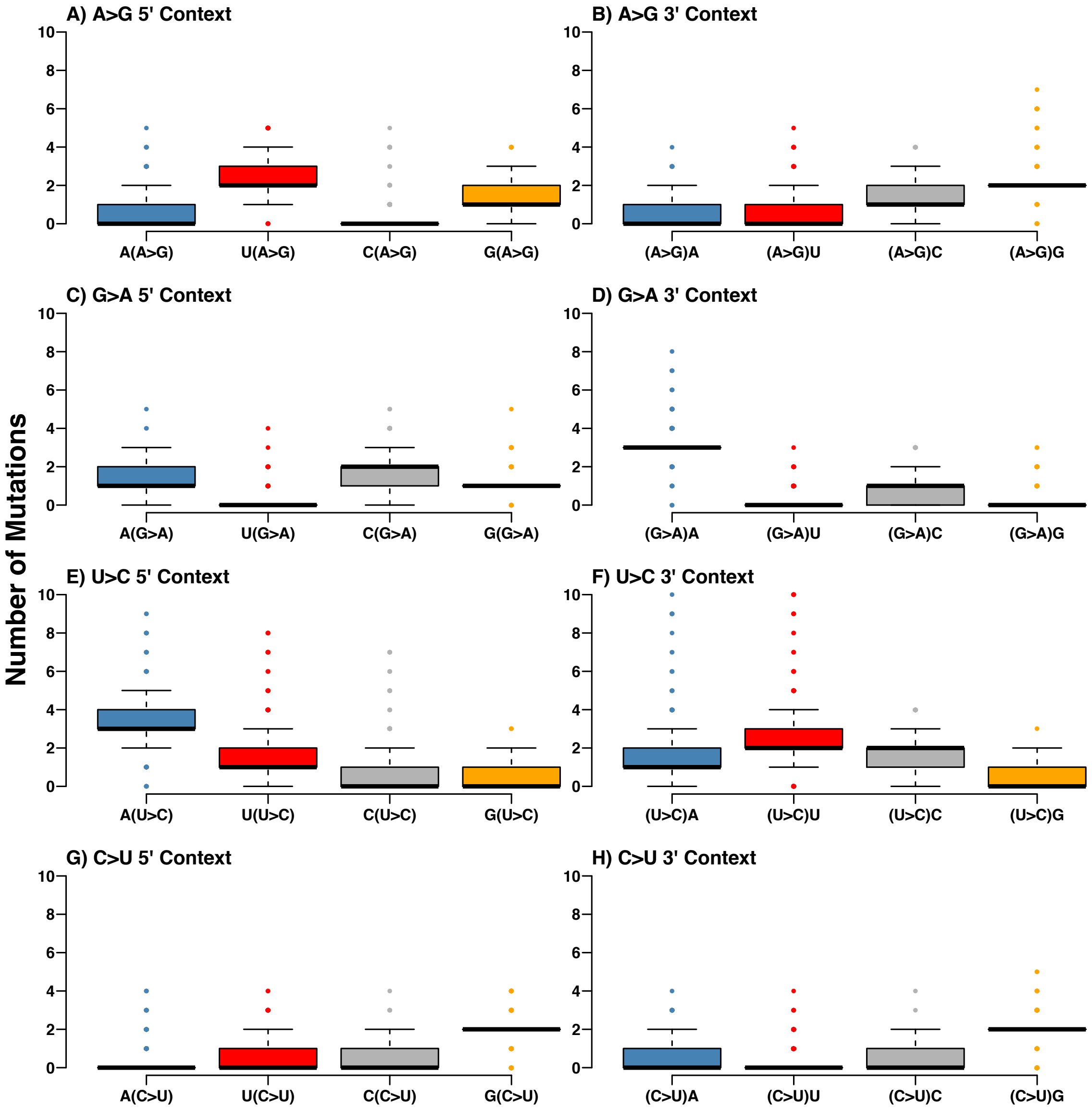


Figure 5.11: Mutation contexts of transition type mutations for all *Ebola virus* samples compared to the early West Africa outbreak reference sample. **A)** 5’ A>G mutation contexts. **B)** 3’ A>G mutation contexts. **C)** 5’ G>A mutation contexts. **D)** 3’ G>A mutation contexts. **E)** 5’ U>C mutation contexts. **F)** 3’ U>C mutation contexts. **G)** 5’ C>U mutation contexts. **H)** 3’ C>U mutation contexts.

Figure 5.12 shows the protein coding outcomes of the intra-outbreak mutation types for each individual sample. The numbers of intra-outbreak synonymous and non-synonymous outcomes for each mutation type are harder to analyse compared to the mutation set in reference to the 1976 reference genome, because there are so few mutations in each category. Therefore, there is little to be inferred from these comparisons, but for the four transition types synonymous outcomes do not appear to be consistently more frequent than non-synonymous outcomes. Synonymous outcomes are generally more common for U>C mutations than non-synonymous outcomes, but this trend is reversed for G>A mutations, and the trend is more mixed for A>G and C>U mutations (Figure 5.12). A similar trend is observed for the unique set of mutations across all samples (Appendix 4 Figure 42). As with mutations compared to the 1976 reference genome, there is no clear difference between transitions and transversions for each starting base in the BLOSUM62 scores for non-synonymous mutations, i.e. for starting C bases, C>U transitions show little difference compared to C>A and C>G transversions. See Figure 5.13 for mutations per sample, and Appendix 4 Figure 43 for the unique set of mutations.

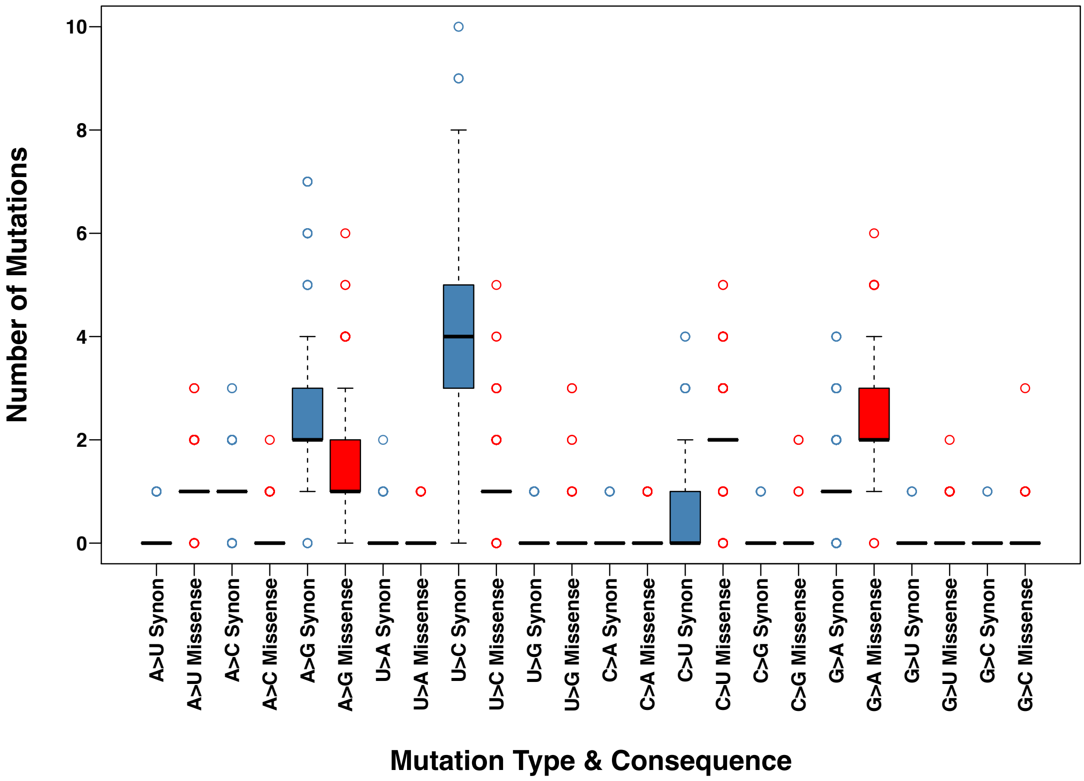


Figure 5.12: Protein coding consequences of each sample's mutations compared to the early West Africa outbreak reference sample, separated by mutation type.

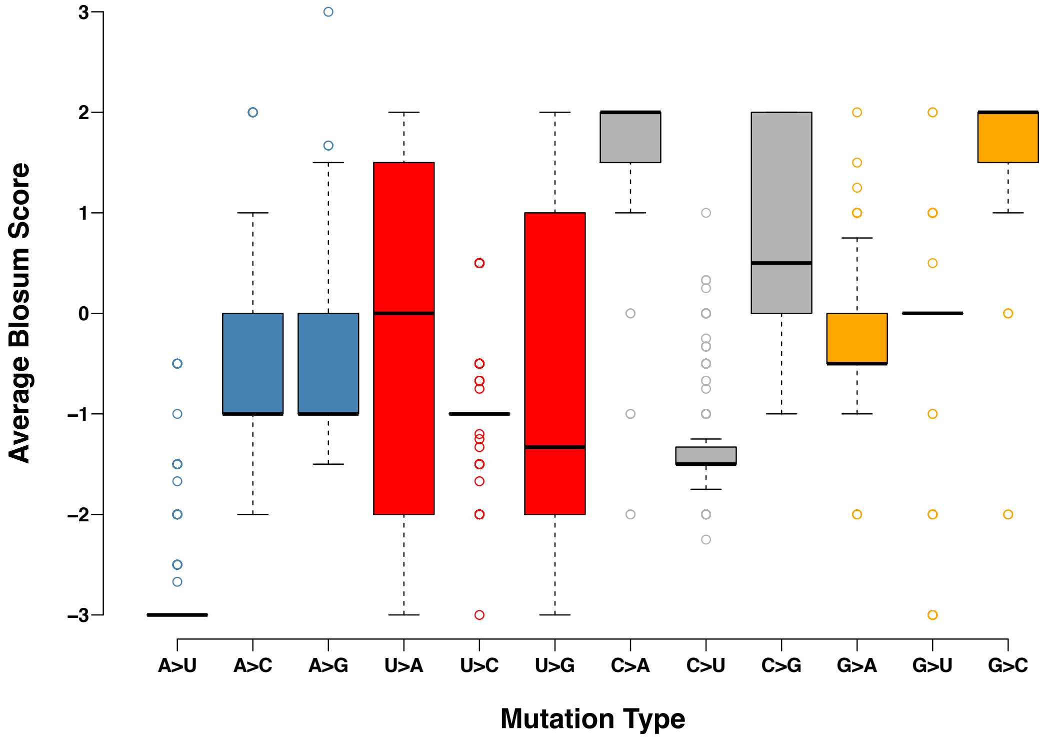


Figure 5.13: Average BLOSUM62 matrix scores for each sample’s mutations compared to the early West Africa outbreak reference sample, separated by mutation type.

Over the course of the West Africa outbreak several distinct lineages of *Ebola virus* emerged, which were characterised by key mutations, such as A82V in the GP protein that occurred early on in the outbreak and defined two broad lineages (Urbanowicz et al. 2016). Using the lineages defined by Urbanowicz et al., we compared the mutation type patterns between lineages, but observed little variation (Appendix 4 Figures 44-47). The most variation was observed within lineage 6, the largest of all of the lineages, especially for U>C mutations where there are multiple high-end outlier samples (Appendix 4 Figure 47).

5.5 Discussion

Ti:Tv bias is a common phenomenon that has been observed across distantly related species, but is still in search of a clear mechanism. This study is the first to our knowledge that has analysed Ti:Tv bias within and between *Ebola virus* outbreaks. When comparing samples to the 1976 reference *Ebola virus* genome we found a clear bias of the four transition type mutations (A>G, G>A, C>U, and U>C) being more common than the eight transversion type mutations. This trend was observed across almost all coding and non-coding regions of the genome, although the extent of the bias varied by region. The observation that the transition preference was observed to be more extreme for the set of unique mutations across all samples suggests a mutational cause, even if damaging transitions are selected against.

This preference for transitions translated to a mean Ti:Tv of 5.4 across the genome, 6.7 in coding regions, and 4.1 in non-coding regions. The Ti:Tv across the genome is higher than the general Ti:Tv of 2-4 observed in studies of other organisms (Kumar 1996; Rosenberg et al. 2003; Stoltzfus & Norris 2016). Some regions of the *Ebola virus* genome have Ti:Tv values consistent with the highest observed Ti:Tv values in other studies, suggesting local hypermutation. Non-coding region 4 was observed to have a mean Ti:Tv of 15.3, with a maximum Ti:Tv of 25, and non-coding region 8 had a maximum Ti:Tv of 22. This level of Ti:Tv bias is closer to that observed for the control region of human mitochondrial DNA (~15) or Ti:Tv bias over cytochrome b phylogeny (~19) (Tamura & Nei 1993; Purvis & Bromham 1997; Meyer et al. 1999).

A high Ti:Tv bias within non-coding regions again suggests a mutational cause, as selective forces on these regions are lower than coding regions, but there could still be selection for regulatory regions, non-coding RNAs, or other conserved non-coding elements. Samples from the 2014 outbreak in the Democratic Republic of the Congo were found to have a more extreme Ti:Tv bias. The mean Ti:Tv across the genome was 9.5, with a mean Ti:Tv of 29.4 for the NP gene, and a maximum Ti:Tv of 31 for the NP gene. The mean and maximum NP Ti:Tv for West Africa outbreak samples were 6.00 and 7.14, respectively. These very high Ti:Tv rates in specific genome regions again suggest possible localised hypermutation by a mutational process, rather than a mutational process acting generally across the genome.

The numbers of transition mutations compared to the 1976 reference genome appear to have steadily risen over time for the whole genome, but there are differences between regions and individual transition types. Different transition types have increased at different times, and increases are seen in different regions at different times, suggesting hypermutation events of specific mutation types. The numbers of transitions have increased more than transversions between outbreaks of *Ebola virus*, indicating that there is also a transition bias in the *Ebola virus* reservoir species. Mutations that occur within an outbreak are unlikely to be retained in the reservoir population, as it is extremely unlikely that viruses are re-entering the reservoir species from human hosts, though not impossible. When an outbreak occurs in humans, the samples are a snapshot of a sub-population of the virus within the reservoir species. Sequence sampling of sub-populations in the reservoir species would provide useful insight in to Ti:Tv bias of the virus, especially when compared to samples from human outbreaks.

Within the 2013-2016 West Africa outbreak the number of transitions also grew more rapidly than transversions, with many transition type high-end outliers, particularly for U>C mutations. Analysis of the coding consequences of transitions vs transversions suggested that transitions result in synonymous outcomes more often than non-synonymous outcomes, possibly due to selection. Non-synonymous outcomes of transitions are still more common than synonymous outcomes of transversions, suggesting a mutational cause. For the set of non-synonymous changes, there was no clear pattern of transitions being more conservative than transversions, based on BLOSUM62 scores of the resulting amino acid substitutions. This suggests a minimal role for the selective hypothesis in *Ebola virus* Ti:Tv bias, but this is a very crude measure of how conservative the changes are. Ideally the fitness effects of each mutation would be experimentally characterised, and the fitness effects of transitions and transversions compared (Stoltzfus & Norris 2016; Lyons & Lauring 2017).

Our data suggest that a mutational cause for the observed Ti:Tv bias is likely, even if selection is also playing a role. This could be entirely due to mutation bias in the polymerase protein, but without experimental analysis of the Ti:Tv bias of the *Ebola virus* RNA polymerase L it is difficult to know if additional mutational effects are present. We were interested in potential editing by host ADAR and APOBEC enzymes, both of which have reported sequence context preferences.

Analysis of the 5’ and 3’ contexts of the mutations did not identify a context preference for transition type mutations. The lack of sequence context preference could indicate that ADAR and APOBEC proteins are having a limited effect on Ti:Tv bias in *Ebola virus*, consistent with the observation that APOBEC restricts Coronaviral growth but does not cause hypermutation of the genome (Milewska et al. 2018). However, the context preferences of these enzymes are only preferences, and the enzymes are able to edit in other sequence contexts (Bishop et al. 2004; Eggington et al. 2011). These context preferences are also mostly measured for DNA, and particularly for APOBEC enzymes the preferences vary by the individual APOBEC family member (Pham et al. 2003; Bishop et al. 2004; Beale et al. 2004; Henry et al. 2009; Thielen et al. 2010; Roberts et al. 2012; Alexandrov et al. 2013). Crucially, the reservoir species of ebolaviruses remains unknown (Leendertz 2016), and this is where the majority of the evolution of these viruses is occurring. This means that the context preferences of host enzymes in the reservoir is also unknown.

Therefore, ADAR and APOBEC editing remains a likely contributor to ebolavirus mutation, as previously suggested in other studies (Park et al. 2015; Whitmer et al. 2018). It is also possible that some individual virions are hyperedited to the point that they are no longer viable, and these would not be observed in the data. Specific local hyperediting seems plausible given the patterns of individual mutations observed over time within different genome regions. Analysis of the raw sequencing data could be used in the future to identify clearer signals of editing by host enzymes. Finally, APOBEC and ADAR enzymes could be working against one another, with the target bases of APOBEC being the same as the endpoint bases of ADAR editing, and vice versa (Figure 5.4).

Overall our data suggest that there has been a strong Ti:Tv bias during the evolution of *Ebola virus*. This is likely due to a biased mutational process, with a smaller contribution from selective pressure, and a plausible role for host ADAR and APOBEC enzymes.

Chapter 6: Discussion

This thesis has presented four different projects all on the broad theme of genetic variation, in addition to two projects presented in Appendices 5 and 6. This chapter considers the implications of the work presented in Chapters 2-5 of this thesis, and the future work necessary to further advance our understanding of genetic variation.

6.1 Precision Medicine for Cystinuria

While rarely life-threatening, cystinuria patients frequently suffer from symptoms that reduce their quality of life. Cystinuria is a rare disease, with a worldwide incidence of 1 in 7,000 (Barbosa et al. 2012), but despite low incidence rates rare diseases are collectively a significant burden on health services (Angelis et al. 2015). Therefore, improvements in cystinuria treatment options have the potential to both increase quality of life for patients and reduce the burden on worldwide health services.

The work presented in Chapter 2 is an important step in advancing precision medicine in cystinuria. Firstly, the comprehensive characterisation of cystinuria associated variants provides a tool for future studies, and highlights some of the key functional features of cystinuria associated mutations. Secondly, the analysis of clinical data demonstrates the potential of computational tools for the improvement of cystinuria treatment in predicting a patient’s disease severity based on the patient’s genotype.

The method used to sort the samples in to high and low risk groups was limited by: A) The simplicity of the scoring system B) The lack of consideration of possible functional variants outside of SLC3A1 and SLC7A9 C) The use of pathogenicity predictors primarily designed to distinguish neutral variants from deleterious variants rather than mildly deleterious variants from highly deleterious variants. However, when compared to clinical outcomes the sample groupings broadly corresponded to the observed phenotypes. Many of these phenotypic differences between the predicted severity groups were shown to be statistically significant, and for the remaining analyses the expected trend was mostly observed, e.g. predicted low-severity patients were on average diagnosed at a later age compared to predicted high-severity patients, even if not statistically significant.

Despite these limitations, the sample grouping results are promising, and with a larger sample set a specialised model could be trained to better group samples. Additional improvements could be made with better functional annotation of rBAT and b(0+)AT. One key unanswered question is how the two proteins dimerise to form the final transporter. This information could allow the identification of variants that are likely to impact the formation of the transporter complex, and may explain why no clear structural impacts could be identified for some of the cystinuria associated variants.

The utility of an improved disease severity predictor would be in directing treatment options for cystinuria patients. As discussed in Section 1.5 of this thesis, mild forms of the disease can be treated with simple dietary changes, without the need for other therapeutic agents (Zee et al. 2017). This is the preferred treatment option for patients, as it is cheaper, easier for the patient, and less likely to cause adverse side effects. However, it will not work for patients with a severe form of cystinuria, and starting these patients out on this treatment option is a waste of valuable treatment time. Early identification of these patients, using their genotypic information, would allow high-severity patients to be given more advanced treatment options before they show symptoms – again improving patient care.

The ongoing clinical trial to assess the efficacy of alpha-lipoic acid in the treatment of cystinuria could result in a major advancement of the treatment of high-severity cystinuria (ClinicalTrials.gov identifier: NCT02910531, https://clinicaltrials.gov/ct2/show/record/NCT02910531), as current treatment options are limited in their effectiveness (Zee et al. 2017). If alpha-lipoic acid is proved to be an effective treatment option for cystinuria it could be combined with early genotypic screening to provide precision medicine for high-severity patients.

6.2 The Future Threat from Ebolaviruses

Ebolaviruses represent a past, present, and future threat to public health. Since 1976, the first time an ebolavirus was observed to infect a human, the four human pathogenic species have caused repeated, severe epidemiological crises, and resulted in devastating loss of life (Table 1.1). While the *Reston virus* is not known to cause EVD in humans, previous work has suggested that a small set of mutations could make *Reston virus* pathogenic in humans (Pappalardo et al. 2016; Pappalardo et al. 2017), a suggestion supported by the data presented in Chapter 4 of this thesis. These combined data suggest that *Reston virus* also represents a threat to public health and must be carefully monitored in order to detect signs of evolution towards human pathogenicity.

6.2.1 The Future of Virus Epidemiology

Monitoring of viruses for epidemiological purposes is already a key strategy for maintaining public health, for example it has long been applied in order to pre-emptively design seasonal flu vaccines and predict risk groups, based on which influenza virus strain is predicted to predominate the next flu season (The World Health Organisation 2018).

Improvements in our understanding of ebolavirus pathogenicity also have the potential to help improve our ability to monitor and control the threat of ebolaviruses. For example, the results presented in Chapter 4 support the theory that sequence changes in VP24 are a key determinant of ebolavirus pathogenicity, and would be essential in driving the emergence of a human pathogenic *Reston virus* (Volchkov et al. 2000; Ebihara et al. 2006; Reid et al. 2006; Reid et al. 2007; Zhang et al. 2012; Dowall et al. 2014; Xu et al. 2014; de La Vega et al. 2015; Pappalardo et al. 2016; Guito et al. 2017; Pappalardo et al. 2017). This and other predicted pathogenicity determinants could be used to monitor reservoir ebolavirus populations, and also help identify drug and vaccine targets.

The data in Chapter 4 also show that the SDP-based approach applied in the project, and also in previous work (Pappalardo et al. 2016), is a robust method even when limited by the number of available sequences. This approach can now more confidently be applied to other virus species, and is a valuable tool for rapid functional characterisation of known viruses and also in the early characterisation of emergent virus species.

6.2.2 The *Bombali virus* Emerges

On the 27th of August 2018, Goldstein *et al*. published work identifying a new species of ebolavirus – the *Bombali virus* (BOMV) (Goldstein et al. 2018). The BOMV species was identified in two species of bats in Sierra Leone: *Chaerephon pumilus* and *Mops condylurus*. These bats were found roosting inside houses, highlighting the threat of human transmission from close human-bat contact (Goldstein et al. 2018). The key question raised by the identification of BOMV is whether this new species of ebolavirus is pathogenic in humans.

In their work, Goldstein *et al*. demonstrated that the BOMV GP protein is able to mediate entry in to human cells, a necessary step for human pathogenicity, but not a determining step as the non-pathogenic *Reston virus* is also able to gain entry in to human cells (Miranda & Miranda 2011; Goldstein et al. 2018). Goldstein *et al* next compared the protein sequences of BOMV to *Ebola virus* (EBOV) and *Reston virus* (RESTV), specifically looking at key motifs in VP35 and VP24, as well as some previously identified SDPs (Pappalardo et al. 2016; Goldstein et al. 2018). This showed that for some positions BOMV is more similar to EBOV, but for others it is more similar to RESTV. Therefore, further studies are needed to investigate the human pathogenicity of BOMV. A comprehensive comparison of the SDP set defined in Chapter 4 of this thesis could provide additional clues as to the pathogenicity of BOMV. The discovery of BOMV also reemphasises the need for computational tools to rapidly characterise viruses and identify likely functional differences, in order to fully leverage the advances made in our ability to sequence viruses.

The discovery of BOMV in bats supports the hypothesis of bats being the reservoir species for ebolaviruses (Goldstein et al. 2018). As discussed in Chapter 4 of this thesis, if the reservoir species of ebolaviruses were identified this could be used to better monitor ebolaviruses, crucially before they spill-over in to human populations. This finding would also have implications for the work presented in

Chapter 5 of this thesis. Monitoring of ebolaviruses within the reservoir would serve as a useful comparison to the mutational patterns observed when the virus is within human populations. This would better elucidate the roles of APOBEC and ADAR enzymes in the evolution of ebolaviruses, and the different mutational pressures of each host species, especially if combined with the analysis of raw sequencing data instead of the analysis of the final assembled genome.

6.3 Complexity in Genetic Variation

Over the course of the last century much progress has been made in generating, analysing, and leveraging genetic variation data to advance scientific research and improve healthcare. However, much progress is still needed to unpick the complexity involved in many genetic conditions. Since Mendel's laws of inheritance were broken with the discovery of epistasis, the study of genetics has repeatedly identified hidden layers of genotypic complexity. From complex structural rearrangements, to epigenetics, to functional non-coding variation, to deleterious synonymous coding variants, and finally to the focus of the work in Chapter 3 of this thesis – the combined effects of non-synonymous coding variants.

6.3.1 Gestalt Variants

As discussed in Chapter 1 of this thesis, the way that genetic variation is viewed is changing. With ever-growing genome sequence data sets, hidden complexities of disease-associated variants are uncovered (M Lek et al. 2016; Martell et al. 2017; Minikel et al. 2016; Walsh et al. 2017), with some disease-associated variants found to be natural sequence variation, but only in the genetic context of other compensatory variants. This has led to the active search for "buffering" variants that determine individual resilience to genetic conditions (Chen et al. 2016).

The work presented in Chapter 3 of this thesis suggests that compensatory non-synonymous variants may be a common compensatory mechanism for deleterious non-synonymous variants, helping to preserve protein structure and function. This has many implications for the future study of genetic variation. For example, the analysis of cystinuria-associated variants in Chapter 2 of this thesis could be improved in the future by taking in to account the combined effects of variants. Similarly, some of the SDPs associated with human pathogenic ebolaviruses identified in Chapter 4 of this thesis may have combined effects. This could be particularly interesting for the newly discovered *Bombali virus*, initial analysis of which suggested that at some SDP positions it is more similar to the human pathogenic *Ebola virus* and at others it is more similar to the human non-pathogenic *Reston virus* (see Section 6.2.2). How combinations or subsets of SDPs affect ebolavirus pathogenicity remains unknown, but the discovery of *Bombali virus* presents a useful test case, as well as the opportunity to further refine the set of SDPs associated with human pathogenicity.

6.3.2 Implications for Precision Medicine

Precision medicine has the potential to advance healthcare from “Pre-Womb to Tomb” (Topol 2014), improving the accuracy and efficacy of treatments at every stage of life. As discussed in Chapter 1, there are a number of ongoing large-scale sequencing projects designed to advance precision medicine, and many findings have already been implemented for the treatment of a variety of human diseases (Church 2005; Sudlow et al. 2015; Chen et al. 2016; Deciphering Developmental Disorders Study 2017; National Institutes of Health 2018; Davies 2017).

However, there are still many practical challenges to overcome for precision medicine in the treatment of more complex genotypes. Although doubt has been cast on the importance of the majority of protein isoforms (Tress et al. 2017), it will be important in the future to improve on the work presented in Chapter 3 by expanding the analysis to non-canonical protein isoforms, especially those with tissue-specificity or disease-association (Petryszak et al. 2016). Additionally, the consideration of combinations of coding and non-coding variants, and combined variation at multiple genomic loci will be key.

Such analyses will require generation of new data sets. Firstly, it will be essential to generate more publicly available non-aggregate sequencing data sets, like the 1,000 Genomes Project data set (1000 Genomes Project Consortium 2015). Secondly, two limiting factors of the work presented in Chapter 3 were the lack of structural information for many human proteins and the lack of high-quality clinical significance annotations for the majority of variants. Generation of new experimental structures, improvements in *de novo* structural modelling, and growth of clinical annotation databases in the future will be needed to overcome these obstacles.

It is likely that computational predictions can never replace traditional experimental screening in healthcare, due to the high number of false positives and negatives in complex genotype analyses (Berg et al. 2017). In the future, approaches that combine computational screening with more traditional screening approaches, such as tandem mass spectrometry screening, will offer the most practical and powerful solutions to the analysis of genetic variation in humans (Berg et al. 2017).

References

1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. *Nature*, 526, pp.68–74.

Adams, K.L. & Wendel, J.F., 2005. Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology*, 8(2), pp.135–141.

Adhikari, B. & Cheng, J., 2016. Protein Residue Contacts and Prediction Methods. *Methods Mol Biol.*, 1415, pp.463–476.

Adzhubei, I.A. et al., 2010. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4), pp.248–249.

Del Álamo, M. & Mateu, M.G., 2005. Electrostatic repulsion, compensatory mutations, and long-range non-additive effects at the dimerization interface of the HIV capsid protein. *Journal of Molecular Biology*, 345(4), pp.893–906.

Alexandrov, L.B. et al., 2013. Signatures of mutational processes in human cancer. *Nature*, 500(7463), pp.415–421.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215, pp.403–410.

Altshuler, D.M. et al., 2010. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311), pp.52–58.

Andersson, L. et al., 2015. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biology*, 16(1), pp.4–9.

Angelis, A., Tordrup, D. & Kanavos, P., 2015. Socio-economic burden of rare diseases: A systematic review of cost of illness evidence. *Health Policy*, 119(7), pp.964–979.

Arias, A. et al., 2016. Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evolution*, 2(1).

Ashkenazy, H. et al., 2016. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res. Oxford University Press*, 44.

Auer, P.L. et al., 2016. Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project. *American Journal of Human Genetics*, 99(4), pp.791–801.

Auton, A. et al., 2015. A global reference for human genetic variation. *Nature*, 526(7571), pp.68–74.

Azevedo, L. et al., 2009. Epistatic interactions modulate the evolution of mammalian mitochondrial respiratory complex components. , 12, pp.1–12.

Azevedo, L. et al., 2017. Improving the in silico assessment of pathogenicity for compensated variants. *European journal of human genetics : EJHG*, 25(1), pp.2–7.

Baize, S. et al., 2014. Emergence of Zaire Ebola Virus Disease in Guinea. *New England Journal of*

Medicine, 371(15), pp.1418–1425.

Barbosa, M. et al., 2012. Clinical, biochemical and molecular characterization of cystinuria in a cohort of 12 patients. *Clin Genet Blackwell Publishing Ltd*, 81.

Barrette, R.W. et al., 2009. Discovery of swine as a host for the reston ebolavirus. *Science*, 325(5937), pp.204–206.

Bartoccioni, P. et al., 2008. Distinct classes of trafficking rBAT mutants cause the type I cystinuria phenotype. *Hum Mol Genet*, 17.

Basler, C.F. et al., 2000. The Ebola virus VP35 protein functions as a type I IFN antagonist. *Proceedings of the National Academy of Sciences*, 97(22), pp.12289–12294.

Basler, C.F. et al., 2003. The Ebola virus VP35 protein inhibits activation of interferon regulatory factor 3. *Journal of virology*, 77(14), pp.7945–56.

Bateman, A. et al., 2015. UniProt: A hub for protein information. *Nucleic Acids Research*, 43(D1), pp.D204–D212.

Beale, R.C.L. et al., 2004. Comparison of the Differential Context-dependence of DNA Deamination by APOBEC Enzymes: Correlation with Mutation Spectra in Vivo. *Journal of Molecular Biology*, 337(3), pp.585–596.

Bell, C.J. et al., 2011. Carrier Testing for Severe Childhood Recessive Diseases by Next-Generation Sequencing. *Science Translational Medicine*, 3(65), pp.1–26.

Belmont, J.W. et al., 2005. A haplotype map of the human genome. *Nature*, 437(7063), pp.1299–1320.

Bennett-Lovsey, R.M. et al., 2008. Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. *Proteins*, 70(2), pp.311–319.

Berg, J.S. et al., 2017. Newborn Sequencing in Genomic Medicine and Public Health. *Pediatrics*, 139(2).

Berman, H.M. et al., 2000. The protein data bank. *Nucleic acids research*, 28(1), pp.235–242.

Bhattacharya, D. et al., 2016. FRAGSION: Ultra-fast protein fragment library generation by IOHMM sampling. *Bioinformatics*, 32(13), pp.2059–2061.

Birney, E. et al., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), pp.799–816.

Bisceglia, L. et al., 1997. Localization, by linkage analysis, of the cystinuria type III gene to chromosome 19q13.1. Am. J. Hum. Genet. *Elsevier*, 60.

Bishop, K.N. et al., 2004. Cytidine Deamination of Retroviral DNA by Diverse APOBEC Proteins. *Current Biology*, 14, pp.1392–1396.

Bornholdt, Z.A. et al., 2013. Structural basis for ebolavirus matrix assembly and budding; protein plasticity allows multiple functions. *Cell*, 154(4), pp.763–774.

Brister, J.R. et al., 2015. NCBI viral Genomes resource. *Nucleic Acids Research*, 43(D1), pp.D571–D577.

Bromberg, Y., 2013. *Building a genome analysis pipeline to predict disease risk and prevent disease*, Biol: J. Mol.

Buslje, C.M. et al., 2009. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics*, 25(9), pp.1125–1131.

Calonge, M.J. et al., 1994. Cystinuria caused by mutations in rBAT, a gene involved in the transport of cystine. *Nat Genet*, 6.

Camacho, C. et al., 2009. BLAST+: architecture and applications. *BMC bioinformatics*, 10, p.421.

Capra, J.A. & Singh, M., 2007. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15), pp.1875–1882.

Carpenter, J.A. et al., 2009. Evidence for ADAR-induced hypermutation of the Drosophila sigma virus (Rhabdoviridae). *BMC Genetics*, 10, pp.1–7.

Carroll, M.W. et al., 2015. Temporal and spatial analysis of the 2014 – 2015 Ebola virus outbreak in West Africa. *Nature*, 524, pp.97–101.

Cattaneo, R., 1994. Biased (A→I) hypermutation of animal RNA virus genomes. *Current Opinion in Genetics and Development*, 4(6), pp.895–900.

CDC, 2018. CDC - National Center for Health Statistics - Cases and Outbreaks of EVD by Year. Available at: https://www.cdc.gov/vhf/ebola/history/chronology.html [Accessed July 11, 2018].

Chairoungdua, A. et al., 1999. Identification of an amino acid transporter associated with the cystinuria-related type II membrane glycoprotein. *J Biol Chem*, 274.

Chatr-Aryamontri, A. et al., 2015. The BioGRID interaction database: 2015 update. *Nucleic Acids Research*, 43(D1), pp.D470–D478.

Chen, R. et al., 2016. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nature Biotechnology*, 34, pp.531–38.

Chillarón, J. et al., 2010. Pathophysiology and treatment of cystinuria. *Nat Rev Nephrol*, 6.

Chothia, C. & Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4), pp.823–6.

Chun, S. & Fay, J.C., 2009. Identification of deleterious mutations within three human genomes. *Identification of deleterious mutations within three human genomes.*, 19(9), pp.1553–1561.

Church, G.M., 2005. The Personal Genome Project. *Molecular Systems Biology*, 1(1), pp.E1–E3.

Collins, F.S. et al., 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), pp.931–945.

Comai, L., 2005. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics*, 6(11), pp.836–846.

Dagan, T., Talmor, Y. & Graur, D., 2002. Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection. *Molecular Biology and Evolution*, 19(7), pp.1022–1025.

Damas, J. et al., 2017. Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. *Genome Research*, 27, pp.1–10.

Danecek, P. et al., 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15), pp.2156–2158.

Daneshjou, R. et al., 2017. Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Human Mutation*, 38(9), pp.1182–1192.

David, A. et al., 2012. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Human Mutation*, 33(2), pp.359–363.

Davies, S.C., 2017. *Annual Report of the Chief Medical Officer 2016, Generation Genome*,

Deciphering Developmental Disorders Study, 2017. Prevalence and architecture of de novo mutations in developmental disorders. *Nature*, 542.

Denver, D.R. et al., 2009. A genome-wide view of Caenorhabditis elegans base-substitution mutation processes. *Proceedings of the National Academy of Sciences*, 106(38), pp.16310–16314.

Denver, D.R. et al., 2004. High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome. *Nature*, 430(August), pp.679–682.

Dib, L., Salamin, N. & Gfeller, D., 2018. Polymorphic sites preferentially avoid co-evolving residues in MHC class I proteins. *PLoS Computational Biology*, 14(5), pp.1–19.

Dowall, S.D. et al., 2014. Elucidating variations in the nucleotide sequence of Ebola virus associated with increasing pathogenicity. *Genome biology*, 15(11), p.540.

Dunham, I. et al., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp.57–74.

Dutta, P. et al., 2017. A survey on Ebola genome and current trends in computational research on the Ebola virus. *Briefings in Functional Genomics*, (October), pp.1–7.

Duy, J. et al., 2018. Virus-encoded miRNAs in Ebola virus disease. *Scientific Reports*, 8(1), pp.1–14.

Ebihara, H. et al., 2006. Molecular determinants of Ebola virus virulence in mice. *PLoS Pathogens*, 2(7), pp.0705–0711.

Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), pp.1792–1797.

Eggermann, T., Venghaus, A. & Zerres, K., 2012. Cystinuria: an inborn cause of urolithiasis. *Orphanet J Rare Dis.*, 7(19).

Eggington, J.M., Greene, T. & Bass, B.L., 2011. Predicting sites of ADAR editing in double-stranded RNA. *Nature Communications*, 2(1), p.319.

Ekeberg, M. et al., 2013. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 87(1), pp.1–19.

Fehrholz, M. et al., 2012. The innate antiviral factor APOBEC3G targets replication of measles, mumps and respiratory syncytial viruses. *Journal of General Virology*, 93(3), pp.565–576.

Feingold, E.A. et al., 2004. The ENCODE (ENCyclopedia of DNA Elements) Project. *Science*, 306(5696), pp.636–640.

Feldman, H. & Geisbert, T., 2011. Ebola haemorrhagic fever. *Lancet*, 377(9768), pp.849–862.

Feliubadaló, L. et al., 1999. Non-type I cystinuria caused by mutations in SLC7A9, encoding a subunit (bo,+AT) of rBAT. *Nat Genet*, 23.

Fernández, E. et al., 2002. rBAT-b(0,+)AT heterodimer is the main apical reabsorption system for cystine in the kidney. *Am J Physiol Renal Physiol American Physiological Society*, 283.

Fischer, K. et al., 2017. Serological Evidence for the Circulation of Ebolaviruses in Pigs From Sierra Leone. *The Journal of Infectious Diseases*, (July), pp.1–7.

Fowler, D.M. & Fields, S., 2014. Deep mutational scanning: A new style of protein science. *Nature Methods*, 11(8), pp.801–807.

Franca, R. et al., 2005. Heterodimeric amino acid transporter glycoprotein domains determining functional subunit association. *Biochem J*, 388.

Frazer, K.A. et al., 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164), pp.851–861.

Fryxell, K.J., 1996. The coevolution of gene family trees. *Trends in Genetics*, 12(9), pp.364–369.

Fu, W. et al., 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431), pp.216–220.

Fu, Y. et al., 2014. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome biology*, 15(10), p.480.

Gao, L. & Zhang, J., 2003. Why are some human disease-associated mutations fixed in mice? , 19(12), pp.678–681.

Gasperini, M., Starita, L. & Shendure, J., 2016. The power of multiplexed functional analysis of genetic variants. *Nature Protocols*, 11(10), pp.1782–1787.

Gelinas, J.-F. et al., 2011. Enhancement of Replication of RNA Viruses by ADAR1 via RNA Editing and Inhibition of RNA-Activated Protein Kinase. *Journal of Virology*, 85(17), pp.8460–8466.

Gibbs, R.A. et al., 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428(6982), pp.493–520.

Gingold, H. & Pilpel, Y., 2011. Determinants of translation efficiency and accuracy. *Molecular Systems Biology*, 7(481), pp.1–13.

Gire, S.K. et al., 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202), pp.1369–72.

Goh, C.S. et al., 2000. Co-evolution of proteins with their interaction partners. *Journal of Molecular Biology*, 299(2), pp.283–293.

Gojobori, T., Li, W.H. & Graur, D., 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. *Journal of Molecular Evolution*, 18(5), pp.360–369.

Goldstein, T. et al., 2018. The discovery of Bombali virus adds further support for bats as hosts of ebolaviruses. *Nature Microbiology*.

Gonzalez-Perez, A. & Lopez-Bigas, N., 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score. *Condel Am J Hum Genet Elsevier*, 88.

González-Pérez, A. & López-Bigas, N., 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *American Journal of Human Genetics*, 88(4), pp.440–449.

Gottesman, M.M., Fojo, T. & Bates, S.E., 2002. Multidrug Resistance in Cancer: Role of Atp-Dependent Transporters. *Nature Reviews Cancer*, 2(1), pp.48–58.

Grimm, D.G. et al., 2015. The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. *Human Mutation*, 36(5), pp.513–523.

Groenen, M.A.M. et al., 2012. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*, 491(7424), pp.393–8.

Gront, D. et al., 2011. Generalized fragment picking in rosetta: Design, protocols and applications. *PLoS ONE*, 6(8).

Guito, J.C. et al., 2017. Novel activities by ebolavirus and marburgvirus interferon antagonists revealed using a standardized in vitro reporter system. *Virology*, 501(September 2016), pp.147–165.

Haller, O., Kochs, G. & Weber, F., 2006. The interferon response circuit: Induction and suppression by pathogenic viruses. *Virology*, 344(1), pp.119–130.

Harewood, L. et al., 2017. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biology*, 18(1), pp.1–11.

Harnevik, L. et al., 2003. Mutation analysis of SLC7A9 in cystinuria patients in Sweden. *Genet Test*, 7.

Harrison, S.M. et al., 2016. *Using ClinVar as a resource to support variant interpretation*,

Haussler, D. et al., 2009. Genome 10K: A proposal to obtain whole-genome sequence for 10000 vertebrate species. *Journal of Heredity*, 100(6), pp.659–674.

Hecht, M., Bromberg, Y. & Rost, B., 2015. Better prediction of functional effects for sequence variants. *BMC Genomics*, 16(Suppl 8.

Henry, M. et al., 2009. Genetic editing of HBV DNA by monodomain human APOBEC3 cytidine deaminases and the recombinant nature of APOBEC3G. *PLoS ONE*, 4(1), pp.1–12.

Hoenen, T. et al., 2016. Nanopore sequencing as a rapidly deployable Ebola outbreak tool. *Emerging Infectious Diseases*, 22(2), pp.331–334.

Hoenen, T. et al., 2015. Soluble Glycoprotein Is Not Required for Ebola Virus Virulence in Guinea Pigs. *Journal of Infectious Diseases*, 212(Suppl 2), pp.S242–S246.

Hoenen, T. et al., 2005. VP40 Octamers Are Essential for Ebola Virus Replication VP40 Octamers Are Essential for Ebola Virus Replication. *Journal of Virology*, 79(3), pp.1898–1905.

Hopf, T.A. et al., 2017. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2), pp.128–135.

Hoskins, R.A. et al., 2017. Reports from CAGI: The Critical Assessment of Genome Interpretation. *Human Mutation*, 38(9), pp.1039–1041.

Houben, A., 2017. B Chromosomes – A Matter of Chromosome Drive. *Frontiers in Plant Science*, 08(February), pp.1–6.

Huang, Y. et al., 2002. The assembly of Ebola virus nucleocapsid requires virion-associated proteins 35 and 24 and posttranslational modification of nucleoprotein. *Molecular Cell*, 10(2), pp.307–316.

Hulo, C. et al., 2011. ViralZone: A knowledge resource to understand virus diversity. *Nucleic Acids Research*, 39(SUPPL. 1), pp.576–582.

Ingram, V.M., 1959. Abnormal human haemoglobins. III the chemical difference between normal and sickle cell haemoglobins. *Biochimica et Biophysica Acta*, 36(2), pp.402–411.

Innis, C.A., Anand, A.P. & Sowdhamini, R., 2004. Prediction of functional sites in proteins using conserved functional group analysis. *Journal of Molecular Biology*, 337(4), pp.1053–1068.

International Chicken Genome Sequencing Consortium, 2004. Sequencing and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432, pp.695–777.

International Genome Sample Resource, 2018. IGSR and the 1000 Genomes Project. Available at: http://www.internationalgenome.org/sites/1000genomes.org/files/images/1000g_map.png [Accessed August 18, 2018].

Ioannidis, N.M. et al., 2016. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *American Journal of Human Genetics*, 99(4), pp.877–885.

Ionita-Laza, I. et al., 2016. A Spectral Approach Integrating Functional Genomic Annotations for Coding and Noncoding Variants. *Nature Genetics*, 48(2), pp.214–220.

Jeong, C.S. & Kim, D., 2012. Reliable and robust detection of coevolving protein residues. *Protein Engineering, Design and Selection*, 25(11), pp.705–713.

Jiang, C. & Zhao, Z., 2006. Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms. *Genomics*, 88(5), pp.527–534.

Jiang, Y. et al., 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, 17(1), pp.1–19.

Jones, D.T. et al., 2015. MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31(7), pp.999–1006.

Joosten, R.P. et al., 2011. A series of PDB related databases for everyday needs. *Nucleic Acids Research*, 39, pp.411–419.

Jordan, D.M. et al., 2015. Identification of cis-suppression of human disease mutations by comparative genomics. *Nature*, 524(7564), pp.225–230.

Juan, D., Pazos, F. & Valencia, A., 2008. Co-evolution and co-adaptation in protein networks. *FEBS Letters*, 582(8), pp.1225–1230.

Kaján, L. et al., 2014. FreeContact: Fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics*, 15(1), pp.1–6.

Kamisetty, H., Ovchinnikov, S. & Baker, D., 2013. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences*, 110(39), pp.15674–15679.

Karczewski, K.J. et al., 2017. The ExAC browser: Displaying reference data information from over 60 000 exomes. *Nucleic Acids Research*, 45(D1), pp.D840–D845.

Kay, K.M. et al., 2005. Rapid speciation and the evolution of hummingbird pollination in neotropical Costus subgenus Costus (Costaceae): Evidence from nrDNA its and ETS sequences. *American Journal of Botany*, 92(11), pp.1899–1910.

Keller, I., Bensasson, D. & Nichols, R.A., 2007. Transition-transversion bias is not universal: A counter example from grasshopper pseudogenes. *PLoS Genetics*, 3(2), pp.0185–0191.

Kelley, L.A. et al., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*, 10.

Kelly, L.A. et al., 2015. The Phyre2 web portal for protein modelling, prediction, and analysis. *Nature Protocols*, 10(6), pp.845–858.

Kimchi-Sarfaty, C. et al., 2007. A ‘silent’ polymorphism in the MDR1 gene changes substrate specificity. *Science (New York, N.Y.)*, 315(5811), pp.525–8.

Kimura, M., 1985. The role of compensatory neutral mutations in molecular evolution. *Journal of Genetics*, 64(1), pp.7–19.

Kircher, M. et al., 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3), pp.310–315.

Kondrashov, A.S., Sunyaev, S. & Kondrashov, F.A., 2002. Dobzhansky-Muller incompatibilities in protein evolution. *Proc Natl Acad Sci U S A*, 99(23), pp.14878–14883.

Koonin, E. V., Wolf, Y.I. & Karev, G.P., 2002. The structure of the protein universe and genome evolution. *Nature*, 420(6912), pp.218–223.

Kowalczyk, L. et al., 2011. Molecular basis of substrate-induced permeation by an amino acid antiporter. *Proc Natl Acad Sci U S A*, 108.

Kuhn, J.H. et al., 2014. Nomenclature- and database-compatible names for the Two Ebola virus variants that emerged in guinea and the Democratic Republic of the Congo in 2014. *Viruses*, 6(11), pp.4760–4799.

Kumar, P., Henikoff, S. & Ng, P.C., 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols*, 4(8), pp.1073–1081.

Kumar, S., 1996. Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. *Genetics*, 143(1), pp.537–548.

de La Vega, M.-A. et al., 2015. The Multiple Roles of sGP in Ebola Pathogenesis. *Viral Immunology*, 28(1), pp.3–9.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921.

Landrum, M.J. et al., 2018. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1), pp.D1062–D1067.

Latour, S. et al., 2001. Bidirectional Negative Regulation of Human T and Dendritic Cells by CD47 and Its Cognate Receptor Signal-Regulator Protein- : Down-Regulation of IL-12 Responsiveness and Inhibition of Dendritic Cell Activation. *The Journal of Immunology*, 167(5), pp.2547–2554.

Leendertz, S.A.J., 2016. Testing new hypotheses regarding ebolavirus reservoirs. *Viruses*, 8(2).

Lek, M. et al., 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), pp.285–291.

Lek, M. et al., 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536.

Lemon, J., 2006. Plotrix: a package in the red light district of R. *R-News*, 6(4), pp.8–12.

Lensink, M.F. et al., 2016. Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins: Structure, Function and Bioinformatics*, (April), pp.323–348.

Lensink, M.F., Velankar, S. & Wodak, S.J., 2017. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins: Structure, Function and Bioinformatics*, 85(3), pp.359–377.

Levy, S. et al., 2007. The Diploid Genome Sequence of an Individual Human. *PLoS Biology*, 5(10).

Lewin, H.A. et al., 2018. Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17), pp.4325–4333.

Li, J. et al., 2018. Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Research*, 46(15), pp.7793–7804.

Liu, S.Q. et al., 2015. Identifying the pattern of molecular evolution for Zaire ebolavirus in the 2014 outbreak in West Africa. *Infection, Genetics and Evolution*, 32, pp.51–59.

Lopez, G. et al., 2011. Firestar--advances in the prediction of functionally important residues. *Nucleic Acids Res*, 39.

Lundgren, R., Nordle, O. & Josefsson, K., 1995. Immediate estrogen or estramustine phosphate therapy versus deferred endocrine treatment in nonmetastatic prostate cancer: a randomized multicenter study with 15 years of followup. The South Sweden prostate cancer study group. *J Urol*, 153.

Lynch, M., 2010. Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences*, 107(3), pp.961–968.

Lyons, D.M. & Lauring, A.S., 2017. Evidence for the Selective Basis of Transition-to-Transversion Substitution Bias in Two RNA Viruses. *Molecular Biology and Evolution*, 34(12), pp.3205–3215.

Ma, D. et al., 2012. Structure and mechanism of a glutamate-GABA antiporter. *Nature*, 483.

Maganga, G.D. et al., 2014. Ebola Virus Disease in the Democratic Republic of Congo. *New England Journal of Medicine*, 371(22), pp.2083–2091.

Mahmood, K. et al., 2017. Variant effect prediction tools assessed using independent , functional assay-based datasets : implications for discovery and diagnostics. *Human Genomics*, 11, pp.1–8.

Marcotte, E.M. et al., 1999. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757), pp.83–86.

Maretty, L. et al., 2017. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature*, 548(7665), pp.87–91.

Marks, D.S. et al., 2011. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE*, 6(12).

Marsh, G.A. et al., 2011. Ebola reston virus infection of pigs: Clinical significance and transmission potential. *Journal of Infectious Diseases*, 204(SUPPL. 3).

Martell, H.J. et al., 2017. Associating mutations causing cystinuria with disease severity with the aim of providing precision medicine. *BMC Genomics*, 18(Suppl 5).

Mascher, M. et al., 2017. A chromosome conformation capture ordered sequence of the barley genome. *Nature*, 544(7651), pp.427–433.

Mateu, M.G. & Fersht, a R., 1999. Mutually compensatory mutations during evolution of the tetramerization domain of tumor suppressor p53 lead to impaired hetero-oligomerization. *Proceedings of the National Academy of Sciences of the United States of America*, 96(7), pp.3595–3599.

McCarthy, D.J. et al., 2014. Choice of transcripts and software has a large effect on variant annotation. *Genome medicine*, 6(3), p.26.

McKenna, A. et al., 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20, pp.1297–1303.

McLaren, W. et al., 2016. The Ensembl Variant Effect Predictor. *Genome Biology*, p.042374.

McWilliam, H. et al., 2013. Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res. Oxford University Press*, 41.

Mehedi, M. et al., 2011. A New Ebola Virus Nonstructural Glycoprotein Expressed through RNA Editing. *Journal of Virology*, 85(11), pp.5406–5414.

Messaoudi, I., Amarasinghe, G.K. & Basler, C.F., 2015. Filovirus pathogenesis and immune evasion: insights from Ebola virus and Marburg virus. *Nature Publishing Group*, 13(11), pp.663–676.

Meyer, S., Weiss, G. & von Haeseler, A., 1999. Pattern of nucleotide susbtitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics*, 152, pp.1103–1110.

Milewska, A. et al., 2018. APOBEC3-mediated restriction of RNA virus replication. *Scientific Reports*, 8(1), p.5960.

Minikel, E.V. et al., 2016. Quantifying prion disease penetrance using large population control cohorts. *Science Translational Medicine*, 8(322), p.322ra9-322ra9.

Miosge, L.A. et al., 2015. Comparison of predicted and actual consequences of missense mutations.

Miranda, M.E. et al., 1999. Epidemiology of Ebola (Subtype Reston) Virus in the Philippines, 1996. *The Journal of Infectious Diseases*, 179(s1), pp.S115–S119.

Miranda, M.E.G. & Miranda, N.L.J., 2011. Reston Ebolavirus in humans and animals in the Philippines: A review. *Journal of Infectious Diseases*, 204(SUPPL. 3), pp.757–760.

Miyata, T., Miyazawa, S. & Yasunaga, T., 1979. Two types of amino acid substitutions in protein evolution. *Journal of Molecular Evolution*, 12(3), pp.219–236.

Mizoguchi, K. et al., 2001. Human cystinuria-related transporter: localization and functional characterization. *Kidney Int*, 59.

Modrof, J. et al., 2002. Phosphorylation of VP30 impairs Ebola virus transcription. *Journal of Biological Chemistry*, 277(36), pp.33099–33104.

Moll, K.M. et al., 2017. Strategies for optimizing BioNano and Dovetail explored through a second reference quality assembly for the legume model, Medicago truncatula. *BMC Genomics*, 18(1), pp.1–16.

Morcos, F. et al., 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49), pp.E1293–E1301.

Mosca, R., Céol, A. & Aloy, P., 2013. Interactome3D : adding structural details to protein networks. *Nature Methods*, 10(1), pp.47–53.

Moult, J. et al., 2016. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins: Structure, Function and Bioinformatics*, 84(February), pp.4–14.

Moult, J. et al., 2018. Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins: Structure, Function and Bioinformatics*, 86(August 2017), pp.7–15.

Murphy, D.G., Dimock, K. & Kang, C.Y., 1991. Numerous Transitions in Human Parainfluenza

Virus 3 RNA Recovered from Persistently Infected Cells. *Virology*, 181, pp.760–763.

National Institutes of Health, 2018. About the All of Us Research Program. Available at: https://allofus.nih.gov/about/about-all-us-research-program [Accessed August 28, 2018].

Ng, P.C. & Henikoff, S., 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31.

Nishida, C. et al., 2008. Characterization of chromosome structures of Falconinae (Falconidae, Falconiformes, Aves) by chromosome painting and delineation of chromosome rearrangements during their differentiation. *Chromosome Research*, 16(1), pp.171–181.

Ochoa, D. & Pazos, F., 2010. Studying the co-evolution of protein families with the Mirrortree web server. *Bioinformatics*, 26(10), pp.1370–1371.

De Oliveira, S.H.P. et al., 2018. Sequential search leads to faster, more efficient fragment-based de novo protein structure prediction. *Bioinformatics*, 34(7), pp.1132–1140.

De Oliveira, S.H.P. & Deane, C.M., 2018. Combining co-evolution and secondary structure prediction to improve fragment library generation. *Bioinformatics*, 34(13), pp.2219–2227.

de Oliveira, S.H.P., Shi, J. & Deane, C.M., 2015. Building a Better Fragment Library for De Novo Protein Structure Prediction. *Plos One*, 10(4), p.e0123998.

OMIM, 2018a. Online Mendelian Inheritance in Man, OMIM®. Johns Hopkins University, Baltimore, MD. MIM Number:143100:15/06/2018.

OMIM, 2017a. Online Mendelian Inheritance in Man, OMIM®. Johns Hopkins University, Baltimore, MD. MIM Number:190685:02/15/2017. Available at: https://omim.org/ [Accessed August 24, 2018].

OMIM, 2017b. Online Mendelian Inheritance in Man, OMIM®. Johns Hopkins University, Baltimore, MD. MIM Number:603903:12/05/2017. Available at: https://omim.org/ [Accessed August 5, 2018].

OMIM, 2018b. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD).

Ovchinnikov, S. et al., 2017. Protein structure determination using metagenome sequence data. *Science*, 355(6322), pp.294–298.

Pagani, L. et al., 2016. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, 538(7624), pp.238–242.

Pagès, S. et al., 1997. Species-specificity of the cohesin-dockerin interaction between Clostridium thermocellum and Clostridium cellulolyticum: Prediction of specificity determinants of the dockerin domain. *Proteins: Structure, Function and Genetics*, 29(4), pp.517–527.

Palacín, M. et al., 2001. The amino acid transport system b(o,+) and cystinuria. *Mol Membr Biol*, 18.

Pappalardo, M. et al., 2017. Changes associated with Ebola virus adaptation to novel species. *Bioinformatics*, 33(December), pp.1911–1915.

Pappalardo, M. et al., 2016. Conserved differences in protein sequence determine the human pathogenicity of Ebolaviruses. *Scientific Reports*, 6(March), p.23743.

Pappalardo, M. & Wass, M.N., 2014. VarMod: Modelling the functional effects of non-synonymous variants. *Nucleic Acids Research*, 42(W1), pp.1–6.

Park, D.J. et al., 2015. Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell*, 161(7), pp.1516–1526.

Pauly, M.D., Procario, M.C. & Lauring, A.S., 2017. A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza A viruses. *eLife*, 6, pp.1–18.

Pazos, F. & Valencia, A., 2001. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Engineering, Design and Selection*, 14(9), pp.609–614.

Pechmann, S. & Frydman, J., 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature structural & molecular biology*, 20(2), pp.237–43.

Pellegrini, M. et al., 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc Natl Acad Sci USA*, 96, pp.4285–4288.

Perry, G.H. et al., 2007. Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, 39(10), pp.1256–1260.

Petryszak, R. et al., 2016. Expression Atlas update - An integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research*, 44(D1), pp.D746–D752.

Pfeiffer, R. et al., 1999. Luminal heterodimeric amino acid transporter defective in cystinuria. *Mol Biol Cell*, 10.

Pham, P. et al., 2003. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature*, 424(6944), pp.103–107.

Pickett, B.E. et al., 2012. ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research*, 40(D1), pp.593–598.

Pires, D.E. V, Ascher, D.B. & Blundell, T.L., 2014. mCSM: Predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30(3), pp.335–342.

Piton, A., Redin, C. & Mandel, J.L., 2013. XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing. *American Journal of Human Genetics*, 93(2), pp.368–383.

Plotkin, J.B. & Kudla, G., 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews*, 12(1), pp.32–42.

Prins, K.C., Cardenas, W.B. & Basler, C.F., 2009. Ebola Virus Protein VP35 Impairs the Function of Interferon Regulatory Factor-Activating Kinases IKK and TBK-1. *Journal of Virology*, 83(7), pp.3069–3077.

Purvis, A. & Bromham, L., 1997. Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny. *Journal of Molecular Evolution*, 44(1), pp.112–119.

Quick, J. et al., 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589), pp.228–32.

R Core Team, 2015. R: A Language and Environment for Statistical Computing.

Ranea, J.A.G. et al., 2010. Finding the ‘dark matter’ in human and yeast protein network prediction and modelling. *PLoS Computational Biology*, 6(9).

Rausell, A. et al., 2010. Protein interactions and ligand binding: From protein subfamilies to functional specificity. *Proceedings of the National Academy of Sciences*, 107(5), pp.1995–2000.

Reid, S.P. et al., 2006. Ebola Virus VP24 Binds Karyopherin 1 and Blocks STAT1 Nuclear Accumulation. *Journal of Virology*, 80(11), pp.5156–5167.

Reid, S.P. et al., 2007. Ebola Virus VP24 Proteins Inhibit the Interaction of NPI-1 Subfamily Karyopherin   Proteins with Activated STAT1. *Journal of Virology*, 81(24), pp.13469–13477.

Remmert, M. et al., 2012. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2), pp.173–175.

Reva, B., Antipin, Y. & Sander, C., 2011. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research*, 39(17), pp.37–43.

Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007. Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. *Science*, 316.

Rice, P., Longden, I. & Bleasby, A., 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(1), pp.276–277.

Riggi, N. et al., 2007. Sarcomas: genetics, signalling, and cellular origins. Part I: The fellowship of TET. *The Journal of pathology*, 213, pp.4–20.

Rimmer, A., 2018. New Ebola outbreak declared in Democratic Republic of the Congo. *BMJ*, 361.

Ritchie, G.R.S. et al., 2014. Functional annotation of non-coding sequence variants. *Nature Methods*, 11(3), pp.294–296.

Roberts, R.J., Carneiro, M.O. & Schatz, M.C., 2013. The advantages of SMRT sequencing. *Genome Biology*, 14(6), pp.2–5.

Roberts, S.A. et al., 2012. Clustered Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions. *Molecular Cell*, 46(2), pp.424–435.

Robinson, G.E. et al., 2011. Creating a Buzz About Insect Genomes. *Science*, 331, pp.1386–1386.

Rose, K.M. et al., 2004. The viral infectivity factor (Vif) of HIV-1 unveiled. *Trends in Molecular Medicine*, 10(6), pp.291–297.

Rosenberg, M.S., Subramanian, S. & Kumar, S., 2003. Patterns of transitional mutation biases within and among mammalian genomes. *Molecular Biology and Evolution*, 20(6), pp.988–993.

Rotkiewicz, P. & Skolnick, J., 2008. Fast procedure for reconstruction of full-atom protein models from reduced representations. *Journal of Computational Chemistry*, 29(9), pp.1460–1465.

Rueda, P., García-Barreno, B. & Melero, J.A., 1994. Loss of conserved cysteine residues in the attachment (G) glycoprotein of two human respiratory syncytial virus escape mutants that contain multiple A- G substitutions (hypermutations). *Virology*, 198(1), pp.653–662.

Sandmann, S. et al., 2017. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Scientific Reports*, 7, pp.1–12.

Sasidharan Nair, P. & Vihinen, M., 2012. VariBench: A Benchmark Database for Variations. *Human Mutation*, 34(1), pp.42–49.

Sauna, Z.E. & Kimchi-sarfaty, C., 2011. Understanding the contribution of synonymous mutations to human disease. *Nature Publishing Group*, 12(10), pp.683–691.

Saunders, R. & Deane, C.M., 2010. Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Research*, 38(19), pp.6719–6728.

Schneider, M. & Brock, O., 2014. Combining physicochemical and evolutionary information for protein contact prediction. *PLoS ONE*, 9(10).

Schwarz, T.M. et al., 2017. VP24-Karyopherin Alpha Binding Affinities Differ between Ebolavirus Species, Influencing Interferon Inhibition and VP24 Stability. *Journal of virology*, 91(4), pp.1–16.

Schymkowitz, J. et al., 2005. The FoldX web server: An online force field. *Nucleic Acids Research*, 33(SUPPL. 2), pp.382–388.

Seemayer, S., Gruber, M. & Soding, J., 2014. CCMpred - Fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, 30(21), pp.3128–3130.

Shaffer, P.L. et al., 2009. Structure and mechanism of a Na+–independent amino acid transporter. *Science*, 325.

Shen, Y. et al., 2013. Detecting protein candidate fragments using a structural alphabet profile comparison approach. *PLoS ONE*, 8(11).

Sherry, S.T., 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1), pp.308–311.

Shi, Y., 2014. A glimpse of structural biology through X-ray crystallography. *Cell*, 159(5), pp.995–1014.

Shihab, H.A. et al., 2015. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, 31(10), pp.1536–1543.

Shihab, H.A. et al., 2013. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat*, 34.

Shihab, H.A. et al., 2013. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation*, 34(1), pp.57–65.

Sievers, F. et al., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7(1), p.539.

Sievers, F. et al., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. *Mol Syst Biol*, 7.

Sim, N.L. et al., 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*, 40.

Simon-Loriere, E. et al., 2015. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature*, 524(7563), pp.102–104.

Smedley, D. et al., 2016. A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *American Journal of Human Genetics*, 99(3), pp.595–606.

Stoltzfus, A. & Norris, R.W., 2016. On the Causes of Evolutionary Transition:Transversion Bias. *Molecular biology and evolution*, 33(3), pp.595–602.

Strologo Dello, L. et al., 2002. Comparison between SLC3A1 and SLC7A9 cystinuria patients and carriers: a need for a new classification. *J Am Soc Nephrol*, 13.

Stumpf, M.P.H. et al., 2008. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 105(19), pp.6959–64.

Sudlow, C. et al., 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3), pp.1–10.

Sugiki, T., Kobayashi, N. & Fujiwara, T., 2017. Modern Technologies of Solution Nuclear Magnetic

Resonance Spectroscopy for Three-dimensional Structure Determination of Proteins Open Avenues for Life Scientists. *Computational and Structural Biotechnology Journal*, 15, pp.328–339.

Supek, F. et al., 2014. Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, 156(6), pp.1324–1335.

Szklarczyk, D. et al., 2015. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1), pp.D447–D452.

Tamura, K. & Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpazees. *Molecular Biology and Evolution*, 10(3), pp.512–526.

Telenti, A. et al., 2016. Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 113(42), pp.11901–11906.

The Bovine Genome Sequencing and Analysis Consortium, 2009. The Genome Sequence of Taurin Cattle: A window to ruminant biology and evolution. *Science*, 324, pp.522–528.

The C. elegans Sequencing Consortium, 1998. Genome Sequence of the Nematode C . elegans : A Platform for Investigating Biology. *Science*, 282(5396), pp.2012–2018.

The Deciphering Developmental Disorders Study, 2015. Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, 519(7542), pp.223–228.

The International HapMap Consortium, 2003. The International HapMap Project. *Nature*, 426(6968), pp.789–796.

The World Health Organisation, 2018. Influenza: Surveillance and Monitoring. Available at: http://www.who.int/influenza/surveillance_monitoring/en/ [Accessed September 8, 2018].

The World Health Organisation, 2014. Virological analysis: no link between Ebola outbreaks in west Africa and Democratic Republic of Congo. Available at: http://www.who.int/mediacentre/news/ebola/2-september-2014/en/.

Thielen, B.K. et al., 2010. Innate immune signaling induces high levels of TC-specific deaminase activity in primary monocyte-derived cells through expression of APOBEC3A isoforms. *Journal of Biological Chemistry*, 285(36), pp.27753–27766.

Thomas, K. et al., 2014. Cystinuria-a urologist's perspective. *Nat Rev Urol*, 11.

Tomczak, K., Czerwińska, P. & Wiznerowicz, M., 2015. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Wspolczesna Onkologia*, 1A, pp.A68–A77.

Tong, Y.-G. et al., 2015. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature*, 524(7563), pp.93–96.

Topol, E.J., 2014. INDIVIDUALIZED MEDICINE From Pre-Womb to Tomb. *Cell*, 157(1), pp.241–253.

Tress, M.L., Abascal, F. & Valencia, A., 2017. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends in Biochemical Sciences*, 42(2), pp.98–110.

Trevizani, R. et al., 2017. Critical features of fragment libraries for protein structure prediction. *PLoS ONE*, 12(1), pp.1–22.

Tsai, C.-J. et al., 2008. Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *Journal of molecular biology*, 383(2), pp.281–91.

Tuller, T. et al., 2010. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell*, 141(2), pp.344–354.

UK Biobank, 2018. Regeneron announces major collaboration to exome sequence UK Biobank genetic data more quickly. Available at: http://www.ukbiobank.ac.uk/2018/01/regeneron-announces-major-collaboration-to-exome-sequence-uk-biobank-genetic-data-more-quickly/ [Accessed September 6, 2018].

Urbanowicz, R.A. et al., 2016. Human Adaptation of Ebola Virus during the West African Outbreak. *Cell*, 167(4), pp.1079–1087.

Velankar, S. et al., 2013. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research*, 41(D1), pp.483–489.

De Vivo, M. et al., 2016. Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Medicinal Chemistry*, 59(9), pp.4035–4061.

Vogel, F. & Kopun, M., 1977. Higher of transitions among point mutations. *Journal of Molecular Evolution*, 9, pp.159–180.

Volchkov, V.E. et al., 1999. Characterization of the L gene and 5' trailer region of Ebola virus. *Journal of General Virology*, 80(2), pp.355–362.

Volchkov, V.E. et al., 1995. GP mRNA of Ebola Virus Is Edited by the Ebola Virus Polymerase and by T7 and Vaccinia Virus Polymerases. *Virology*, 214(2), pp.421–430.

Volchkov, V.E. et al., 2000. Molecular characterization of guinea pig-adapted variants of Ebola virus. *Virology*, 277(1), pp.147–155.

Wagner, C.A., Lang, F. & Bröer, S., 2001. Function and structure of heterodimeric amino acid transporters. Am. J. Physiol. *Cell Physiol*, 281.

Wakeley, J., 1996. The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. *Trends in Ecology and Evolution*, 11(4), pp.158–163.

Walsh, R. et al., 2017. Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genetics in Medicine*, 19(2), pp.192–203.

Walters-Sen, L.C. et al., 2015. Variability in pathogenicity prediction programs: impact on clinical diagnostics. *Molecular Genetics & Genomic Medicine*, 3(2), pp.99–110.

Wan, W. et al., 2017. Structure and assembly of the Ebola virus nucleocapsid. *Nature*, 551(7680), pp.394–397.

Wang, F. -x. et al., 2008. APOBEC3G upregulation by alpha interferon restricts human immunodeficiency virus type 1 infection in human peripheral plasmacytoid dendritic cells. *Journal of General Virology*, 89(3), pp.722–730.

Wang, K., Li, M. & Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16), pp.1–7.

Wang, T. et al., 2017. LRFragLib: An effective algorithm to identify fragments for de novo protein structure prediction. *Bioinformatics*, 33(5), pp.677–684.

Wang, X. et al., 2012. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature Biotechnology*, 30(2), pp.159–164.

Wass, M.N., Kelley, L.A. & Sternberg, M.J.E., 2010. 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res*, 38.

Wass, M.N. & Sternberg, M.J.E., 2009. Prediction of ligand binding sites using homologous structures and conservation at CASP8. *Proteins*, 77.

Watanabe, K. et al., 1997. The refined crystal structure of Bacillus cereus oligo-1,6-glucosidase at 2.0 A resolution: structural characterization of proline-substitution sites for protein thermostabilization. *J. Mol. Biol*, 269, pp.142–53.

Waterston, R.H. et al., 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915), pp.520–562.

Wei, L. et al., 2015. MAC: identifying and correcting annotation for multi-nucleotide variations. *BMC genomics*, 16(1), p.569.

Weik, M. et al., 2002. Ebola Virus VP30-Mediated Transcription Is Regulated by RNA Secondary Structure Formation Ebola Virus VP30-Mediated Transcription Is Regulated by RNA Secondary Structure Formation. *Virology*, 76(17), pp.8532–8539.

Wetterstrand, K., 2018. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcostsdata [Accessed August 18, 2018].

Whitmer, S.L.M. et al., 2018. Active Ebola Virus Replication and Heterogeneous Evolutionary Rates in EVD Survivors Report Active Ebola Virus Replication and Heterogeneous Evolutionary Rates in EVD Survivors. *Cell Reports*, 22, pp.1159–1168.

Wong, K.A. et al., 2015. The genetic diversity of cystinuria in a UK population of patients. *BJU International*, 116(1), pp.109–116.

Wong, K.A., Wass, M. & Thomas, K., 2016. The Role of Protein Modelling in Predicting the Disease Severity of Cystinuria. *European Urology*, 69(3), pp.541–546.

Wong, T.C. et al., 1989. Generalized and Localized Biased Hypermutation Affecting the Matrix Gene of a Measles Virus Strain That Causes Cubacute Sclerosing Panencephalitis. *Journal of Virology*, 63(12), pp.5464–5468.

World Health Organization, 2009. *WHO experts consultation on Ebola Reston pathogenicity in humans*,

Wright, C.F., FitzPatrick, D.R. & Firth, H. V., 2018. Paediatric genomics: Diagnosing rare disease in children. *Nature Reviews Genetics*, 19(5), pp.253–268.

Xu, W. et al., 2014. Ebola virus VP24 targets a unique NLS binding site on karyopherin alpha 5 to selectively compete with nuclear import of phosphorylated STAT1. *Cell Host and Microbe*, 16(2), pp.187–200.

Xue, Y. et al., 2012. Deleterious- and disease-allele prevalence in healthy individuals: Insights from current predictions, mutation databases, and population-scale resequencing. *American Journal of Human Genetics*, 91(6), pp.1022–1032.

Yampolsky, L.Y. & Stoltzfus, A., 2005. The exchangeability of amino acids in proteins. *Genetics*, 170(4), pp.1459–1472.

Yang, J. et al., 2014. The I-TASSER suite: Protein structure and function prediction. *Nature Methods*, 12(1), pp.7–8.

Yates, A. et al., 2016. Ensembl 2016. *Nucleic Acids Research*, 44(D1), pp.D710–D716.

Yates, C.M. et al., 2014. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J Mol Biol*, 426.

Zahn, R.C. et al., 2007. A-to-G Hypermutation in the Genome of Lymphocytic Choriomeningitis Virus. *Journal of Virology*, 81(2), pp.457–464.

Zappala, Z. & Montgomery, S.B., 2016. Non-coding loss-of-function variation in human genomes. *Human Heredity*, 81(2), pp.78–87.

Zee, T. et al., 2017. α-Lipoic acid treatment prevents cystine urolithiasis in a mouse model of cystinuria. *Nature medicine*, 23(3), pp.288–290.

Zhang, A.P.P. et al., 2012. The ebolavirus VP24 interferon antagonist: Know your enemy. *Virulence*, 3(5), pp.440–445.

Zhang, F. & Lupski, J.R., 2015. Non-coding genetic variants in human disease. *Human Molecular Genetics*, 24(R1), pp.R102–R110.

Zhang, G., 2015. Genomics: Bird sequencing project takes off. *Nature*, 522(7554), p.34.

Zhang, G., Hubalewska, M. & Ignatova, Z., 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature structural & molecular biology*, 16(3), pp.274–280.

Zhang, Z. & Gerstein, M., 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Research*, 31(18), pp.5338–5348.

Zhou, J. & Troyanskaya, O.G., 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10), pp.931–934.

Zhou, Z.H., 2011. Atomic Resolution Cryo Electron Microscopy of Macromolecular Complexes. *Adv Protein Chem Struct Biol*, 82, pp.1–35.

Zhu, Y.O. et al., 2014. Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences*, 111(22), pp.E2310–E2318.

Zimmermann, M.T. et al., 2017. Molecular modeling and molecular dynamic simulation of the effects of variants in the TGFBR2 kinase domain as a paradigm for interpretation of variants obtained by next generation sequencing. *PLoS ONE*, 12(2), pp.1–21.

Appendix 1: Chapter 2 Supplementary Material

Appendix 1 Figures

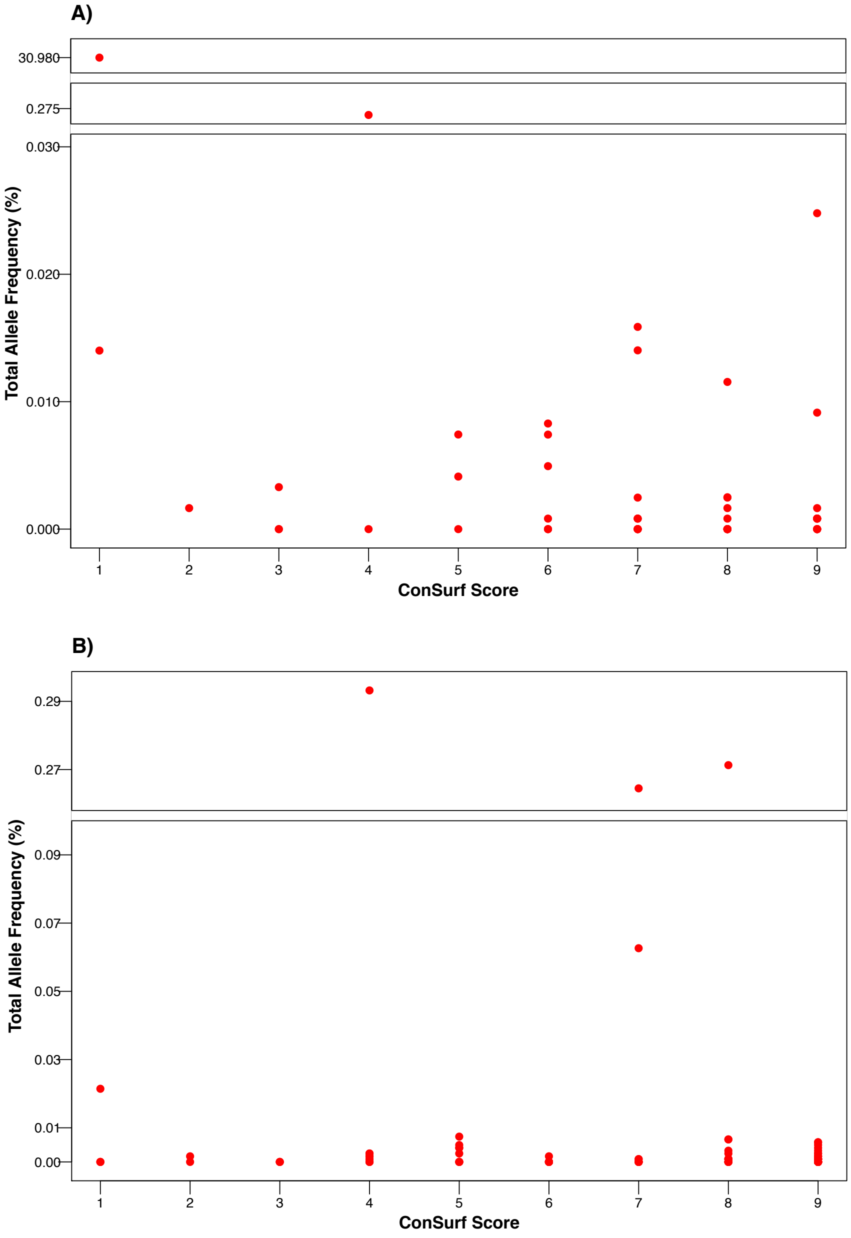


Figure 1: Comparison of allele frequency (based on ExAC) and evolutionary conservation (based on ConSurf scores). **A)** Cystinuria associated mutations of b(0+)AT. **B)** Cystinuria associated mutations of rBAT.

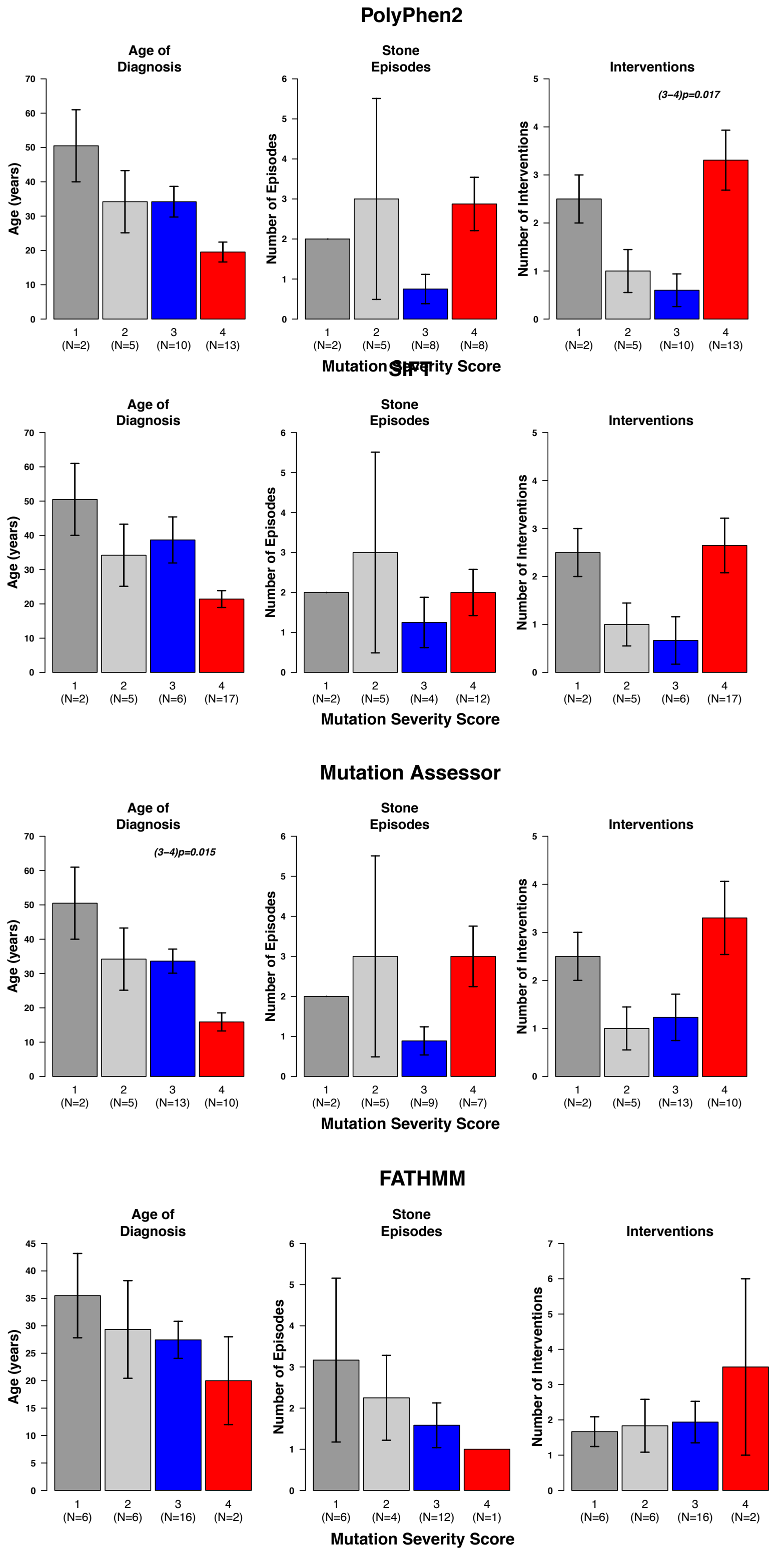


Figure 2: Comparison of other clinical parameters between the different severity scoregroups, for individuals with b(o+)AT mutations. There is one plot per prediction method (PolyPhen2, SIFT Mutation Assessor, and FATHMM). The group numbers are given at the bottom of the plots, with the sample number given in brackets underneath the group name. Where significant differences between groups occur (*p*<0.05) the *p*-value is displayed on the plot, e.g. (1-2)*p*=0.001 means a significant difference between groups 1 and 2.

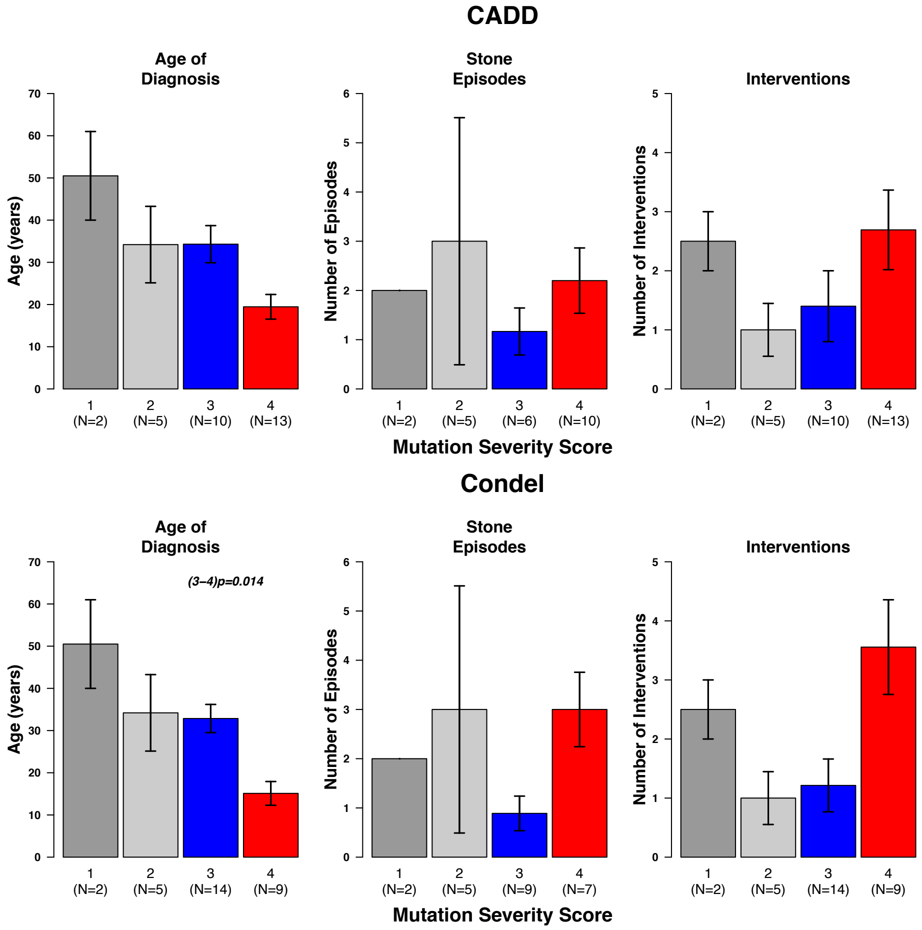


Figure 3: Comparison of other clinical parameters between the different severity score groups, for individuals with b(o+)AT mutations. There is one plot per integrated prediction method (CADD and Condel). The group numbers are given at the bottom of the plots, with the sample number given in brackets underneath the group name. Where significant differences between groups occur (p<0.05) the p-value is displayed on the plot, e.g. (1-2)p=0.001 means a significant difference between groups 1 and 2.

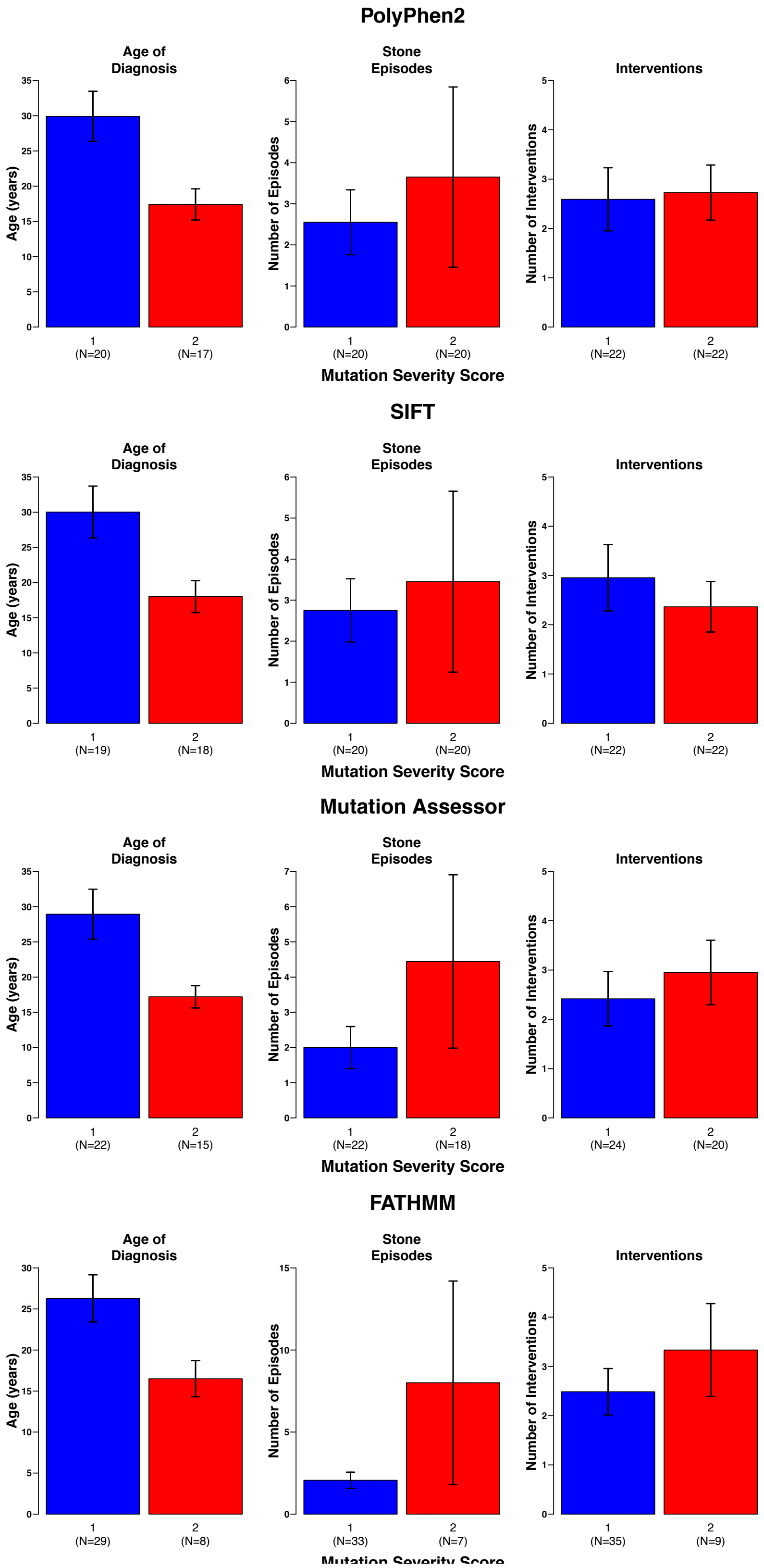


Figure 4: Comparison of other clinical parameters between the different severity score groups, for individuals with rBAT mutations. There is one plot per prediction method (PolyPhen2, SIFT Mutation Assessor, and FATHMM). The group numbers are given at the bottom of the plots, with the sample number given in brackets underneath the group name. Where significant differences between groups occur (*p*<0.05) the *p*-value is displayed on the plot, e.g. (1-2)*p*=0.001 means a significant difference between groups 1 and 2.

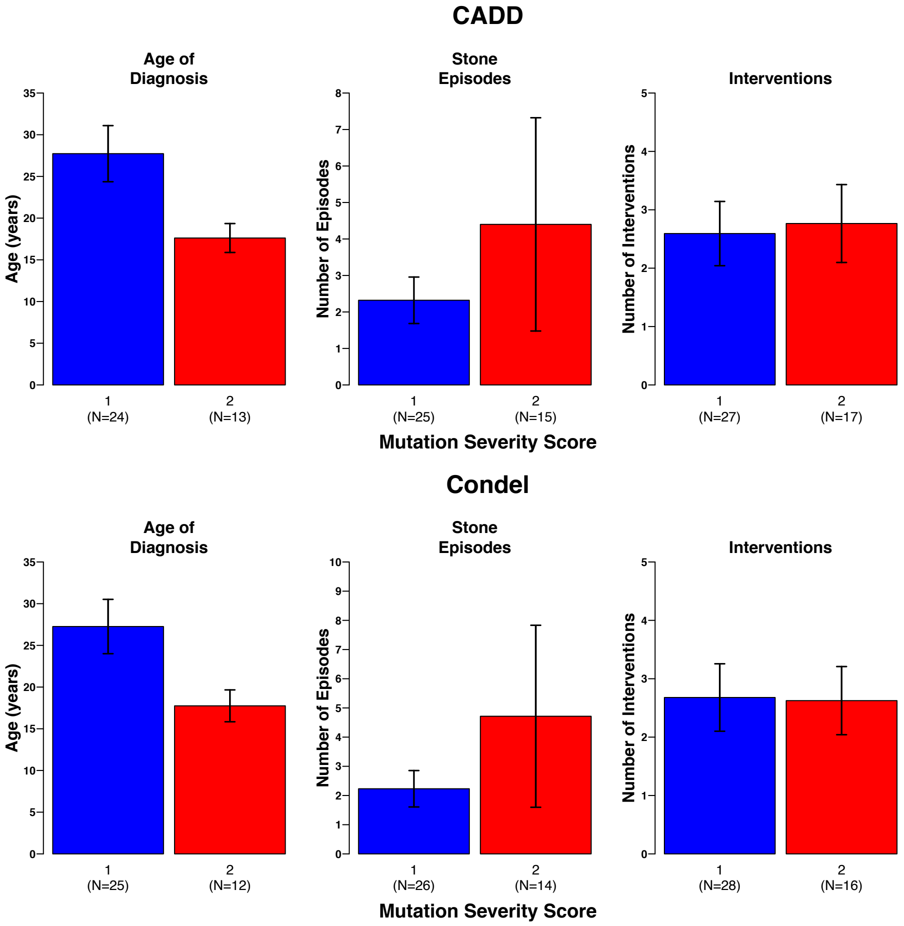


Figure 5: Comparison of other clinical parameters between the different severity score groups, for individuals with rBAT mutations. There is one plot per integrated prediction method (CADD and Condel). The group numbers are given at the bottom of the plots, with the sample number given in brackets underneath the group name. Where significant differences between groups occur (p<0.05) the p-value is displayed on the plot, e.g. (1-2)p=0.001 means a significant difference between groups 1 and 2.

Appendix 1 Tables

Table 1: Proteins used in the multiple sequence alignment for b(0+)AT. The final column indicates whether or not proteins from the same species were used in the alignment of rBAT.

Table 1 is too large to be printed here but is available as a digital version. This can be found on the USB stick provided with this thesis.

Table 2: Proteins used in the multiple sequence alignment for rBAT. The final column indicates whether or not proteins from the same species were used in the alignment of b(0+)AT.

Table 2 is too large to be printed here but is available as a digital version. This can be found on the USB stick provided with this thesis.

Table 3: Results of sample grouping for thresholding of prediction methods, for patients with mutations in SLC3A1

Table 3 is too large to be printed here but is available as a digital version. This can be found on the USB stick provided with this thesis.

Table 4: Results of sample grouping for thresholding of prediction methods, for patients with mutations in SLC7A9

Table 4 is too large to be printed here but is available as a digital version. This can be found on the USB stick provided with this thesis.

Table 5: Results of mCSM and ConSurf predictions for all SLC3A1 variants present in ExAC that are not known to be associated with cystinuria.

Table 5 is too large to be printed here but is available as a digital version. This can be found on the USB stick provided with this thesis.

Table 6: Results of mCSM and ConSurf predictions for all SLC7A9 variants present in ExAC that are not known to be associated with cystinuria.

Table 6 is too large to be printed here but is available as a digital version. This can be found on the USB stick provided with this thesis.

Table 7: Structural analysis of b(0+)AT nsSNV associated with cystinuria.

Table 7 is too large to be printed here but is available as a digital version. This can be found on the USB stick provided with this thesis.

Table 8: Structural analysis of rBAT nsSNV associated with cystinuria.

Table 8 is too large to be printed here but is available as a digital version. This can be found on the USB stick provided with this thesis.

Table 9: Non synonymous mutation in b(0+)AT associated with cystinuria. Summary of the results for all computational methods used to predict the effects of the SLC7A9 mutations already known to be associated with cystinuria, as well as the total allele count, total allele frequency and total homozygote count in the ExAC data set for each mutation. Predictions are given for: PolyPhen2 (HumDiv and HumVar), MutationAssessor, SIFT, FATHMM, Condel, CADD, mCSM, and ConSurf. See methods for details of the prediction programs used.

Table 9 is too large to be printed here but is available as a digital version. This can be found on the USB stick provided with this thesis.

Table 10: Non-synonymous mutations in rBAT associated with cystinuria. Summary of the results for all computational methods used to predict the effects of the SLC3A1 mutations already known to be associated with cystinuria, as well as the total allele count, total allele frequency and total homozygote count in the ExAC data set for each mutation. Predictions are given for: PolyPhen2 (HumDiv and HumVar), MutationAssessor, SIFT, FATHMM, Condel, CADD, mCSM, and ConSurf. See methods for details of the prediction programs used.

Table 10 is too large to be printed here but is available as a digital version. This can be found on the USB stick provided with this thesis.

Appendix 2: Chapter 3 Supplementary Material

Appendix 2 Figures

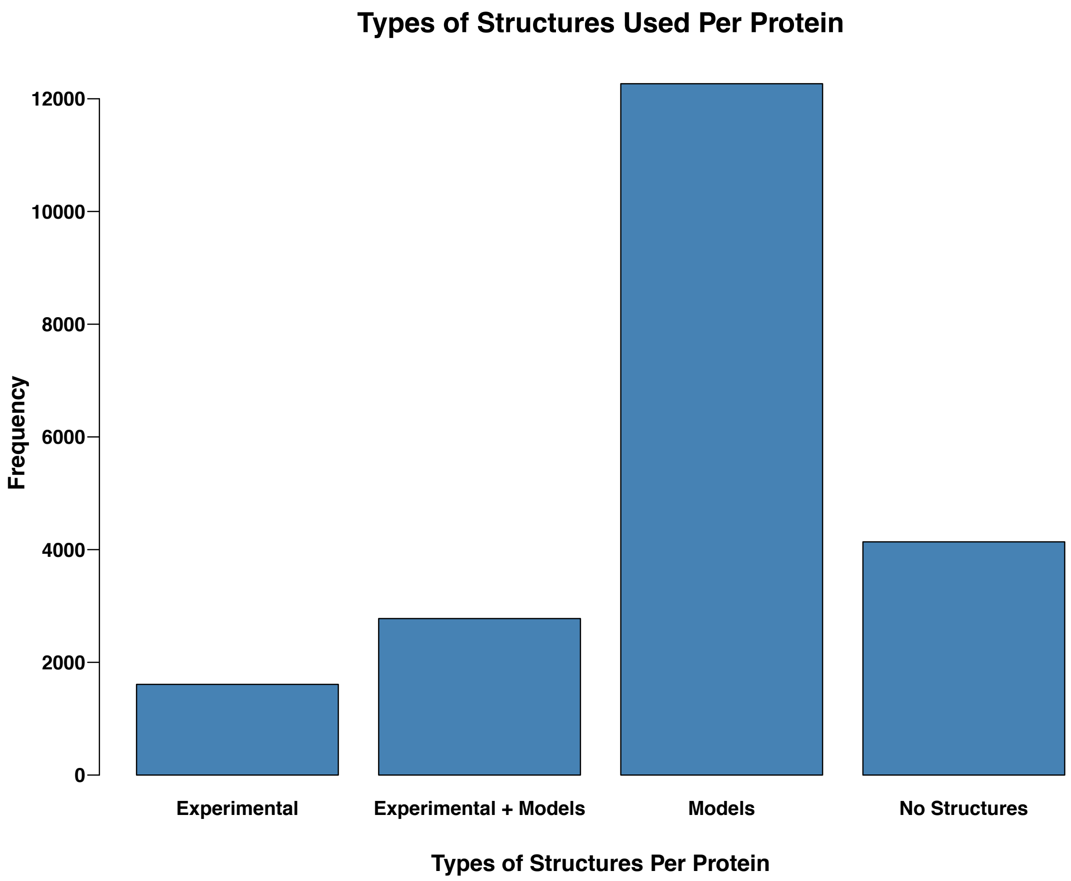


Figure 1: Types of structures used per protein. Experimental – proteins with corresponding experimental structures, with no sequence gaps ≥50 residues. Experimental + Models - proteins with corresponding experimental structures, but with sequence gaps ≥50 residues for which structural modelling was performed. Models – proteins for which there were no available experimental structures and structural modelling was performed. No structures – proteins without experimental models and for which no high-quality structural modelling template was identified.

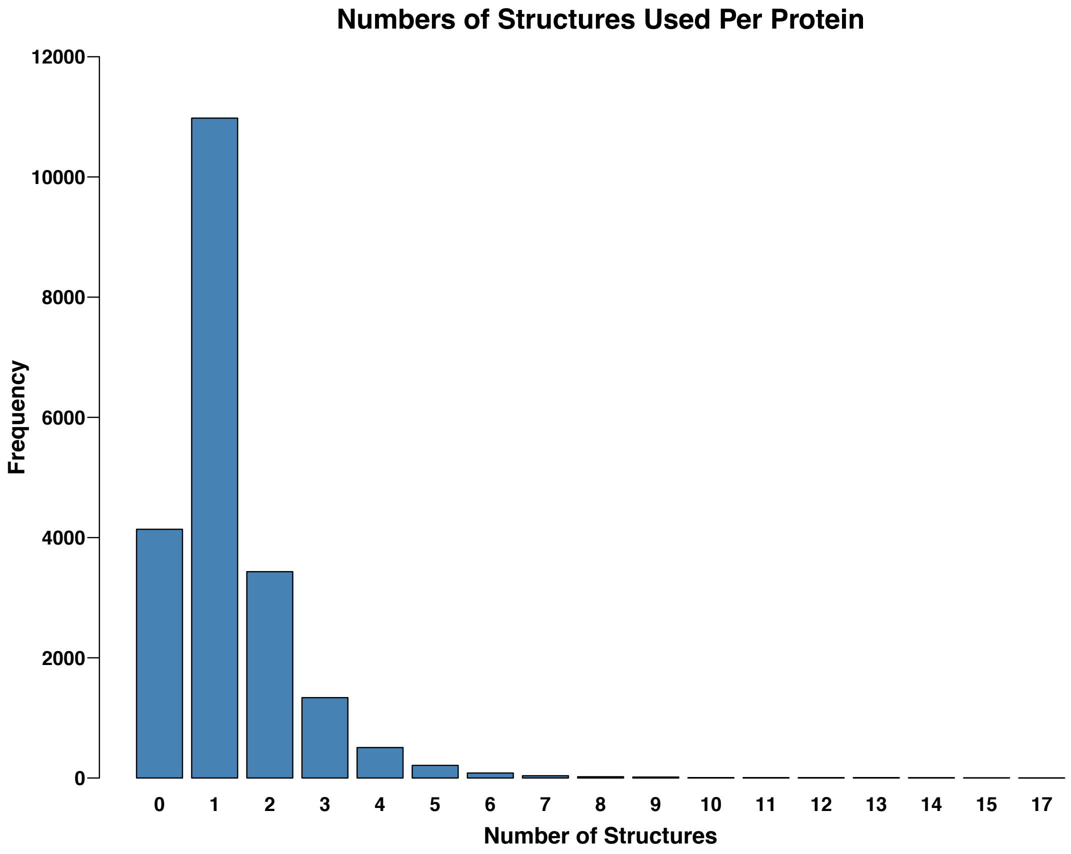


Figure 2: Numbers of non-overlapping structures used per protein

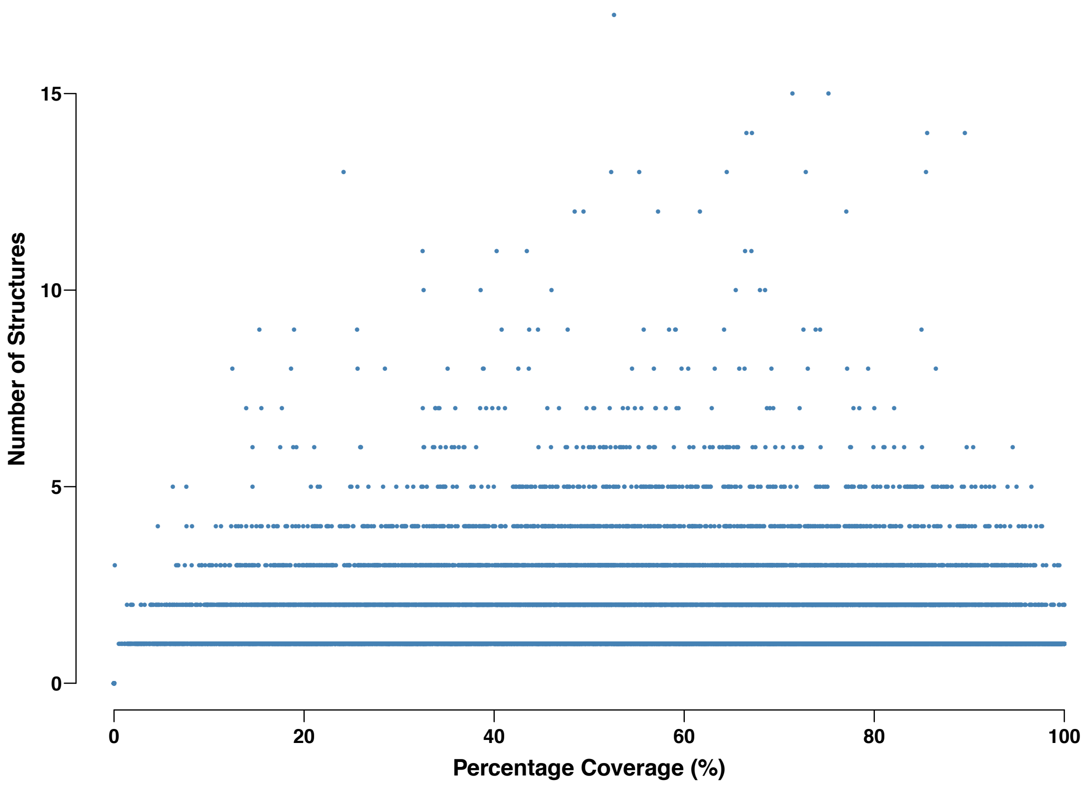


Figure 3: Percentage sequence coverage by structures for each protein vs the number of structures per protein. Each point is a protein, with the x-axis position indicating the sequence coverage achieved and y-axis indicating the number of structures used.

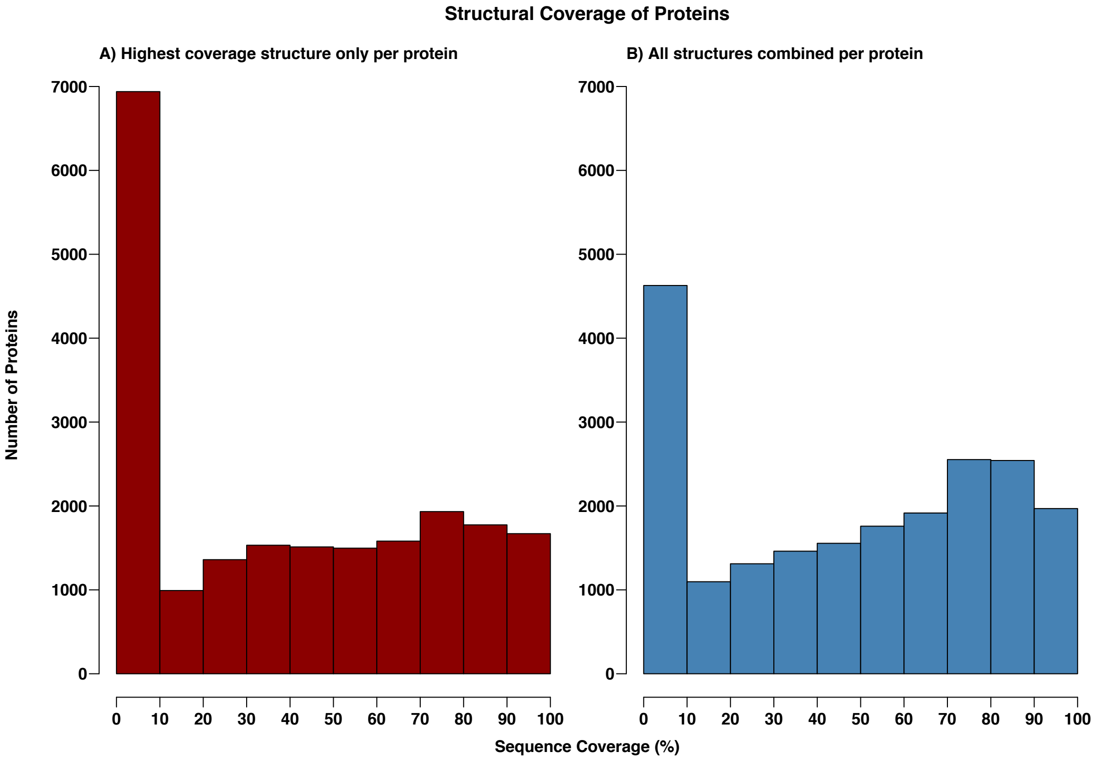


Figure 4: Percentage protein sequence coverage per protein. **A)** Using only the top coverage structure per protein. **B)** Using all non-overlapping structures per protein.

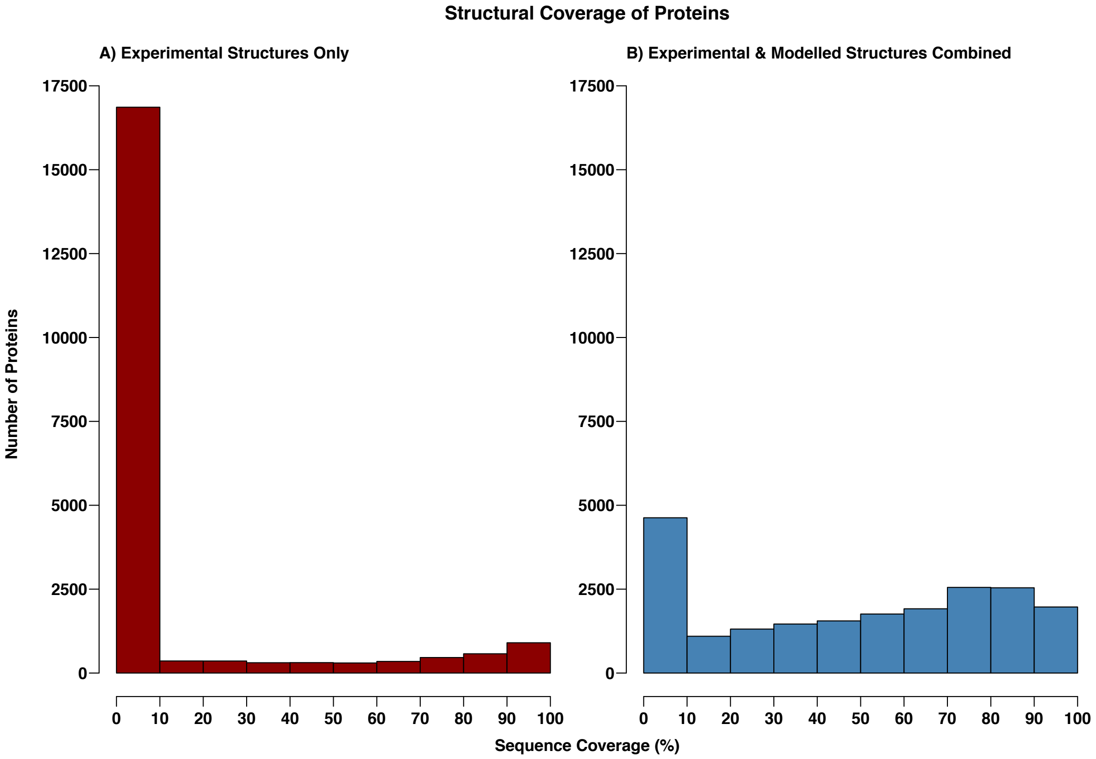


Figure 5: Percentage protein sequence coverage per protein. **A)** Using only experimental structures. **B)** Using experimental and modelled structures.

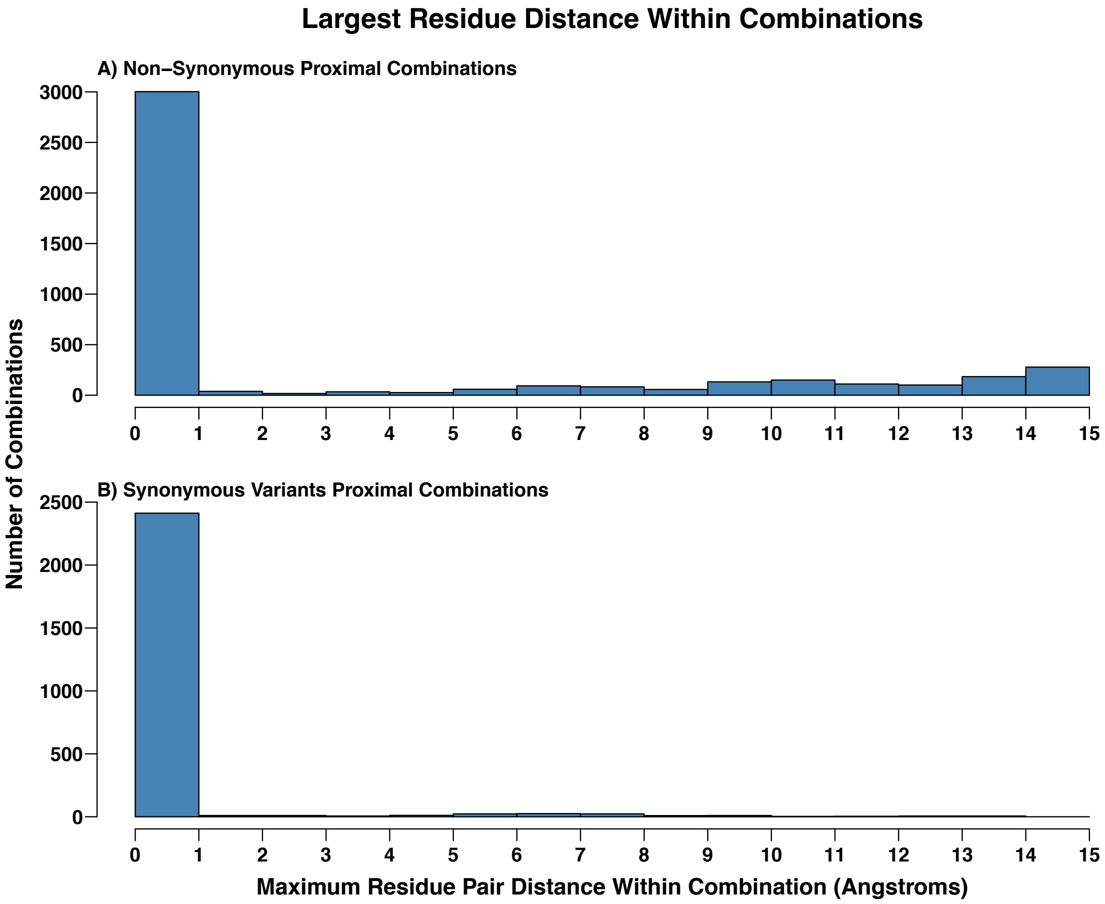


Figure 6: Maximum distances observed between any two variants within variant combinations. **A)** Non-Synonymous Proximal Combinations. **B)** Synonymous Proximal Combinations.

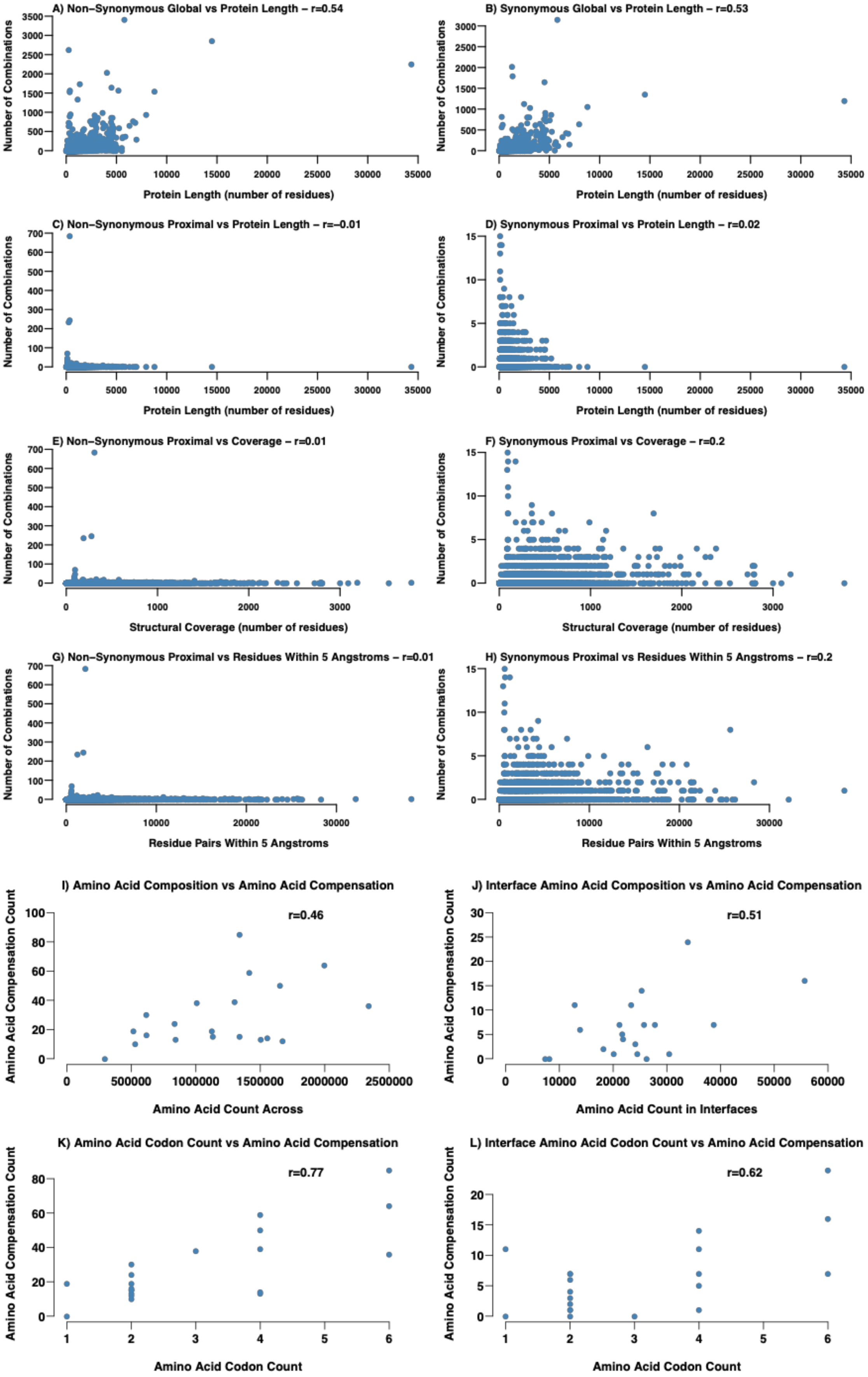


Figure 7: (see legend on next page)

Figure 7: Normalisation of combination numbers per protein and compensation types. **A)** Non-Synonymous Global Combination number vs protein length. **B)** Synonymous Global Combination number vs protein length. **C)** Non-Synonymous Proximal Combination number vs protein length. **D)** Synonymous Proximal Combination number vs protein length. **E)** Non-Synonymous Proximal Combination number vs structural coverage. **F)** Synonymous Proximal Combination number vs structural coverage. **G)** Non-Synonymous Proximal Combination number vs residue pairs within 5 angstroms. **H)** Synonymous Proximal Combination number vs residue pairs within 5 angstroms. **I)** Numbers of each amino acid across proteins vs numbers of amino acid compensations per amino acid. **J)** Numbers of each amino acid across interfaces vs numbers of amino acid compensations per amino acid. **K)** Numbers of amino acid compensations per amino acid vs numbers of codons per amino acid – Proximal Combinations. **L)** Numbers of amino acid compensations per amino acid vs numbers of codons per amino acid – Interface Variant Combinations.

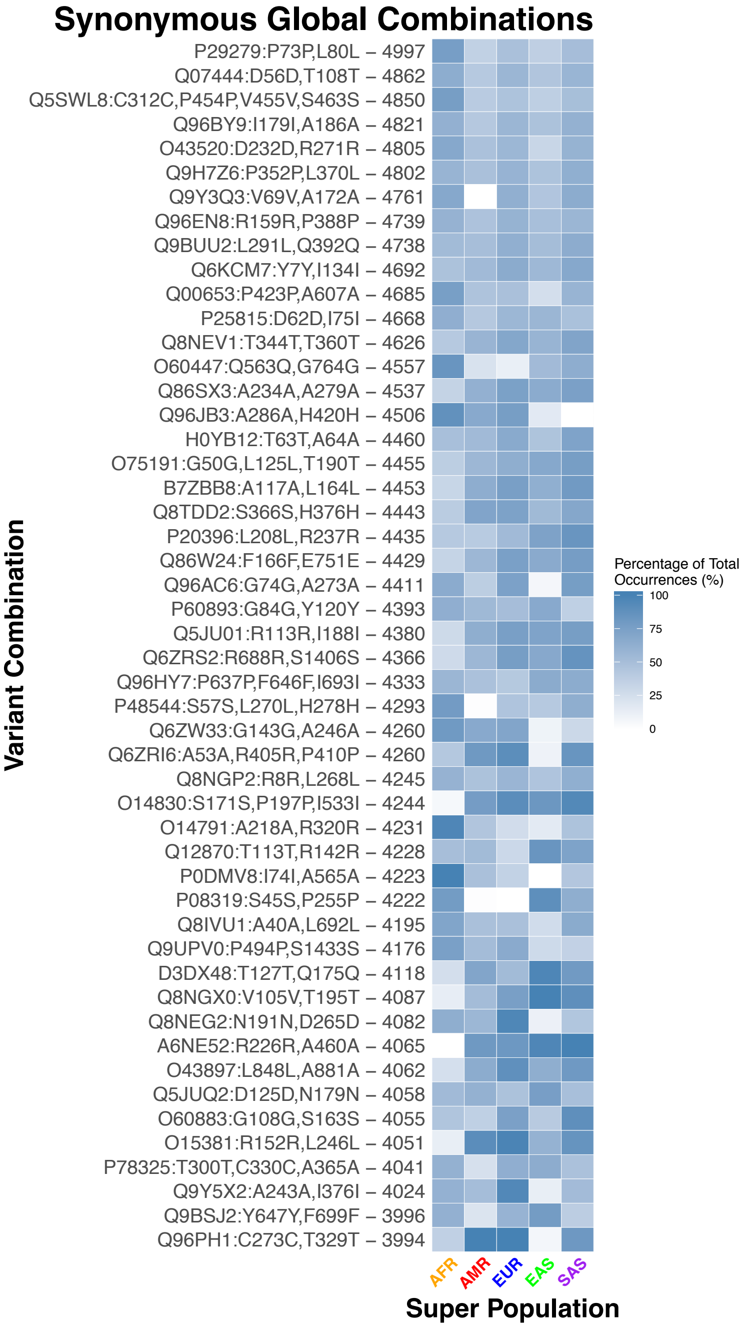


Figure 8: Heatmap of occurrences within super populations of the 50 most common synonymous Global Combinations. Combinations are given on the y-axis, see Methods for a description of the variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

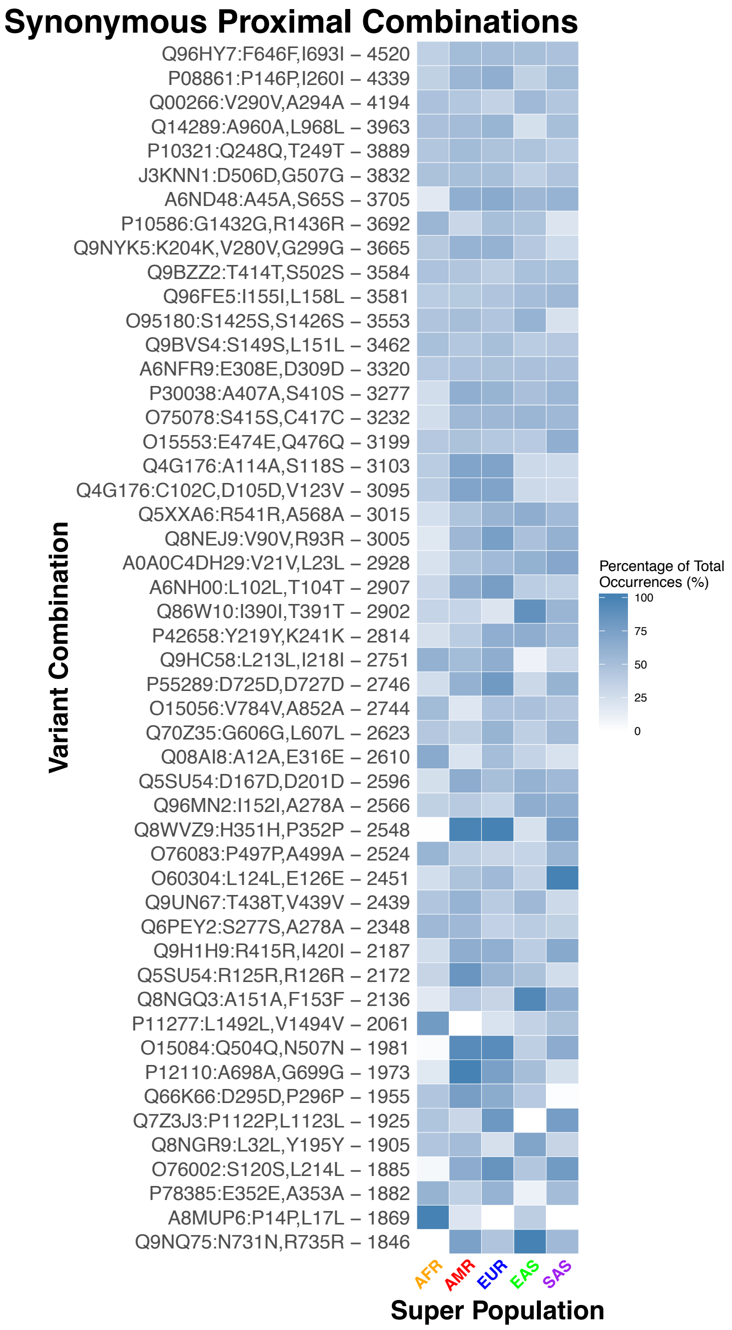


Figure 9: Heatmap of occurrences within individual populations of the 50 most common synonymous Proximal Combinations. Combinations are given on the y-axis, see Methods for a description of the variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

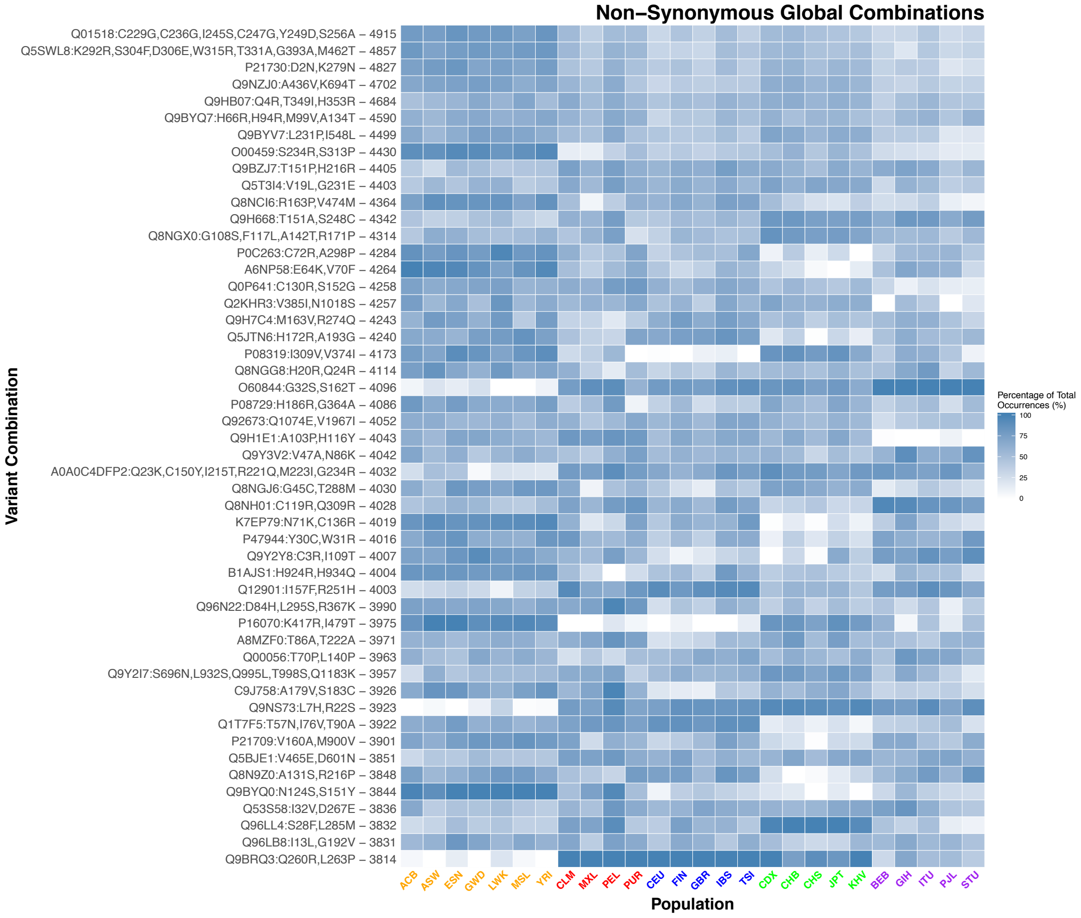


Figure 10: Heatmap of occurrences within individual populations of the 50 most common non-synonymous Global Combinations. Combinations are given on the y-axis, see Methods for a description of the variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. The population names on the x-axis are coloured to correspond to the super population colours in the super population heatmaps (see Appendix 2 Tables 1&2).

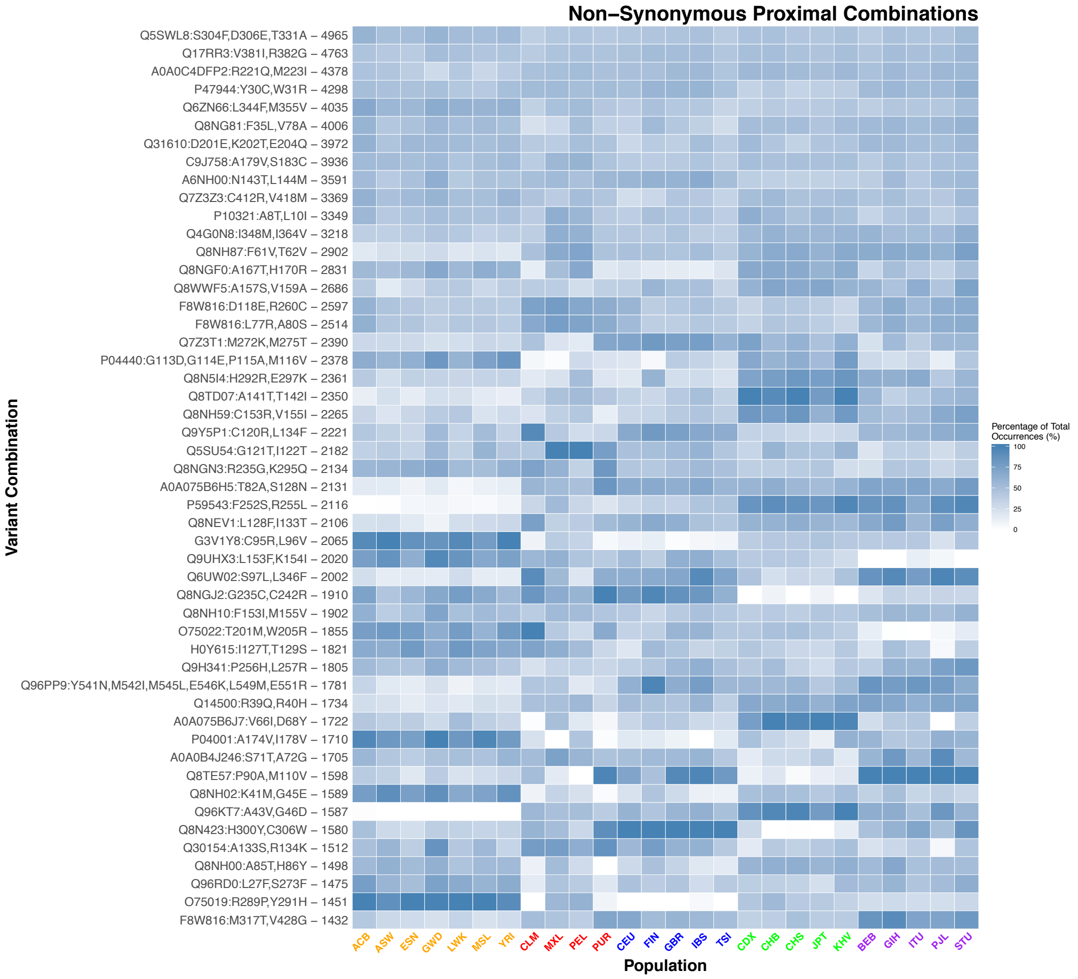


Figure 11: Heatmap of occurrences within individual populations of the 50 most common non-synonymous Proximal Combinations. Combinations are given on the y-axis, see Methods for a description of the variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. The population names on the x-axis are coloured to correspond to the super population colours in the super population heatmaps (see Appendix 2 Tables 1&2).

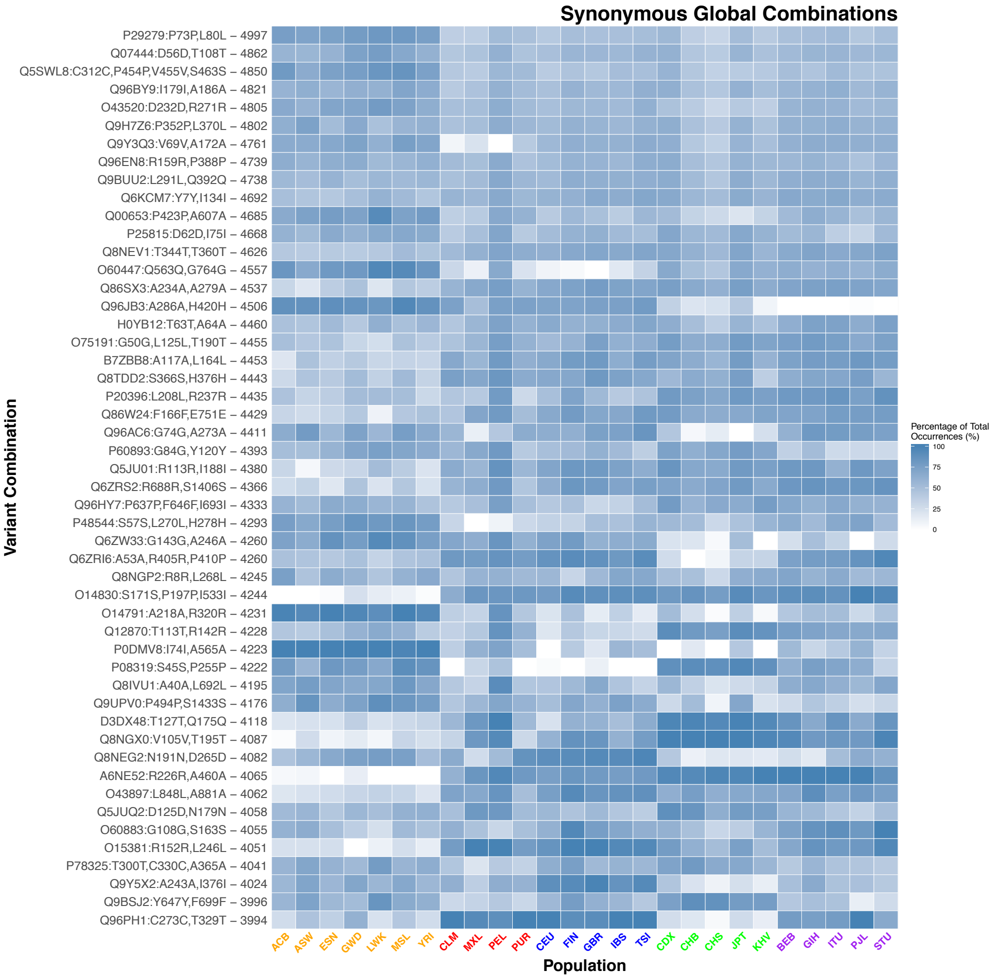


Figure 12: Heatmap of occurrences within individual populations of the 50 most common synonymous Global Combinations. Combinations are given on the y-axis, see Methods for a description of the variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. The population names on the x-axis are coloured to correspond to the super population colours in the super population heatmaps (see Appendix 2 Tables 1&2).

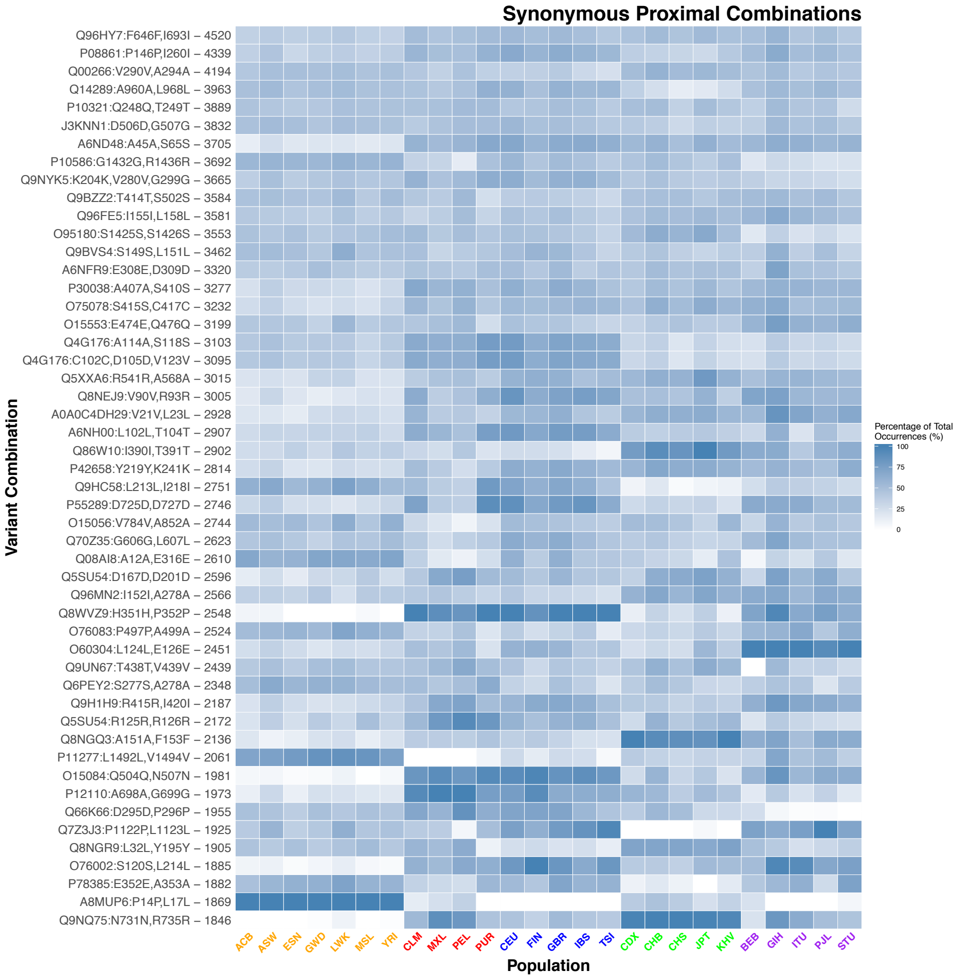


Figure 13: Heatmap of occurrences within individual populations of the 50 most common synonymous Proximal Combinations. Combinations are given on the y-axis, see Methods for a description of the variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. The population names on the x-axis are coloured to correspond to the super population colours in the super population heatmaps (see Appendix 2 Tables 1&2).

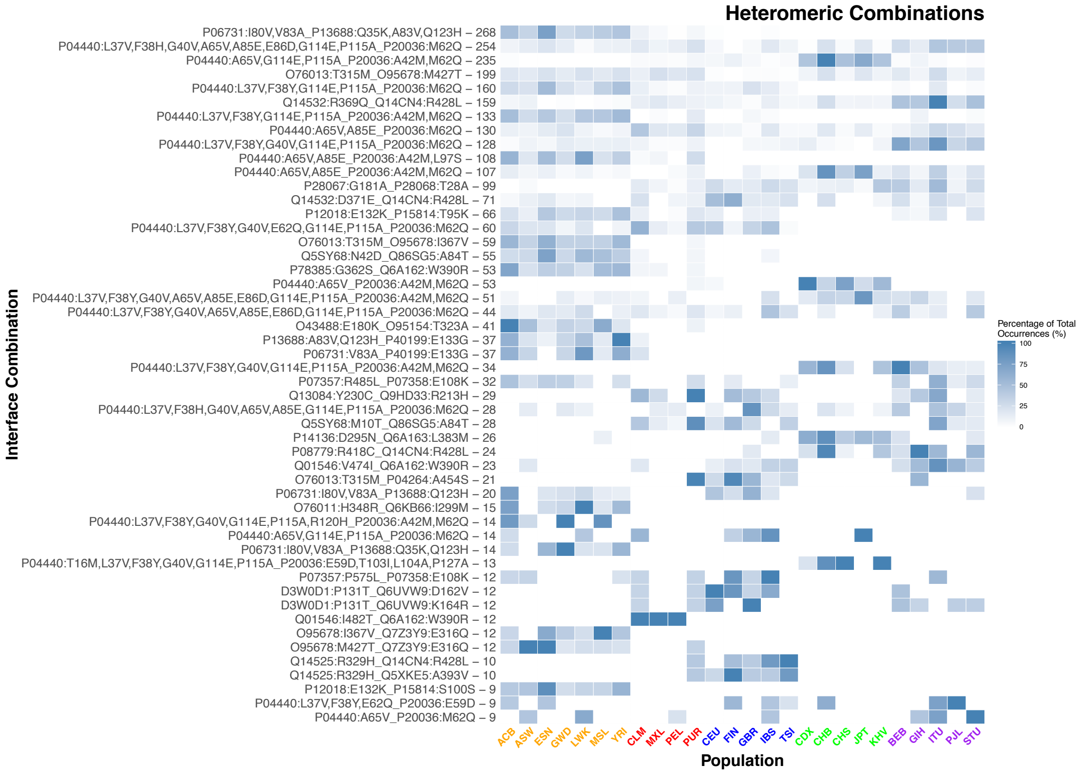


Figure 14: Heatmap of occurrences within individual populations of the 50 most common Heteromeric non-synonymous interface variant combinations. Combinations are given on the y-axis, see Chapter 3 Methods for a description of the interface variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. The population names on the x-axis are coloured to correspond to the super population colours in the super population heatmaps (see Appendix 2 Tables 1&2).

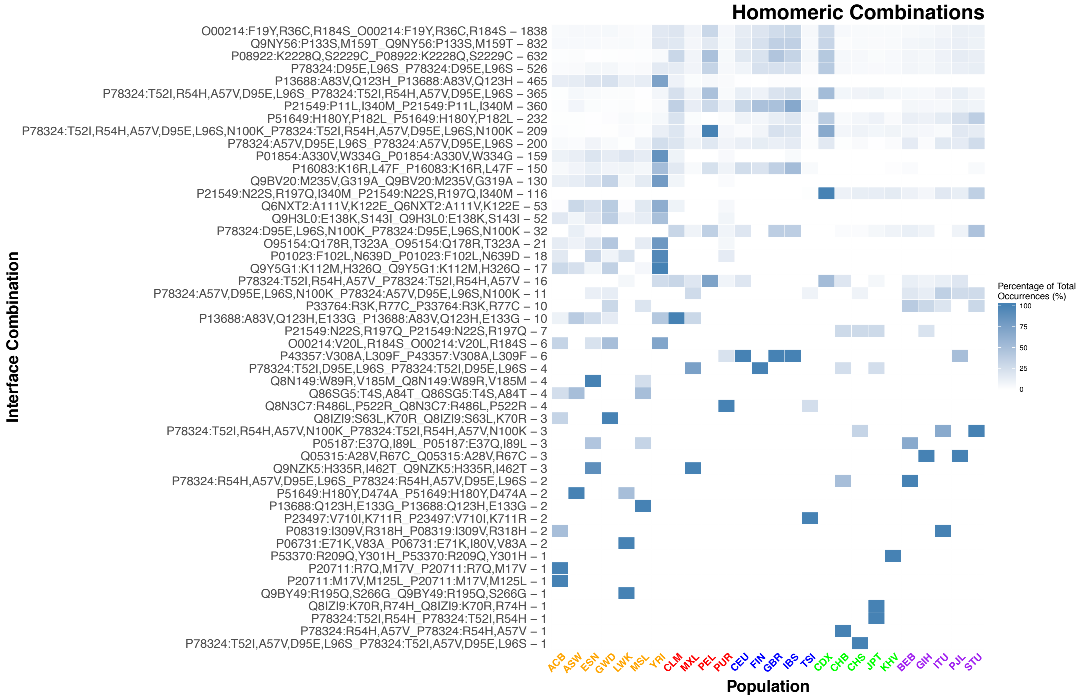


Figure 15: Heatmap of occurrences within individual populations of the 50 most common Homomeric non-synonymous interface variant combinations. Combinations are given on the y-axis, see Chapter 3 Methods for a description of the interface variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. The population names on the x-axis are coloured to correspond to the super population colours in the super population heatmaps (see Appendix 2 Tables 1&2).

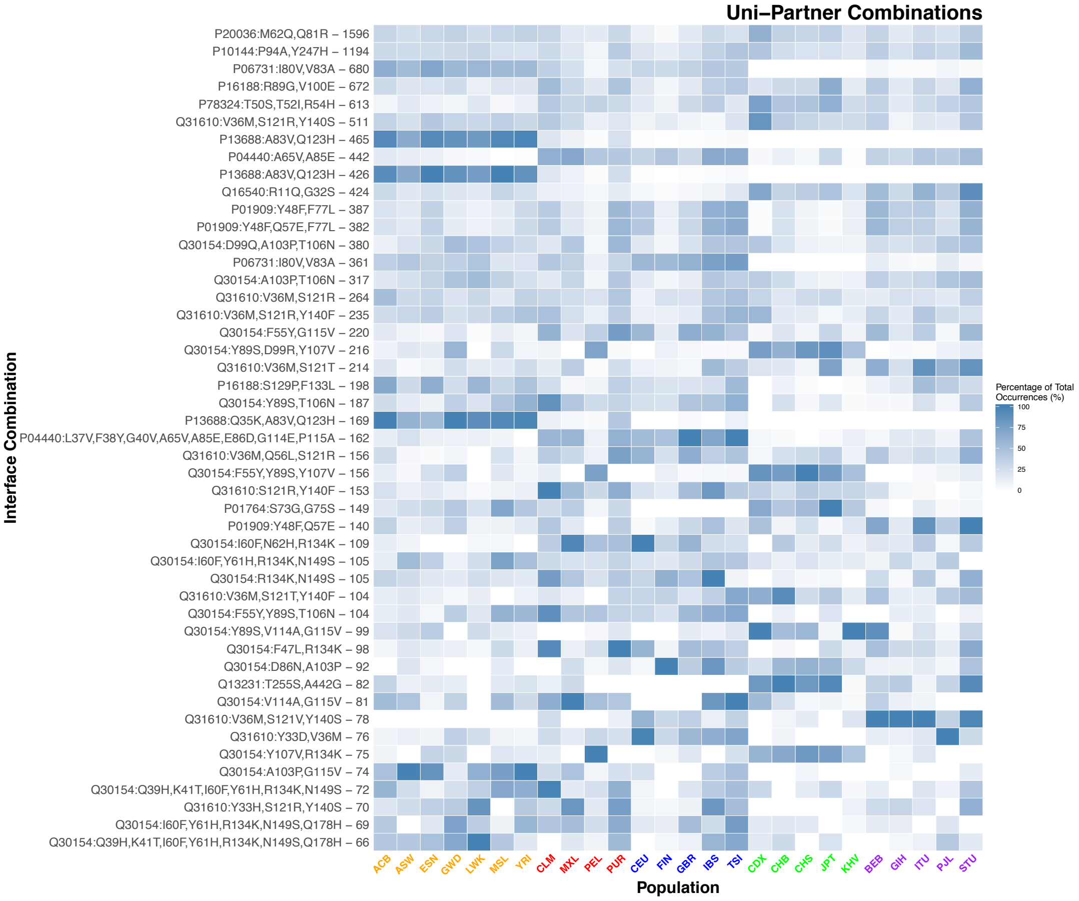


Figure 16: Heatmap of occurrences within individual populations of the 50 most common Uni-Partner non-synonymous interface variant combinations. Combinations are given on the y-axis, see Chapter 3 Methods for a description of the interface variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. The population names on the x-axis are coloured to correspond to the super population colours in the super population heatmaps (see Appendix 2 Tables 1&2).

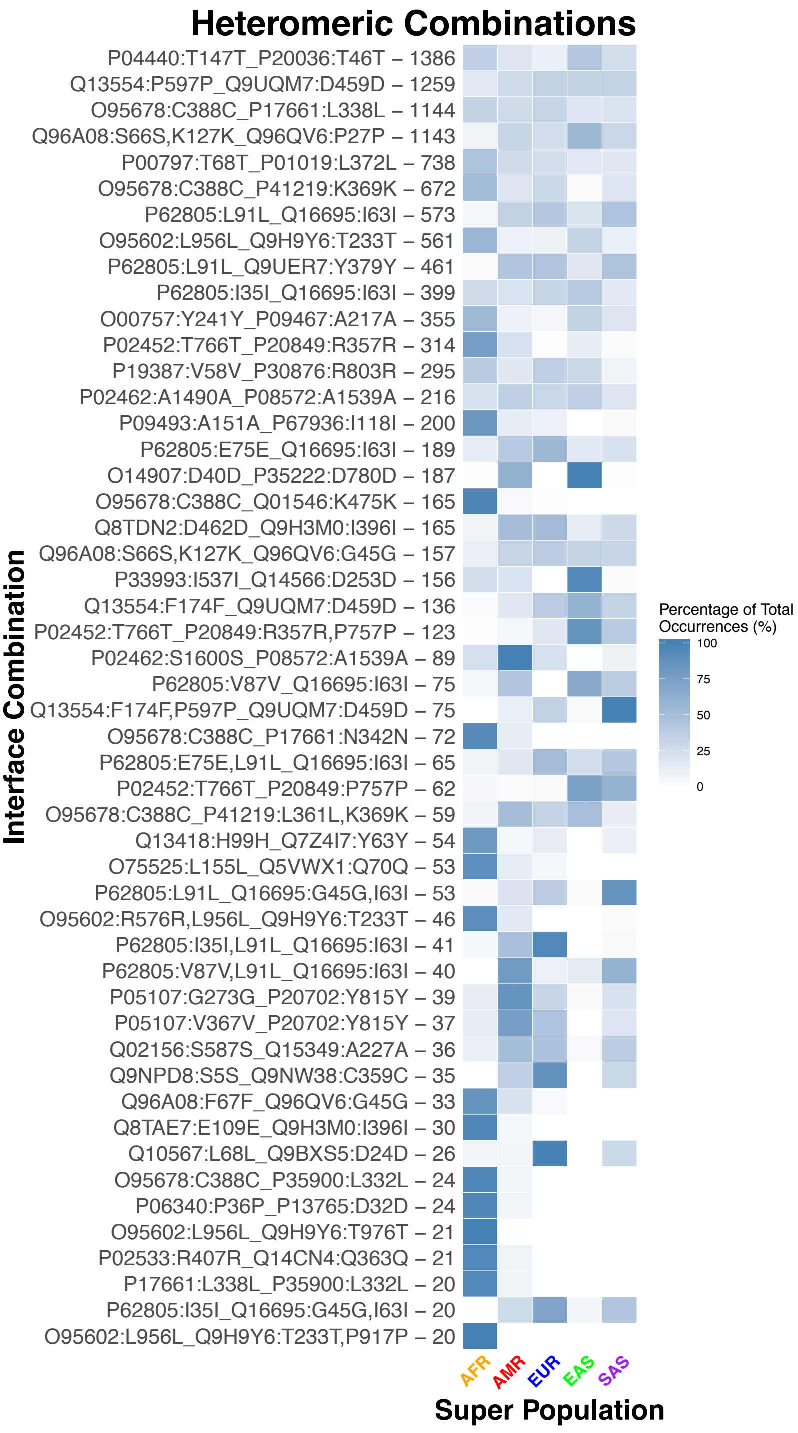


Figure 17: Heatmap of occurrences within super populations of the 50 most common Heteromeric synonymous interface variant combinations. Combinations are given on the y-axis, see Methods for a description of the interface variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

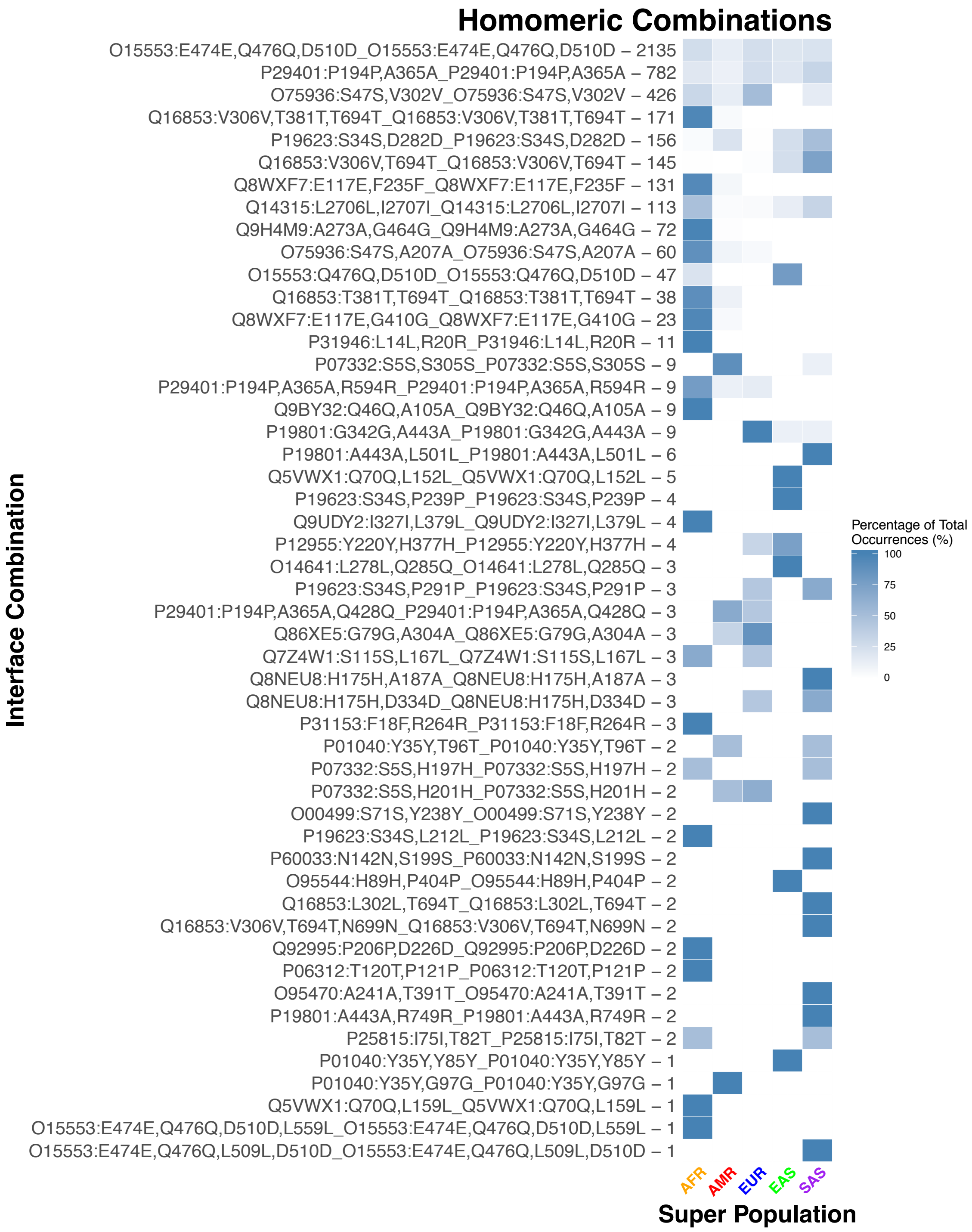


Figure 18: Heatmap of occurrences within super populations of the 50 most common Homomeric synonymous interface variant combinations. Combinations are given on the y-axis, see Methods for a description of the interface variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

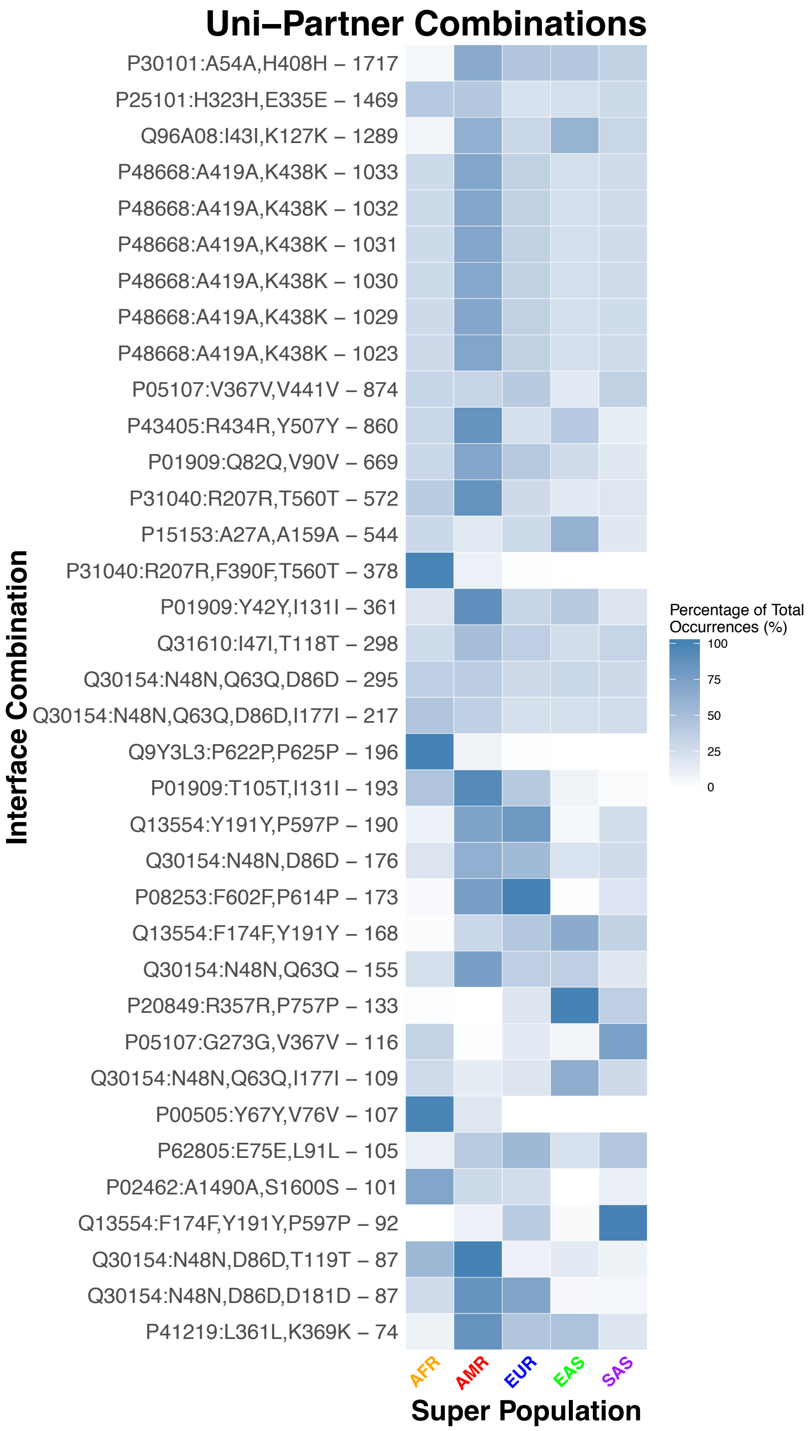


Figure 19: Heatmap of occurrences within super populations of the 50 most common Uni-Partner synonymous interface variant combinations. Combinations are given on the y-axis, see Methods for a description of the interface variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

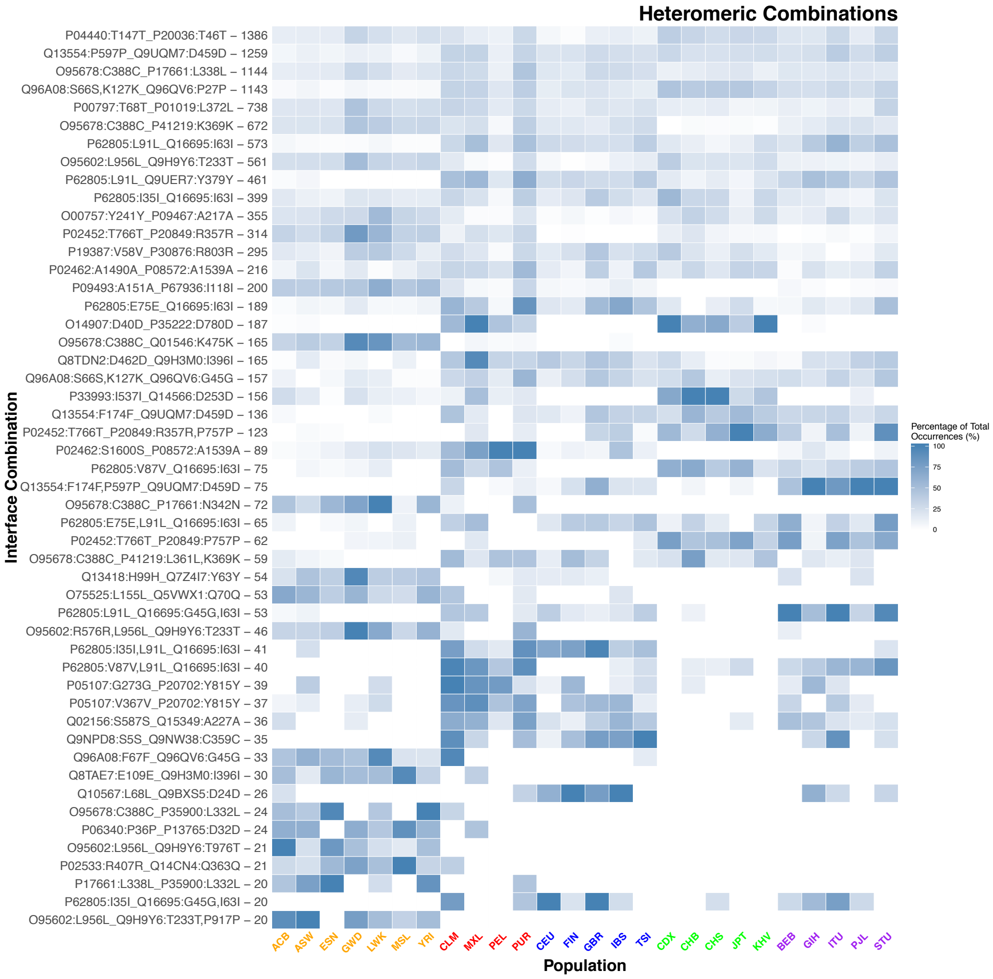


Figure 20: Heatmap of occurrences within individual populations of the 50 most common Heteromeric synonymous interface variant combinations. Combinations are given on the y-axis, see Chapter 3 Methods for a description of the interface variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. The population names on the x-axis are coloured to correspond to the super population colours in the super population heatmaps (see Appendix 2 Tables 1&2).

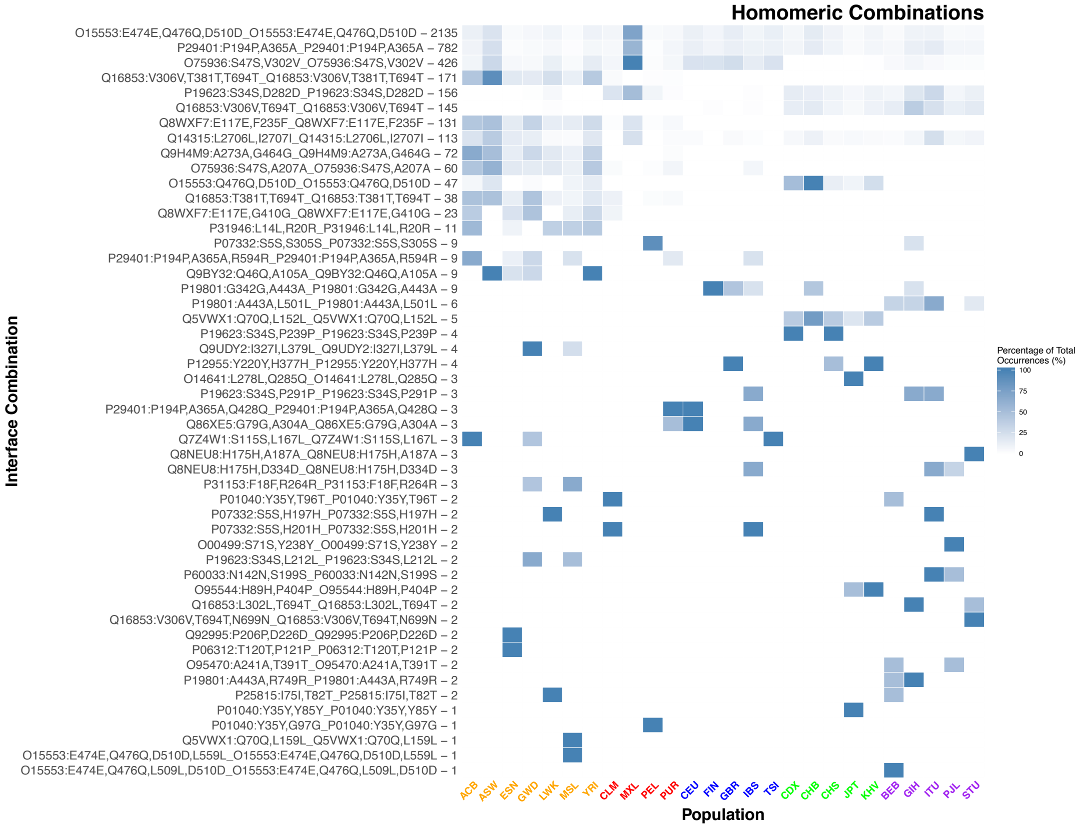


Figure 21: Heatmap of occurrences within individual populations of the 50 most common Homomeric synonymous interface variant combinations. Combinations are given on the y-axis, see Chapter 3 Methods for a description of the interface variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. The population names on the x-axis are coloured to correspond to the super population colours in the super population heatmaps (see Appendix 2 Tables 1&2).

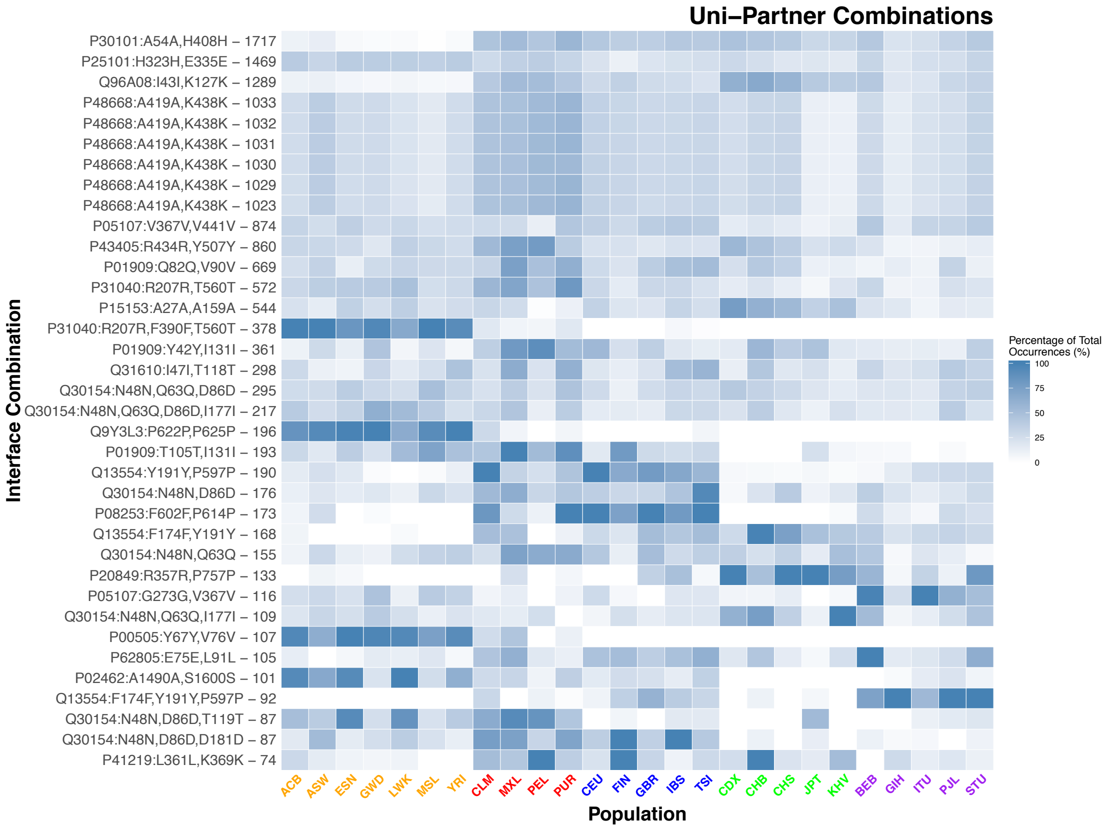


Figure 22: Heatmap of occurrences within individual populations of the 50 most common Uni-Partner synonymous interface variant combinations. Combinations are given on the y-axis, see Chapter 3 Methods for a description of the interface variant combination notation format. The number given after the variant combination is the total number of occurrences of the combination for the whole sample set. The population names on the x-axis are coloured to correspond to the super population colours in the super population heatmaps (see Appendix 2 Tables 1&2).

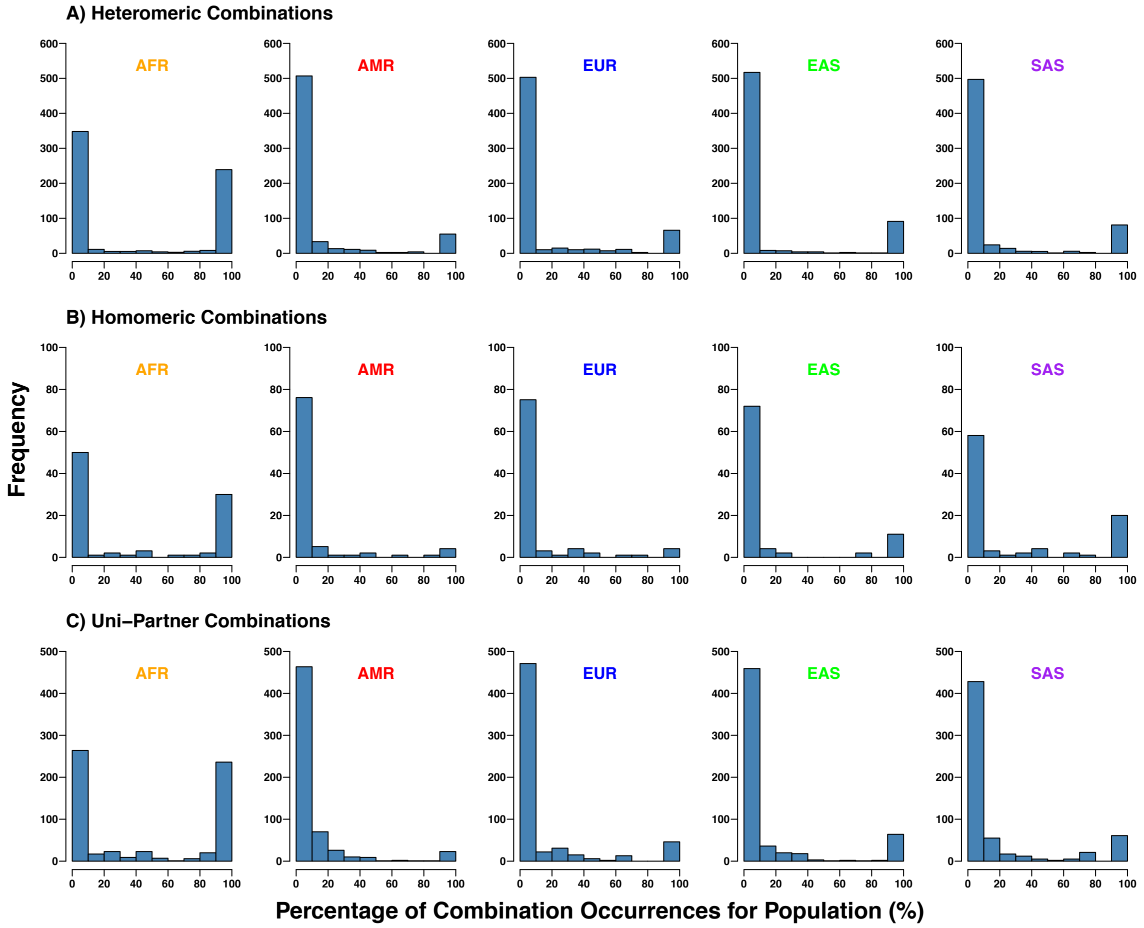


Figure 23: Distributions of percentages of total occurrences of each unique synonymous interface variant combination from each super population. **A)** Heteromeric Combinations. **B)** Homomeric Combinations. **C)** Uni-Partner Combinations. AFR – African super population, AMR – American super population, EAS – East Asian super population, EUR – European super population, SAS – South Asian super population.

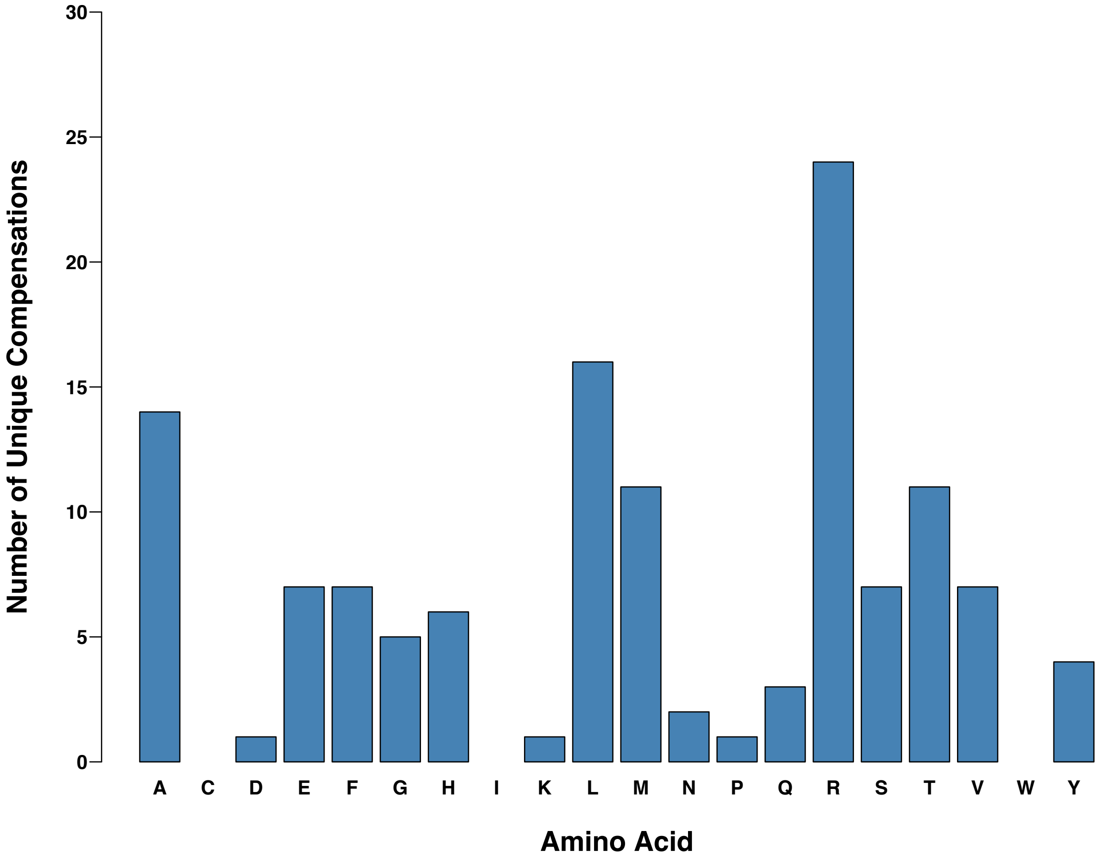


Figure 24: Numbers of unique occurrences of potential direct amino acid compensations within non-synonymous interface variant combinations

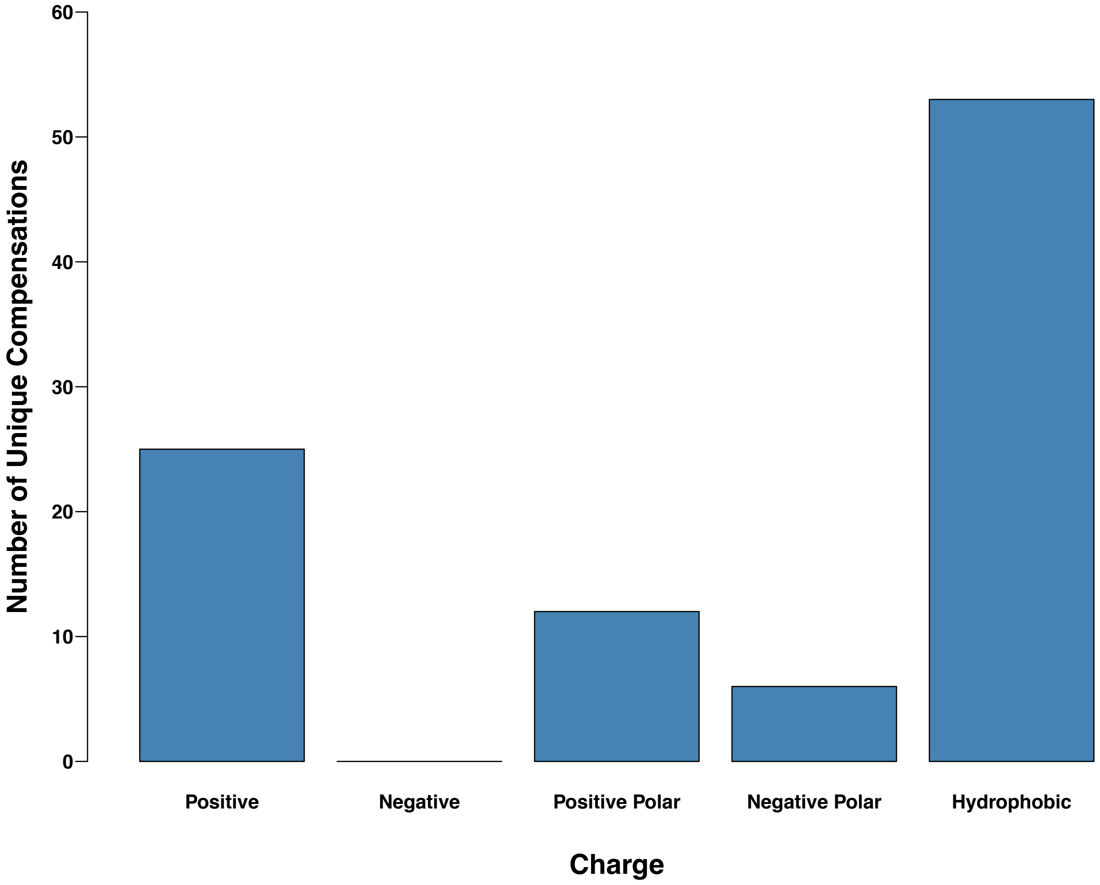


Figure 25: Numbers of unique occurrences of potential charge compensations within non-synonymous interface variant combinations

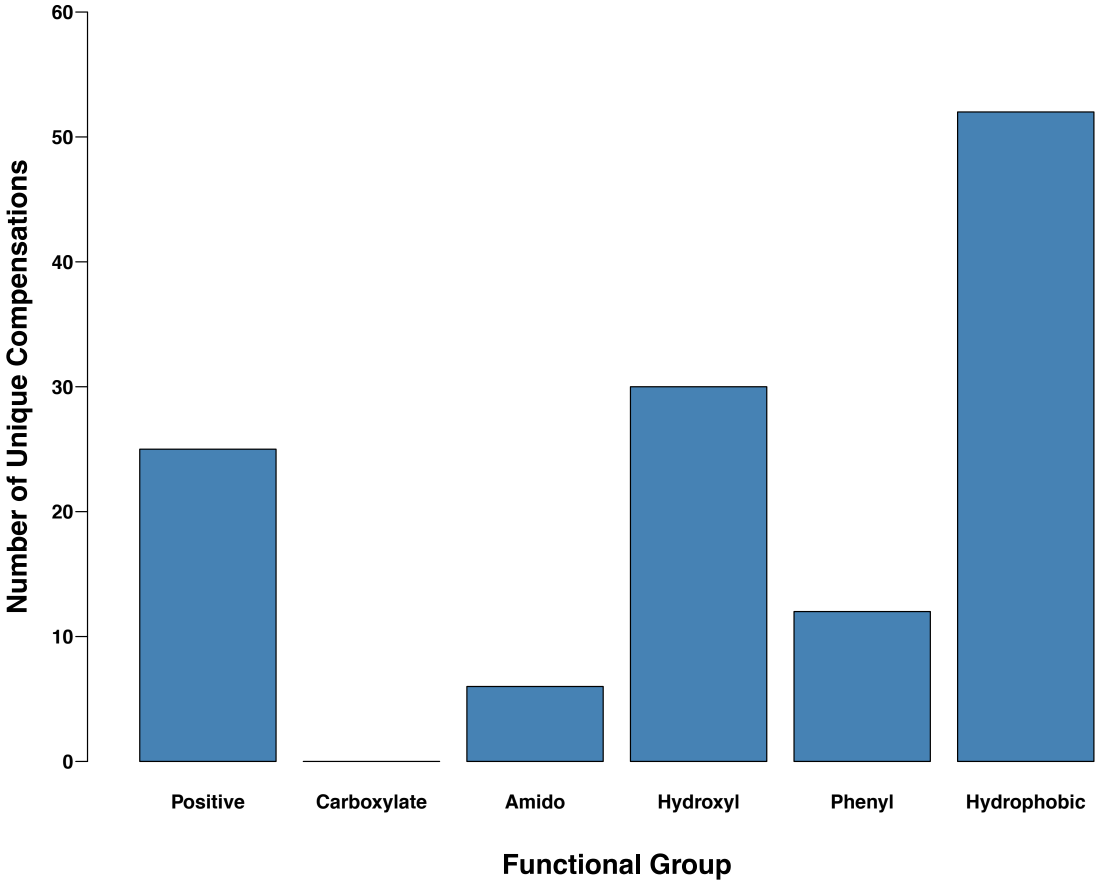


Figure 26: Numbers of unique occurrences of potential functional group compensations within non-synonymous interface variant combinations

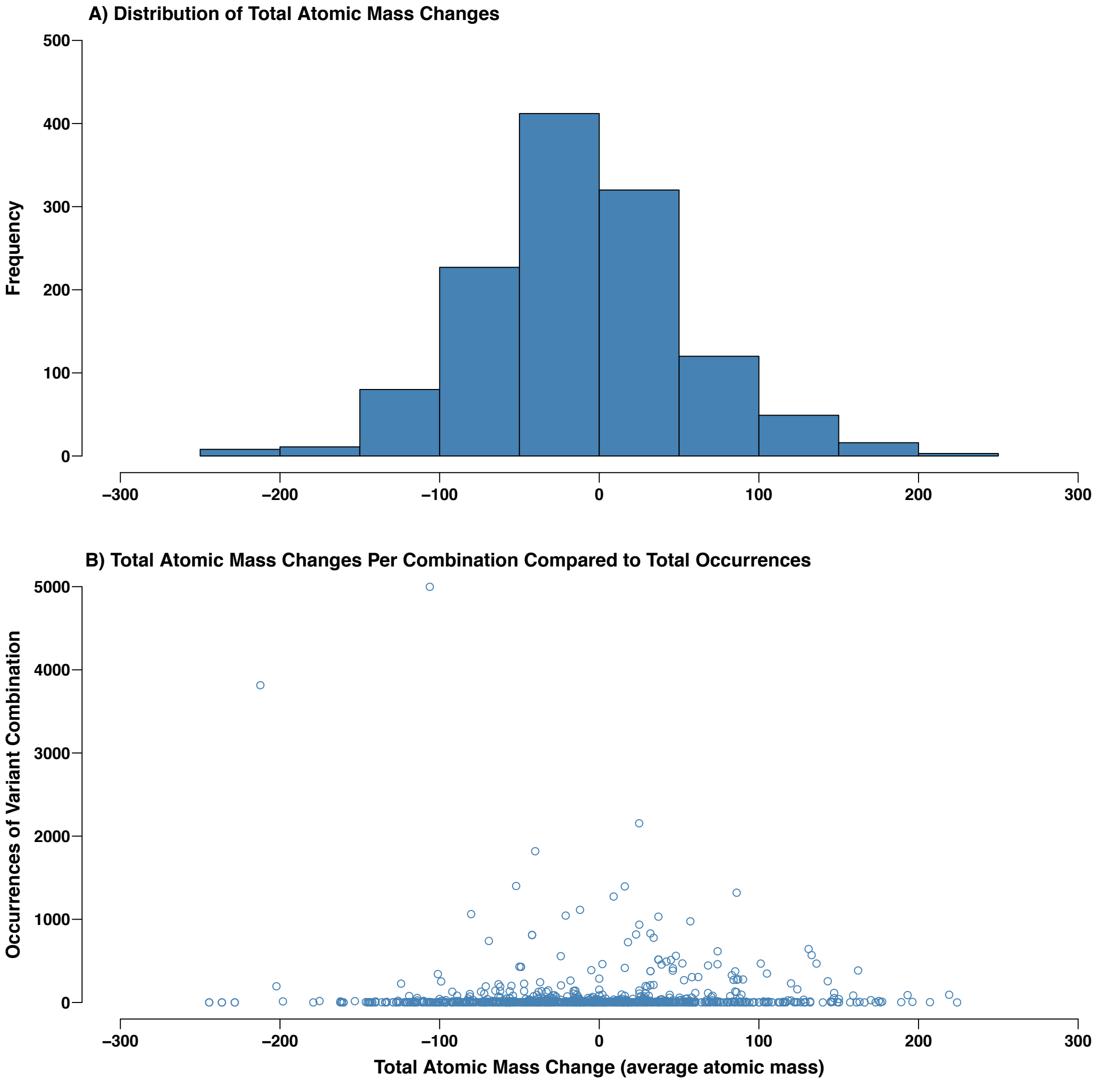


Figure 27: Total mass changes for non-synonymous interface variant combinations. **A)** The distribution of total mass changes. **B)** Comparison of the total mass change of each combination to the total number of occurrences of the combination.

Appendix 2 Tables

Table 1: Super population information for all 2,504 samples in the 1,000 Genomes Project

Super Population Code	Super Population Description	Number of Individuals
AFR	African	661
AMR	Ad Mixed American	347
EAS	East Asian	504
EUR	European	503
SAS	South Asian	489

Table 2: Population information for all 2,504 samples in the 1,000 Genomes Project

Population Code	Population Description	Super Population Code	Number of individuals
CHB	Han Chinese in Beijing, China	EAS	103
JPT	Japanese in Tokyo, Japan	EAS	104
CHS	Southern Han Chinese	EAS	105
CDX	Chinese Dai in Xishuangbanna, China	EAS	93
KHV	Kinh in Ho Chi Minh City, Vietnam	EAS	99
CEU	Utah Residents (CEPH) with Northern and Western Ancestry	EUR	99
TSI	Toscani in Italia	EUR	107
FIN	Finnish in Finland	EUR	99
GBR	British in England and Scotland	EUR	91
IBS	Iberian Population in Spain	EUR	107
YRI	Yoruba in Ibadan, Nigeria	AFR	108
LWK	Luhya in Webuye, Kenya	AFR	99
GWD	Gambian in Western Divisions in the Gambia	AFR	113
MSL	Mende in Sierra Leone	AFR	85
ESN	Esan in Nigeria	AFR	99
ASW	Americans of African Ancestry in SW USA	AFR	61
ACB	African Caribbeans in Barbados	AFR	96
MXL	Mexican Ancestry from Los Angeles USA	AMR	64
PUR	Puerto Ricans from Puerto Rico	AMR	104
CLM	Colombians from Medellin, Colombia	AMR	94
PEL	Peruvians from Lima, Peru	AMR	85
GIH	Gujarati Indian from Houston, Texas	SAS	103
PJL	Punjabi from Lahore, Pakistan	SAS	96
BEB	Bengali from Bangladesh	SAS	86
STU	Sri Lankan Tamil from the UK	SAS	102
ITU	Indian Telugu from the UK	SAS	102

Appendix 3: Chapter 4 Supplementary Material

Appendix 3 Tables

Table 1: Meta information for all ebolavirus samples used in this study, including their genome identifiers, the *Ebolavirus* species, date of collection, location of collection, and source database.

Table 1 is too large to be printed here but is available as a digital version. This can be found on the USB stick provided with this thesis.

Table 2: SDPs identified for the gene NP. For each species the amino acid residue at the alignment position is given followed by the sequence position, e.g. A1 for Alanine at sequence position 1, and the number of samples with this amino acid is given in brackets after. Where there is more than one amino acid at a position for a species these are separated by semi-colons. A 'Lost' status indicates the SDP was only found in the original analysis, a 'Retained' status indicates that the SDP was found in the old and new analysis, and a 'Gained' status indicates that the SDP is only found in the new analysis.

Alignment Position	EBOV	SUDV	BDBV	TAFV	RESTV	EBOV REF	Status
4	R4(1356)	R4(14)	R4(8)	R4(3)	G4(27)	R4	Retained
16	X16(1);E16(1355)	E16(14)	E16(8)	G16(3)	D16(27)	E16	Retained
30	S30(1356)	S30(14)	S30(8)	S30(3)	T30(27)	S30	Retained
39	R39(1356)	R39(14)	R39(8)	R39(3)	K39(27)	R39	Retained
56	I56(1356)	I56(14)	I56(8)	I56(3)	V56(27)	I56	Retained
64	V64(1356)	V64(14)	V64(8)	V64(3)	I64(27)	V64	Retained
105	R105(1354); X105(2)	R105(14)	R105(8)	R105(3)	K105(27)	R105	Retained
137	M137(1354); X137(2)	M137(14)	M137(8)	M137(3)	L137(27)	M137	Retained
212	X212(1);F212(1355)	F212(14)	F212(8)	F212(3)	Y212(27)	F212	Retained
274	K274(1355); X274(1)	K274(14)	K274(8)	K274(3)	R274(27)	K274	Retained
279	X279(1);S279(1355)	S279(14)	S279(8)	S279(3)	A279(27)	S279	Retained
416	X416(1);K416(1355)	K416(14)	K416(8)	K416(3)	N416(27)	K416	Retained
421	X421(1);Y421(1355)	Y421(14)	Y421(8)	Y421(3)	Q421(27)	Y421	Retained
426	D426(1356)	D426(14)	D426(8)	D426(3)	E426(27)	D426	Retained
435	D435(1356)	D435(14)	D435(8)	D435(3)	N435(27)	D435	Retained
443	D443(1356)	D443(14)	D443(8)	D443(3)	E443(27)	D443	Retained
453	T453(1356)	T453(14)	T453(8)	T453(3)	I453(27)	T453	Retained
497	P497(1316); S497(40)	P497(14)	P497(8)	R497(3)	A497(27)	P497	Retained
571	T563(1348);X563(7);--(1)	T563(14)	T563(8)	T563(3)	S563(27)	T563	Retained
573	--(1);X565(7);I565(1348)	I565(14)	I565(8)	I565(3)	V565(27)	I565	Retained
610	X602(24);P602(1332)	P602(14)	P602(8)	N602(3)	T602(27)	P602	Retained
650	X641(5);N641(1351)	N641(14)	N641(8)	K641(3)	Q641(27)	N641	Retained
714	A705(1356)	A705(14)	A705(8)	A705(3)	R705(27)	A705	Retained
726	G717(1354); X717(2)	G717(14)	G717(8)	D717(3)	N717(27)	G717	Retained

42	Q42(9);S42(1);P42(1346)	P42(14)	P42(8)	Q42(3)	S42(27)	P42	Lost
374	R374(1);K374(1355)	K374(14)	K374(8)	K374(3)	R374(27)	K374	Lost
492	X492(57);D492(1299)	D492(14)	D492(8)	D492(3)	E492(27)	D492	Lost
530	P526(1356)	V526(14)	G524(4);S524(4)	N524(3)	V530(1);A530(26)	P526	Lost
725	D716(1354);N716(1);X716(1)	D716(14)	D716(8)	D716(3)	N716(27)	D716	Lost

Table 3: SDPs identified for the gene VP35. For each species the amino acid residue at the alignment position is given followed by the sequence position, e.g. A1 for Alanine at sequence position 1, and the number of samples with this amino acid is given in brackets after. Where there is more than one amino acid at a position for a species these are separated by semi-colons. A 'Lost' status indicates the SDP was only found in the original analysis, a 'Retained' status indicates that the SDP was found in the old and new analysis, and a 'Gained' status indicates that the SDP is only found in the new analysis.

Alignment Position	EBOV	SUDV	BDBV	TAFV	RESTV	EBOV REF	Status
59	X26(1);S7(1);S59(2);S26(1352)	S15(14)	S27(8)	S27(3)	T15(27)	S26	Retained
81	E81(2);E48(1353);E29(1)	E37(14)	E49(8)	E49(3)	D37(27)	E48	Retained
109	G76(3);X76(1);D76(1349);D109(2);D57(1)	D65(14)	D77(8)	D77(3)	E65(27)	D76	Retained
117	X117(1);E65(1);E84(1346);E117(1);G84(1);X84(6)	E73(14)	A85(8)	E85(3)	K73(27)	E84	Gained
118	X85(7);D85(1);X118(1);E66(1);E85(1345);E118(1)	E74(14)	E86(8)	D86(3)	K74(27)	E85	Retained
125	S92(1348);X92(5);S73(1);S125(1);X125(1)	S81(14)	S93(8)	S93(3)	M81(27)	S92	Retained
130	V130(2);V97(1350);V78(1);X97(3)	V86(14)	V98(8)	I98(3)	T86(27)	V97	Retained
134	T101(1351);X101(2);T134(2);T82(1)	T90(14)	T102(8)	A102(3)	N90(27)	T101	Retained
139	S106(1352);S87(1);X106(1);S139(2)	S95(14)	S107(8)	S107(3)	A95(27)	S106	Retained
145	T112(1352);T145(2);T93(1);X112(1)	A101(3);T101(11)	T113(8)	I113(3)	S101(27)	T112	Gained
154	V154(2);X121(2);V102(1);V121(1351)	V110(14)	V122(8)	M122(3)	I110(27)	V121	Retained
158	A106(1);V125(1);A158(2);A125(1352)	A114(14)	T126(8)	A126(3)	G114(27)	A125	Gained

187	A154(1353);A135(1);A187(2)	A143(14)	A155(8)	A155(3)	S143(27)	A154	Retained
192	T140(1);T159(1353);T192(2)	T148(14)	T160(8)	T160(3)	V148(27)	T159	Retained
193	E141(1);E160(1353);E193(2)	E149(14)	E161(8)	E161(3)	D149(27)	E160	Retained
200	G200(2);X167(1);G148(1);G167(1352)	G156(14)	G168(8)	G168(3)	K156(27)	G167	Retained
207	S207(2);S155(1);S174(1353)	S163(14)	S175(8)	S175(3)	A163(27)	S174	Retained
214	I181(1353);I162(1);I214(2)	I170(14)	I182(8)	I182(3)	L170(27)	I181	Retained
302	X269(2);E269(1351);E302(1);X302(1);E250(1)	E258(14)	E270(8)	E270(3)	D258(27)	E269	Retained
323	A323(2);A271(1);X290(3);A290(1350)	A279(14)	A291(8)	A291(3)	V279(27)	A290	Retained
347	X314(3);V314(1350);V295(1);V347(2)	V303(14)	V315(8)	V315(3)	A303(27)	V314	Retained
362	Q329(1351);Q362(2);--(1);Q310(1);X329(1)	Q318(14)	Q330(8)	Q330(3)	K318(27)	Q329	Retained

Table 4: SDPs identified for the gene VP40. For each species the amino acid residue at the alignment position is given followed by the sequence position, e.g. A1 for Alanine at sequence position 1, and the number of samples with this amino acid is given in brackets after. Where there is more than one amino acid at a position for a species these are separated by semi-colons. A 'Lost' status indicates the SDP was only found in the original analysis, a 'Retained' status indicates that the SDP was found in the old and new analysis, and a 'Gained' status indicates that the SDP is only found in the new analysis.

Alignment Position	EBOV	SUDV	BDBV	TAFV	RESTV	EBOV REF	Status
4	X4(5);V4(1348);--(2);I4(1)	V4(14)	A4(8)	I4(3)	G4(27)	V4	Gained
46	I46(1);T33(2);T46(1353)	T46(14)	T46(8)	T46(3)	V46(27)	T46	Retained
85	P85(1352);P72(2);X85(2)	P85(14)	P85(8)	P85(3)	T85(27)	P85	Retained
105	T92(2);T105(1352);X105(2)	M105(1);T105(13)	T105(8)	T105(3)	I105(27)	T105	Gained
122	X122(3);I122(1351);I109(2)	I122(14)	I122(8)	I122(3)	V122(27)	I122	Retained
128	X128(1);A128(1353);A115(2)	A128(14)	T128(8)	T128(3)	I128(27)	A128	Gained
201	G201(1353);G188(2);X201(1)	G201(14)	G201(8)	G201(3)	N201(27)	G201	Retained
209	F196(2);X209(3);F209(1351)	F209(14)	F209(8)	F209(3)	L209(27)	F209	Retained
244	L244(1354);L231(2)	M244(14)	L244(8)	L244(3)	I244(27)	L244	Gained
245	Q245(1353);Q232(2);X245(1)	Q245(14)	Q245(8)	Q245(3)	P245(27)	Q245	Retained
259	M246(2);X259(1);M259(1353)	I259(14)	M259(8)	M259(3)	V259(27)	M259	Gained
269	R269(1);X269(1);H256(2);H269(1352)	H269(14)	H269(8)	H269(3)	Q269(27)	H269	Retained
277	T264(2);T277(1354)	S277(14)	T277(8)	T277(3)	Q277(27)	T277	Gained
293	I280(2);I293(1353);X293(1)	I293(14)	I293(8)	I293(3)	V293(27)	I293	Retained
323	V310(2);M323(1);A323(1);V323(1352)	L323(14)	V323(8)	V323(3)	H323(27)	V323	Gained
325	E312(2);E325(1354)	E325(14)	E325(8)	E325(3)	D325(27)	E325	Retained

Table 5: SDPs identified for the gene GP. For each species the amino acid residue at the alignment position is given followed by the sequence position, e.g. A1 for Alanine at sequence position 1, and the number of samples with this amino acid is given in brackets after. Where there is more than one amino acid at a position for a species these are separated by semi-colons. A 'Lost' status indicates the SDP was only found in the original analysis, a 'Retained' status indicates that the SDP was found in the old and new analysis, and a 'Gained' status indicates that the SDP is only found in the new analysis.

Alignment Position	EBOV	SUDV	BDBV	TAFV	RESTV	EBOV REF	Status
2	M1(1356)	M1(14)	M1(8)	M1(3)	G2(27)	M1	Retained
32	F31(1355);S31(1)	F31(14)	F31(8)	F31(3)	I32(27)	F31	Retained
38	V37(1356)	V37(14)	V37(8)	V37(3)	I38(27)	V37	Retained
46	V45(1356)	V45(14)	V45(8)	V45(3)	A46(27)	V45	Retained
76	V75(1349);A75(7)	V75(14)	V75(8)	V75(3)	I76(27)	V75	Retained
197	S196(1356)	S196(14)	S196(8)	S196(3)	A197(27)	S196	Retained
261	I260(1356)	I260(14)	I260(8)	I260(3)	L261(27)	I260	Retained
270	T269(1356)	T269(14)	T269(8)	T269(3)	S270(27)	T269	Retained
308	X307(1);S307(1355)	S307(14)	S307(8)	S307(3)	H308(27)	S307	Retained
497	X476(4);S476(1352)	S476(14)	S476(8)	L476(3)	P477(27)	S476	Gained
519	R498(1353);X498(3)	R498(14)	R498(8)	R498(3)	K499(27)	R498	Retained
521	X500(2);R500(1354)	R500(14)	R500(8)	R500(3)	K501(27)	R500	Retained
535	N514(1354);X514(2)	N514(14)	N514(8)	N514(3)	D515(27)	N514	Retained
542	X521(1);Q521(1355)	Q521(14)	Q521(8)	L521(3)	V522(27)	Q521	Retained
605	I584(1356)	I584(14)	I584(8)	I584(3)	L585(27)	I584	Retained
628	D607(1354);X607(2)	D607(14)	D607(8)	D607(3)	S608(27)	D607	Retained
643	X622(1);K622(1355)	K622(14)	K622(8)	K622(3)	E623(27)	K622	Retained
659	Q638(1352);L638(1);R638(1);X638(2)	Q638(14)	Q638(8)	Q638(3)	H639(27)	Q638	Retained
665	X644(2);W644(1354)	W644(14)	W644(8)	W644(3)	L645(27)	W644	Retained
680	A659(1);T659(1354);X659(1)	T659(14)	T659(8)	T659(3)	I660(27)	T659	Retained
3	G2(1356)	G2(11);E2(3)	V2(8)	G2(3)	S3(27)	G2	Lost

208	X207(57);E207(1299)	E207(14)	T207(8)	T207(3)	D208(27)	E207	Lost
211	S210(1299);X210(57)	S210(14)	S210(8)	S210(3)	T211(27)	S210	Lost
326	R325(1354);X325(2)	R325(14)	V325(8)	V325(3)	G326(27)	R325	Lost
355	H354(1356)	H354(14)	R354(8)	Q354(3)	L355(27)	H354	Lost
417	X403(10);Q403(1346)	S412(14)	A409(8)	T409(3)	E412(27)	Q403	Lost
432	S418(1339);X418(15);X417(1);A417(1)	T427(14)	S419(8)	T419(3)	T422(27)	S418	Lost
468	T448(1345);X448(8);A448(3)	-(14)	T451(8)	K451(3)	-(27)	T448	Lost
537	H516(1355);X516(1)	H516(14)	H516(8)	H516(3)	H517(14);Y517(13)	H516	Lost
568	L547(1352);X547(4)	L547(14)	I547(8)	I547(3)	V548(27)	L547	Lost
663	D642(1354);X642(2)	D642(14)	D642(8)	S642(3)	L643(27)	D642	Lost

Table 6: SDPs identified for the gene VP30. For each species the amino acid residue at the alignment position is given followed by the sequence position, e.g. A1 for Alanine at sequence position 1, and the number of samples with this amino acid is given in brackets after. Where there is more than one amino acid at a position for a species these are separated by semi-colons. A 'Lost' status indicates the SDP was only found in the original analysis, a 'Retained' status indicates that the SDP was found in the old and new analysis, and a 'Gained' status indicates that the SDP is only found in the new analysis.

Alignment Position	EBOV	SUDV	BDBV	TAFV	RESTV	EBOV REF	Status
40	H39(2);X39(1);Y39(1353)	Y39(14)	Y39(8)	Y39(3)	R40(27)	Y39	Gained
53	X52(1);T52(1355)	T52(14)	T52(8)	T52(3)	N53(27)	T52	Retained
54	X53(1);V53(1355)	V53(14)	V53(8)	V53(3)	L54(27)	V53	Retained
64	X63(3);T63(1353)	T63(14)	T63(8)	T63(3)	I64(27)	T63	Retained
94	E93(1356)	E93(14)	E93(8)	E93(3)	D94(27)	E93	Retained
97	T96(1355);X96(1)	T96(14)	T96(8)	T96(3)	N97(27)	T96	Retained
99	R98(1355);X98(1)	R98(14)	R98(8)	R98(3)	H99(27)	R98	Retained
108	K107(1354); X107(2)	K107(14)	K107(8)	K107(3)	R108(27)	K107	Retained
112	S111(1356)	S111(14)	S111(8)	S111(3)	I112(27)	S111	Retained
117	X116(1);L116(1355)	L116(14)	L116(8)	L116(3)	S117(27)	L116	Retained
118	N117(1356)	N117(14)	N117(8)	S117(3)	Q118(27)	N117	Gained
121	A120(1356)	A120(14)	A120(8)	A120(3)	S121(27)	A120	Retained
151	T150(1355);X150(1)	T150(14)	T150(8)	T150(3)	I151(27)	T150	Retained
158	X157(1);Q157(1355)	Q157(14)	Q157(8)	Q157(3)	R158(27)	Q157	Retained
160	X159(1);I159(1355)	I159(14)	I159(8)	I159(3)	L160(27)	I159	Retained
206	E205(1356)	E205(14)	E205(8)	E205(3)	D206(27)	E205	Retained
263	R262(1356)	R262(14)	R262(8)	R262(3)	A263(27)	R262	Retained
269	S268(1356)	S268(14)	S268(8)	S268(3)	Q269(27)	S268	Retained
272	E271(1356)	E271(14)	T271(8)	T271(3)	S272(27)	E271	Gained
279	X278(1);G278(1355)	G278(14)	E278(8)	E278(3)	N279(27)	G278	Gained
197	H196(1);R196(1354);X196(1)	R196(14)	R196(8)	R196(3)	H197(27)	R196	Lost

Table 7: SDPs identified for the gene VP24. For each species the amino acid residue at the alignment position is given followed by the sequence position, e.g. A1 for Alanine at sequence position 1, and the number of samples with this amino acid is given in brackets after. Where there is more than one amino acid at a position for a species these are separated by semi-colons. A 'Lost' status indicates the SDP was only found in the original analysis, a 'Retained' status indicates that the SDP was found in the old and new analysis, and a 'Gained' status indicates that the SDP is only found in the new analysis.

Alignment Position	EBOV	SUDV	BDBV	TAFV	RESTV	EBOV REF	Status
17	X17(1);L17(1355)	L17(14)	L17(8)	L17(3)	M17(27)	L17	Retained
22	X22(1);V22(1355)	V22(14)	V22(8)	V22(3)	I22(27)	V22	Retained
31	V31(1356)	V31(14)	V31(8)	V31(3)	I31(27)	V31	Retained
102	I102(1354);V102(2)	I102(14)	I102(8)	I102(3)	L102(27)	I102	Gained
131	T131(1356)	T131(14)	T131(8)	T131(3)	S131(27)	T131	Retained
132	N132(1356)	N132(14)	N132(8)	N132(3)	T132(27)	N132	Retained
136	I136(15);M136(1341)	M136(14)	M136(8)	M136(3)	L136(27)	M136	Retained
139	Q139(1356)	Q139(14)	Q139(8)	Q139(3)	R139(27)	Q139	Retained
140	R140(1356)	R140(14)	H140(8)	Q140(3)	S140(27)	R140	Gained
226	X226(4);T226(1352)	T226(14)	T226(8)	T226(3)	A226(27)	T226	Retained
248	S248(1356)	S248(14)	S248(8)	S248(3)	L248(27)	S248	Retained

Table 8: SDPs identified for the gene L. For each species the amino acid residue at the alignment position is given followed by the sequence position, e.g. A1 for Alanine at sequence position 1, and the number of samples with this amino acid is given in brackets after. Where there is more than one amino acid at a position for a species these are separated by semi-colons. A 'Lost' status indicates the SDP was only found in the original analysis, a 'Retained' status indicates that the SDP was found in the old and new analysis, and a 'Gained' status indicates that the SDP is only found in the new analysis.

Alignment Position	EBOV	SUDV	BDBV	TAFV	RESTV	EBOV REF	Status
67	V59(1);V66(1355)	V67(14)	V66(8)	V66(3)	T66(27)	V66	Retained
137	I136(1355);I129(1)	I137(14)	I136(8)	I136(3)	L136(27)	I136	Retained
147	L139(1);L146(1355)	L147(14)	L146(8)	L146(3)	V146(27)	L146	Retained
203	T202(1346);S202(9);T195(1)	T203(14)	T202(8)	T202(3)	I202(27)	T202	Gained
222	A221(1355);A214(1)	A222(14)	A221(8)	A221(3)	S221(27)	A221	Retained
224	Q223(1355);Q216(1)	Q224(14)	Q223(8)	Q223(3)	L223(27)	Q223	Retained
227	T219(1);T226(1354);A226(1)	T227(14)	T226(8)	T226(3)	S226(27)	T226	Gained
228	H227(1355);H220(1)	H228(14)	H227(8)	H227(3)	Q227(27)	H227	Retained
237	V236(1355);V229(1)	V237(14)	V236(8)	V236(3)	I236(27)	V236	Gained
284	L283(1355);L276(1)	L284(14)	L283(8)	L283(3)	V283(27)	L283	Retained
331	T323(1);T330(1355)	T331(14)	T330(8)	T330(3)	D330(27)	T330	Retained
351	E343(1);E350(1355)	E351(14)	E350(8)	E350(3)	D350(27)	E350	Retained
362	M361(1);T361(1353);T354(1);X361(1)	T362(14)	T361(8)	T361(3)	S361(27)	T361	Retained
366	L358(1);L365(1354);X365(1)	L366(14)	L365(8)	L365(3)	F365(27)	L365	Retained
380	V379(1353);V372(1);X379(2)	V380(14)	V379(8)	V379(3)	I379(27)	V379	Retained
448	X447(4);Q447(1351);Q440(1)	Q448(14)	Q447(8)	Q447(3)	H447(27)	Q447	Retained
451	P450(1351);P443(1);X450(4)	P451(14)	P450(8)	P450(3)	S450(27)	P450	Retained

466	D465(1355);D458(1)	D466(14)	D465(8)	D465(3)	N465(27)	D465	Retained
848	X847(1);S847(1354);S840(1)	S848(14)	S847(8)	S847(3)	A847(27)	S847	Retained
869	S861(1);S868(1355)	S869(14)	S868(8)	S868(3)	A868(27)	S868	Retained
1025	T1017(1);T1024(1355)	T1025(14)	T1024(8)	T1024(3)	N1024(27)	T1024	Retained
1074	R1066(1);R1073(1355)	R1074(14)	R1073(8)	R1073(3)	K1073(27)	R1073	Retained
1120	A1112(1);A1119(1355)	A1120(14)	A1119(8)	A1119(3)	S1119(27)	A1119	Retained
1164	P1156(1);P1163(1355)	P1162(14)	P1163(8)	P1163(3)	A1161(27)	P1163	Retained
1190	D1189(1355);D1182(1)	D1188(14)	D1189(8)	D1189(3)	S1187(27)	D1189	Retained
1215	A1214(1355);A1207(1)	A1213(14)	A1214(8)	A1214(3)	S1212(27)	A1214	Retained
1218	R1210(1);R1217(1355)	R1216(14)	R1217(8)	R1217(3)	K1215(27)	R1217	Retained
1238	D1237(1355);D1230(1)	D1236(14)	D1237(8)	D1237(3)	E1235(27)	D1237	Retained
1355	R1354(1355);R1347(1)	R1353(14)	R1354(8)	R1354(3)	K1352(27)	R1354	Retained
1367	T1359(1);T1366(1355)	T1365(14)	T1366(8)	T1366(3)	A1364(27)	T1366	Retained
1409	I1408(1355);I1401(1)	I1407(14)	I1408(8)	I1408(3)	M1406(27)	I1408	Retained
1415	I1407(1);I1414(1355)	I1413(14)	I1414(8)	I1414(3)	L1412(27)	I1414	Retained
1437	S1429(1);S1436(1355)	S1435(14)	S1436(8)	S1436(3)	N1434(27)	S1436	Retained
1474	S1466(1);X1473(2);S1473(1353)	S1472(14)	S1473(8)	S1473(3)	C1471(27)	S1473	Retained
1489	L1488(1355);L1481(1)	L1487(14)	L1488(8)	L1488(3)	Y1486(27)	L1488	Retained
1500	I1499(1355);I1492(1)	I1498(14)	I1499(8)	I1499(3)	L1497(27)	I1499	Retained
1507	S1506(1355);S1499(1)	S1505(14)	S1506(8)	S1506(3)	A1504(27)	S1506	Retained
1510	I1509(1355);I1502(1)	I1508(14)	I1509(8)	I1509(3)	V1507(27)	I1509	Retained
1627	L1617(1);L1624(1355)	L1623(14)	L1624(8)	L1624(3)	Y1624(27)	L1624	Retained
1631	C1628(1355);C1621(1)	C1627(14)	C1628(8)	C1628(3)	S1628(27)	C1628	Retained
1786	V1755(1);V1762(1355)	V1759(14)	V1762(8)	V1762(3)	I1760(27)	V1762	Retained
1874	V1843(1);V1850(1355)	V1847(14)	V1850(8)	V1850(3)	T1848(27)	V1850	Retained

1897	I1873(1);T1866(1);T1873(1354)	T1870(14)	T1873(8)	T1873(3)	S1871(27)	T1873	Retained
1941	R1909(1);R1916(1355)	R1913(14)	R1916(8)	R1916(3)	N1915(3);N1914(24)	R1916	Retained
1966	E1941(1354);X1941(1);E1934(1)	E1938(14)	E1941(8)	E1941(3)	R1939(24);R1940(3)	E1941	Retained
2069	L2044(1355);L2037(1)	L2041(14)	L2044(8)	L2044(3)	I2043(3);I2042(24)	L2044	Retained
2102	S2077(1355);S2070(1)	S2074(14)	S2077(8)	S2077(3)	T2075(24);T2076(3)	S2077	Retained
2123	E2091(1);E2098(1355)	E2095(14)	E2098(8)	E2098(3)	D2096(24);D2097(3)	E2098	Retained
2182	L2157(1353);X2157(2);L2150(1)	L2154(14)	L2157(8)	L2157(3)	V2155(24);V2156(3)	L2157	Retained
2193	R2168(1355);R2161(1)	R2165(14)	R2168(8)	R2168(3)	H2167(3);H2166(24)	R2168	Retained
2200	R2168(1);R2175(1355)	R2172(14)	R2175(8)	R2175(3)	K2173(24);K2174(3)	R2175	Retained
2202	X2177(1);L2177(1354);L2170(1)	L2174(14)	L2177(8)	L2177(3)	F2175(24);F2176(3)	L2177	Retained
2211	X2186(2);M2179(1);M2186(1353)	M2183(14)	M2186(8)	M2186(3)	L2185(3);L2184(24)	M2186	Retained
110	Q109(1298);X109(57);Q102(1)	Q110(14)	Q109(8)	Q109(3)	R109(2);H109(25)	Q109	Lost
277	L276(1355);X269(1)	L277(14)	L276(8)	L276(3)	I276(27)	L276	Lost
313	Y312(1354);X305(1);X312(1)	Y313(14)	Y312(8)	Y312(3)	F312(27)	Y312	Lost
327	A319(1);X326(1);A326(1354)	A327(14)	A326(8)	A326(3)	S326(27)	A326	Lost
690	E689(1353);E682(1);X689(2)	E690(14)	E689(8)	E689(3)	S689(27)	E689	Lost
897	X896(58);F896(1297);F889(1)	F897(14)	F896(8)	F896(3)	Y896(27)	F896	Lost
926	L925(1352);X925(3);L918(1)	L926(14)	L925(8)	L925(3)	F925(27)	L925	Lost
955	X954(2);A954(1353);A947(1)	A955(14)	A954(8)	A954(3)	S954(27)	A954	Lost
996	X995(2);S995(1353);S988(1)	S996(14)	S995(8)	S995(3)	T995(27)	S995	Lost

1256	V1255(1);I1248(1);I1255(1354)	I1254(14)	I1255(8)	I1255(3)	V1253(27)	I1255	Lost
1396	A1395(1);S1395(1353);S1388(1);X1395(1)	S1394(14)	S1395(8)	S1395(3)	T1393(27)	S1395	Lost
1462	X1461(1);K1454(1);K1461(1354)	K1460(14)	K1461(8)	K1461(3)	Q1459(27)	K1461	Lost
1539	X1538(1);A1538(1354);A1531(1)	A1537(14)	A1538(8)	A1538(3)	S1536(27)	A1538	Lost
2033	X2008(57);L2001(1);L2008(1298)	L2005(14)	L2008(8)	L2008(3)	I2007(3);I2006(24)	L2008	Lost
2130	X2105(2);Q2098(1);Q2105(1353)	Q2102(14)	Q2105(8)	Q2105(3)	L2104(3);L2103(24)	Q2105	Lost
2133	Q2108(1353);Q2101(1);X2108(2)	Q2105(14)	Q2108(8)	Q2108(3)	E2107(3);E2106(24)	Q2108	Lost
2156	Y2124(1);Y2131(1354);X2131(1)	Y2128(14)	Y2131(8)	Y2131(3)	F2129(24);F2130(3)	Y2131	Lost

Appendix 4: Chapter 5 Supplementary Material

Appendix 4 Figures

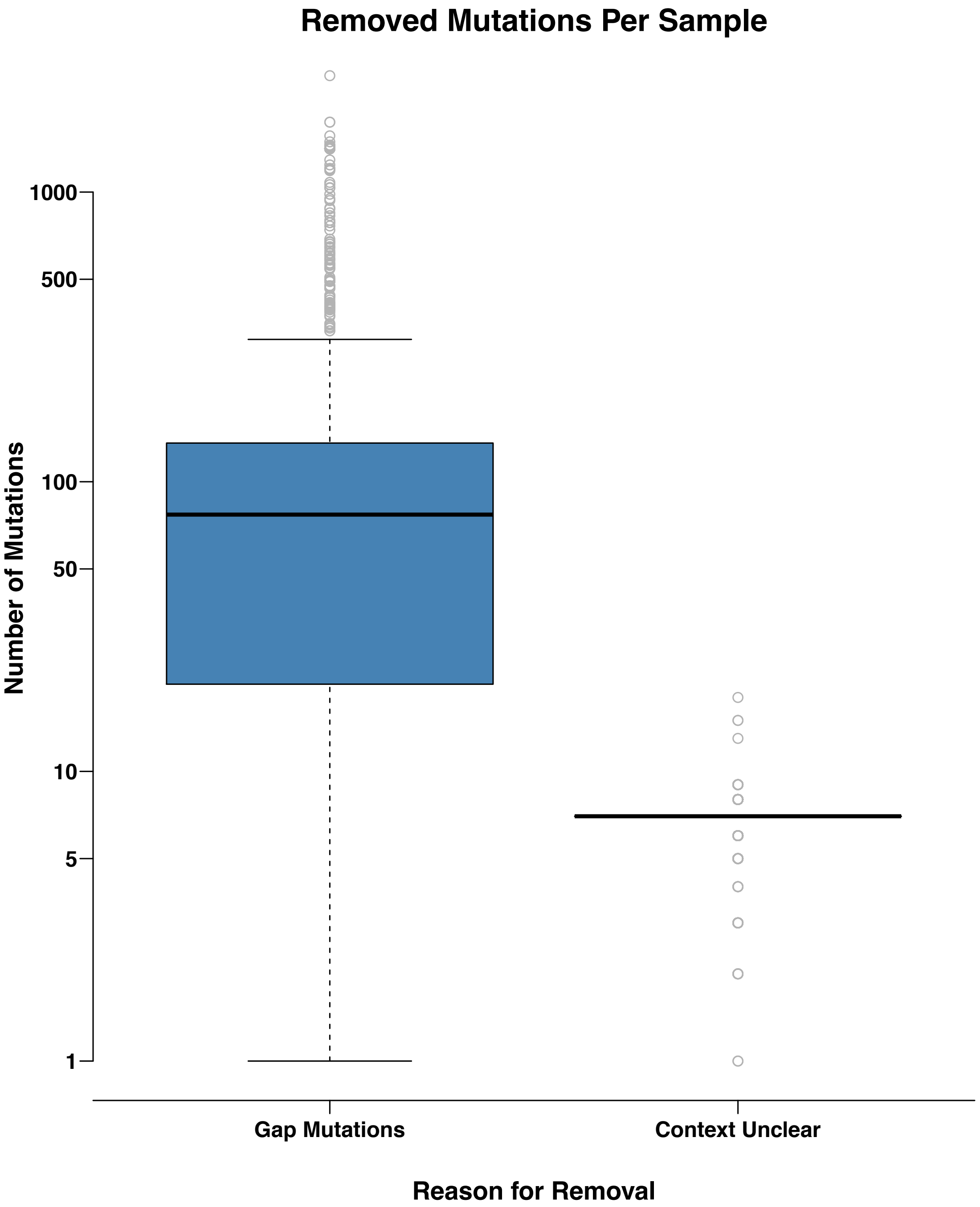


Figure 1: Reasons for removal of variants per sample. Gap Mutations are those where differences compared to the reference were caused by gaps in the alignment, and Context Unclear mutations are those where either the 5’ or 3’ residue to the mutated residue are a gap in the alignment.

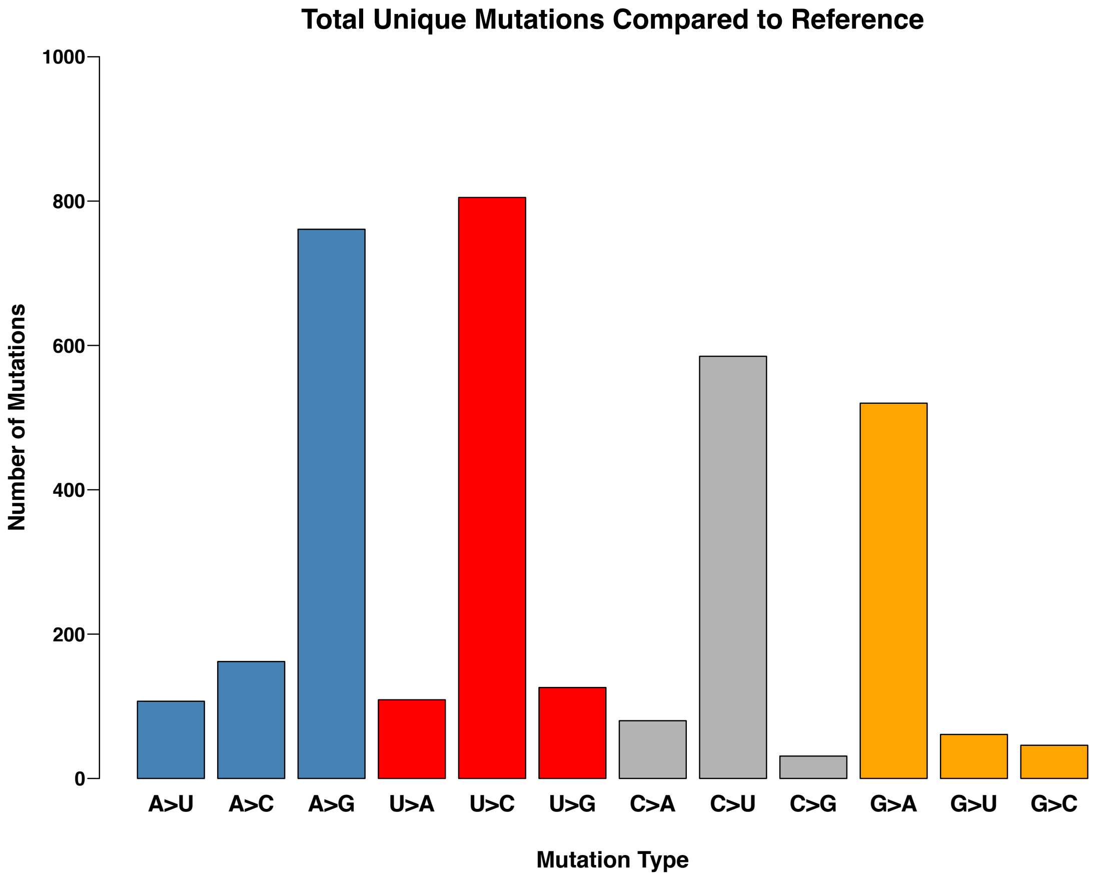


Figure 2: Observed mutation types compared to the reference genome for all unique mutations across the sample set. Colours indicated the starting base of the mutation – blue A, red U, grey C, and orange G.

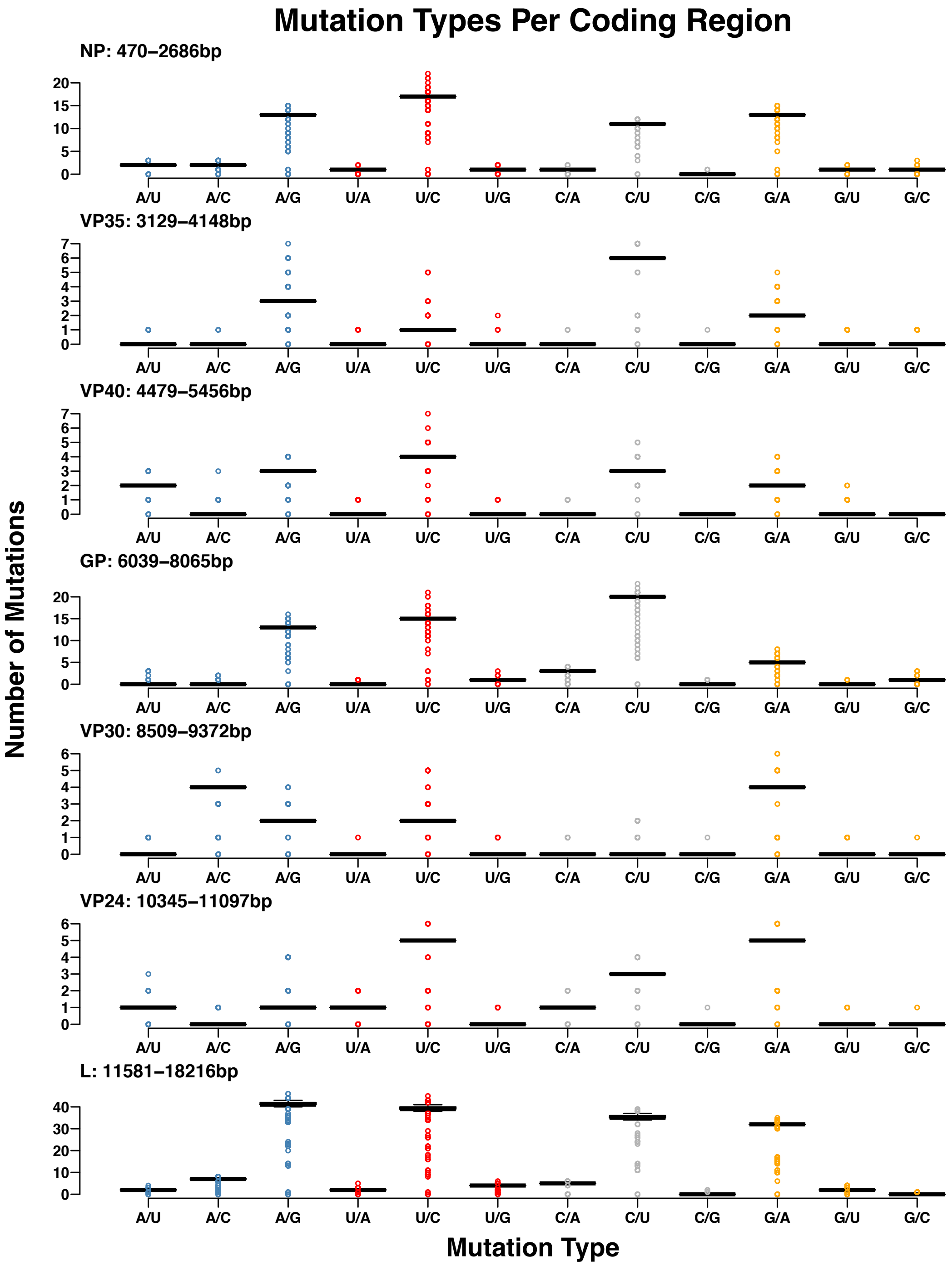


Figure 3: Observed mutation types for each sample compared to the reference genome, for each of the coding regions of the genome. Colours indicated the starting base of the mutation – blue A, red U, grey C, and orange G.

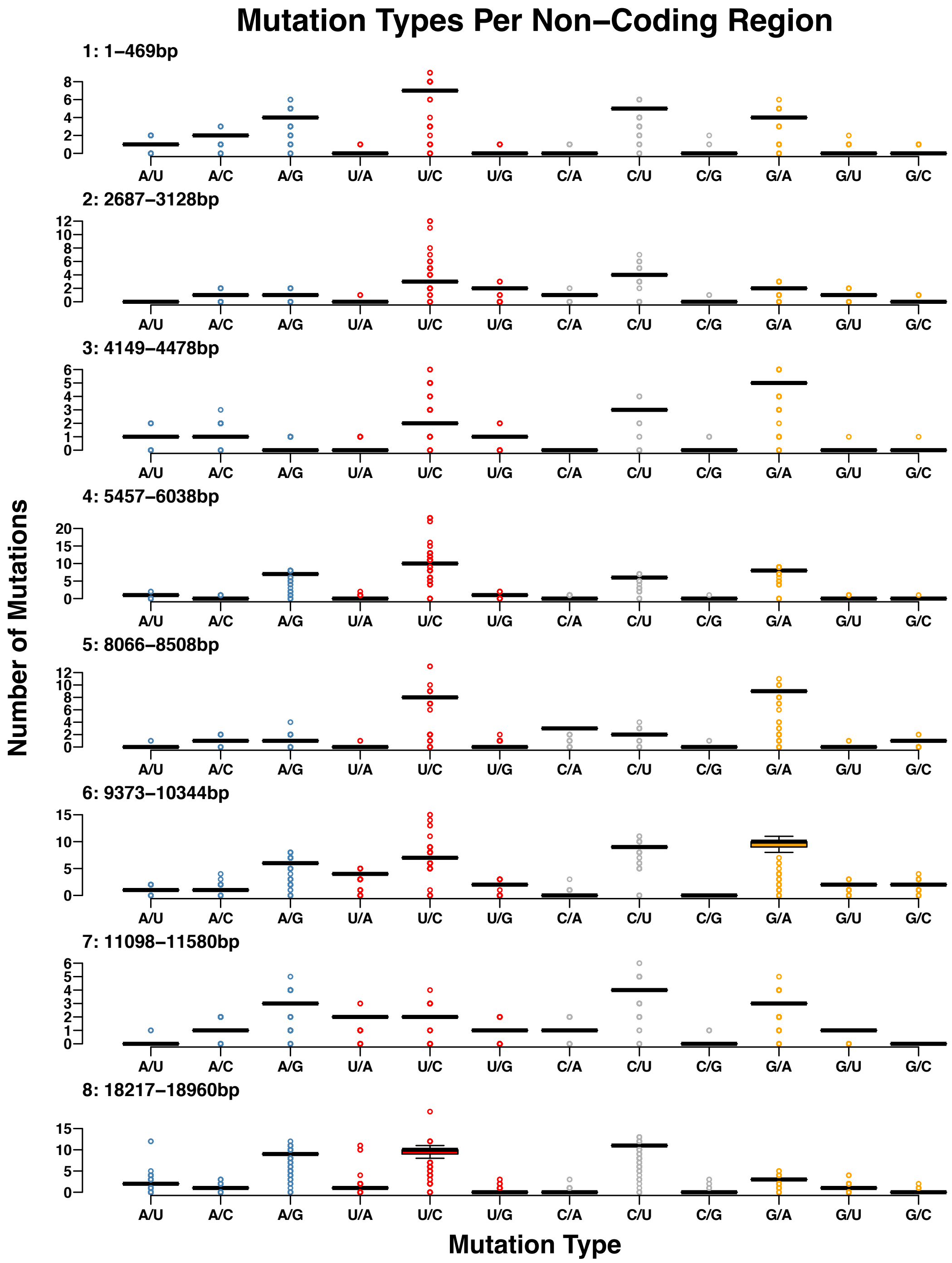


Figure 4: Observed mutation types for each sample compared to the reference genome, for each of the non-coding regions of the genome. Colours indicated the starting base of the mutation – blue A, red U, grey C, and orange G.

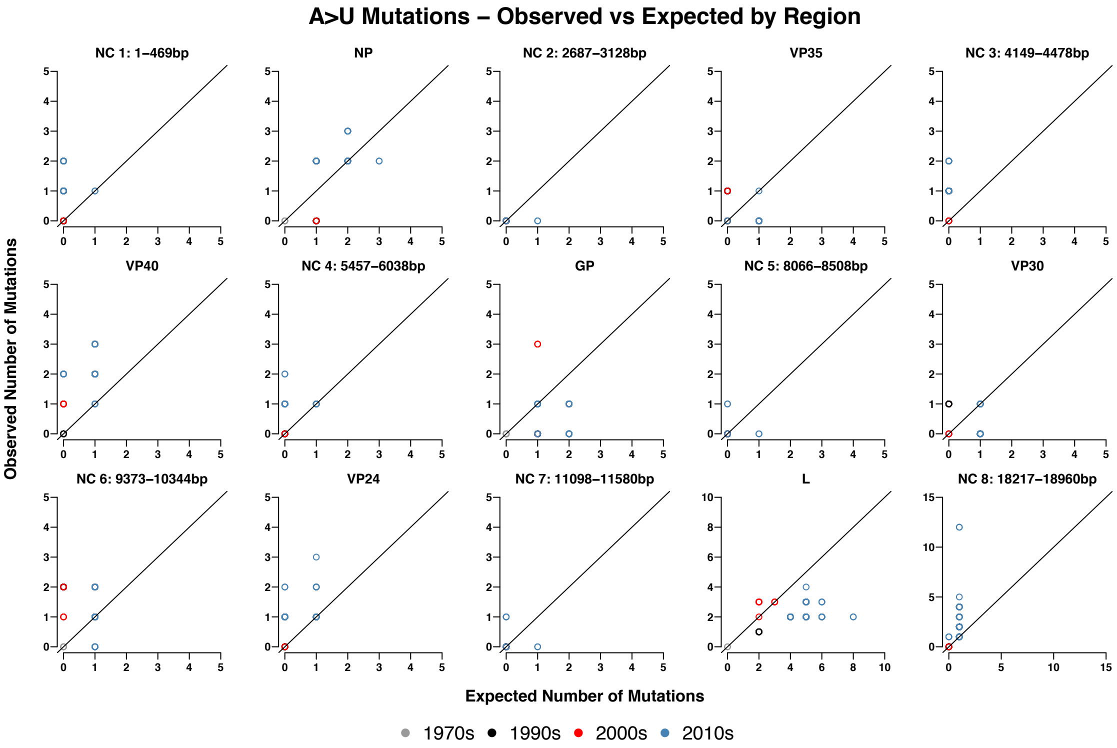


Figure 5: Observed vs Expected numbers of A>U mutations, for each region of the genome, compared to the 1976 reference genome. The colour of the point indicates the decade that the sample is from, see legend. Note, x- and y-axes differ between the subplots.

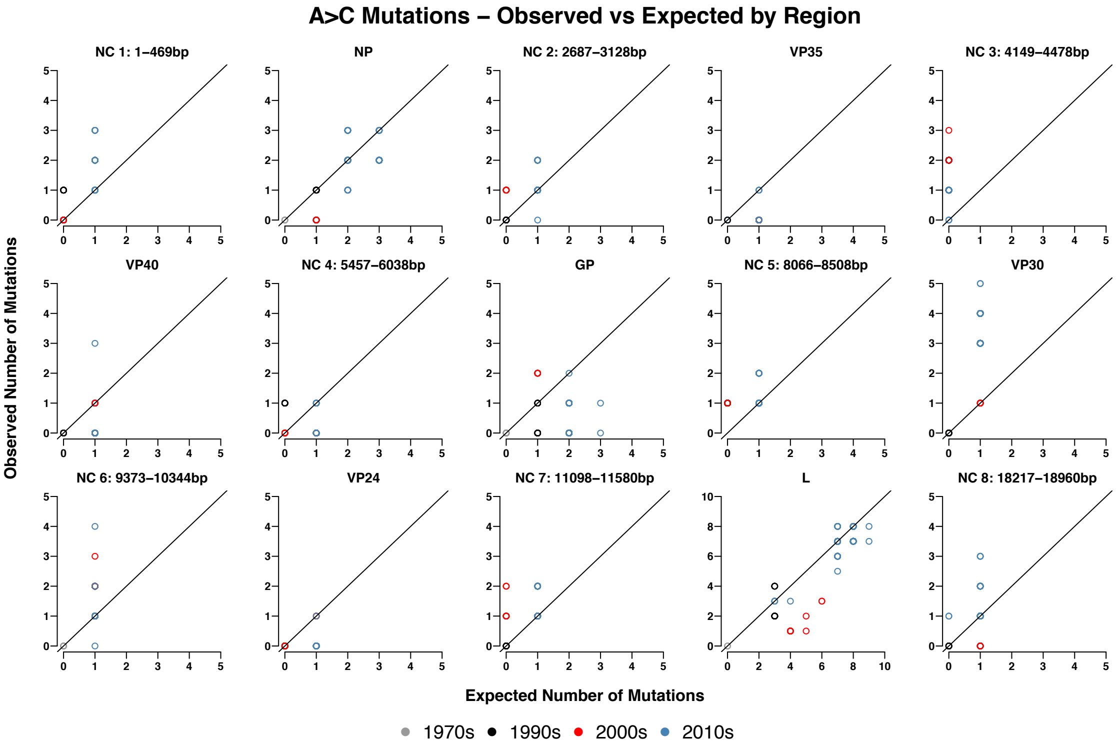


Figure 6: Observed vs Expected numbers of A>C mutations, for each region of the genome, compared to the 1976 reference genome. The colour of the point indicates the decade that the sample is from, see legend. Note, x- and y-axes differ between the subplots.

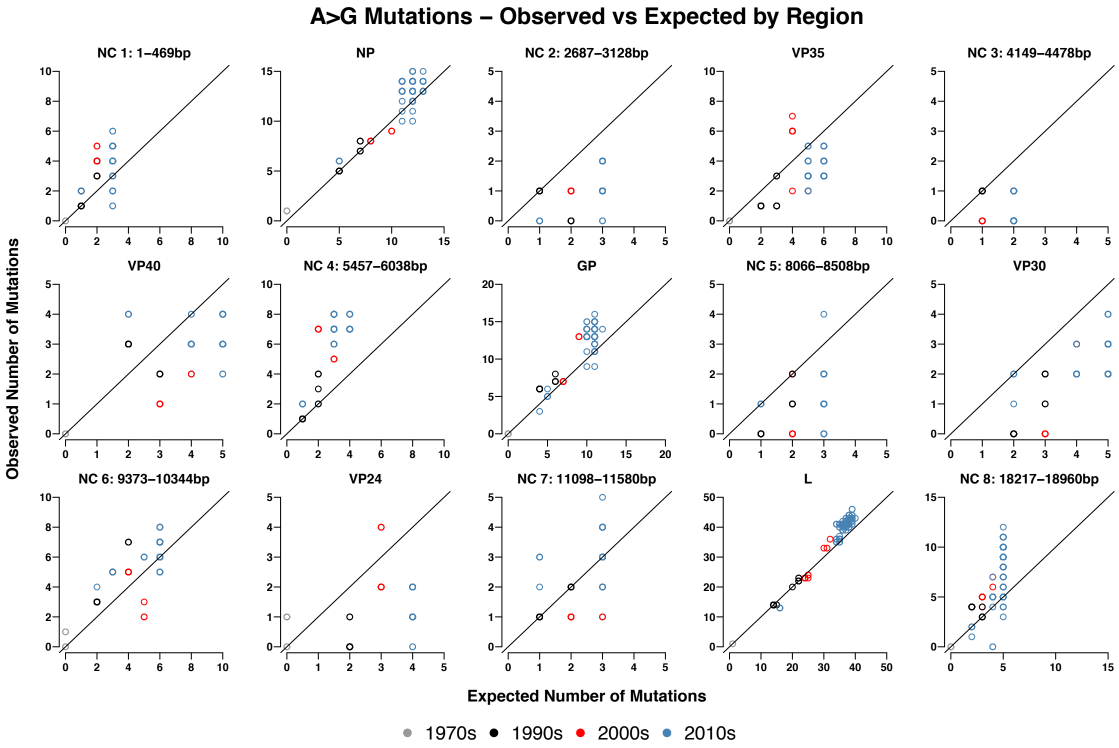


Figure 7: Observed vs Expected numbers of A>G mutations, for each region of the genome, compared to the 1976 reference genome. The colour of the point indicates the decade that the sample is from, see legend. Note, x- and y-axes differ between the subplots.

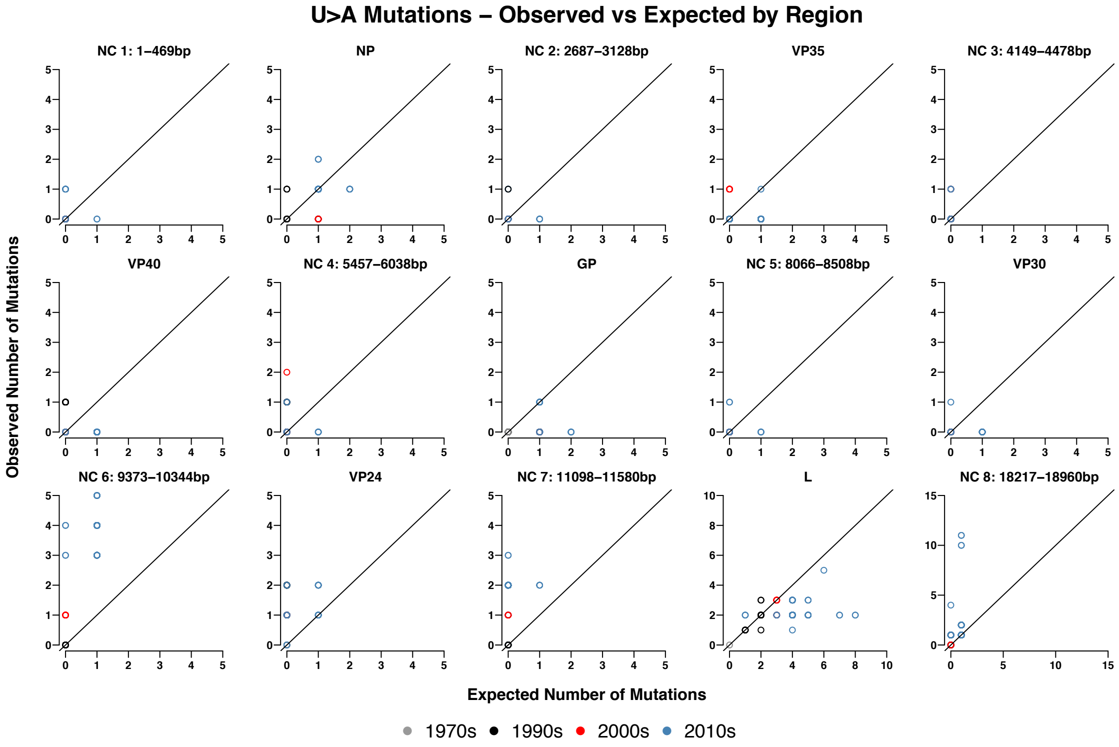


Figure 8: Observed vs Expected numbers of U>A mutations, for each region of the genome, compared to the 1976 reference genome. The colour of the point indicates the decade that the sample is from, see legend. Note, x- and y-axes differ between the subplots.

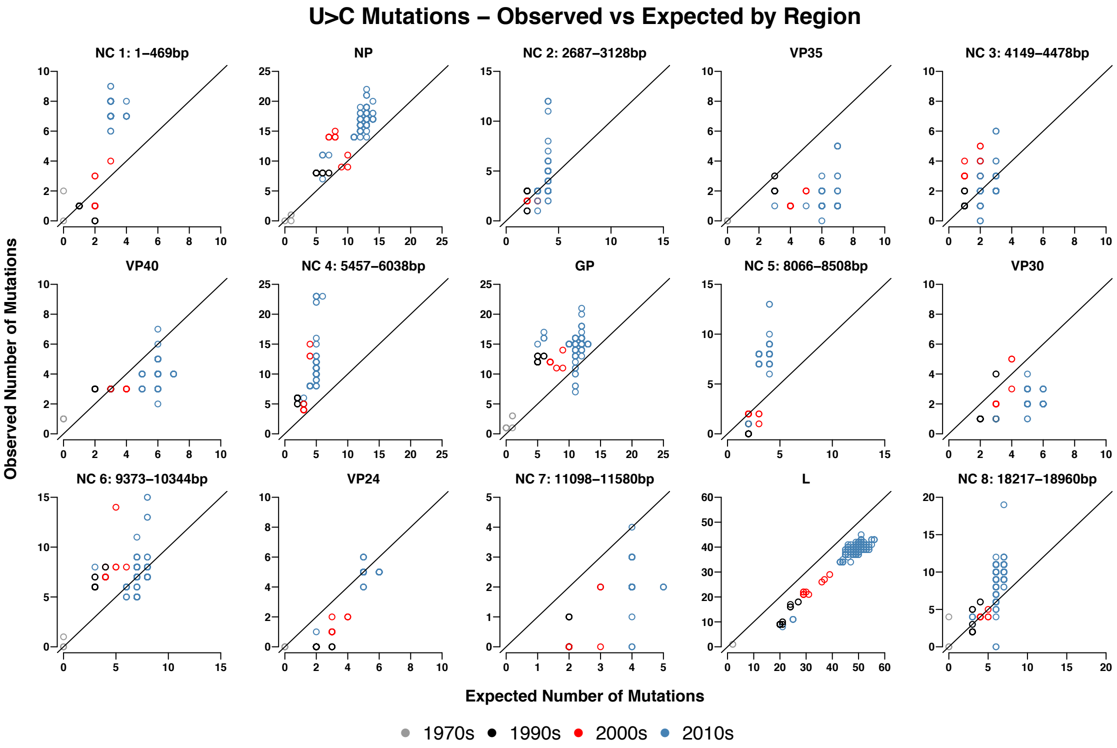


Figure 9: Observed vs Expected numbers of U>C mutations, for each region of the genome, compared to the 1976 reference genome. The colour of the point indicates the decade that the sample is from, see legend. Note, x- and y-axes differ between the subplots.

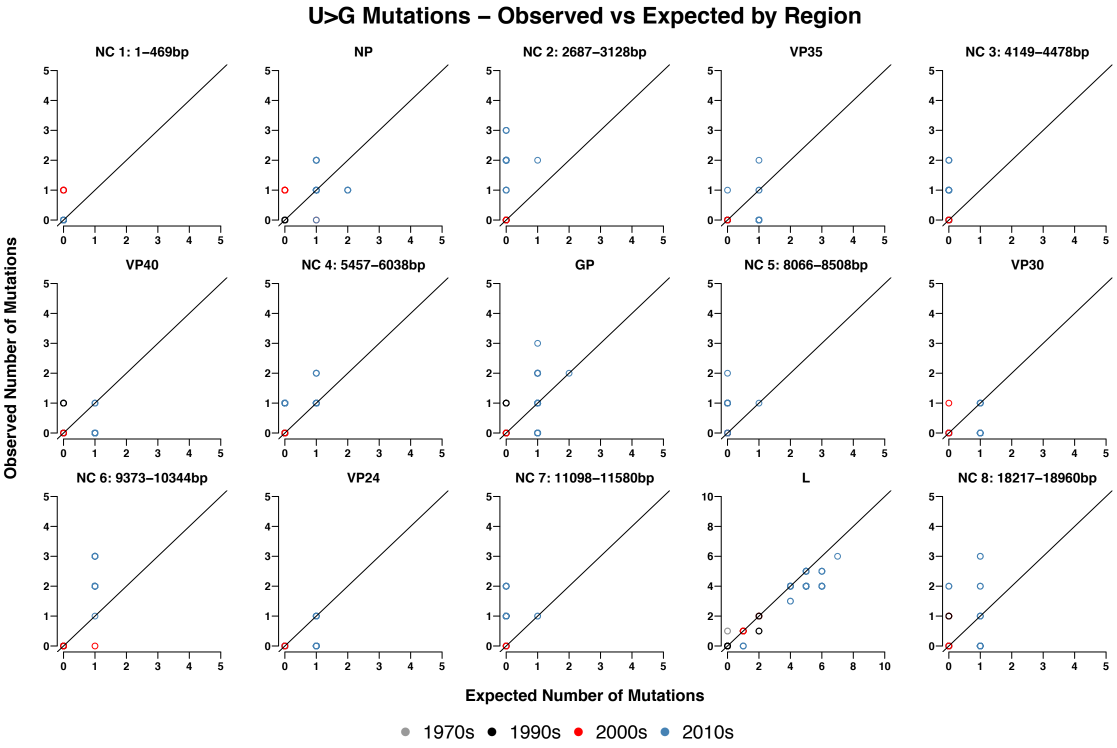


Figure 10: Observed vs Expected numbers of U>G mutations, for each region of the genome, compared to the 1976 reference genome. The colour of the point indicates the decade that the sample is from, see legend. Note, x- and y-axes differ between the subplots.

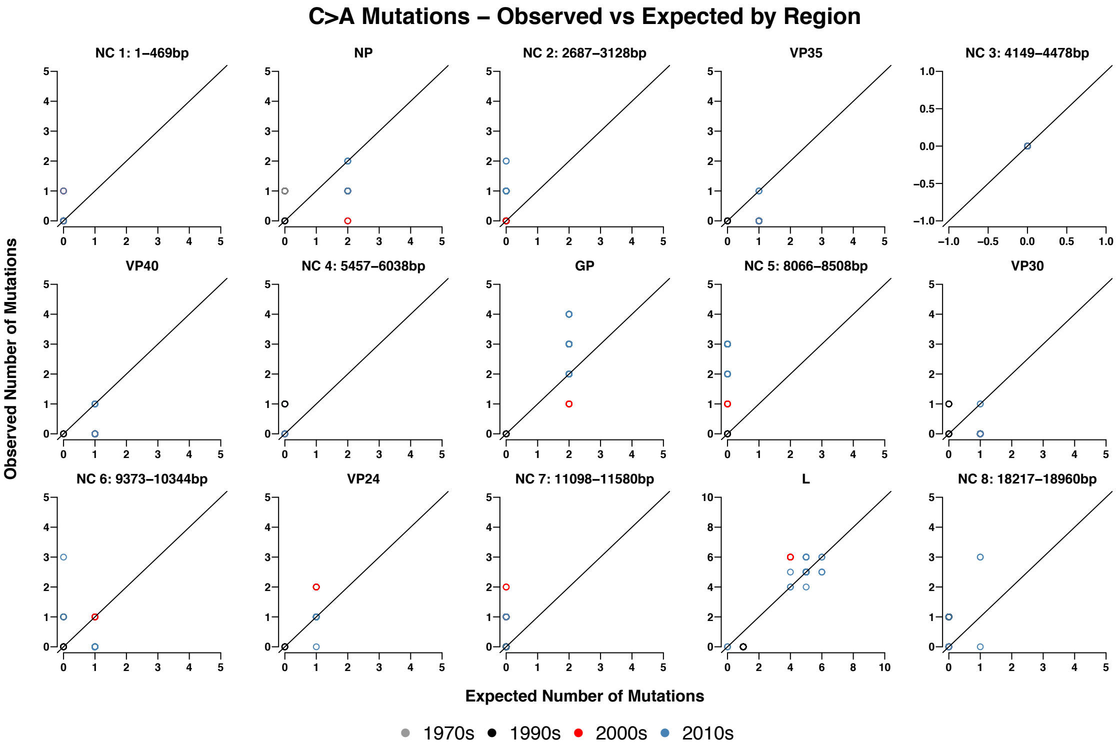


Figure 11: Observed vs Expected numbers of C>A mutations, for each region of the genome, compared to the 1976 reference genome. The colour of the point indicates the decade that the sample is from, see legend. Note, x- and y-axes differ between subplots.

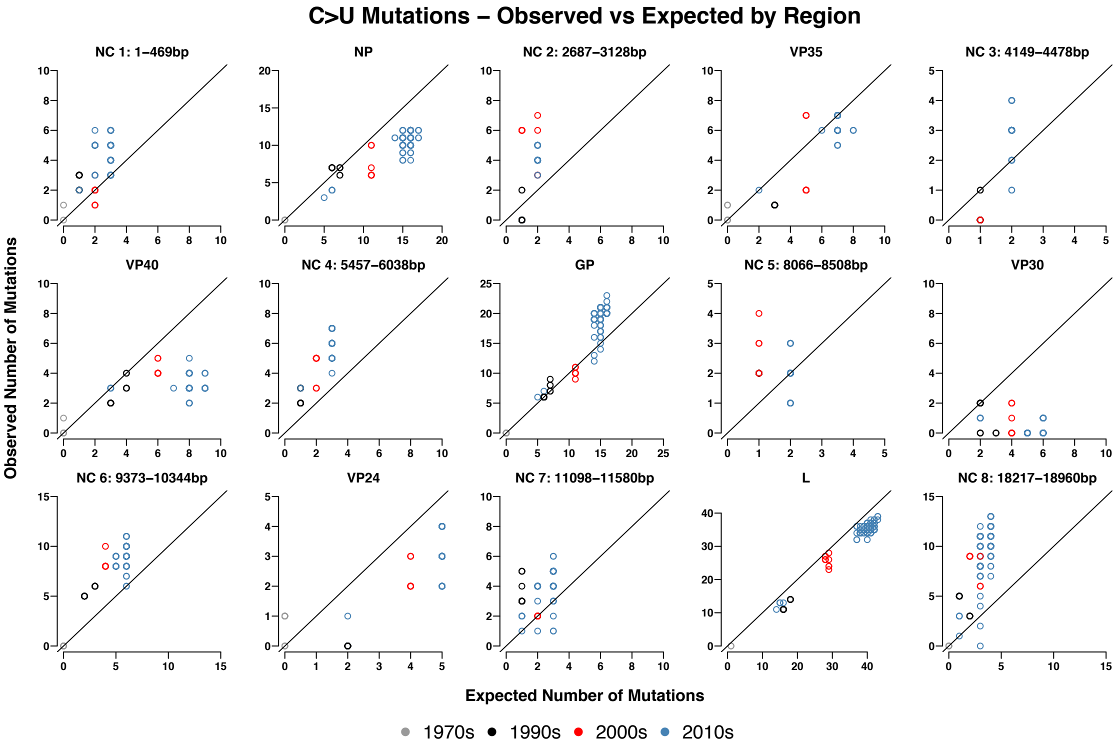


Figure 12: Observed vs Expected numbers of C>U mutations, for each region of the genome, compared to the 1976 reference genome. The colour of the point indicates the decade that the sample is from, see legend. Note, x- and y-axes differ between the subplots.

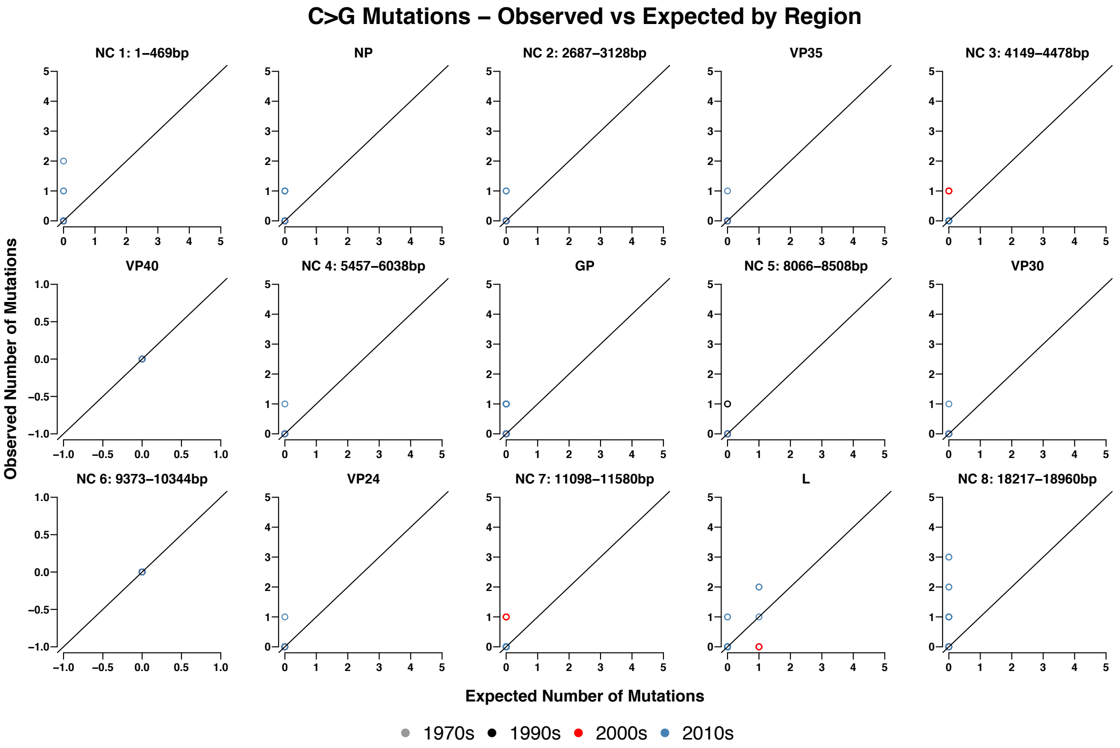


Figure 13: Observed vs Expected numbers of C>G mutations, for each region of the genome, compared to the 1976 reference genome. The colour of the point indicates the decade that the sample is from, see legend. Note, x- and y-axes differ between the subplots.

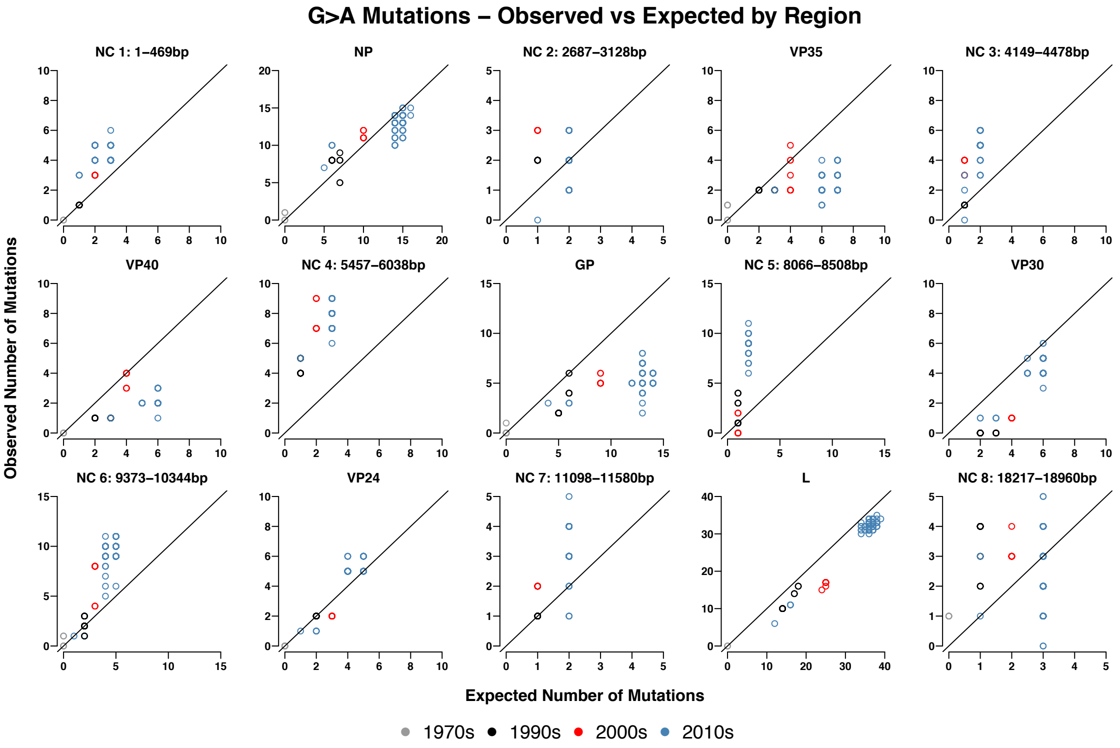


Figure 14: Observed vs Expected numbers of G>A mutations, for each region of the genome, compared to the 1976 reference genome. The colour of the point indicates the decade that the sample is from, see legend. Note, x- and y-axes differ between the subplots.

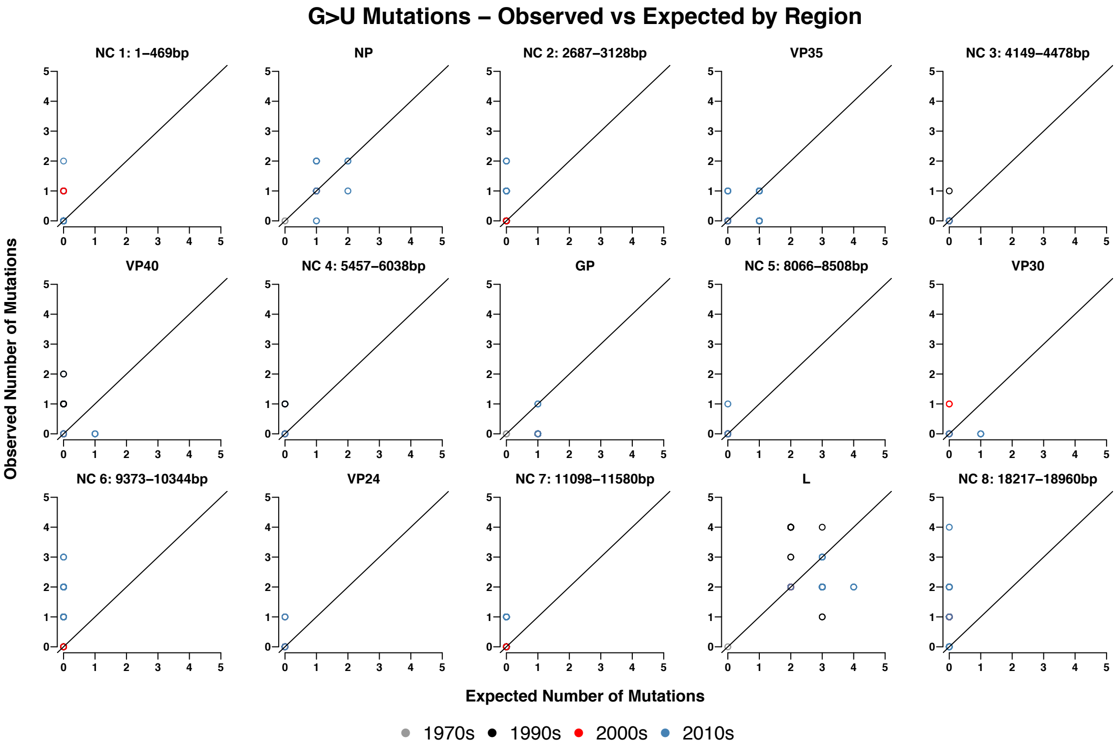


Figure 15: Observed vs Expected numbers of G>U mutations, for each region of the genome, compared to the 1976 reference genome. The colour of the point indicates the decade that the sample is from, see legend. Note, x- and y-axes differ between the subplots.

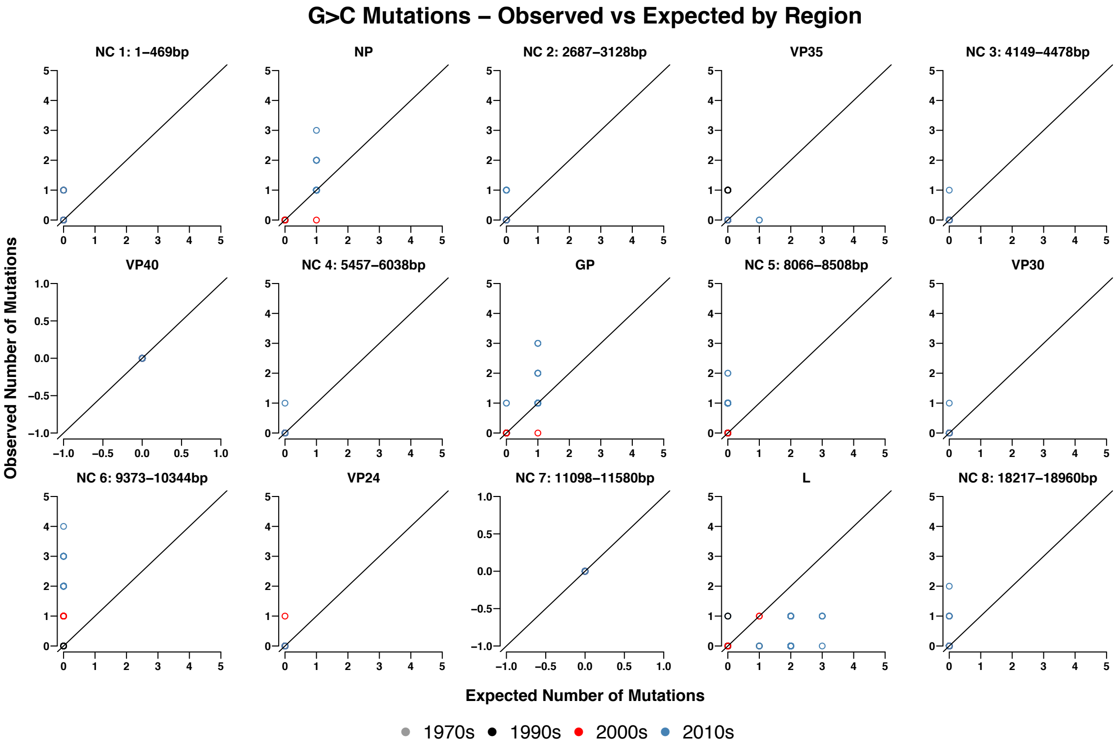


Figure 16: Observed vs Expected numbers of G>C mutations, for each region of the genome, compared to the 1976 reference genome. The colour of the point indicates the decade that the sample is from, see legend. Note, x- and y-axes differ between the subplots.

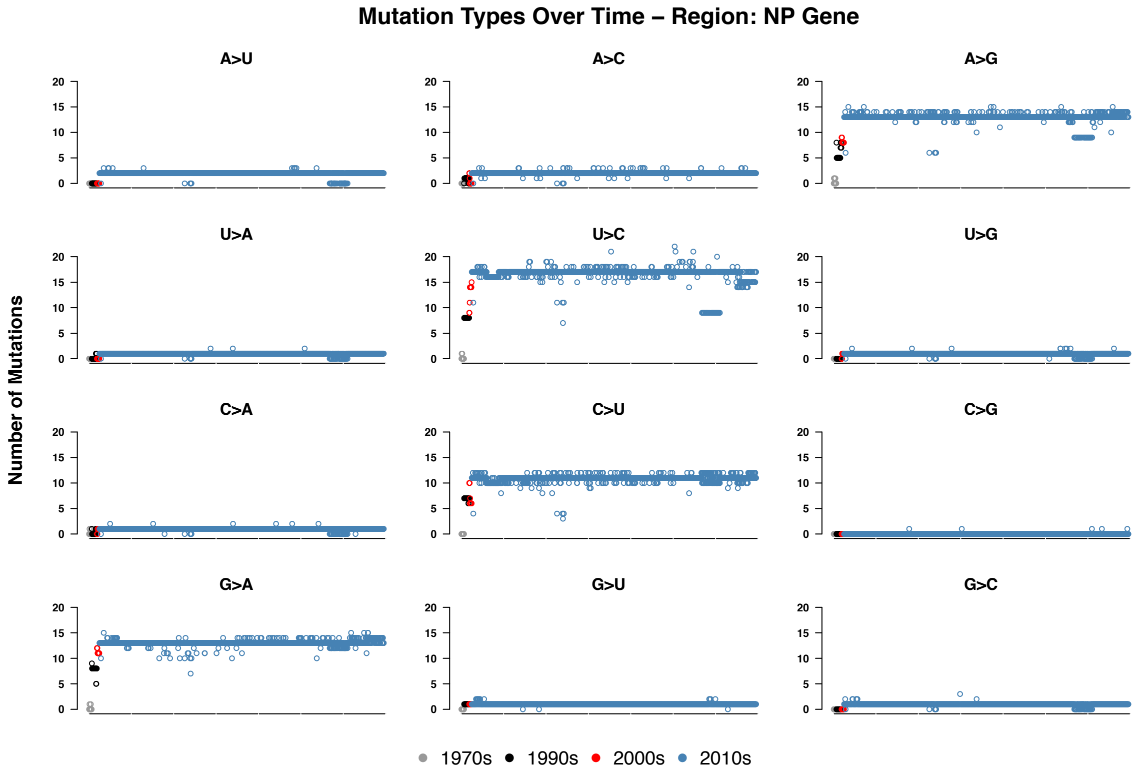


Figure 17: Observed mutation types over time for each sample compared to the reference genome, for the gene NP. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

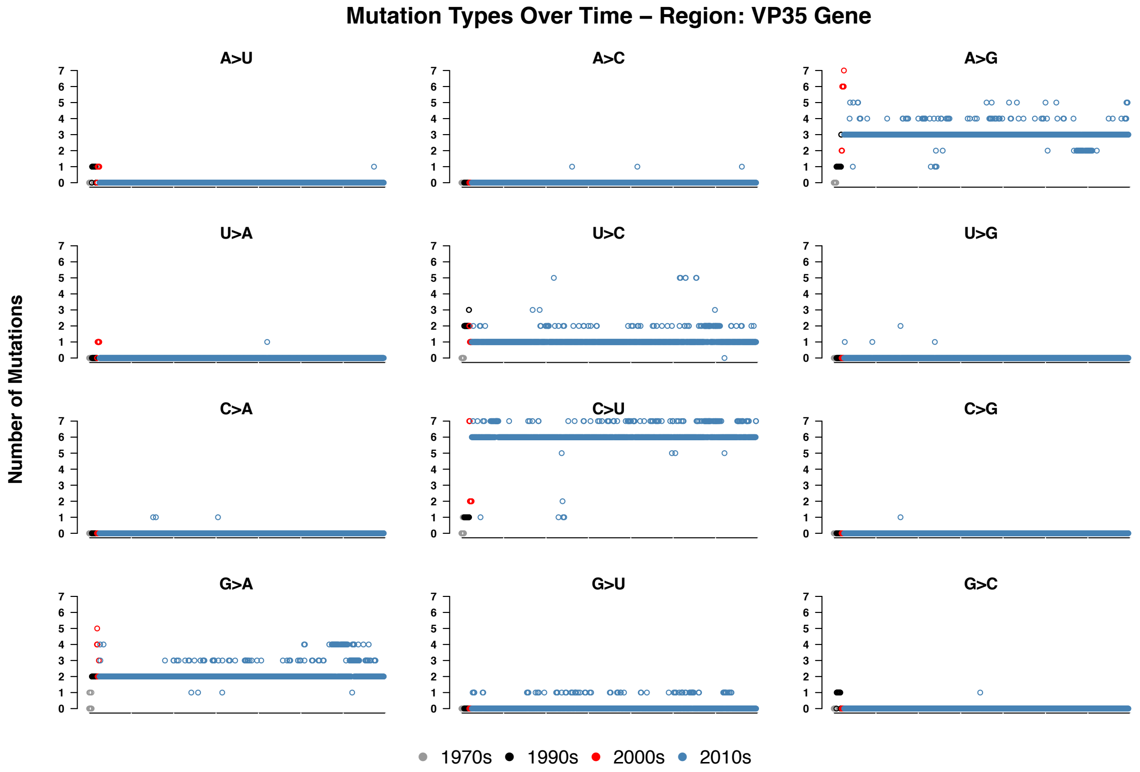


Figure 18: Observed mutation types over time for each sample compared to the reference genome, for the gene VP35. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

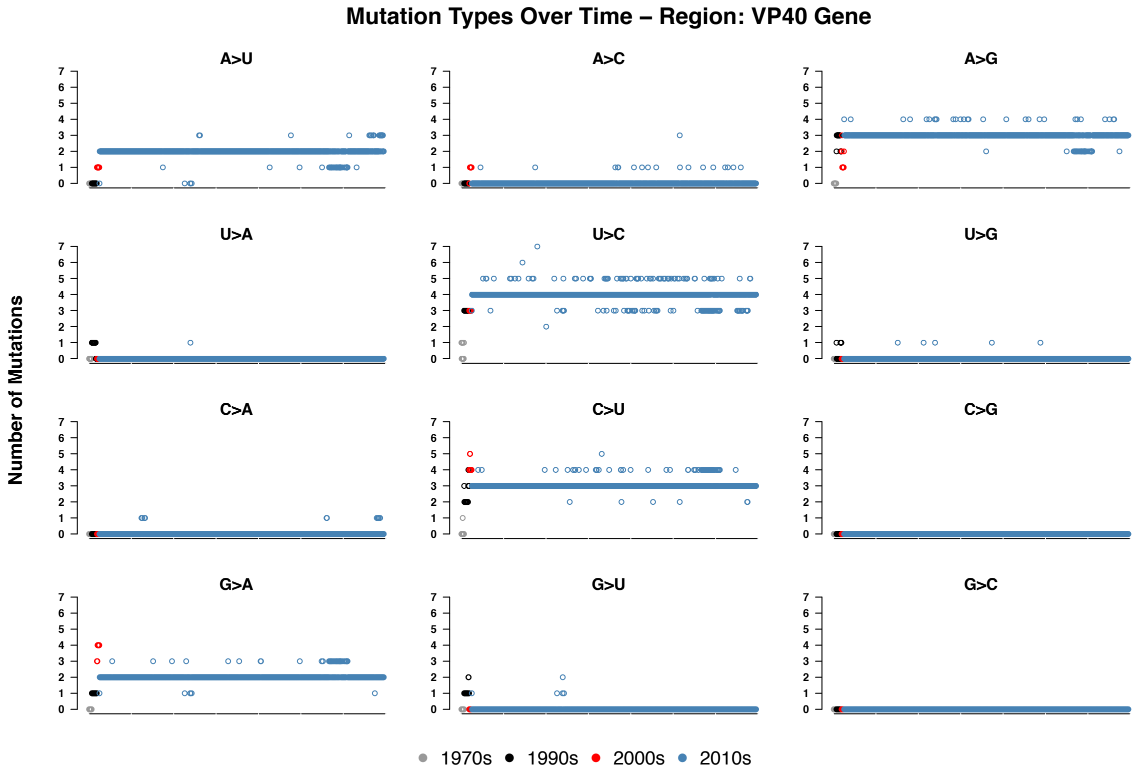


Figure 19: Observed mutation types over time for each sample compared to the reference genome, for the gene VP40. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

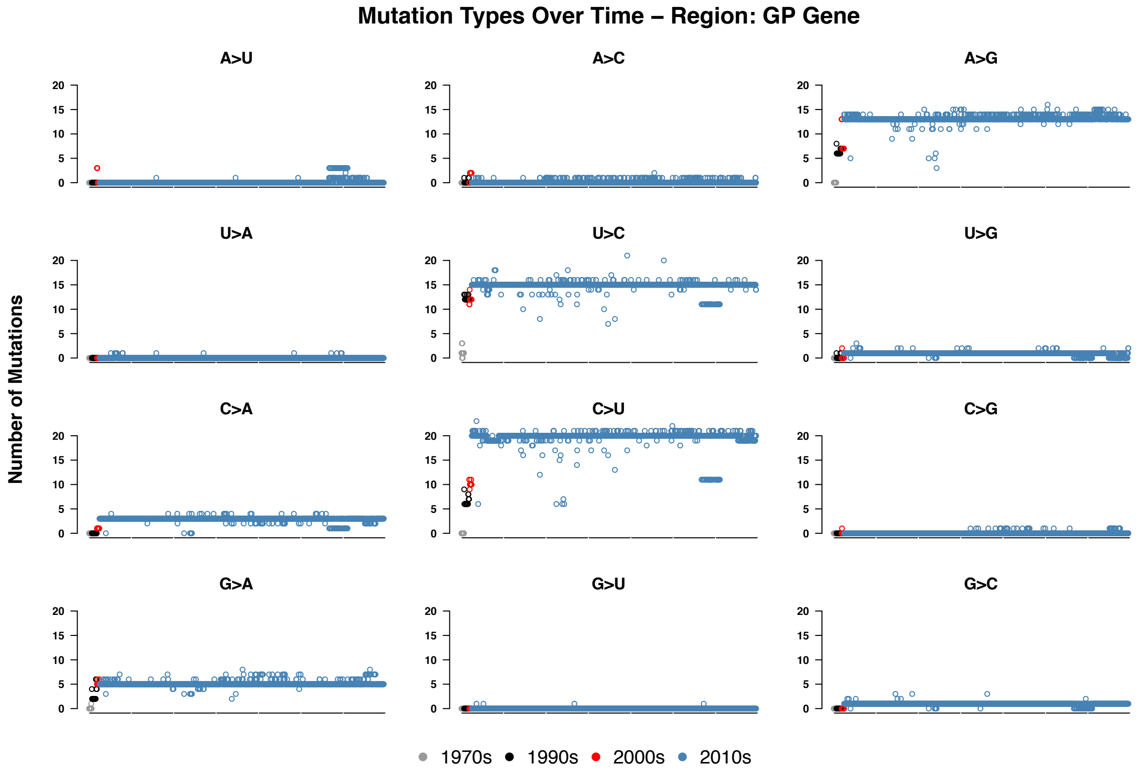


Figure 20: Observed mutation types over time for each sample compared to the reference genome, for the gene GP. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

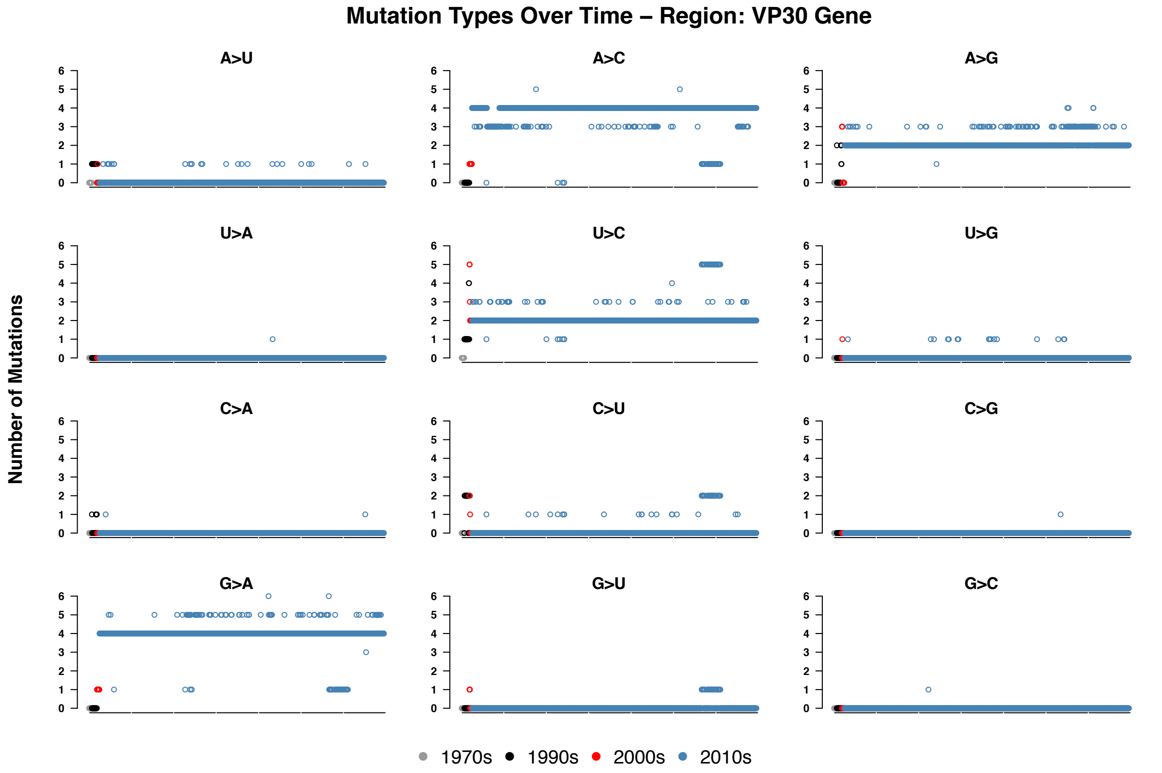


Figure 21: Observed mutation types over time for each sample compared to the reference genome, for the gene VP30. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

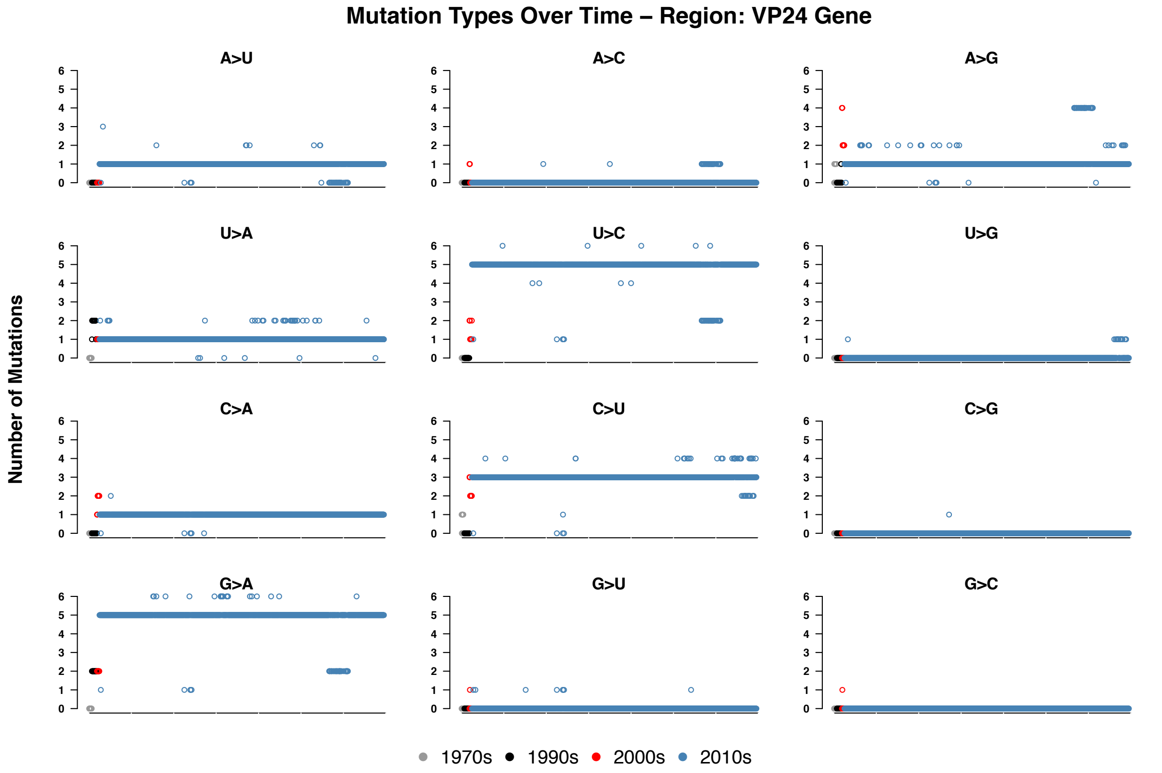


Figure 22: Observed mutation types over time for each sample compared to the reference genome, for the gene VP24. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

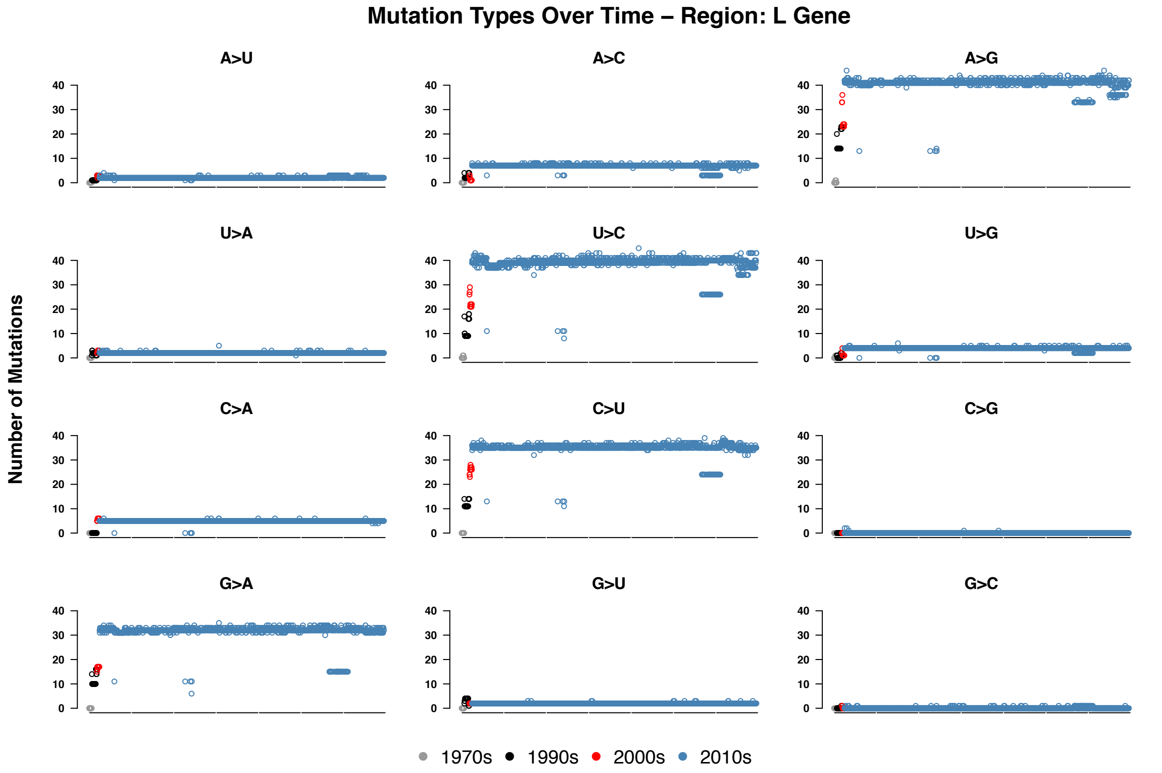


Figure 23: Observed mutation types over time for each sample compared to the reference genome, for the gene L. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

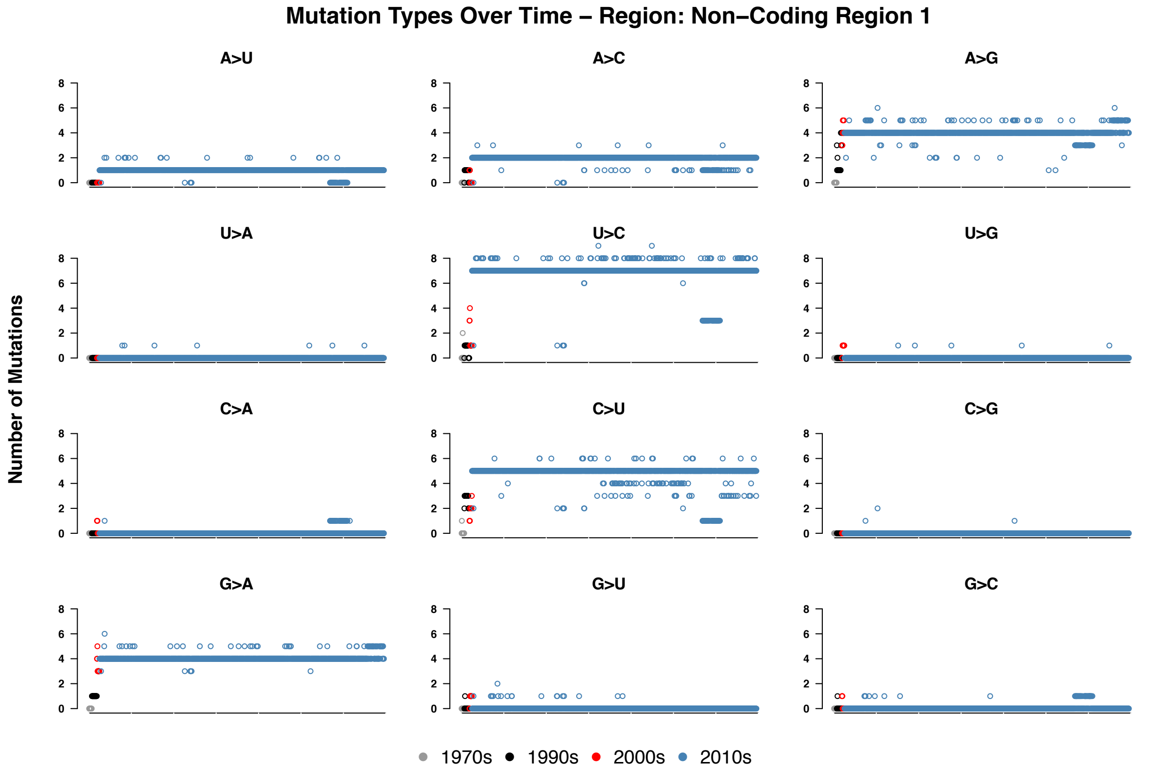


Figure 24: Observed mutation types over time for each sample compared to the reference genome, for non-coding region 1. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

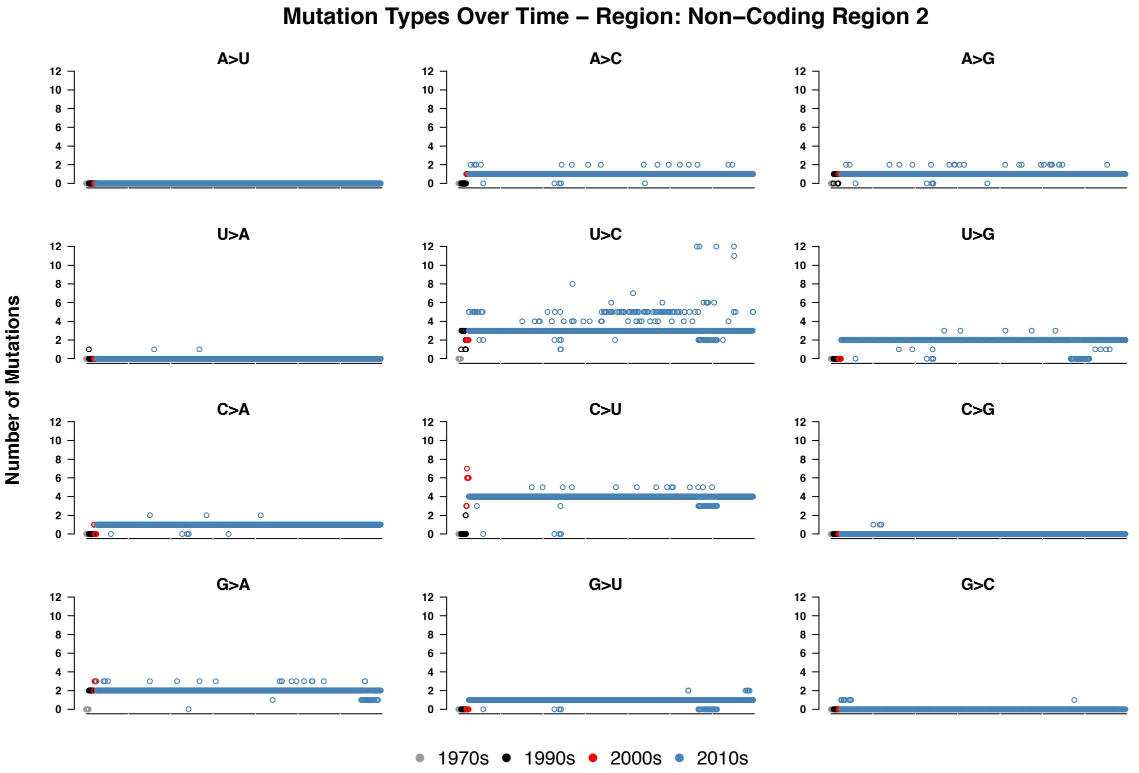


Figure 25: Observed mutation types over time for each sample compared to the reference genome, for non-coding region 2. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

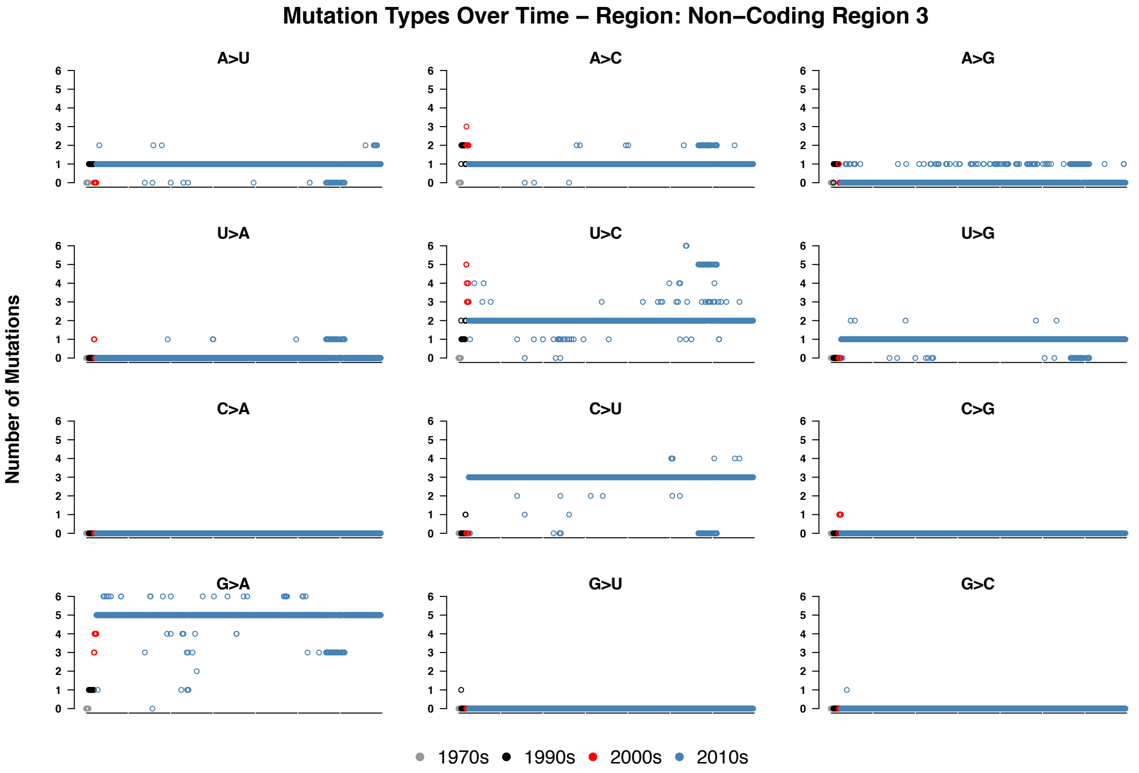


Figure 26: Observed mutation types over time for each sample compared to the reference genome, for non-coding region 3. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

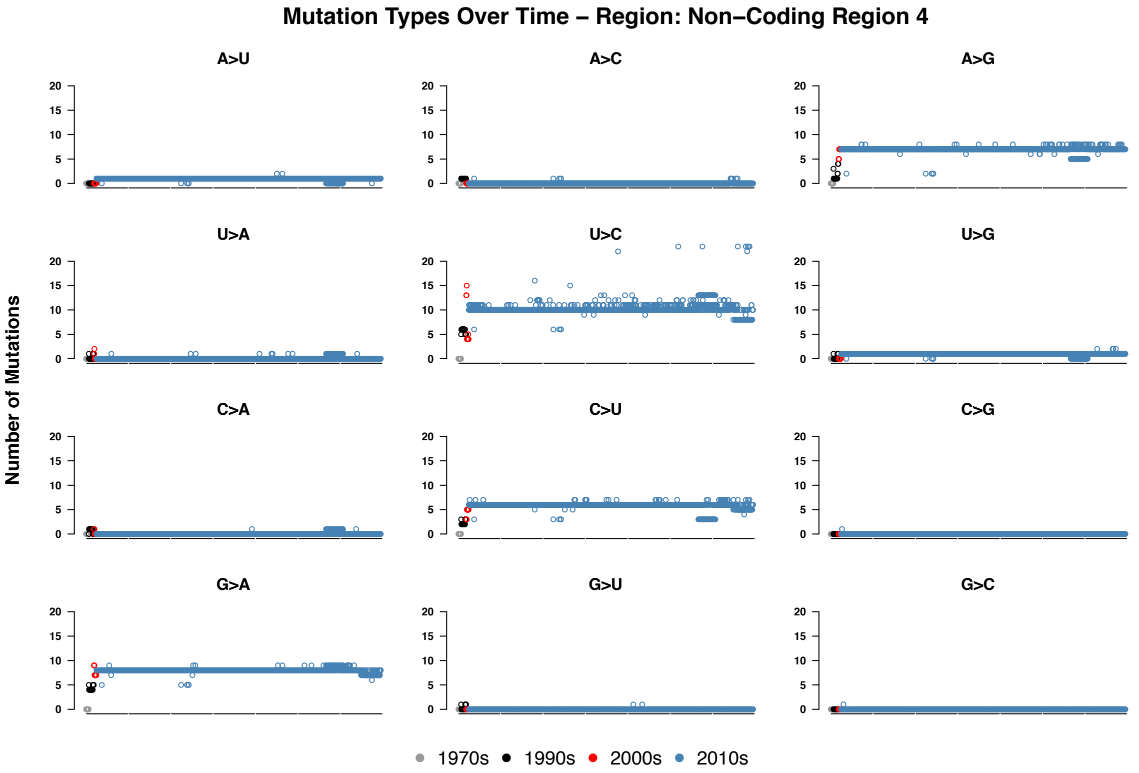


Figure 27: Observed mutation types over time for each sample compared to the reference genome, for non-coding region 4. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

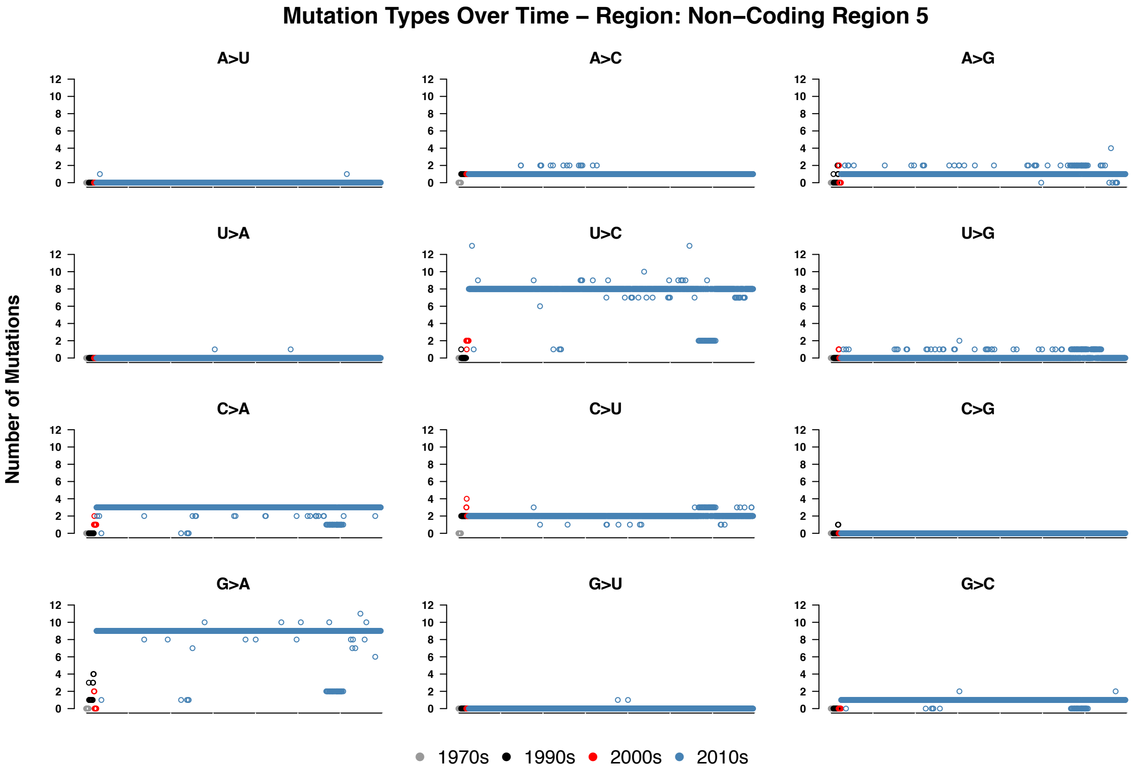


Figure 28: Observed mutation types over time for each sample compared to the reference genome, for non-coding region 5. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

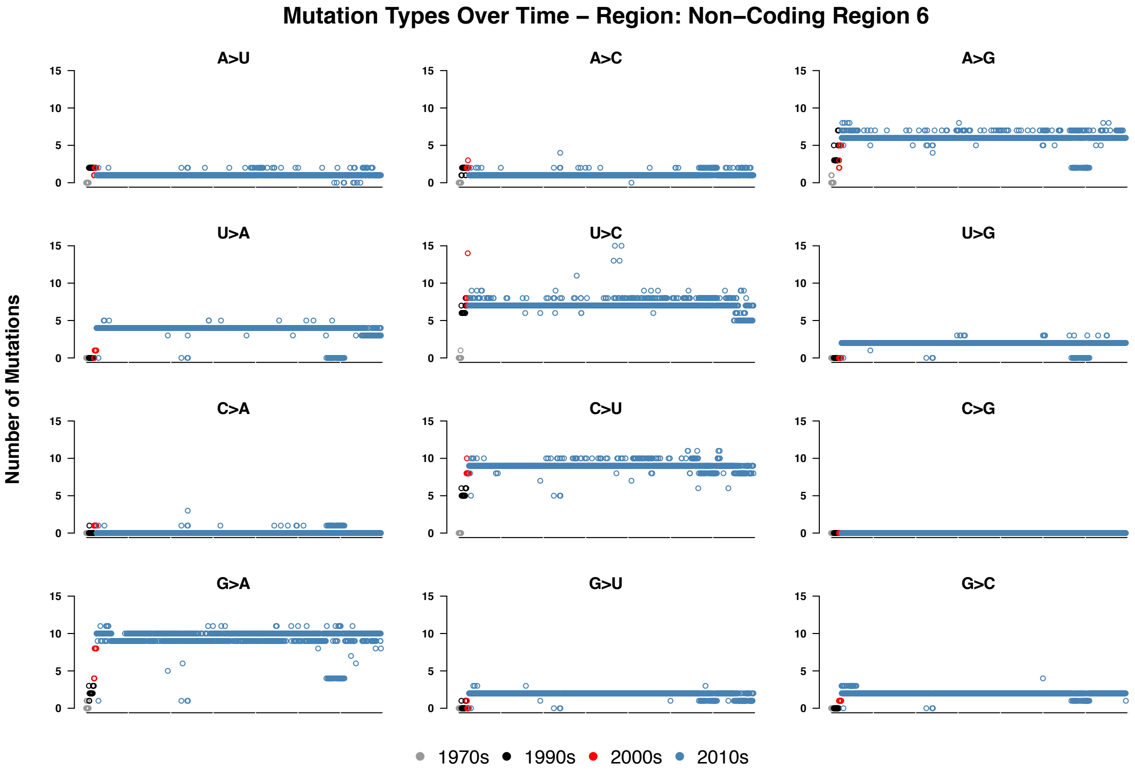


Figure 29: Observed mutation types over time for each sample compared to the reference genome, for non-coding region 6. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

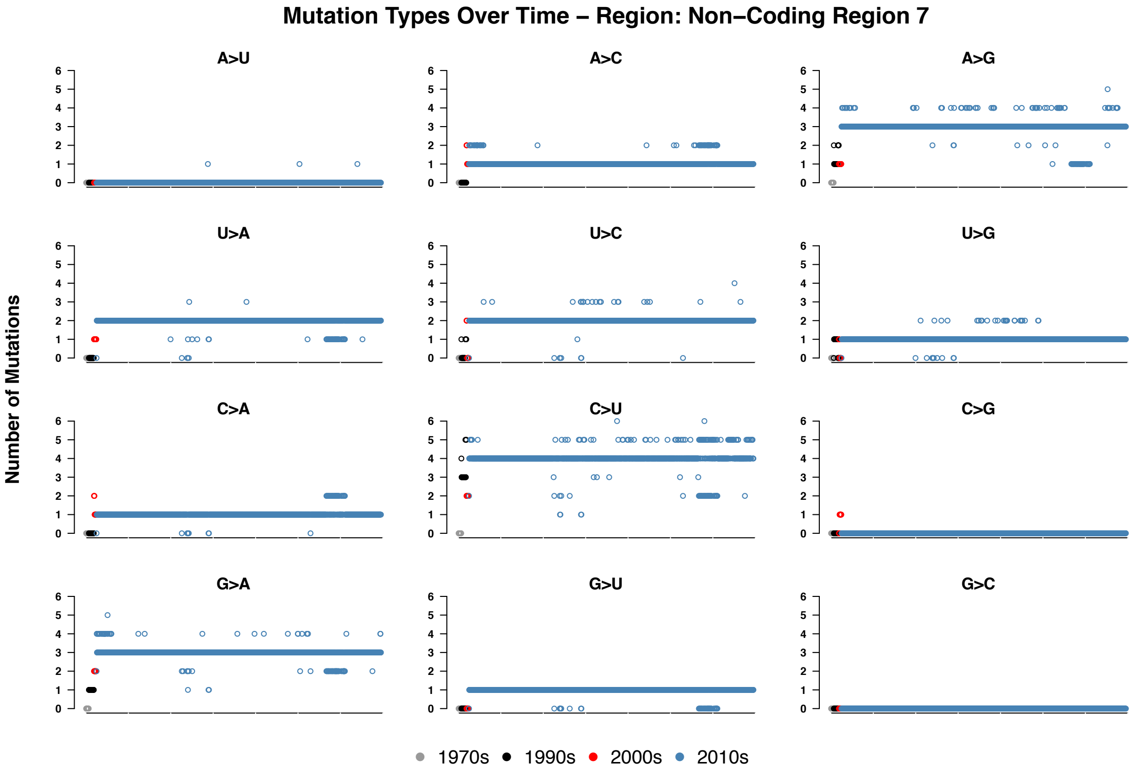


Figure 30: Observed mutation types over time for each sample compared to the reference genome, for non-coding region 7. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

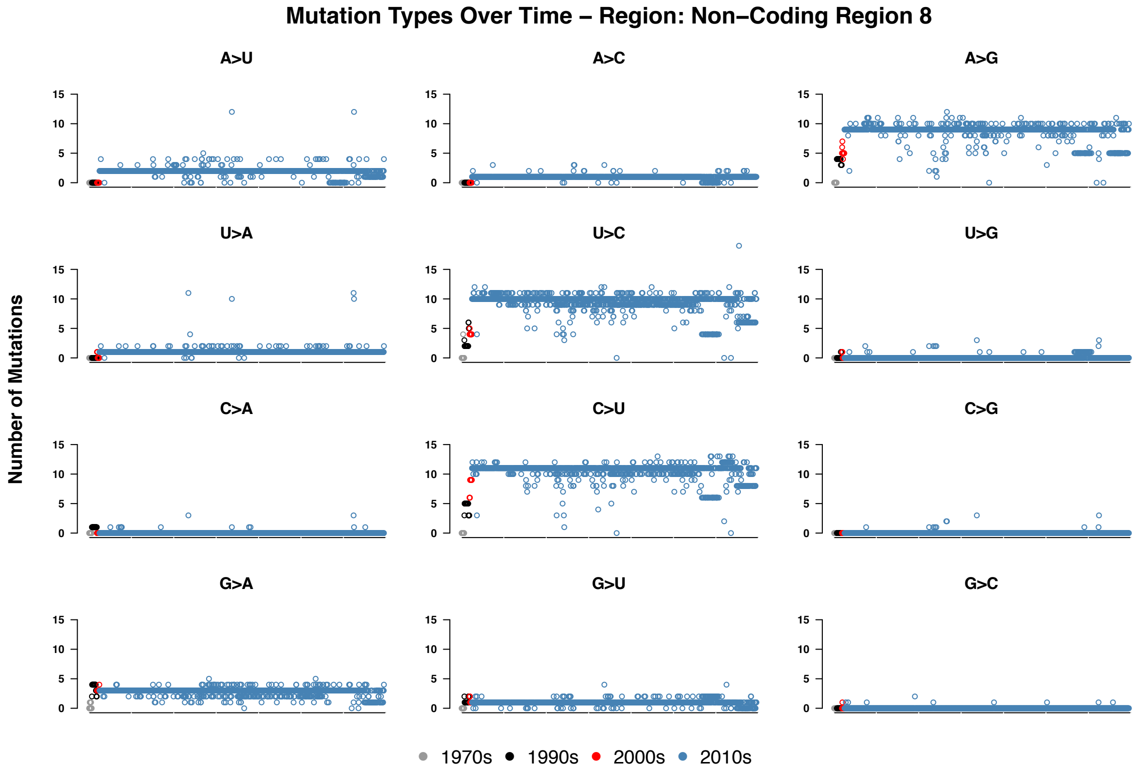


Figure 31: Observed mutation types over time for each sample compared to the reference genome, for non-coding region 8. Samples are ordered left to right from oldest to newest. The colour of the point indicates the decade that the sample is from, see legend.

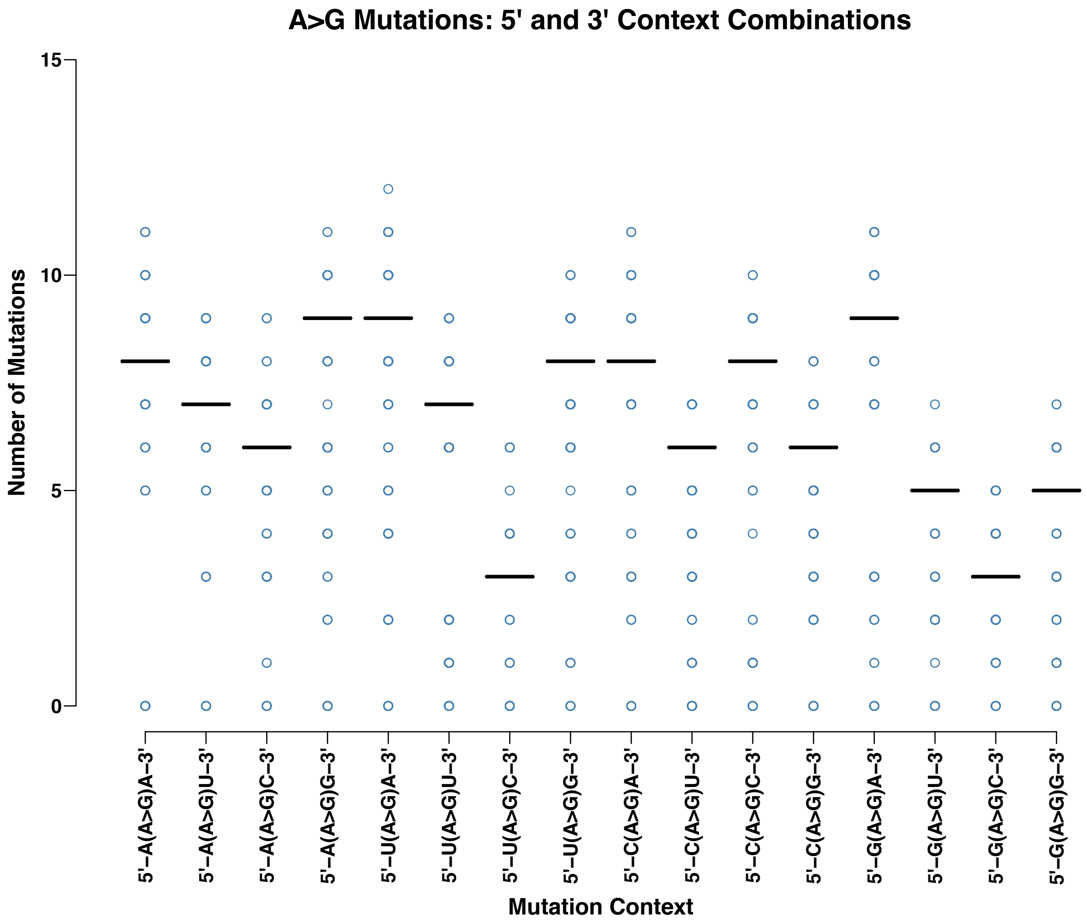


Figure 32: Combined mutation contexts of A>G mutations for all *Ebola virus* samples compared to the 1976 reference genome.

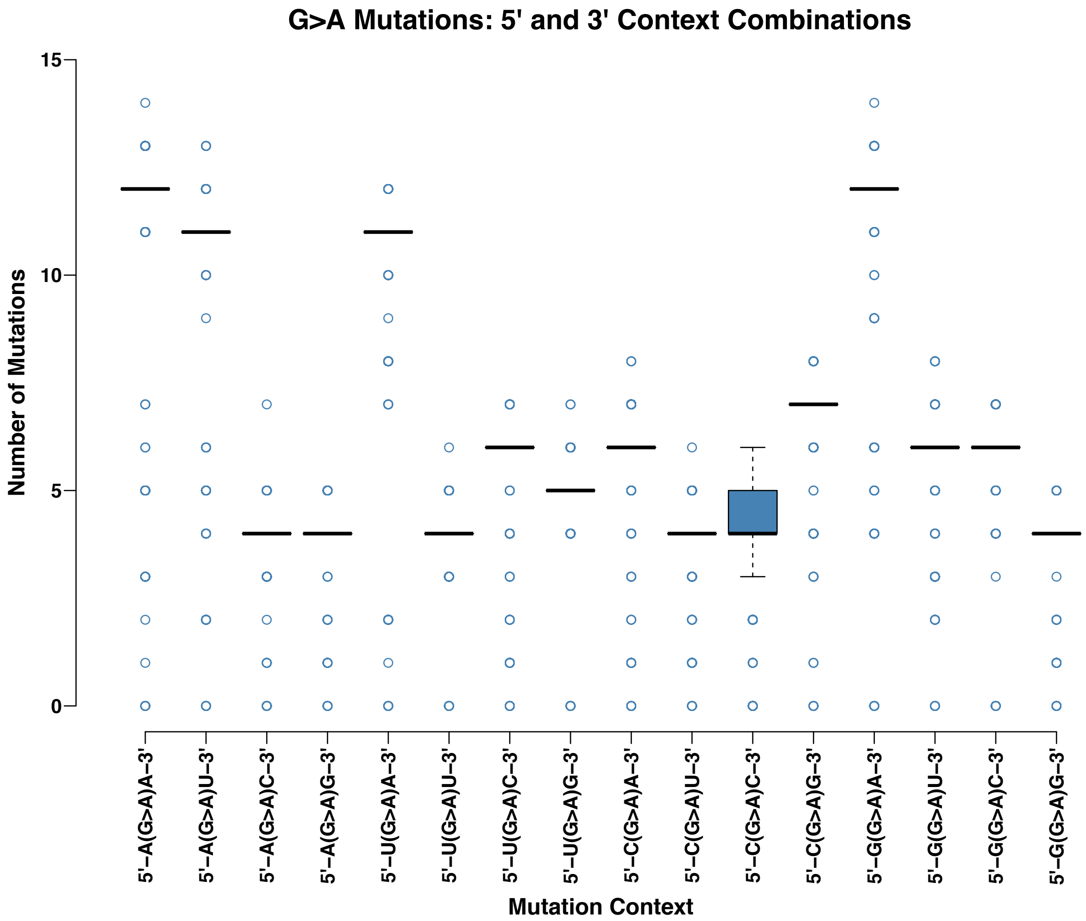


Figure 33: Combined mutation contexts of G>A mutations for all *Ebola virus* samples compared to the 1976 reference genome.

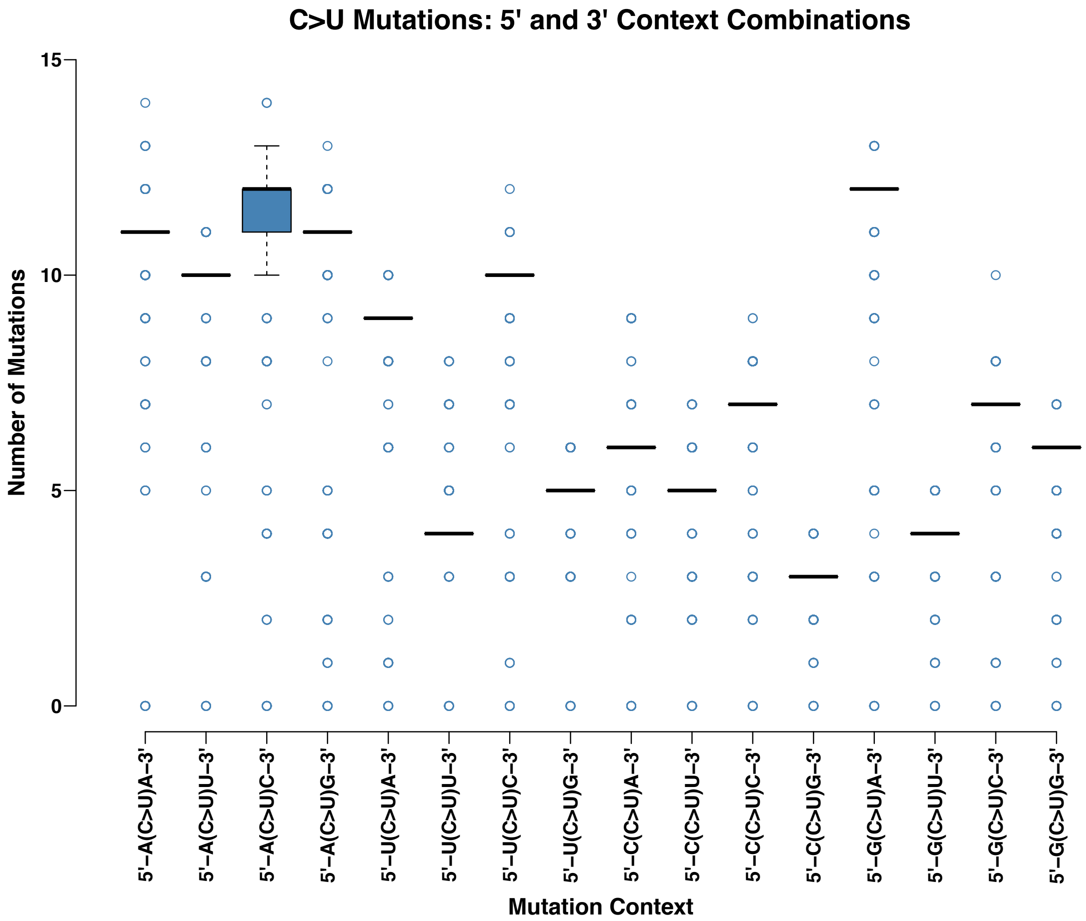


Figure 34: Combined mutation contexts of C>U mutations for all *Ebola virus* samples compared to the 1976 reference genome.

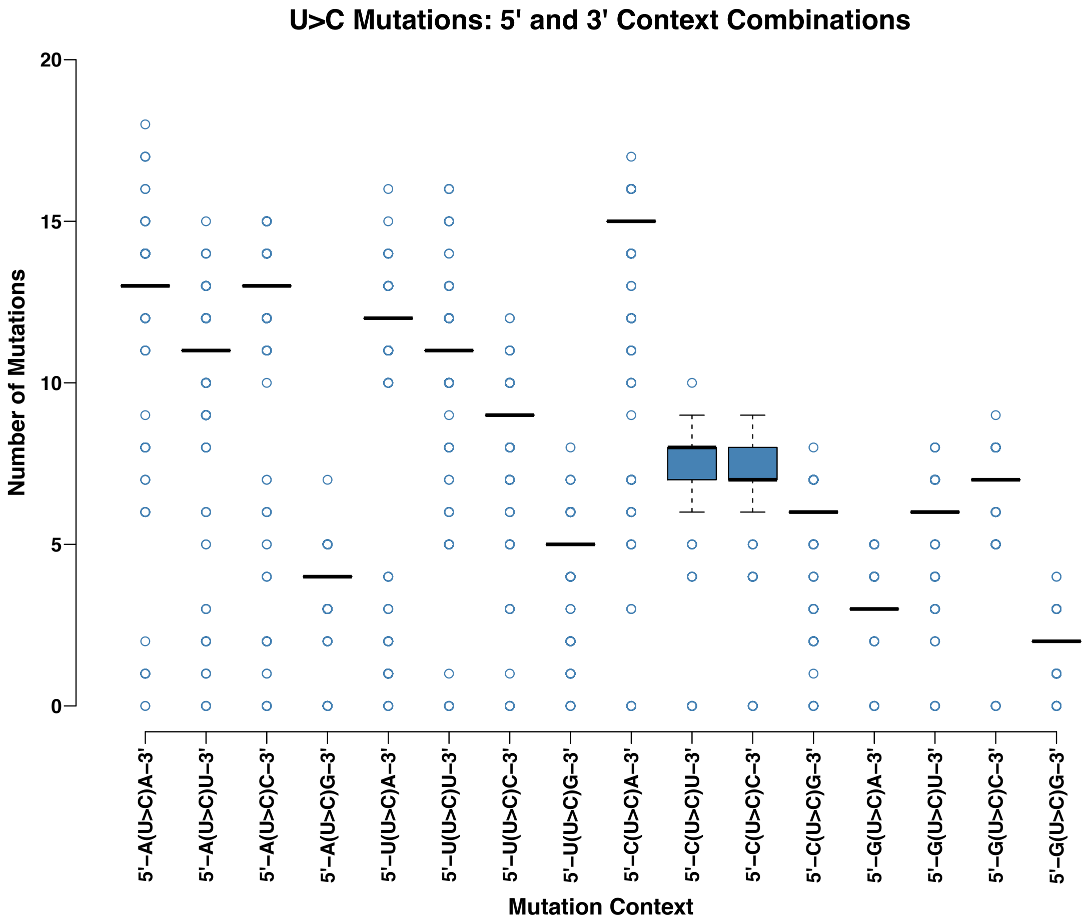


Figure 35: Combined mutation contexts of U>C mutations for all *Ebola virus* samples compared to the 1976 reference genome.

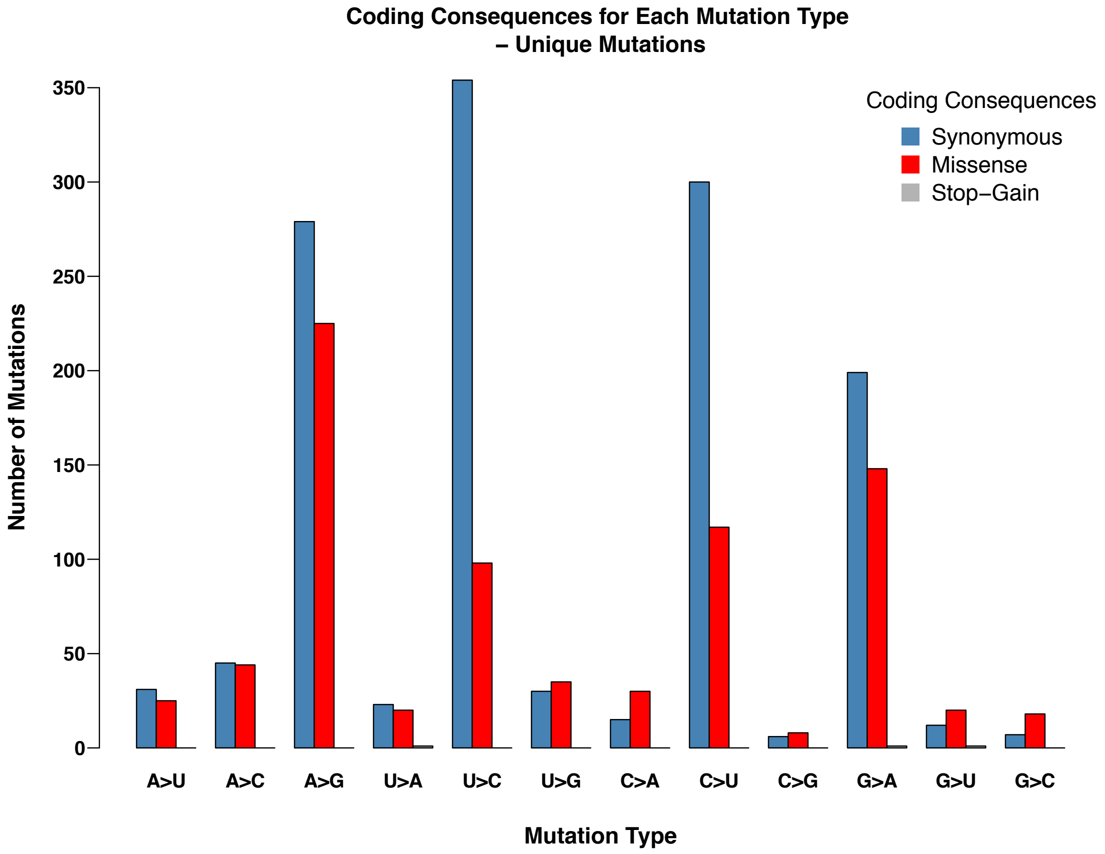


Figure 36: Protein coding consequences of each mutation type for the unique set of mutations across all samples compared to the 1976 reference genome.

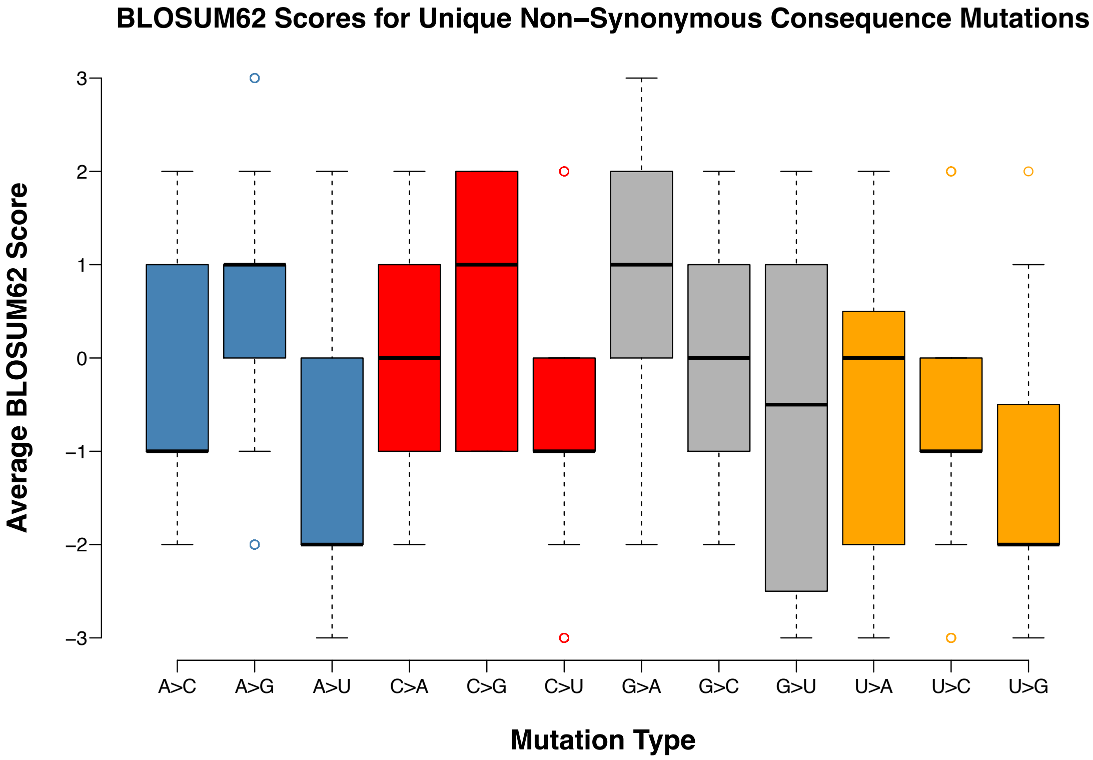


Figure 37: Average BLOSUM62 matrix scores for the unique set of mutations across all samples compared to the 1976 reference genome, separated by mutation type.

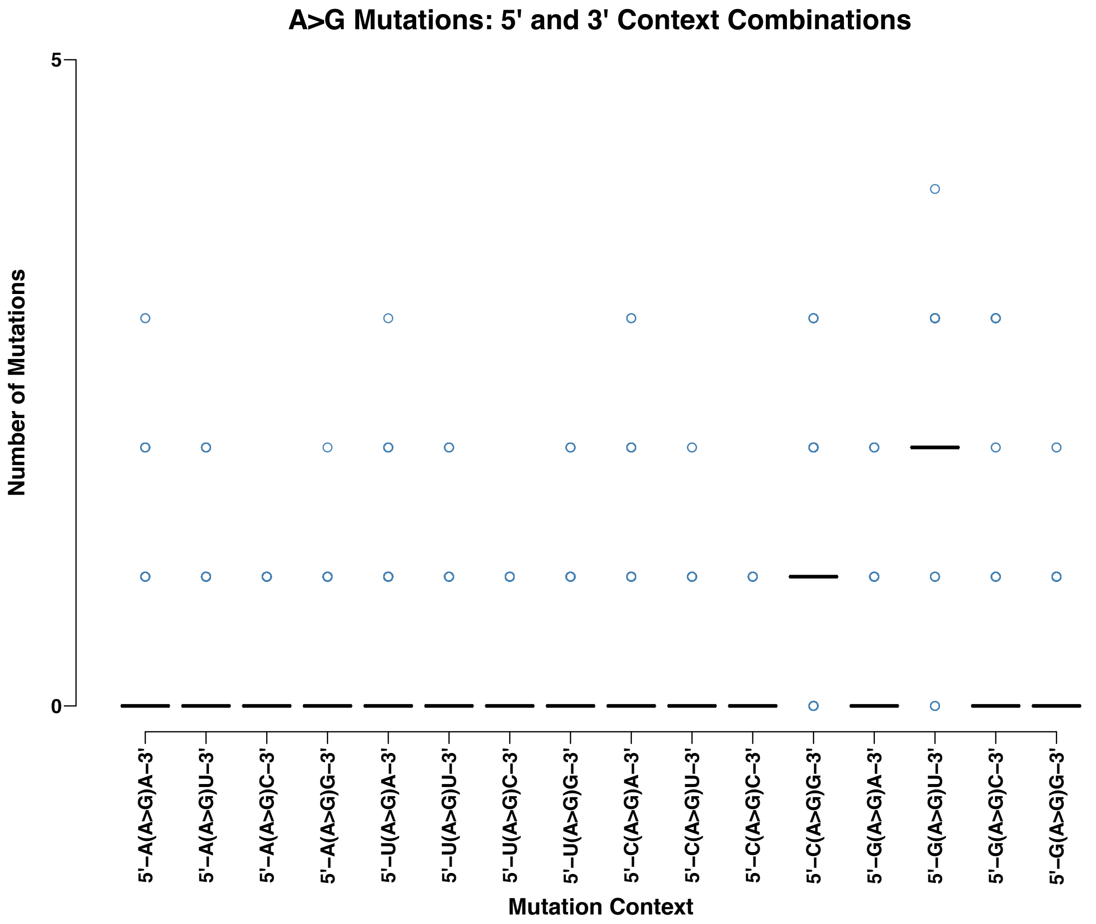


Figure 38: Combined mutation contexts of A>G mutations compared to the early West African outbreak reference sample.

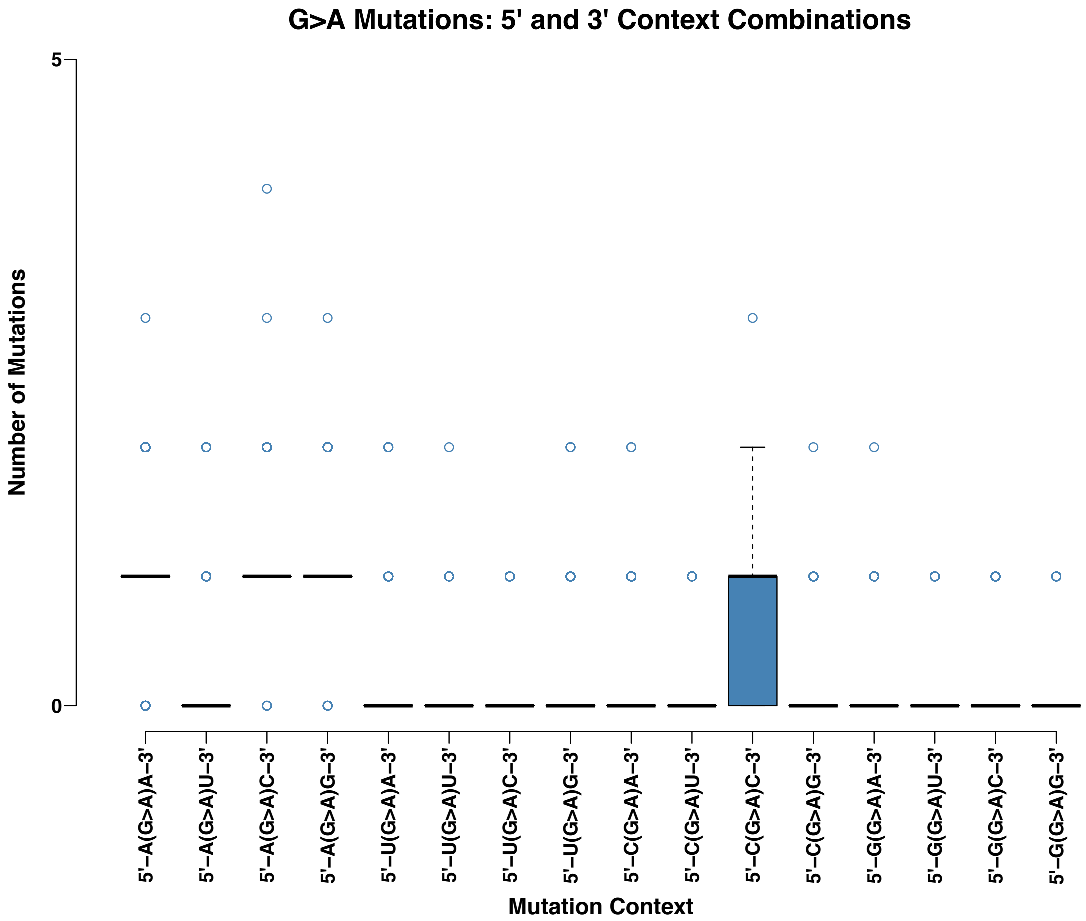


Figure 39: Combined mutation contexts of G>A mutations compared to the early West African outbreak reference sample.

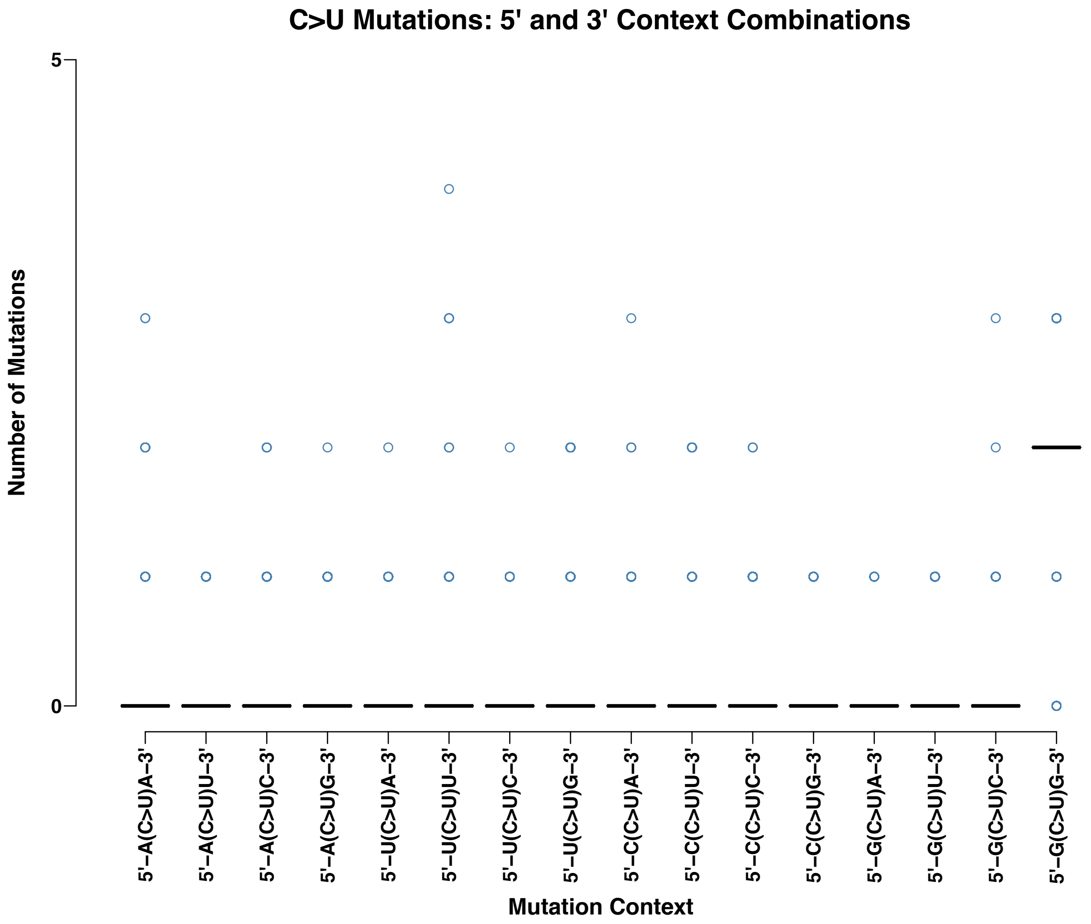


Figure 40: Combined mutation contexts of C>U mutations compared to the early West African outbreak reference sample.

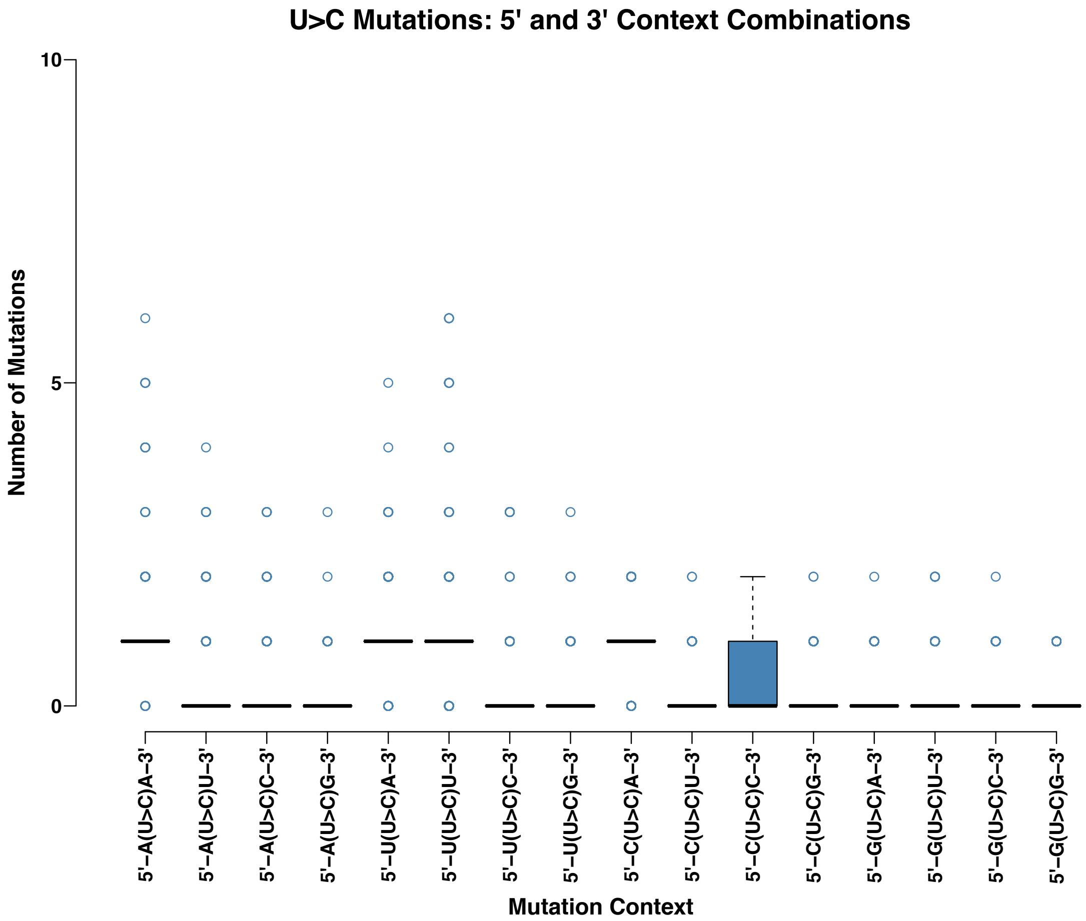


Figure 41: Combined mutation contexts of U>C mutations compared to the early West African outbreak reference sample.

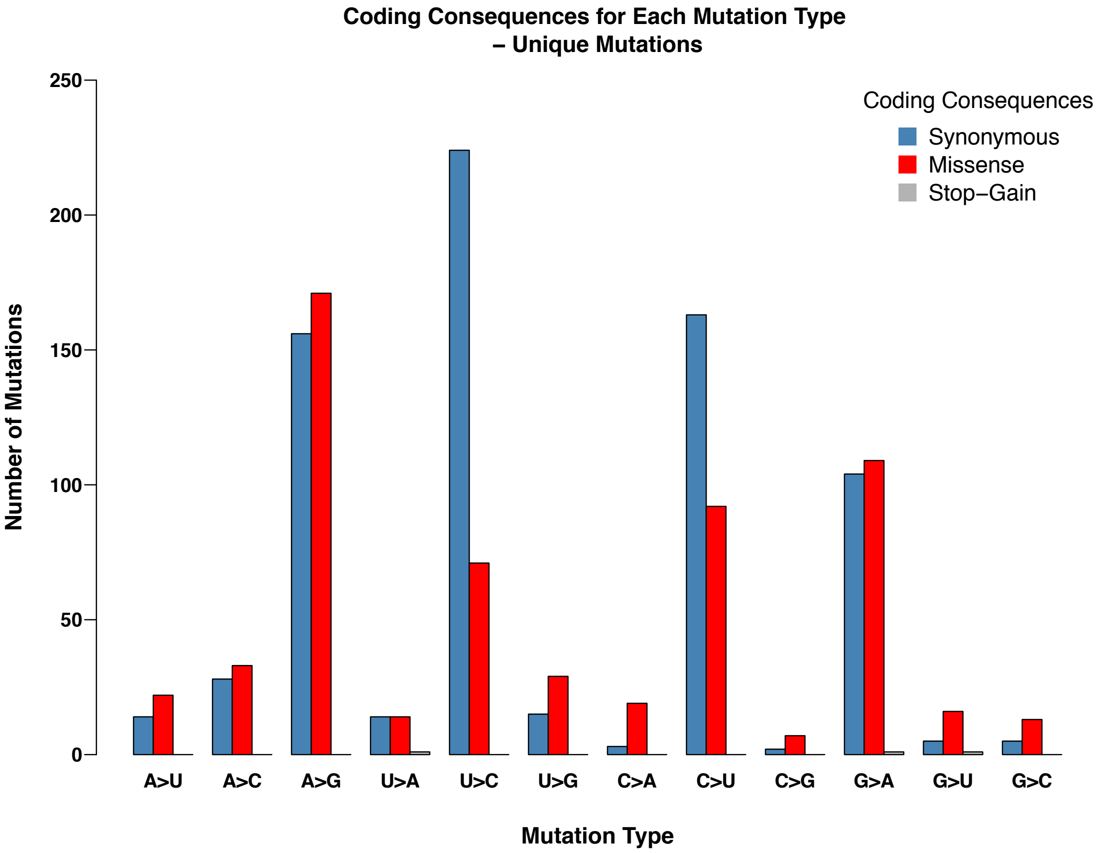


Figure 42: Protein coding consequences of each mutation type for the unique set of mutations across all samples compared to the early West Africa outbreak reference sample.

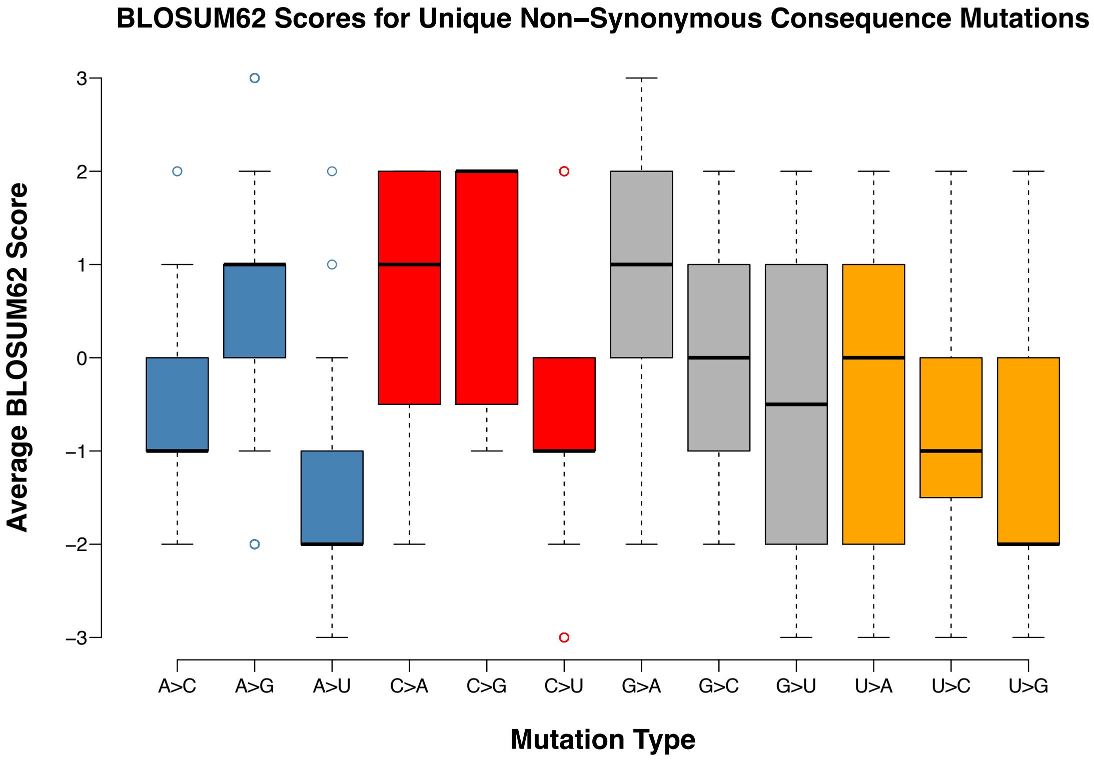


Figure 43: Average BLOSUM62 matrix scores for the unique set of mutations across all samples compared to the early West African outbreak reference sample, separated by mutation type.

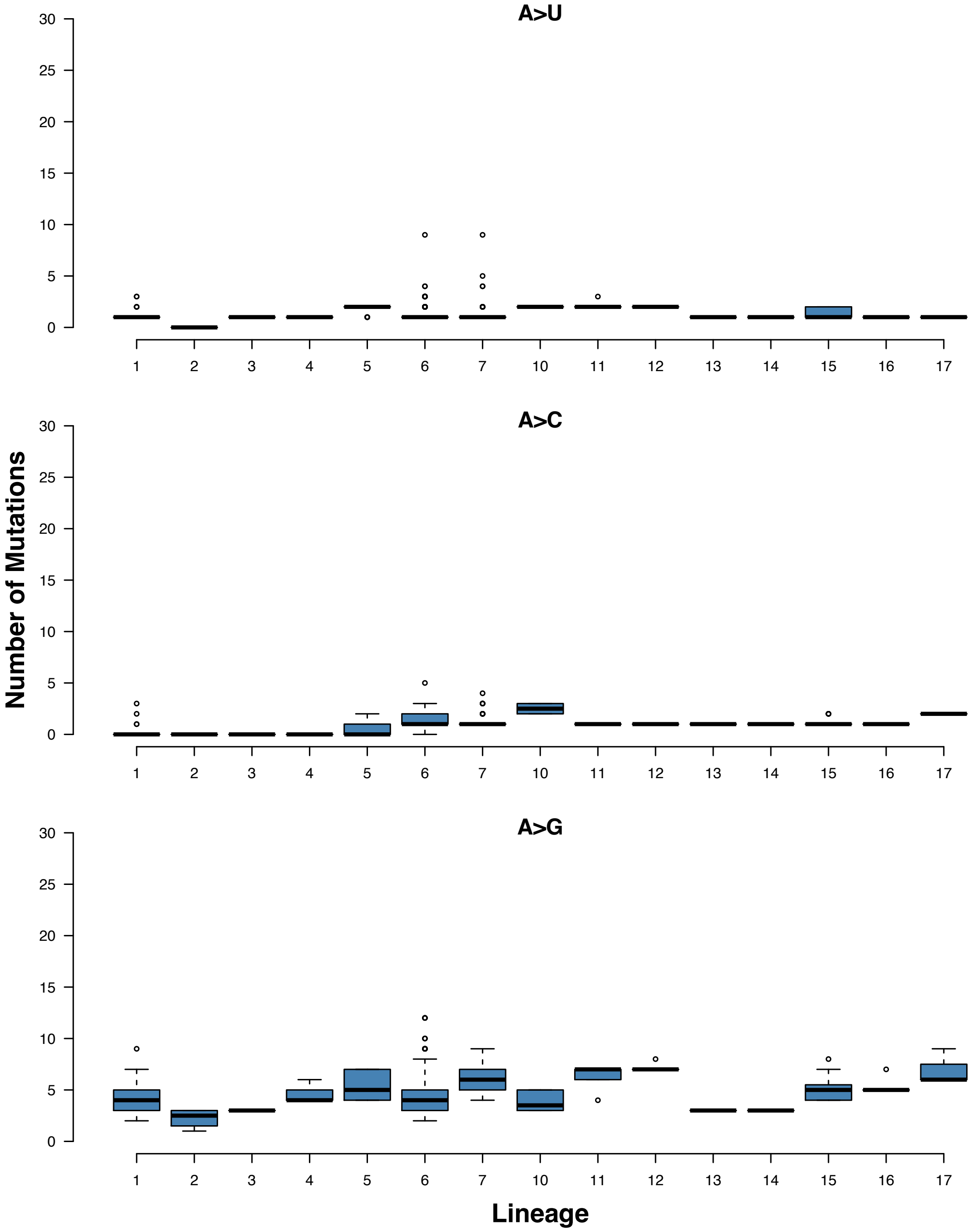


Figure 44: Observed numbers of mutation types per sample compared to the early West African outbreak reference sample, for all A>X mutations. Samples are divided in to the outbreak lineages defined by Urbanowicz et al.

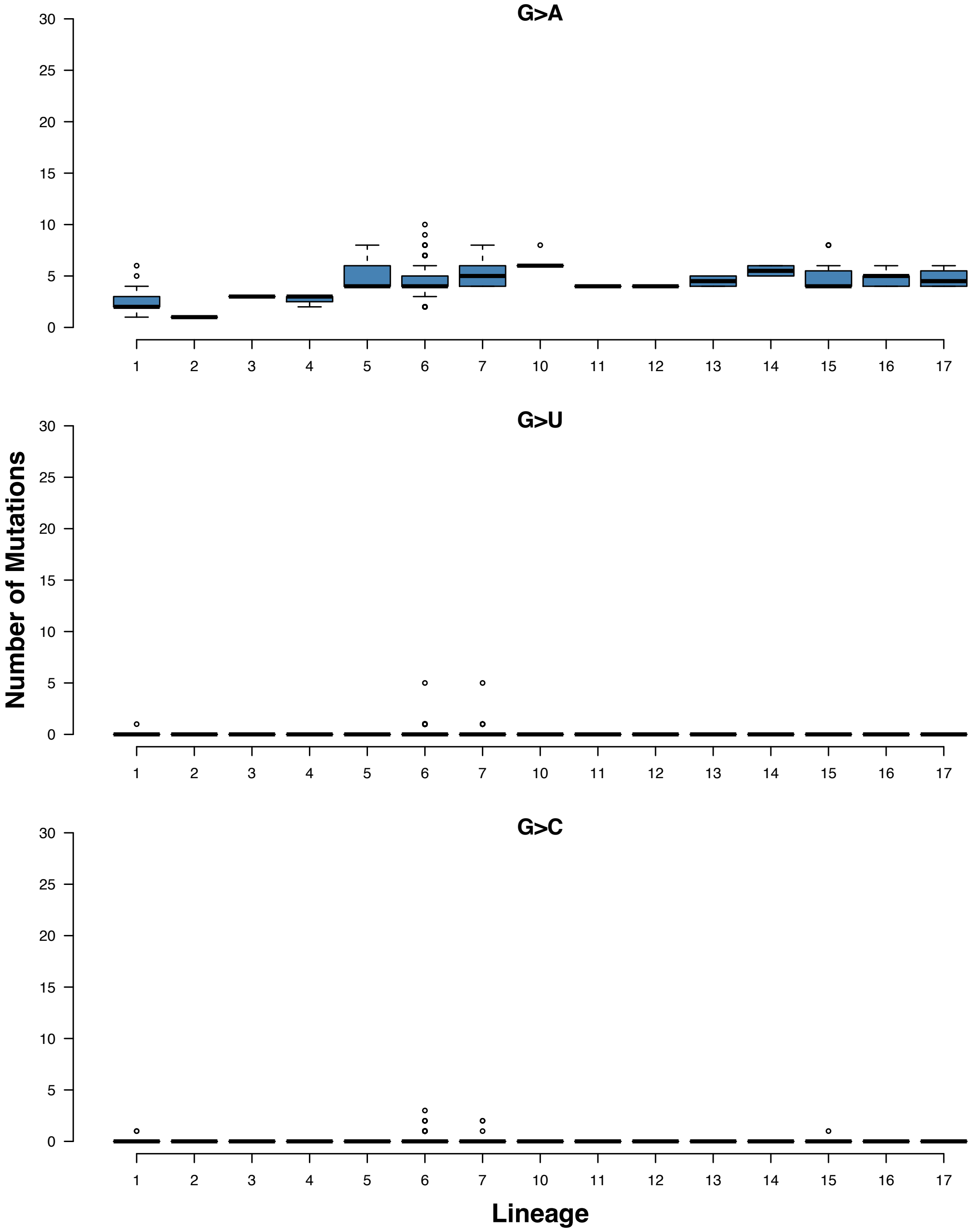


Figure 45: Observed numbers of mutation types per sample compared to the early West African outbreak reference sample, for all G>X mutations. Samples are divided in to the outbreak lineages defined by Urbanowicz et al.

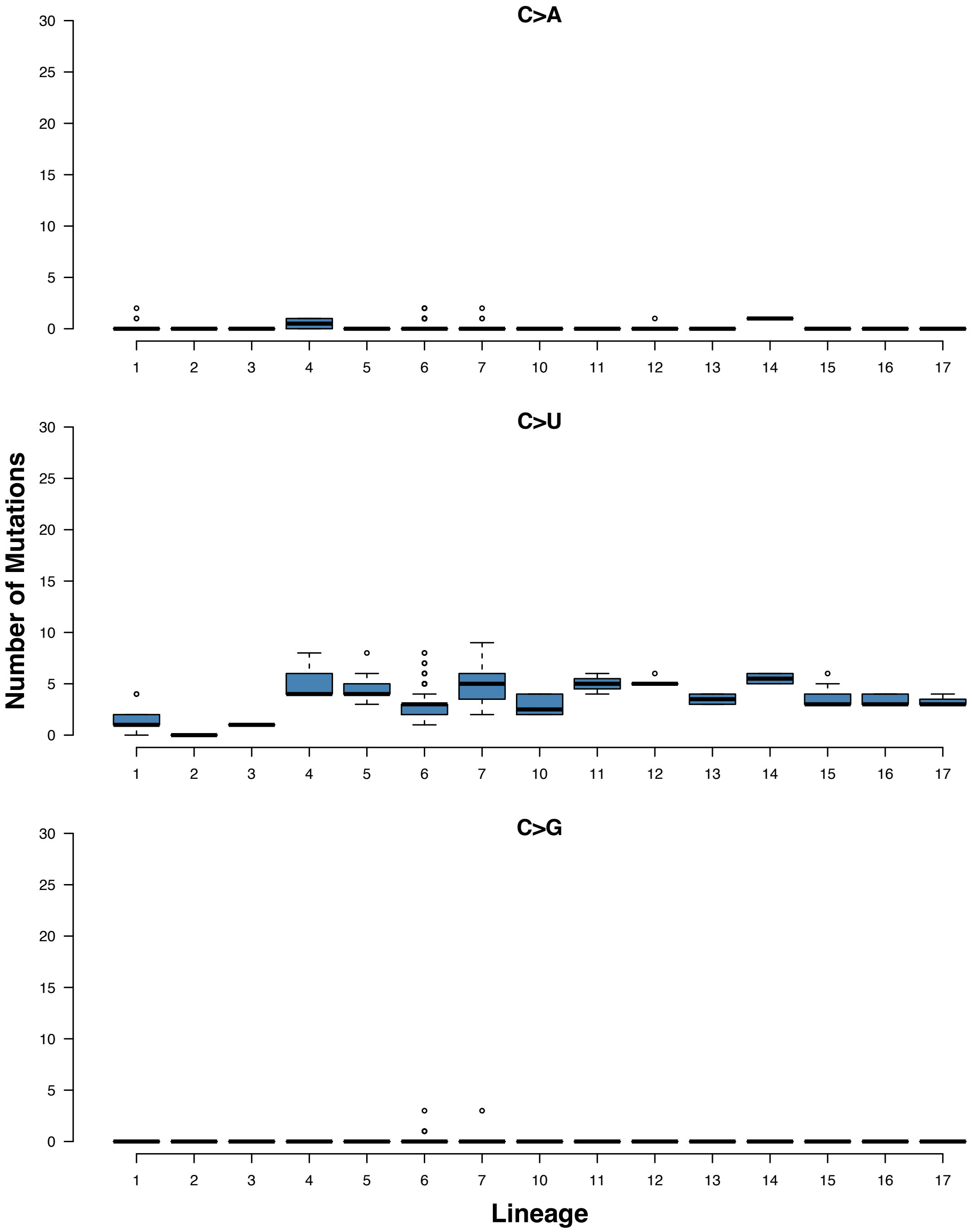


Figure 46: Observed numbers of mutation types per sample compared to the early West African outbreak reference sample, for all C>X mutations. Samples are divided in to the outbreak lineages defined by Urbanowicz et al.

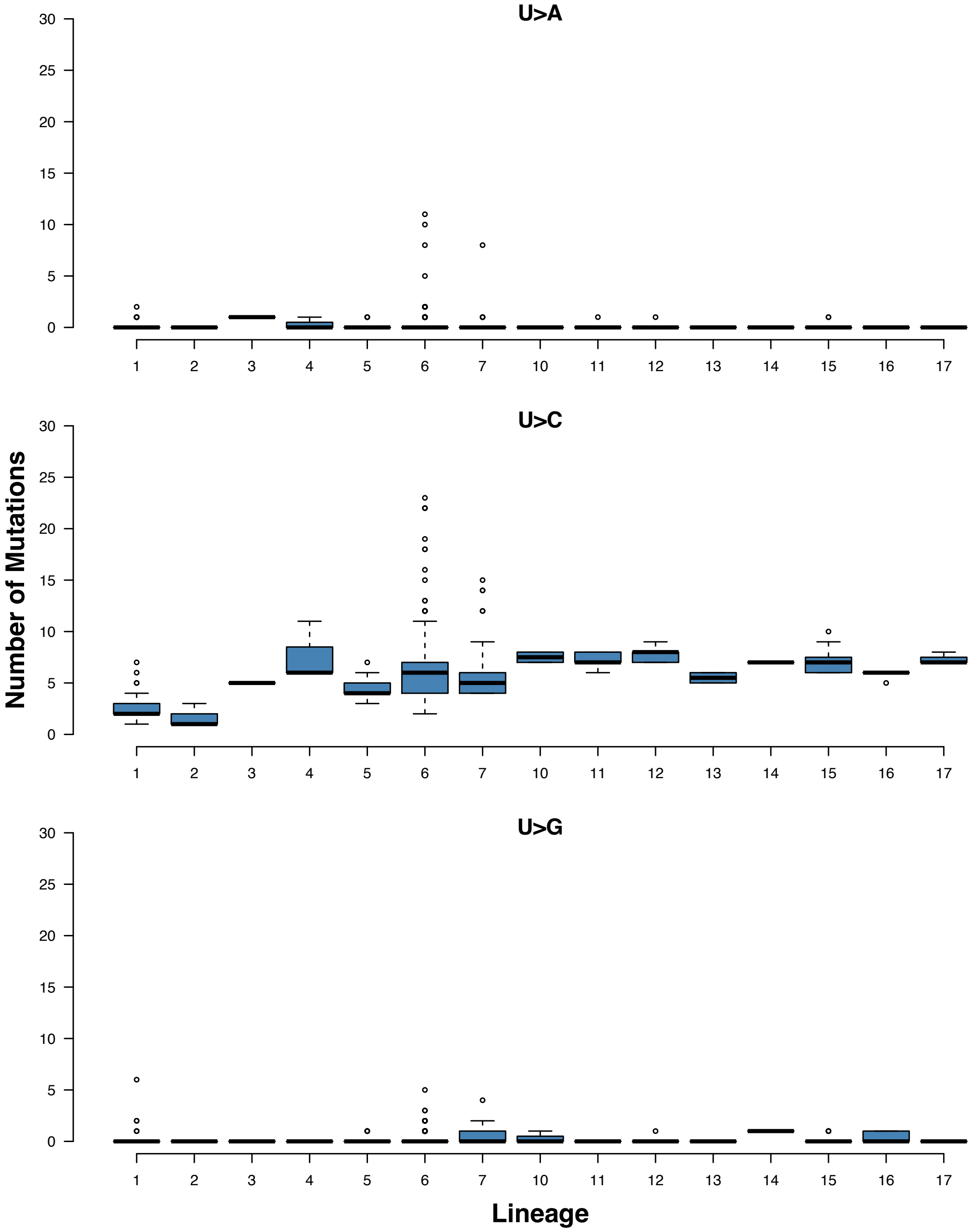


Figure 47: Observed numbers of mutation types per sample compared to the early West African outbreak reference sample, for all U>X mutations. Samples are divided in to the outbreak lineages defined by Urbanowicz et al.

Appendix 4 Tables

Table 1: Meta information for all *Ebola virus* samples used in this study, including their genome identifiers, date of collection, location of collection, and source database.

Table 1 is too large to be printed here but is available as a digital version. This can be found on the USB stick provided with this thesis.

Appendix 5: Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set

Damas, J., O'Connor, R., Farré, M., Lenis, V. P. E., **Martell, H. J.**, Mandawala, A., Fowler, K., Joseph, S., Swain, M. T., Griffin, D. K., Larkin, D. M., 2017. Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. *Genome Research*, 27, pp.1–10.

My contribution to the work was as follows:

1. Helped in the creation and testing of the universal FISH probe set, with Rebecca O'Connor, Anjali Mandawala, and Katie Fowler
2. Helped in culturing of bird tissue samples and creation of metaphase preparations, with Rebecca O'Connor, and Sunitha Joseph
3. Helped in the use of FISH probes in mapping PCFs to chromosomes, with Rebecca O'Connor, Anjali Mandawala, and Katie Fowler

1.0 Abstract

Most recent initiatives to sequence and assemble new species' genomes de novo fail to achieve the ultimate endpoint to produce contigs, each representing one whole chromosome. Even the best-assembled genomes (using contemporary technologies) consist of subchromosomal-sized scaffolds. To circumvent this problem, we developed a novel approach that combines computational algorithms to merge scaffolds into chromosomal fragments, PCR-based scaffold verification, and physical mapping to chromosomes. Multigenome-alignment-guided probe selection led to the development of a set of universal avian BAC clones that permit rapid anchoring of multiple scaffolds to chromosomes on all avian genomes. As proof of principle, we assembled genomes of the pigeon (Columbia livia) and peregrine falcon (Falco peregrinus) to chromosome levels comparable, in continuity, to avian reference genomes. Both species are of interest for breeding, cultural, food, and/or environmental reasons. Pigeon has a typical avian karyotype (2n = 80), while falcon (2n = 50) is highly rearranged compared to the avian ancestor. By using chromosome breakpoint data, we established that avian interchromosomal breakpoints appear in the regions of low density of conserved noncoding elements (CNEs) and that the chromosomal fission sites are further limited to long CNE "deserts." This corresponds with fission being the rarest type of rearrangement in avian genome evolution. High-throughput multiple hybridization and rapid capture strategies using the current BAC set provide the basis for assembling numerous avian (and possibly other reptilian) species, while the overall strategy for scaffold assembly and mapping provides the basis for an approach that (provided metaphases can be generated) could be applied to any animal genome.

2.0 Introduction

The ability to sequence complex animal genomes quickly and inexpensively has initiated numerous genome projects beyond those of agricultural/medical importance (e.g., Hu et al. 2009; Groenen et al. 2012) and inspired ambitious undertakings to sequence thousands of species (Zhang et al. 2014a; Koepfli et al. 2015). De novo genome assembly efforts ultimately aim to create a series of contigs, each representing a single chromosome, from p- to q-terminus ("chromosome-level" assembly). Assembling genomes using next-generation sequencing (NGS) technologies, however, typically relies on integration of the NGS data with a pre-existing chromosome-level reference assembly built with previous sequencing/ mapping technologies (Larkin et al. 2012). Indeed, use of short-read NGS data rarely produces assemblies at a similar level of integrity as those provided by traditional methodologies because of (1) an inability of NGS to generate long error-free contigs or scaffolds to cover chromosomes completely and (2) a paucity of inexpensive mapping technologies to upgrade NGS genomes to chromosome level. Even for projects with sufficient read-depths and long insert libraries, software algorithms at best produce subchromosomal-sized "scaffolds" requiring physical mapping to assemble chromosomes. Newer technologies such as optical mapping (Teague et al. 2010) including BioNano (Mak et al. 2016), Dovetail (Putnam et al. 2016), and Pacific Biosciences (PacBio) long-read sequencing (Rhoads and Au 2015) provide a long-term solution to this problem. To date, however, such approaches suffer from multiple limitations; e.g., BioNano contigs do not extend across multiple DNA nick site regions, centromeres, or large heterochromatin blocks, while PacBio sequencing requires hundreds of micrograms of high-molecular-weight DNA, which is often not easy to obtain.

Bioinformatic approaches, e.g., the Reference-Assisted Chromosome Assembly (RACA) algorithm (Kim et al. 2013), were developed to approximate near chromosome-sized fragments for a de novo assembled NGS genome. RACA use requires a genome from the same clade (e.g., Order for mammals) of the target species being assembled to chromosomes (Kim et al. 2013) and sequencing of long-insert libraries. RACA produces, at best, subchromosome-sized predicted chromosome fragments (PCFs) that require further verification and subsequent chromosome assembly. It is worth mentioning that, unlike RACA, other reference-assisted assembly algorithms, e.g., Ragout (Kolmogorov et al. 2014) or Chromosomer (Tamazian et al. 2016), do not use the target genome short- and long-range paired-read data to verify synteny breaks in/between scaffolds, meaning that the target species-specific rearrangements could be missed from the reconstructed PCFs/pseudochromosomes, making the reconstructed target chromosome structures more heavily biased to the reference genome(s) than when using RACA. RACA algorithm applied to the Tibetan antelope and blind mole rat genomes significantly improved continuities of these assemblies, but they still contain more than one large PCF for most chromosomes (Kim et al. 2013; Fang et al. 2014). Therefore, a novel, integrative approach that would allow de novo assembled genomes to retain structures of the target species karyotypes is a necessity.

A dearth of chromosome-level assemblies for nearly all newly sequenced genomes limits their use for critical aspects of evolutionary and applied genomics. Chromosome-level assemblies are essential for species that are regularly bred (e.g., for food or conservation) because a known order of DNA markers facilitates establishment of phenotype-to-genotype associations for gene-assisted selection and breeding (Andersson and Georges 2004). While such assemblies are established for popular livestock species, they are not available for those species widely used in developing countries (e.g., camels, yaks, buffalo, ostrich, quail) or species bred for conservation reasons (e.g., falcons). Chromosome-level information is essential for addressing basic biological questions pertaining to overall genome (karyotype) evolution and speciation (Lewin et al. 2009). Karyotype differences between species arise from DNA aberrations in germ cells that were fixed throughout evolution. These are associated with repetitive sequences used for nonallelic homologous recombination (NAHR) in evolutionary breakpoint regions (EBRs) where ancestral chromosomes break and/or combine in descendant species genomes (Murphy et al. 2005). An alternative theory, however, suggests that proximity of DNA regions in chromatin is the main driver of rearrangements and repetitive sequences play a minor role (Branco and Pombo 2006). Regardless of the mechanism, comparisons of multiple animal genomes show that between EBRs are evolutionary stable homologous synteny blocks (HSBs). Our studies in mammals (Larkin et al. 2009) and birds (Farré et al. 2016) suggest that at least the largest HSBs are maintained nonrandomly and are highly enriched for conserved noncoding elements (CNEs), many of which are gene regulatory sequences and miRNAs (Zhang et al. 2014b). We recently hypothesized that a higher fraction of elements under negative selection involved in gene regulation and chromosome structure in avian genomes (~7%) (Zhang et al. 2014b) compared with mammals (~4%) (Lindblad-Toh et al. 2011) could contribute to some avian-specific phenotypes and the evolutionary stability of most avian karyotypes (Farré et al. 2016). While a high density of CNEs in avian multispecies (ms) HSBs supports this hypothesis (Farré et al. 2016), a more definitive answer might be obtained by examining the fate of CNEs in the "interchromosomal EBRs" (flanking interchromosomal rearrangements) of an avian genome with a highly rearranged karyotype.

In this study, we focused on two avian genomes. The first, the peregrine falcon (Falco peregrinus), has an atypical karyotype (2n = 50) (Nishida et al. 2008). The falcon's ability to fly at speeds >300 km/h and its enhanced visual acuity make it the fastest predator on earth (Tucker et al. 1998). A prolonged period of extinction risk due to persecution around World War II and secondary poisoning from organochlorine pesticides (e.g., DDT) in the 1950s–1960s (Ferguson-Lees and Christie 2005) led to its placement on the CITES list of endangered species. The second avian genome that was focused on here, the pigeon (Columba livia), has a typical avian karyotype (2n = 80) similar to those of reference avian genomes: chicken, turkey, and zebra finch. Pigeon is one of the earliest examples of domestication in birds (Driscoll et al. 2009) contemporarily used as food and in sporting circles (Price 2002). Pigeon breeds can vary significantly in appearance with color, pattern, head crest, body shape, feathers, tails, vocalization, and flight display variations (Price 2002), inspiring considerable interest in identifying the genetic basis for these variations (Stringham et al. 2012; Shapiro et al. 2013). For the above reasons, both species' genomes were sequenced (Shapiro et al. 2013; Zhan et al. 2013); however, their assemblies are highly fragmented and chromosome-level assemblies are thus essential.

The objective of this study was therefore to develop a novel, inexpensive, transferrable approach to upgrade fragmented genome assemblies (i.e., pigeon

and falcon) to the chromosome level and to use them to address novel biological questions related to avian genome evolution. The method combines computational algorithms for ordering scaffolds into PCFs, retaining local structures of the target genome chromosomes after verification of a limited number of scaffolds, and physical mapping of PCFs directly to chromosomes with a universal set of avian bacterial artificial chromosome (BAC) probes. Studying a highly rearranged genome (falcon) compared with the avian ancestor sheds light on why interchromosomal rearrangements are infrequent in bird evolution.

3.0 Results

Our method involves (1) the construction of PCFs for fragmented assemblies based on the comparative and sequence read data implemented in the RACA algorithm, (2) PCR and computational verification of a limited number of scaffolds that are essential for revealing species-specific chromosome structures, (3) creation of a refined set of PCFs using the verified scaffolds and adjusted adjacency thresholds in RACA, and (4) the use of a panel of "universal" BAC clones to anchor PCFs to chromosomes in a high-throughput manner (Fig. 1).

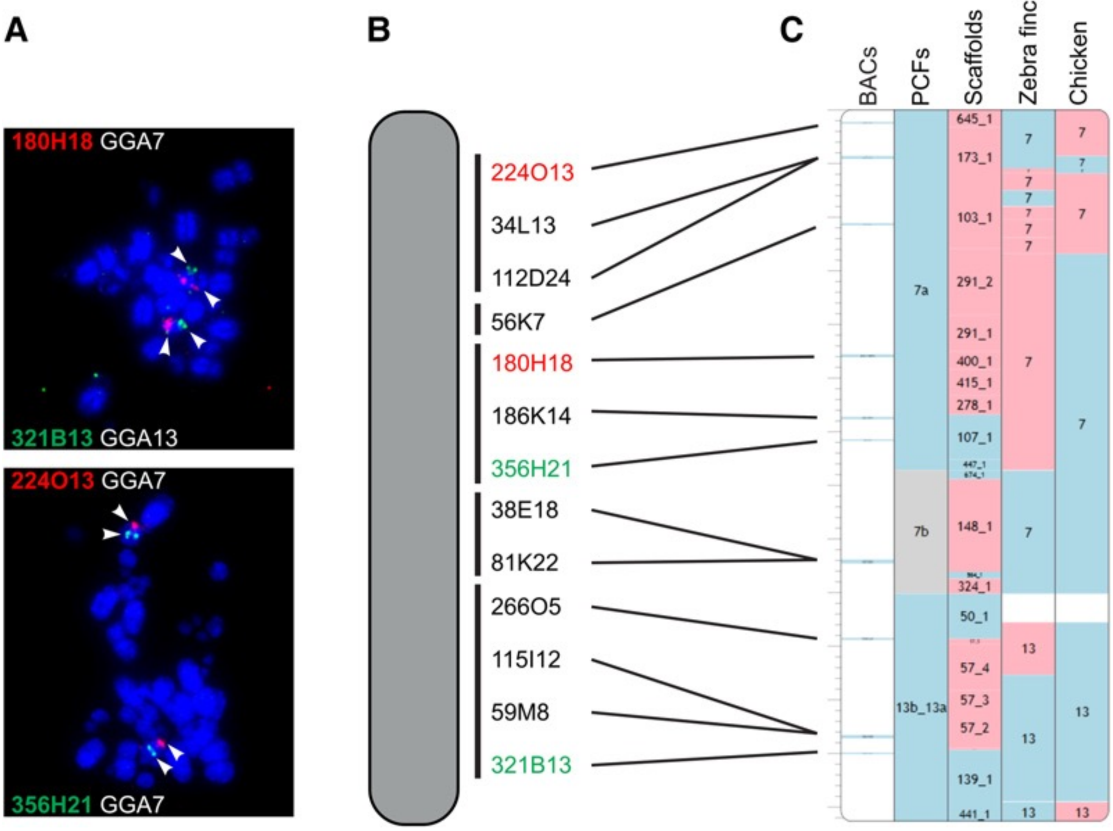


Figure 1: Methodology for the placement of the PCFs on chromosomes. (A) Dual-color FISH of universal BAC clones, (B) cytogenetic map of the falcon chromosome 8 (FPE8) with indication of the relative positions of the BAC clones along the chromosome, and (C) assembled chromosome containing PCFs 7a, 7b, and 13b_13a. Blue blocks indicate positive (+) orientation of tracks compared with the falcon chromosome; red blocks, negative (–) orientation; and gray blocks, unknown (?) orientation.

3.1 Construction of PCFs from fragmented assemblies

PCFs were generated for fragmented falcon and pigeon whole-genome sequences using RACA (Kim et al. 2013). For falcon, the zebra finch chromosome assembly was used as reference (divergence 60 MYA) and the chicken genome as outgroup (divergence 89 MYA). We generated a total of 113 PCFs with N50 of 27.44 Mb (Table 1). For pigeon (≥69 Myr divergence from both the chicken and zebra finch), chicken was used as reference and zebra finch as outgroup because (1) fewer pigeon scaffolds were split in this configuration (Supplemental Table S1) and (2) the high similarity of pigeon and chicken karyotypes (Derjusheva et al. 2004). This resulted in 150 pigeon PCFs with N50 of 34.54 Mb (Table 1). These initial PCF sets contained 72 (15.06%) and 78 (13.64%) scaffolds for falcon and pigeon, respectively, that were split by RACA due to insufficient read and/or comparative evidence to support their structures.

3.2 Verification of scaffolds essential for revealing species-specific chromosome architectures

All scaffolds split by RACA contained structural differences between the target and reference chromosomes, suggesting their importance for revealing the architecture of target species chromosomes. The structures of these scaffolds were tested by PCR amplification across all the split regions defined to <6 kb in the target species scaffolds. Of these, 41 (83.67%) and 58 (84.06%) resulted in amplicons of expected length in pigeon and falcon genomic DNA, respectively (Supplemental Table S2). For the split regions with negative PCR results, we tested an alternative (RACA-suggested) order of the flanking syntenic fragments (SFs). Out of these, amplicons were obtained for 2/4 in falcon and 7/7 in pigeon, confirming the chimeric nature of the original scaffolds properly detected in these cases (Supplemental Table S2). To estimate which of the remaining split regions (>6 kb; 36 in falcon and 40 in pigeon PCFs) were likely to be chimeric, we empirically identified two genome-wide minimum physical coverage (Meyerson et al. 2010) levels, one for falcon and one for pigeon, in the SFs joining regions for which (and higher) the PCR results were most consistent with RACA predictions. If the new thresholds were used in RACA without additional scaffold verification (e.g., by PCR) or mapping data, they would lead to splitting of nearly all scaffolds with large structural misassemblies in falcon, and ~6% of them would still be present in pigeon PCFs. The number of scaffolds containing real structural differences with the reference chromosomes that would still be split by RACA was estimated as ~56% in the falcon and ~43% in pigeon PCFs (Supplemental Table S2). To reduce the number of the real structural differences split in the final PCF set, PCR verification of selected scaffolds and use of independent (cytogenetic) mapping have been introduced.

3.3 Creation of a refined set of pigeon and falcon PCFs

For new reconstructions, the adjusted physical coverage thresholds were used. In addition, we kept intact those scaffolds confirmed by PCR but split those shown to be chimeric and/or disagreeing with the cytogenetic map (see below), resulting in a total of 93 PCFs with N50 25.82 Mb for falcon and 137 PCFs with N50 of 22.17 Mb for pigeon, covering 97.17% and 95.86% of the original scaffold assemblies, respectively (Table 1). The falcon RACA assembly contained six PCFs homeologous to complete zebra finch chromosomes (TGU4A, 9, 11, 14, 17, and 19), while five pigeon PCFs were homeologous to complete chicken chromosomes (GGA11, 13, 17, 22, and 25). Only 3.50% of the original scaffolds used by RACA were split in the pigeon and 3.14% in falcon final PCFs (Table 1). The accuracy for the PCF assembly was estimated as ~85% for falcon and ~89% for pigeon based on the ratio of the number of SFs to the number of scaffolds (Kim et al. 2013).

Table 1: Scaffold-based RACA assemblies for peregrine falcon and pigeon. [a]RACA assembly after the use of adjusted coverage thresholds and post-processing of scaffolds verified by PCR. [b]Percentage of all scaffolds included in the RACA assembly.

	Peregrine Falcon			Pigeon		
Statistics	**Scaffold assembly**	**Default RACA**	**Adjusted RACA[a]**	**Scaffold Assembly**	**Default RACA**	**Adjusted RACA[a]**
No. of scaffolds (≥10 kb)	723	478	478	1081	572	572
No. of PCFs	NA	113	93	NA	150	137
Total length (Gb)	1.17	1.14	1.14	1.10	1.07	1.07
N50 (Mb)	3.94	27.44	25.82	3.15	34.54	22.17
Fraction of scaffold assembly (%)	NA	97.17	97.17	NA	95.86	95.86
No. of scaffolds split by RACA	NA	72 (15.06[b])	15 (3.14[b])	NA	78 (13.64[b])	20 (3.50[b])

3.4 Construction of a panel of comparatively anchored BAC clones designed to hybridize in phylogenetically divergent avian species and link PCFs to chromosomes

Initial experiments on cross-species BAC mapping using fluorescence in-situ hybridization (FISH) on five avian species with divergence times between 28 and 89 Myr revealed highly varying success rates (21%–94%), with hybridizations more likely to succeed on species closely related to that of the BAC origin (Table 2). To minimize the effect of evolutionary distances between species on hybridizations, genomic features that were likely to influence hybridization success were measured in chicken, zebra finch, and turkey BAC clones (Supplemental Tables S3, S4). The classification and regression tree (CART) approach (Loh 2011) was applied to the 101 randomly selected BAC clones (Table 2). The obtained classification shows 87% agreement with FISH results (Supplemental

Fig. S1). Correlating DNA features with actual cross-species FISH results led us to develop the following criteria for the selection of chicken or zebra finch BAC clones very likely to hybridize on metaphase preparations of phylogenetically distant birds (≥69 Myr of divergence; where the hybridization success rate of random BAC clones was <70%): The BAC had to have ≥93% DNA sequence alignable with other avian genomes and contain at least one conserved element (CE) ≥300 bp. Instead of a long CE, the BAC could contain only short repetitive elements (<1290 bp) and CEs of at least 3 bp long (Supplemental Fig. S1; Supplemental Table S4). The hybridization success rate with distant avian species for the set of newly selected clones obeying these criteria was high (71%–94%; Table 2). The success rates for the selected chicken BAC clones only ranged 90%–94%. From these chicken clones, 84% hybridized with chromosomes of all avian species in our set (Supplemental Fig. S2).

Table 2: Comparison of zoo-FISH success rate for random and selected set of BAC clones. Divergence times are the average of the times reported on the ExaML TENT topology from Jarvis et al. (2014).

	Chicken BAC Clones				**Zebra finch BAC Clones**			
		Success rate (%)				**Success rate (%)**		
	Divergence time (MY)	**Random set *N*=53**	**Selected set *NI=99***	**Ratio**	**Divergence time (MY)**	**Random set *N*=53**	**Selected set *NI=99***	**Ratio**
Chicken	NA	NA	NA	NA	89	58.33	75.00	1.29
Turkey	28	88.68	100.00	1.13	89	54.17	83.33	1.54
Pigeon	89	26.42	91.92	3.48	69	68.75	70.83	1.03
Peregrine falcon	89	47.17	93.94	1.99	60	93.75	91.67	0.98
Zebra finch	89	20.75	90.91	4.38	NA	NA	NA	NA

As a final result, we generated a panel of 121 BAC clones spread across the avian genome (GGA 1-28 +Z [except 16]) that successfully hybridized across all species attempted. The collection was supplemented by a further 63 BACs that hybridized on the metaphases of at least one species that was considered phylogenetically distant (i.e., ≥69 Myr; split between Columbea and the remaining Neoavian clades) and a further 33 that hybridized on at least one other species (Fig. 2; Supplemental Table S5).

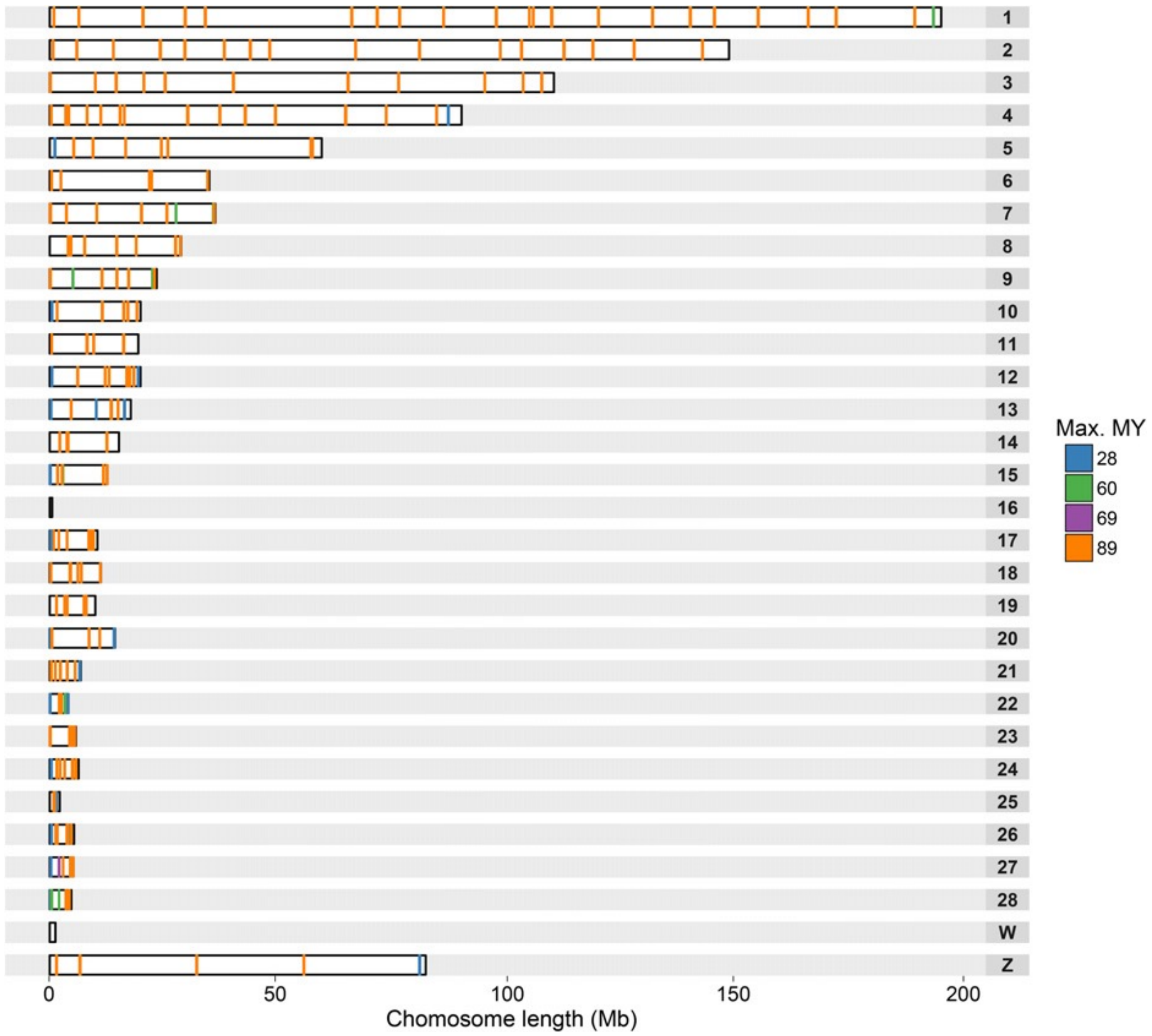


Figure 2: Distribution of universal BAC clones along chicken chromosomes. Each rectangle represents a chicken chromosome; the lines inside, the location of each BAC clone. BAC clones are colored accordingly to the maximum phylogenetic distance of the species they successfully hybridized. The distribution of spacing between all these BAC clones is shown on the Supplemental Figure S3.

3.5 Physical assignment of refined PCFs on the species' chromosomes

In order to place and order PCFs along chromosomes, BAC clones from the panel described above and assigned to PCFs based on alignment results were hybridized to falcon (177 clones) and pigeon (151 clones) chromosomes (Table 3).

The 57 PCFs cytogenetically anchored to the falcon chromosomes represented 1.03 Gb of its genome sequence (88% of the cumulative scaffold length). Of these, 888.67 Mb were oriented on the chromosomes (Table 3; Supplemental Table S6). The pigeon chromosome assembly consisted of 0.91 Gb in 60 pigeon PCFs representing 82% of the combined scaffold length. Of these 687.59 Mb were oriented (Table 3; Supplemental Table S7). Visualizations of both newly assembled genomes are available from the Evolution Highway comparative chromosome browser (see Supplemental Results) and our avian UCSC Genome Browser hub.

Table 3: Statistics for the chromosome assemblies of peregrine falcon and pigeon.

Statistics	Peregrine falcon	Pigeon
No. of informative BAC clones	177	151
No. of PCFs placed on chromosomes	57	60
Combined length (Gb)	1.03	0.91
PCF assembly coverage (%)	90.03	85.23
Scaffold assembly coverage (%)	87.55	81.70
No. of oriented PCFs	32	26
Combined length (Mb)	888.67	687.59

3.6 Pigeon chromosome assembly

No deviations from the standard avian karyotype (2n = 80) were detected for pigeon with each mapped chromosome having an appropriate single chicken and zebra finch homeolog. Compared to chicken, the only interchromosomal rearrangement identified was the ancestral configuration of GGA4 found as two separate chromosomes in the pigeon and other birds (Fig. 3A; Supplemental Fig. S4; Derjusheva et al. 2004; Hansmann et al. 2009; Modi et al. 2009; http://eh-demo.ncsa.uiuc.edu/birds). Nonetheless, 70 intrachromosomal EBRs in the pigeon lineage were identified (Supplemental Table S8).

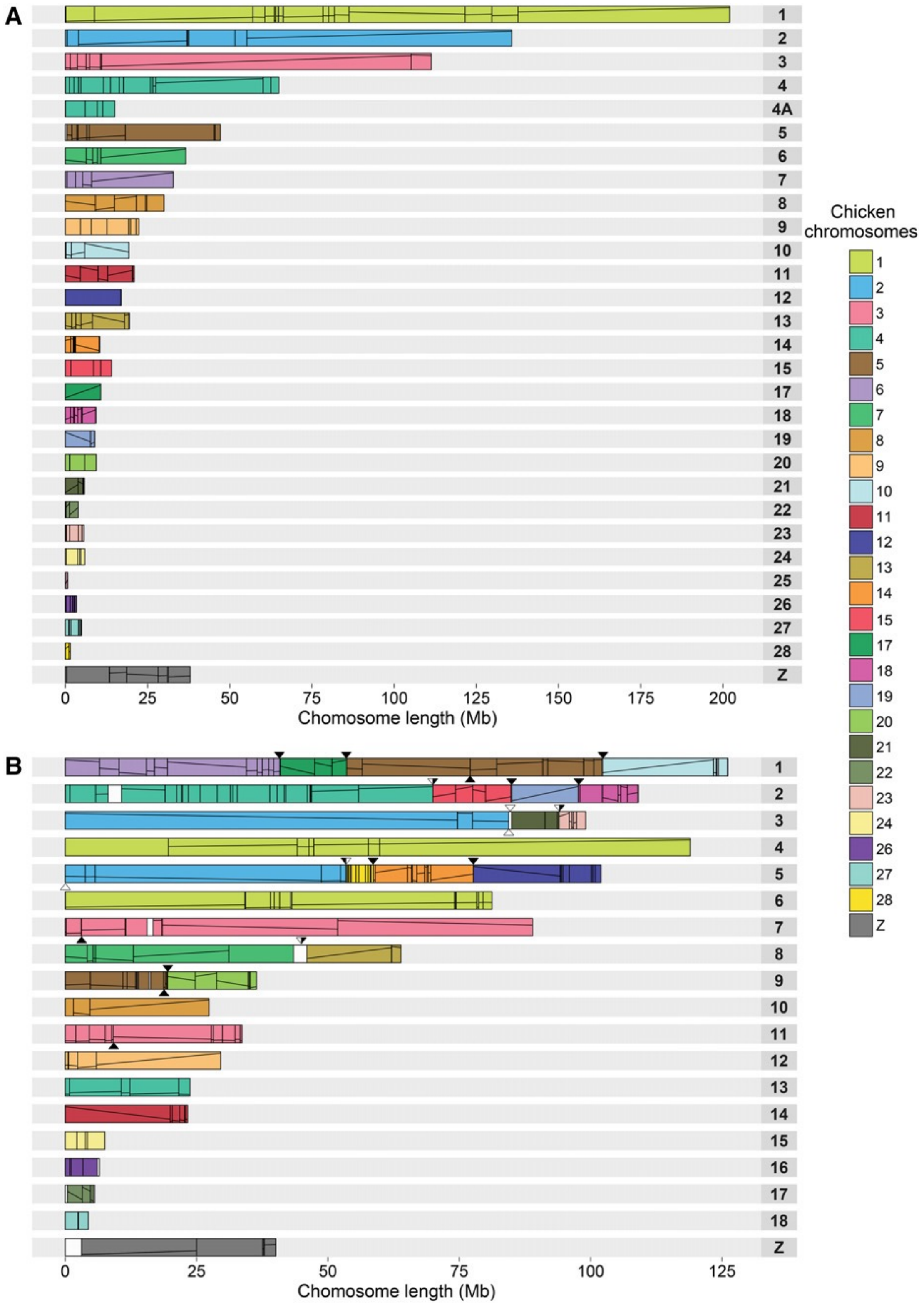


Figure 3: (see legend on next page)

Figure 3: Ideogram of pigeon (A) and peregrine falcon (B) chromosomes. Numbered rectangles represent chromosomes, and colored blocks inside represent regions of homeology with chicken chromosomes. Lines within the colored blocks represent block orientation. Pigeon chromosomes 1–9 and Z were numbered according to the method of Hansmann et al. (2009) and the remaining chromosomes according to their chicken homeologs. Falcon chromosomes 1–13 and Z were numbered accordingly to the method of Nishida et al. (2008). The remaining chromosomes were numbered by decreasing combined length of the placed PCFs. Triangles above the falcon chromosomes point to the positions of falcon-specific fusions; below chromosomes, the positions of fissions. Black filling within the triangles point to the EBR boundaries used in the CNE analysis.

3.7 Falcon chromosome assembly

Homeology between the chicken and the falcon was identified for all mapped chromosomes with the exception of GGA16 and GGA25 (Fig. 3B; Supplemental Fig. S5; http://eh-demo.ncsa.uiuc. edu/birds). In total, 13 falcon-specific fusions and six fissions were detected (Supplemental Table S8). Each of the chicken largest macrochromosome homeologs (GGA1 to GGA5) were split across two falcon chromosomes. Both the GGA6 and GGA7 homeologs were found as single blocks fused with other chicken chromosome material within falcon chromosomes. Among the other chicken macrochromosomes, only GGA8 and GGA9 were represented as individual chromosomes. Of the 17 mapped chicken microchromosomes, 11 were fused with other chromosomes. A total of 69 intrachromosomal EBRs were detected in the falcon lineage (Supplemental Table S8; Supplemental Results). Consistent with our previous report (Farré et al. 2016), falcon intrachromosomal EBRs were found highly enriched for the LTR-ERV1 transposable elements (TEs; t-test $P < 0.05$) (Supplemental Table S9). Both fusion and fission EBRs were not significantly enriched for any type of TEs.

3.8 Fate of CNEs in avian inter- and intrachromosomal EBRs

The falcon chromosome assembly provided us with a set of 19 novel interchromosomal EBRs not previously found in published avian chromosome assemblies (Fig. 3B; Supplemental Table S8). To investigate the fate of CNEs in avian EBRs, we calculated densities of avian CNEs in the chicken chromosome regions corresponding to the chicken, falcon, pigeon, flycatcher, and zebra finch intrachromosomal and interchromosomal EBRs defined to ≤100 kb in the chicken genome (Fig. 4; Supplemental Table S10). Avian EBRs had a significantly lower fraction of CNEs than their two adjacent chromosome intervals of the same size each (up- and downstream; P = 3.35 × 10−7) (Supplemental Table S11). Moreover, the interchromosomal EBRs (fusions and fissions) had, on average, approximately 12 times lower density of CNEs than the intrachromosomal EBRs (P = 2.40 × 10−5) (Supplemental Table S11). The lowest density of CNEs was observed in the fission breakpoints (P = 0.04) (Fig. 4; Supplemental Table S11).

To identify CNE densities and the distribution associated with avian EBRs at the genome-wide level, we counted CNE bases in 1-kb windows overlapping EBRs and avian msHSBs >1.5 Mb (Farré et al. 2016). The average density of CNEs in the EBR windows was lower (0.02) than in msHSBs (0.11). The density of CNEs in the fission EBRs was the lowest observed, without CNE bases (from now on “zero CNE windows”), while intrachromosomal EBRs were the highest among the EBR regions (0.02) (Supplemental Table S12). The genome-wide CNE density was 0.09, closer to the density observed in msHSBs. Of the ~347 Mb of chicken genome found in the “zero CNE windows,” 0.5% was associated with EBRs and 15% with msHSBs. To investigate if these intervals are distributed differently in the breakpoint and synteny regions, we compared distances between the “zero CNE windows” and the closest window with the average msHSB CNE density or higher in EBRs, msHSBs, and genome-wide. The median of the distances between these two types of windows was the lowest in the msHSBs (~4 kb), intermediate in the intrachromosomal (~19 kb) and fusion EBRs (~23 kb), and highest in the fission EBRs (~35 kb) (Supplemental Table S13). All these values were significantly different from the genome-wide average distance of ~6 kb (P<2.2×10−16) and also significantly different from each other (P ≤ 0.004) (Supplemental Table S12; Supplemental Fig. S6).

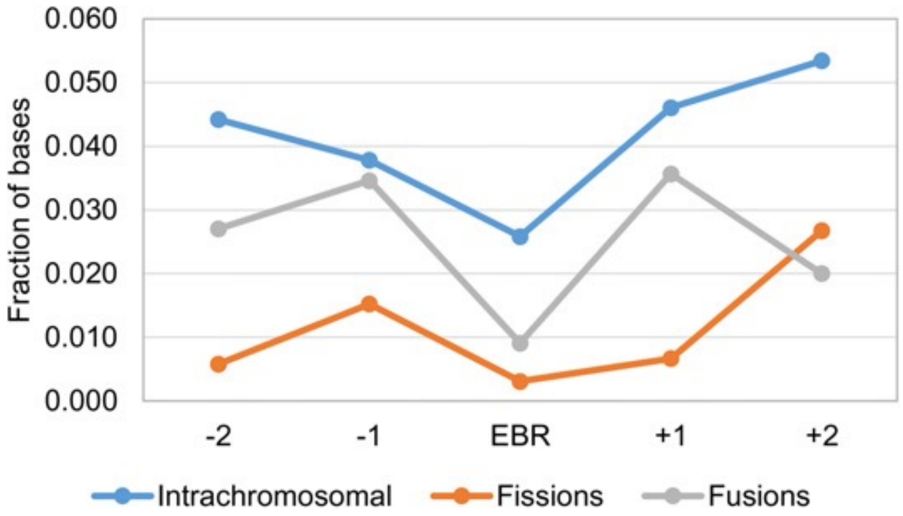


Figure 4: Average fraction of bases within conserved noncoding elements (CNEs) in avian EBRs and two flanking regions upstream (–) and downstream (+).

4.0 Discussion

In this study, we present a novel integrative approach to upgrade sequenced animal genomes to the chromosome level. We have previously reported a limited success with the use of high-gene density and low-repeat content BAC clones for cross-species hybridization (Larkin et al. 2006; Romanov et al. 2011). However, the use of such probes for whole-genome chromosomal assembly has not hitherto been demonstrated. That is, in this study, we made use of the whole-genome sequences from multiple species and applied a systematic approach to design a panel of universally hybridizing BAC probes along the length of each chromosome. By using these probes as a basis and in combination with comparative sequence analysis, targeted PCR, and optimized high-throughput cross-species BAC hybridizations, the approach herein presented thus represents a unique methodology to achieve chromosome-level reconstruction for scaffold-based de novo assemblies that could be applied to any animal genome provided an actively growing population of cells can be obtained to generate metaphase preparations.

In this study, we provide proof of principle for this new approach by generating such assemblies for two previously published, but highly fragmented, avian genomes. The resulting chromosome-level assemblies contain >80% of the genomes (compared with current estimates of genome size) and, in continuity, are comparable to those obtained by combining the traditional sequencing and mapping techniques (Deakin and Ezaz 2014) but require much less cost and resources. Given that it has been suggested that estimates of genome size based on cytology are inaccurate and usually overestimated (Kasai et al. 2012, 2013), techniques such as flow cytometry should be used to estimate genome size more accurately (Kasai et al. 2012, 2013). Flow cytometry will ultimately be able to determine the extent to which the genomes are actually covered by new procedures to upgrade their assemblies and will be invaluable in pointing out any remaining gaps to fill. Indeed, this approach could be augmented further by chromosome-specific DNA sequencing such as has recently been demonstrated in the B chromosomes of two deer species (Makunin et al. 2016).

Molecular and cytogenetic studies to date suggest that the majority of avian genomes remain remarkably conserved in terms of chromosome number (in 60%–

70% of species 2n = ~80) and that interchromosomal changes are relatively rare (Griffin et al. 2007; Schmid et al. 2015). Exceptions include representatives of Psittaciformes (parrots), Sphenisciformes (penguins), and Falconiformes (falcons). This study represents the first reconstruction of a highly rearranged avian karyotype (peregrine falcon). It demonstrates that fusion is the most common mechanism of interchromosomal change in this species, with some resulting chromosomes exhibiting as many as four fused ancestral chromosomes. There was no evidence of reciprocal translocations, and all microchromosomes remained intact, even when fused to larger chromosomes.

Recently, we suggested possible mechanisms why avian genomes, with relatively rare exceptions, remain evolutionarily stable interchromosomally and why microchromosomes represent blocks of conserved synteny (Romanov et al. 2014; Farré et al. 2016). Absence of interchromosomal rearrangement (as seen in most birds) could suggest either an evolutionary advantage to retaining such a configuration or little opportunity for change. A smaller number of transposable elements in avian genomes compared with other animals would indicate that avian chromosomes indeed have fewer opportunities for chromosome merging using NAHR, explaining the presence of multiple microchromosomes.

Our study provides additional support for this hypothesis as in the falcon lineage only intrachromosomal EBRs were significantly enriched in transposable elements, while interchromosomal EBRs (flanking both fusions and fissions) were not found significantly enriched. On the other hand, a strong enrichment for avian CNEs in the regions of interspecies synteny in birds and other reptiles suggests evolutionary advantage of maintaining established synteny (Farré et al. 2016), implying that fission events should be rare in avian evolution. In this study, we present the first analysis of a significant number of interchromosomal EBRs by analysis of the falcon genome, demonstrating that those rare interchromosomal rearrangements that are fixed in the avian lineage-specific evolution did indeed appear in areas of a low density of CNEs. This applies to both fission and fusion events. Our results demonstrate moreover that, to be suitable for chromosomal fission, the sites of interchromosomal EBRs are restricted further as they need to be significantly more distant from the areas with high CNE density than the equivalent intervals found in the regions of ms synteny, other EBR types, or on average in the genome. This might also explain why falcon-specific fission breakpoints appear to be reused in other avian lineages as intrachromosomal EBRs.

The study of intrachromosomal changes in pigeons, falcons (this study), and Passeriform species (Skinner and Griffin 2012; Romanov et al. 2014) suggests that these events might have a less dramatic effect on cis gene regulation than interchromosomal events. Indeed, intrachromosomal EBRs appear in regions of significantly higher CNE density than interchromosomal EBRs. Why, then, do species such as falcons and parrots undergo wholesale interchromosomal rearrangement (previously reported) but (according to this study) with fission restricted to a few events and fusion more common? The absence of positive selection for change in chromosome number (or lack of templates for NAHR) possibly explains why there was little fixation of any interchromosomal change among birds in general (Bush et al. 1977; Fontdevila et al. 1982; Burt et al. 1999; Burt 2002); however, why this positive selection has been reintroduced (or barriers to it have been removed) in selected orders is still a matter of conjecture.

The design and use of a set of BAC probes intended to work equally well on a large number of diverged avian species created a resource for physical mapping that is transferrable to multiple species. In this regard, mammals are the greatest priority as they are the most studied phylogenetic Class of organisms in the scientific literature. Reasons for this include human interest (e.g., clinical studies), biomedical models (e.g., mouse, rat, rabbit, pig), companion animals (e.g., cat, dog), and agricultural mammals (pig, sheep, cattle, etc.). Many are on the CITES threatened/endangered list, and with impending global warming, tools for the study of ecology and conservation of these animals are a priority; many extinct species also still attract considerable interest. Of the more than 5000 extant species, however, only about 20 have genomes assembled to chromosomes (with primates, rodents, and artiodactyls disproportionally overrepresented), with more than 10 of the 26 orders having no chromosome-level assemblies at all.

Recently greater than 50 de novo mammalian assemblies have been produced (more are inevitable); these, however, at best, are collections of subchromosomal-sized scaffolds. Moreover, several hundred are currently being assembled to scaffold level by individual projects or consortia such as Genome10K (Koepfli et al. 2015). Building a mammalian universal BAC set would be a greater challenge than in birds as mammalian genomes have more repetitive sequences and are about three times larger; thus, more BACs would be needed to achieve the same level of mapping resolution. On the other hand, the development of advanced mapping and sequencing techniques (e.g., Dovetail, BioNano or PacBio) will eventually provide an opportunity to replace RACA PCFs with longer and more complete subchromosomal-sized superscaffolds or sequence contigs requiring fewer BACs to anchor them to chromosomes. The availability of large numbers of high-quality mammalian BAC clone libraries from many species makes our approach more applicable to mammals than to any other animal group. If we add the fact that our avian BAC set is showing good success rates on lizard and turtle chromosomes (data not shown), building chromosomal assemblies for all vertebrate and ultimately all animal groups supported by universal collection of BACs is a realistic objective for the near future.

5.0 Methods

5.1 Avian genome assemblies, repeat masking, and gene annotations

The chicken (ICGSC Gallus_gallus 4.0) (Hillier 2004), zebra finch (WUGSC 3.2.4) (Warren et al. 2010), and turkey (TGC Turkey_2.01) (Dalloul et al. 2010) chromosome assemblies were downloaded from the UCSC Genome Browser (Kent et al. 2002). The collared flycatcher (FicAlb1.5) (Ellegren et al. 2012) genome was obtained from NCBI. Scaffold-based (N50 > 2 Mb) assemblies of the pigeon, falcon, and 16 additional avian genomes were provided by the Avian Phylogenomics Consortium (Zhang et al. 2014a). All sequences were repeat-masked using Window Masker (Morgulis et al. 2006) with the -sdust option and Tandem Repeats Finder (Benson 1999). Chicken gene (version of 27/04/2014) and repetitive sequence (version of 11/06/2012) annotations were downloaded from the UCSC Genome Browser (Rosenbloom et al. 2015). Chicken genes with a single ortholog in the human genome were extracted from Ensembl Biomart (v.74) (Kinsella et al. 2011).

5.2 Pairwise and multiple genome alignments, nucleotide evolutionary conservation scores, and CEs

Pairwise alignments using chicken and zebra finch chromosome assemblies as references and all other assemblies as targets were generated with LastZ (v.1.02.00) (Harris 2007) and converted into the UCSC “chains” and “nets” alignment formats with the Kent-library tools (Kent et al. 2003; Supplemental Methods). The evolutionary conservation scores and DNA CEs for all chicken nucleotides assigned to chromosomes were estimated using PhastCons (Siepel et al. 2005) from the multiple alignments of 21 avian genomes (Supplemental Methods). CNEs obtained from the alignments of 48 avian genomes were used (Farré et al. 2016).

5.3 Reference-assisted chromosome assembly of pigeon and falcon genomes

Pigeon and falcon PCFs were generated using RACA (Supplemental Methods; Kim et al. 2013) tool. We chose the zebra finch genome as the reference and chicken as the outgroup for the falcon based on the phylogenetic distances between the species (Jarvis et al. 2014). For the pigeon, both chicken as reference and zebra finch as outgroup and vice versa experiments were performed as the pigeon is phylogenetically distant from the chicken and zebra finch. Two rounds of RACA were done for both species. The initial run was performed using the following parameters: WINDOWSIZE=10 RESOLUTION=150000 MIN_INTRACOV_ PERC=5. Prior to the second run of RACA, we tested the scaffolds split during the initial RACA run using PCR amplification across the split intervals (see below) and adjusted the parameters accordingly (Supplemental Methods).

5.3 PCR testing of adjacent SFs

Primers flanking split SF joints within scaffolds or RACA predicted adjacencies were designed using Primer3 software (v.2.3.6) (Untergasser et al. 2012). To avoid misidentification of EBRs or chimeric joints, we selected primers only within the sequences that had high-quality alignments between the target and reference genomes and were found in adjacent SFs. Due to alignment and SF detection settings, some of the intervals between adjacent SFs could be >6 kb, and primers could not be chosen for a reliable PCR amplification. In such cases, we used CASSIS software (Baudet et al. 2010) and the underlying alignment results to narrow gaps between adjacent SFs where possible. Whole blood was collected aseptically from adult falcons and pigeons. DNA was isolated using DNeasy blood and tissue kit (Qiagen) following standard protocols. PCR amplification was performed according to the protocol described in the Supplemental Methods.

5.4 BAC clone selection

The chromosome coordinates of chicken (CHORI-261), turkey (CHORI-260), and zebra finch (TGMCBA) BAC clones in the corresponding genomes were extracted from NCBI clone database (Schneider et al. 2013). We removed all discordantly placed BAC clones (based on BAC end sequence [BES] mappings) following the NCBI definition of concordant BAC placement. Briefly, a BAC clone placement was considered concordant when the estimated BAC length in the corresponding avian genome is within [library average length ± 3 × standard deviation] and BAC BESs map to the opposite DNA strands in the genome assembly. Turkey and zebra finch BAC clone coordinates were translated into chicken chromosome coordinates using the UCSC Genome Browser liftOver tool (Kent et al. 2002) with a minimum ratio of remapped bases >0.1.

For each BAC clone mapped to the chicken chromosomes, various genomic features selected to estimate the probability of clones to hybridize with metaphase chromosomes in distant avian species were calculated (Supplemental Table S3) using a custom Perl script or extracted from gene, repetitive sequence, CE, and nucleotide conservation score files. The clones selected for mapping experiments were originally obtained from the BACPAC Resource Center at the Children's Hospital Oakland Research Institute and the zebra finch TGMCBa library (Clemson University Genomics Institute).

5.5 Classification tree

The classification tree was created in R (v.3.2.3) (R Core Team 2015) using the CART algorithm included in the rpart package (v.4.1-10) (https://cran.r-project.org/web/packages/rpart). We introduced an adjusted weight matrix setting: The cost of returning a false positive was twice as high as the cost of a false negative. The tree was visualized with rattle package (v.4.1.0) (Williams 2011).

5.6 Cell culture and chromosome preparation

Chromosome preparations were established from fibroblast cell lines generated from collagenase treatment of 5- to 7-d-old embryos or from skin biopsies. Cells were cultured at 40°C and 5% CO2 in Alpha MEM (Fisher), supplemented with 20% fetal bovine serum (Gibco), 2% Pen-Strep (Sigma), and 1% L-glutamine (Sigma). Chromosome suspension preparation followed standard protocols; in brief, mitostatic treatment with colcemid at a final concentration of 5.0 µg/mL for 1 h at 40°C was followed by hypotonic treatment with 75 mM KCl for 15 min at 37°C and fixation with 3:1 methanol:acetic acid.

5.7 Preparation of BAC clones for FISH

BAC clone DNA was isolated using the Qiagen miniprep kit prior to amplification and direct labelling by nick translation. Probes were labeled with Texas red-12-dUTP (Invitrogen) and FITC-fluorescein-12-UTP (Roche) prior to purification using the Qiagen nucleotide removal kit.

5.8 FISH

Metaphase preparations were fixed to slides and dehydrated through an ethanol series (2 min each in 2× SSC, 70%, 85%, and 100% ethanol at room temperature). Probes were diluted in a formamide buffer (Cytocell) with chicken hybloc (Insight Biotech) and applied to the metaphase preparations on a 37°C hotplate before sealing with rubber cement. Probe and target DNA were simultaneously denatured on a 75°C hotplate prior to hybridization in a humidified chamber for 72 h at 37°C. Slides were washed post-hybridization for 30 sec in 2× SSC/0.05% Tween 20 at room temperature and then counterstained using VECTASHIELD anti-fade medium with DAPI (Vector Labs). Images were captured using an Olympus BX61 epifluorescence microscope with cooled CCD camera and SmartCapture (Digital Scientific UK) system. In selected experiments, we used multiple hybridization strategies, making use of the Cytocell octochrome (eight-chamber) and multiprobe (24-chamber) devices. Briefly, labeled probes were air dried on to the device. Probes were rehybridized in standard buffer and applied to the glass slide (which was subdivided to correspond to the hybridization chambers), and FISH continued as above.

5.9 EBR detection and CNE density analysis

The multiple alignments of the chicken, zebra finch, flycatcher, pigeon, and falcon chromosome sequences were obtained using progressiveCactus (Paten et al. 2011) with default parameters. Pairwise synteny blocks were defined using the maf2synteny tool (Kolmogorov et al. 2014) at 100-, 300-, and 500-kb resolution. By using chicken as the reference genome, EBRs were detected and classified using the ad hoc statistical approach described previously (Farré et al. 2016). All well-defined (or flanking oriented PCFs) fusion and fission points were identified from pairwise alignments with the chicken genome. Only the EBRs ≤100 kb were used for the CNE analysis. EBRs <1 kb were extended ±1 kb. For each EBR, we defined two windows upstream of (+1 and +2) and two downstream from (−1 and −2) the same size as the EBR. We calculated the fraction of bases within CNEs in each EBR site and the upstream and downstream windows. Differences in CNE densities were tested for significance using the Kruskall-Wallis test followed by Mann-Whitney U test.

5.10 Comparing CNE densities in EBRs and msHSBs

Chicken chromosomes (excluding GGA16, W, and Z) were divided into 1-kb nonoverlapping intervals. Only windows with >50% of their bases with chicken sequence data available were used in this analysis. All intervals were assigned either to msHSBs >1.5 Mb (Farré et al. 2016); to avian EBRs flanking fusions, fissions, and intrachromosomal EBR; and to the intervals found in the rest of the chicken genome. We estimated the average CNE density for each window type and also the distance, in number of 1-kb windows, between each window with the lowest CNE density (0 bp) and the nearest window with the average msHSB CNE density or higher. CNE densities were obtained using BEDtools (v.2.20-1) (Quinlan and Hall 2010). Differences in distances between the two window types in msHSBs and EBRs were tested for significance using the Kruskall-Wallis test followed by Mann-Whitney U test.

5.11 Densities of TEs in falcon intrachromosomal EBRs, fusions, and fissions

The TE scaffold coordinates reported by Zhan et al. (2013) were translated to falcon chromosome coordinates using a custom Perl script. The densities of TEs (>100 bp on average in the EBR- or non-EBR-containing nonoverlapping 10-kb genome intervals) were compared for the falcon lineage–specific interchromosomal EBRs, EBRs flanking fusion and fission events, and the rest of the genome as previously described (Elsik et al. 2009; Larkin et al. 2009; Groenen et al. 2012; Farré et al. 2016).

5.12 Data access

The falcon and pigeon chromosome assemblies from this study have been submitted to DDBJ/ENA/GenBank (https://www.ncbi. nlm.nih.gov/genbank/) under the accessions numbers MLQY00000000 and MLQZ00000000, respectively. Visualizations of falcon and pigeon genome assemblies are available from the Evolution Highway comparative chromosome browser (http ://eh-demo.ncsa.uiuc.edu/birds) and our UCSC Genome Browser hub (http://sftp.rvc.ac.uk/rvcpaper/birdsHUB/hub.txt).

The supplementary data for this paper can be found on the USB provided with this thesis, or alternatively from *Genome Research* (http://genome.cshlp.org/content/suppl/2017/04/06/gr.213660.116.DC1).

6.0 Acknowledgments

We thank Dr. Yu-Mei Chang for the assistance with statistical analyses. This work was supported in part by the Biotechnology and Biological Sciences Research Council (BB/K008226/1 and BB/ J010170/1 to D.M.L. and BB/K008161/1 to D.K.G.).

7.0 References

Andersson L, Georges M. 2004. Domestic-animal genomics: deciphering the genetics of complex traits. Nat Rev Genet 5: 202–212.

Baudet C, Lemaitre C, Dias Z, Gautier C, Tannier E, Sagot MF. 2010. Cassis: detection of genomic rearrangement breakpoints. Bioinformatics 26: 1897–1898.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573–580.

Branco MR, Pombo A. 2006. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS Biol 4: e138.

Burt DW. 2002. Origin and evolution of avian microchromosomes. Cytogenet Genome Res 96: 97–112.

Burt DW, Bruley C, Dunn IC, Jones CT, Ramage A, Law AS, Morrice DR, Paton IR, Smith J, Windsor D, et al. 1999. The dynamics of chromosome evolution in birds and mammals. Nature 402: 411–413.

Bush GL, Case SM, Wilson AC, Patton JL. 1977. Rapid speciation and chromosomal evolution in mammals. Proc Natl Acad Sci 74: 3942–3946. Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Blomberg Le A, Bouffard P,

Burt DW, Crasta O, Crooijmans RP, et al. 2010. Multi-platform next- generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. PLoS Biol 8: e1000475.

Deakin JE, Ezaz T. 2014. Tracing the evolution of amniote chromosomes. Chromosoma 123: 201–216.

Derjusheva S, Kurganova A, Habermann F, Gaginskaya E. 2004. High chromosome conservation detected by comparative chromosome painting in chicken, pigeon and passerine birds. Chromosome Res 12: 715–723.

Driscoll CA, Macdonald DW, O'Brien SJ. 2009. From wild animals to domestic pets, an evolutionary view of domestication. Proc Natl Acad Sci 106: 9971–9978.

Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, Kawakami T, Kunstner A, Makinen H, Nadachowska-Brzyska K, Qvarnstrom A, et al. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature 491: 756–760.

Elsik CG, Tellam RL, Worley KC. 2009. The genome sequence of a taurine cattle: a window to ruminant biology and evolution. Science 324: 522–528.

Fang X, Nevo E, Han L, Levanon EY, Zhao J, Avivi A, Larkin D, Jiang X, Feranchuk S, Zhu Y, et al. 2014. Genome-wide adaptive complexes to underground stresses in blind mole rats Spalax. Nat Commun 5: 3966.

Farré M, Narayan J, Slavov GT, Damas J, Auvil L, Li C, Jarvis ED, Burt DW, Griffin DK, Larkin DM. 2016. Novel insights into chromosome evolution in birds, archosaurs, and reptiles. Genome Biol Evol 8: 2442–2451.

Ferguson-Lees J, Christie DA. 2005. Raptors of the world. Princeton University Press, Princeton, NJ.

Fontdevila A, Ruiz A, Ocaña J, Alonso G. 1982. Evolutionary history of Drosophila buzzatii. II. How much has chromosomal polymorphism changed in colonization? Evolution 36: 843–851.

Griffin DK, Robertson LBW, Tempest HG, Skinner BM. 2007. The evolution of the avian genome as revealed by comparative molecular cytogenetics. Cytogenet Genome Res 117: 64–77.

Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens H-J, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491: 393–398.

Hansmann T, Nanda I, Volobouev V, Yang F, Schartl M, Haaf T, Schmid M. 2009. Cross-species chromosome painting corroborates microchromosome fusion during karyotype evolution of birds. Cytogenet Genome Res 126: 281–304.

Harris RS. 2007. "Improved pairwise alignment of genomic DNA." PhD thesis, The Pennsylvania State University, State College, PA.

Hillier L. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432: 695–716.

Hu X, Gao Y, Feng C, Liu Q, Wang X, Du Z, Wang Q, Li N. 2009. Advanced technologies for genomic analysis in farm animals and its application for QTL mapping. Genetica 136: 371–386.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346: 1320–1331.

Kasai F, O'Brien PC, Ferguson-Smith MA. 2012. Reassessment of genome size in turtle and crocodile based on chromosome measurement by flow karyotyping: close similarity to chicken. Biol Lett 8: 631–635.

Kasai F, O'Brien PC, Ferguson-Smith MA. 2013. Afrotheria genome; overestimation of genome size and distinct chromosome GC content revealed by flow karyotyping. Genomics 102: 468–471.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res 12: 996–1006.

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci 100: 11484–11489.

Kim J, Larkin DM, Cai Q, Asan, Zhang Y, Ge R-L, Auvil L, Capitanu B, Zhang G, Lewin HA, et al. 2013. Reference-assisted chromosome assembly. Proc Natl Acad Sci 110: 1785–1790.

Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, AlmeidaKing J, Staines D, Derwent P, Kerhornou A, et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database 2011: bar030.

Koepfli K-P, Paten B, Scientists tGKCo, O'Brien SJ. 2015. The Genome 10K Project: a way forward. Ann Rev Anim Biosci 3: 57–111.

Kolmogorov M, Raney B, Paten B, Pham S. 2014. Ragout-a reference-assisted assembly tool for bacterial genomes. Bioinformatics 30: i302–309.

Larkin DM, Prokhorovich MA, Astakhova NM, Zhdanova NS. 2006. Comparative mapping of mink chromosome 8p: in situ hybridization of seven cattle BAC clones. Anim Genet 37: 429–430.

Larkin DM, Pape G, Donthu R, Auvil L, Welge M, Lewin HA. 2009. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. Genome Res 19: 770–777.

Larkin DM, Daetwyler HD, Hernandez AG, Wright CL, Hetrick LA, Boucek L, Bachman SL, Band MR, Akraiko TV, Cohen-Zinder M, et al. 2012. Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. Proc Natl Acad Sci 109: 7693–7698.

Lewin HA, Larkin DM, Pontius J, O'Brien SJ. 2009. Every genome sequence needs a good map. Genome Res 19: 1925–1928.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476–482.

Loh W-Y. 2011. Classification and regression trees. WIREs Data Mining Knowl Discov 1: 14–23.

Mak ACY, Lai YYY, Lam ET, Kwok T-P, Leung AKY, Poon A, Mostovoy Y, Hastie AR, Stedman W, Anantharaman T, et al. 2016. Genome-wide structural variation detection by genome mapping on nanochannel arrays. Genetics 202: 351–362.

Makunin AI, Kichigin IG, Larkin DM, O'Brien PC, Ferguson-Smith MA, Yang F, Proskuryakova AA, Vorobieva NV, Chernyaeva EN, O'Brien SJ, et al. 2016. Contrasting origin of B chromosomes in two cervids (Siberian roe deer and grey brocket deer) unravelled by chromosomespecific DNA sequencing. BMC Genomics 17: 618.

Meyerson M, Gabriel S, Getz G. 2010. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11: 685–696.

Modi WS, Romanov M, Green ED, Ryder O. 2009. Molecular cytogenetics of the California condor: evolutionary and conservation implications. Cytogenet Genome Res 127: 26–32.

Morgulis A, Gertz EM, Schaffer AA, Agarwala R. 2006. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol 13: 1028–1040.

Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L, et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science 309: 613–617.

Nishida C, Ishijima J, Kosaka A, Tanabe H, Habermann FA, Griffin DK, Matsuda Y. 2008. Characterization of chromosome structures of Falconinae (Falconidae, Falconiformes, Aves) by chromosome painting and delineation of chromosome rearrangements during their differentiation. Chromosome Res 16: 171–181.

Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. 2011. Cactus: algorithms for genome multiple sequence alignment. Genome Res 21: 1512–1528.

Price TD. 2002. Domesticated birds as a model for the genetics of speciation by sexual selection. Genetica 116: 311–327.

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res 26: 342–350.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Rhoads A, Au KF. 2015. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13: 278–289.

Romanov MN, Dodgson JB, Gonser RA, Tuttle EM. 2011. Comparative BAC-based mapping in the white-throated sparrow, a novel behavioral genomics model, using interspecies overgo hybridization. BMC Res Notes 4: 211.

Romanov MN, Farré M, Lithgow PE, Fowler KE, Skinner BM, O'Connor R, Fonseka G, Backström N, Matsuda Y, Nishida C, et al. 2014. Reconstruction of gross avian genome structure, organization and evolution suggests that the chicken lineage most closely resembles the dinosaur avian ancestor. BMC Genomics 15: 1–18.

Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. 2015. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res 43: D670–D681.

Schmid M, Smith J, Burt DW, Aken BL, Antin PB, Archibald AL, Ashwell C, Blackshear PJ, Boschiero C, Brown CT, et al. 2015. Third report on chicken genes and chromosomes 2015. Cytogenet Genome Res 145: 78–179.

Schneider VA, Chen HC, Clausen C, Meric PA, Zhou Z, Bouk N, Husain N, Maglott DR, Church DM. 2013. Clone DB: an integrated NCBI resource for clone-associated data. Nucleic Acids Res 41: D1070–D1078.

Shapiro MD, Kronenberg Z, Li C, Domyan ET, Pan H, Campbell M, Tan H, Huff CD, Hu H, Vickrey AI, et al. 2013. Genomic diversity and evolution of the head crest in the rock pigeon. Science 339: 1063–1067.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–2050.

Skinner BM, Griffin DK. 2012. Intrachromosomal rearrangements in avian genome evolution: evidence for regions prone to breakpoints. Heredity 108: 37–41.

Stringham SA, Mulroy EE, Xing J, Record D, Guernsey MW, Aldenhoven JT, Osborne EJ, Shapiro MD. 2012. Divergence, convergence, and the ancestry of feral populations in the domestic rock pigeon. Curr Biol 22: 302–308.

Tamazian G, Dobrynin P, Krasheninnikova K, Komissarov A, Koepfli KP, O'Brien SJ. 2016. Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences. GigaScience 5: 38.

Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, Sarkar D, Valouev A, Churas C, Kidd JM, et al. 2010. High-resolution human genome structure by single-molecule analysis. Proc Natl Acad Sci 107: 10848–10853.

R Core Team. 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.Rproject.org/.

Tucker V, Cade T, Tucker A. 1998. Diving speeds and angles of a gyrfalcon (Falco rusticolus). J Exp Biol 201: 2061–2070.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3-new capabilities and interfaces. Nucleic Acids Res 40: e115.

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, Searle S, White S, Vilella AJ, Fairley S, et al. 2010. The genome of a songbird. Nature 464: 757–762.

Williams G. 2011. Data mining with Rattle and R: the art of excavating data for knowledge discovery. Springer Science & Business Media, New York. Zhan X, Pan S, Wang J, Dixon A, He J, Muller MG, Ni P, Hu L, Liu Y, Hou H,

et al. 2013. Peregrine and saker falcon genome sequences provide in-

sights into evolution of a predatory lifestyle. Nat Genet 45: 563–566. Zhang G, Li B, Li C, Gilbert MT, Jarvis ED, Wang J, Avian Genome C. 2014a. Comparative genomic data of the Avian Phylogenomics Project.

GigaScience 3: 26.

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold

MJ, Meredith RW, et al. 2014b. Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346: 1311–1320.

Appendix 6: A QTL for number of teats shows breed specific effects on number of vertebrae in pigs: Bridging the gap between molecular and quantitative genetics

Maren van Son, Marcos S Lopes, **Henry J Martell**, Martijn F L Derks, Lars Erik Gangsei, Jorgen Kongsro, Mark N Wass, Eli H Grindflek, and Barbara Harlizius, 2018. "A QTL for number of teats shows breed specific effects on number of vertebrae in pigs: Bridging the gap between molecular and quantitative genetics". (this paper is currently under review at the journal *Frontiers in Genetics*)

My contribution to the work was as follows:

1. Established a pipeline to analyse whole genome sequencing data within QTLs associated with sample groups
2. Performed the analysis and mapping of human LTBP2 promoters to the pig genome
3. Performed the analysis of the whole genome sequencing data for the identified QTLs, and identification of potentially functional mutations
4. Helped to write the manuscript with all other authors

1.0 Abstract

Modern breeding schemes for livestock species accumulate a large amount of genotype and phenotype data which can be used for genome-wide association studies. Many chromosomal regions harbouring effects on quantitative traits have been reported from these studies, but the underlying causative mutations remain mostly undetected. In this study, we combine large genotype and phenotype data available from a commercial pig breeding scheme for three different breeds (Duroc, Landrace, and Large White) to pinpoint functional variation for a region on porcine chromosome 7 affecting number of teats (NTE). Our results show that refining trait definition by counting number of vertebrae (NVE) and ribs (RIB) helps to reduce noise from other genetic variation and increases heritability from 0.28 up to 0.62 NVE and 0.78 RIB in Duroc. However, in Landrace, the effect of the same QTL on NTE mainly affects NVE and not RIB, which is reflected in reduced heritability for RIB (0.24) compared to NVE (0.59). Further, differences in allele frequencies and accuracy of rib counting influence genetic parameters. Correction for the top SNP does not detect any other QTL effect on NTE, NVE, or RIB in Landrace or Duroc. At the molecular level, haplotypes derived from 660K SNP data detects a core haplotype of seven SNPs in Duroc. Sequence analysis of 16 Duroc animals shows that two functional mutations of the Vertnin (VRTN) gene known to increase number of thoracic vertebrae (ribs) reside on this haplotype. In Landrace, the linkage disequilibrium extends over a region of more than 3 Mb also containing both VRTN mutations. Here, other modifying loci are expected to cause the breed-specific effect. Additional variants found on the wildtype haplotype surrounding the VRTN region in all sequenced Landrace animals point towards breed specific differences which are expected to be present also across the whole genome. This Landrace specific haplotype contains two missense mutations in the ABCD4 gene, one of which is expected to have a negative effect on the protein function. Together, the integration of largescale genotype, phenotype and sequence data shows exemplarily how population parameters are influenced by underlying variation at the molecular level.

2.0 Introduction

Number of teats (NTE) in pigs is a highly heritable trait and shows considerable variation across breeds (8-21) (Rohrer and Nonneman, 2017) and also within breeds (e.g. 12-20 in Landrace) (Lopes et al., 2014). Several genome-wide association studies (GWAS) have shown that NTE is a polygenic trait influenced by many different quantitative trait loci (QTL) scattered over nearly all chromosomes of the pig genome (Duijvesteijn et al., 2014; Lopes et al., 2014; Arakawa et al., 2015; Verardo et al., 2016; Yang et al., 2016; Tang et al., 2017; Uzzaman et al., 2018). Among these, a QTL on Sus scrofa chromosome 7 (SSC7) has been identified showing a large effect on NTE in several commercial breeding lines and crosses (e.g. Duijvesteijn et al., 2014; Rohrer and Nonneman, 2017; Dall'Olio et al., 2018). In the same region, a QTL was detected affecting number of ribs (RIB) (Mikawa et al., 2011).

After fine mapping, Mikawa et al. (2011) identified a new transcript named Vertnin (VRTN) as being the most promising candidate gene. Several mutations have been detected in the region including a SNP in the promotor region (g.19034A>C) and an insertion of a PRE1-SINE element of 291-basepairs (g.20311_20312ins291) in the first intron of the VRTN gene (Mikawa et al., 2011; Ren et al., 2012; Fan et al., 2013). Just recently, VRTN has been characterized as a novel transcription factor affecting vertebra development (Duan et al., 2018). Duan et al. (2018) showed that these two mutations increase VRTN expression in an early embryonic stage in an additive way. Together these two mutations increase the number of thoracic vertebrae by 1 in the homozygous state. The effect of the PRE1-insertion on RIB has been validated in different commercial crossbred (Rohrer et al., 2015) and purebred lines (Dall'Olio et al., 2018).

However, after correction for the PRE1 insertion allele, an additional negative effect on lumbar vertebrae was detected with a SNP located 500kb further proximal from VRTN and not in high linkage disequilibrium (LD) with the PRE1 insertion in crossbreds (Rohrer et al., 2015). Yang et al. (2016) also observed a pleiotropic effect of the insertion allele on number of vertebrae (NVE) and NTE in Chinese Erhualian, three European commercial purebred populations (Duroc, Landrace, Large White) and an intercross. However, a new candidate gene LTBP2 (latent transforming growth factor binding protein 2) has been pinpointed by Zhang et al. (2016) for RIB in 596 Large-White x Chinese Minzhu F2 intercrosses where the VRTN insertion was not segregating but only the promoter SNP g.19034A>C. Finally, Park et al. (2017) describe a missense mutation in LTBP2 at c.4481A>C associated with thoracic vertebrae number in 1,105 F2 animals of a Landrace cross with Korean native pigs where also the VRTN insertion is increasing thoracic vertebrae number independently.

In this study, we analysed an extended data set of three commercial pig breeds linking genotype data of 20,366 Large White (LW), 23,398 Landrace (L), and 10,044 Duroc (D) animals with phenotypic data on NTE. In addition, NVE, and RIB data were scored from computer tomography (CT) images on 2,756 L and 2,961 D animals. We show that scoring phenotypes closer to the molecular basis of the observed variation (NVE or RIB vs. NTE) increases population genetic parameters such as heritability and explained genetic variance considerably. Furthermore, a detailed analysis combining medium and high-density SNP data with whole-genome sequence (WGS) data with functional parameters of mutations has been performed. Our results show that the two functional mutations analysed by Duan et al. (2018) are in high LD with the most significant SNPs from the GWAS studies in all breeds, however, the phenotypic effect on NVE depends on the genetic background.

3.0 Material and Methods

3.1 Ethics statement

The data used for this study were obtained as part of routine data recording in a commercial breeding program. Samples collected for DNA extraction were only used for routine diagnostic purposes of the breeding program. Data recording and sample collection were conducted strictly in line with the rules given by Dutch and Norwegian Animal Research Authorities.

3.2 Data

In this study, data from the three pig populations LW (Large White-based), L (Landrace) and D (Duroc) were evaluated. The LW population was located in Dutch nucleus farms and were born between 2006 and 2017 (data obtained by Topigs Norsvin, the Netherlands). The Norwegian L and D populations were located in Norwegian nucleus farms and at a boar testing station (Hamar, Norway) and were born between 2010 and 2017 (data obtained by Norsvin, Norway).

The traits evaluated were NTE, recorded at birth on both males and females from all three populations, and NVE and RIB retrieved from CT images only on males from the L and D populations at 120 Kg of live weight. The traits NVE and RIB were recorded using the CT scanner GE Healthcare LightSpeed 32 VCT, and the settings used were 120kV, slice thickness 1.25 mm and dynamic mA (400-500 mA) adjusting for object thickness. Prior to scanning, the boars were sedated using Azaperone (Stresnil Vet®, Janssen-Cilag Ltd., Buckinghamshire, UK), which was injected intramuscularly. The whole skeleton was segmented out by applying a threshold value at 200 Hounsfield units to the full CT volumes.

Individual vertebras and ribs were segmented and identified in accordance with Gangsei & Kongsro (2016). Furthermore, a visual control of the total number of vertebras was performed (Figure 1). This number includes the total number of cervical, thoracic and lumbar vertebras, omitting sacrum and coccyx vertebras. In pigs, the number of cervical vertebras is fixed at 7, whereas the number of thoracic and lumbar vertebras might vary (King and Roberts, 1960). The trait RIB was counted on both right and left sides of each animal. However, due to the similarity of the results when analyzing right and left RIB separately, in this study we will show only the results of the analyses using RIB from the right-hand side.

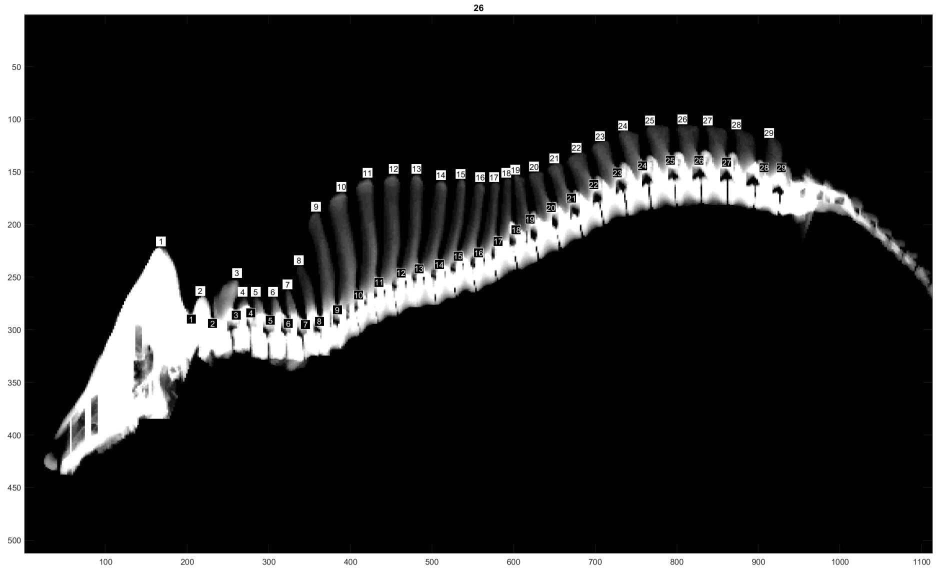


Figure 1: Computer tomography image illustrating the NVE recording.

For each trait, two datasets from each population were used: ALL and GENOTYPED (See Table 1 for descriptive statistics). The dataset ALL consisted of all genotyped animals and their contemporaries that had phenotypes (275,513 LW, 313,475 L and 12,672 D for NTE, 2,756 L and 2,961 D for NVE, and 2,653 L and 2,874 D for RIB). Using ALL, the phenotypes were pre-corrected for all non-genetic effects. The pre-corrected phenotype was used as the response variable in further analysis. The non-genetic effects were estimated by the pedigree-based linear models 1 (NTE) and 2 (NVE and RIB) in ASReml v3.0 (Gilmour et al., 2009):

$$\mathrm{NTEijkl} = \mu + \mathrm{sexi} + \mathrm{hyj} + \mathrm{ak} + \mathrm{litterl} + \mathrm{eijkl} \qquad (1)$$

where NTEijkl was the number of teats of the k animal; μ is the overall mean, sexi was the fixed effect of sex i, hyj was the fixed effect of the herd-year j of birth, ak was the random additive genetic effect of animal k, litterl was the random effect of litter l and eijkl was the random residual effect. The vector of additive genetic effects was assumed to be distributed as ~N(0,Aσ_a^2), which accounted for the (co)variances between animals due to relationships by formation of an A matrix (pedigree-based numerator relationship matrix), σ_a^2 being the additive genetic variance. The vector of litter effects was assumed to be distributed as ~N(0,Iσ_l^2), with I being an identity matrix and σ_l^2 the litter variance. The vector of residual effects was assumed to be distributed as ~N(0,Iσ_e^2), σ_e^2 being the residual variance.

$$\text{NVE/RIBijk} = \mu + \text{yeari} + \text{farmj} + \text{ak} + \text{eijk} \qquad (2)$$

where NVE/RIBijk was either the number vertebrae or ribs of the k animal; μ is the overall mean, yeari was the fixed effect of year I of birth, farmj was the fixed effect of the herd j of birth, ak and eijkl was as described above for model (1).

The dataset GENOTYPED was a subset of ALL consisting of all animals that had both phenotypes and genotypes (20,366 LW, 23,398 L and 10,044 D for NTE, 1,873 L and 2,384 D for NVE, and 1,802 L and 2,322 D for RIB). This dataset was used to perform the GWAS.

Table 1: Summary statistics.

Trait [1]	**Population**	**Dataset** [2]	**N** [3]	**Mean**	**SD** [4]
NTE	Large White	ALL	275,513	15.30	1.08
		GENOTYPED	20,366	15.73	1.02
	Landrace	ALL	313,475	15.84	1.03
		GENOTYPED	23,398	15.99	1.04
	Duroc	ALL	12,672	12.93	1.05
		GENOTYPED	10,044	12.93	1.04
NVE	Landrace	ALL	2,756	29.78	0.53
		GENOTYPED	1,873	29.75	0.54
	Duroc	ALL	2,961	28.72	0.60
		GENOTYPED	2,384	28.71	0.60
RIB	Landrace	ALL	2,653	15.47	0.71
		GENOTYPED	1,802	15.48	0.71
	Duroc	ALL	2,874	14.57	0.62
		GENOTYPED	2,322	14.57	0.61

[1] NTE: number of teats, NVE: total number of vertebrae, RIB: number of ribs on the right side of the animal. [2] ALL: the whole population used in the pre-adjustment of the phenotypes, which includes the animals from GENOTYPED and their contemporaries; GENOTYPED: genotyped and phenotyped animals used for the GWAS. [3] N: number of animals. [4] SD: standard deviation of the traits in each dataset of each population.

3.3 Medium Density Genotypes

All animals from the GENOTYPED dataset were genotyped using a medium density SNP chip. Genotyping was performed at CIGENE (University of Life Sciences, Ås, Norway) and at GeneSeek (Lincoln, NE, USA), mainly using the (Illumina) GeneSeek custom 80K SNP chip (Lincoln, NE, USA). However, a small part of the animals from the three populations were genotyped using the (Illumina) GeneSeek custom 50K SNP chip (Lincoln, NE, USA) and the Illumina Porcine SNP60 Beadchip (Illumina, San Diego, CA, USA).

Quality control consisted of excluding SNPs with GenCall<0.15 (Illumina Inc., 2005), call rate <0.95, minor allele frequency <0.01, strong deviation from Hardy-Weinberg equilibrium (χ^2>600), SNPs located on sex chromosomes and unmapped SNPs. The positions of the SNPs were based on the Sscrofa11.1 assembly of the reference genome. Animals with frequency of missing genotypes ≥0.05 would be removed from the dataset. However, all genotyped animals had a frequency of missing genotypes <0.05 and were therefore kept for further analyses.

After quality control, the remaining missing genotypes of the animals genotyped with the (Illumina) GeneSeek custom 80K SNP chip were imputed within population using Fimpute v2.2 (Sargolzaei et al., 2014). At the same time, the animals genotyped with the other two chips had their genotypes imputed to the set of SNPs on the (Illumina) GeneSeek custom 80K SNP chip that passed the quality control. After quality control and imputation, 50,717 SNPs for LW, 44,961 SNPs for L and 43,309 SNPs for D were available and were used in the imputation towards the high density SNP chip.

3.4 High Density Genotypes

Genotyping of high density genotypes was also performed at CIGENE (University of Life Sciences, Ås, Norway) and GeneSeek (Lincoln, NE, USA). In total, 290 LW, 415 L and 140 D animals from the GENOTYPED dataset were in addition genotyped using the Axiom porcine 660K array from Affymetrix (Affymetrix Inc., Santa Clara, CA, USA). These animals were the most influential sires from each population (sires with the largest number of offspring in the GENOTYPED dataset). Quality control of 660K array was as described above for the medium density genotypes. The imputation from 80K genotypes towards 660K genotypes was performed within population using Fimpute v2.2 (Sargolzaei et al., 2014). After quality control and imputation, there were 527,186 SNPs for LW, 462,414 SNPs for L and 441,288 SNPs for D, which were used in the GWAS.

3.5 Haplotype Analysis and Recombinant Identification

Beagle v.4.1 (Browning and Browning, 2016) was used to phase the medium and high density genotype data. Bcftools v1.5-28-ge9ec882 (Li et al., 2009) was used to extract the phased genotypes in the QTL region SSC7: 97-98 Mb. Next, we build haplotypes in the region SSC7: 97-98 Mb using PyVCF (Casbon, 2012) and report recombinant animals for animals that carry a different haplotype compared to the parent animals.

3.6 Sequence Data

Analysis of WGS data was done to construct a sequence level SNP dataset for the QTL region (SSC7: 85-105 Mb) in L and D. WGS data from 24 L and 23 D boars were available for this purpose. The boars were previously frequently used AI boars and all of them were part of the GENOTYPED dataset. DNA extraction and the sequencing procedure for whole genome re-sequencing is described in detail in van Son et al. (2017). The reads were 2x100 basepair paired-end reads and mapped to Sus scrofa build 11.1, duplicated marked and indexed using the speedseq align module available in SpeedSeq (Chiang et al., 2015). Freebayes v.1.3.1 (Garrison and Marth, 2012) was used to detect variants in the QTL region, using a minimum alternate allele count of 2, and identified 179,241 and 190,260 putative variants in L and D, respectively.

Filtering of variants was done by VCFtools v.0.1.14 (Danecek et al., 2011) and SAMtools bcftools v.1.3.1 (Li et al., 2009). The filtering criteria were that both reference and alternate allele must be present on both strands, a minimum quality score of 25, a mapping quality of >10 for both alleles at a SNP position and a sequencing depth >6 and <2000. Variants with more than one unique non-reference allele were removed for imputation purposes and a distance of at least 4 and 10 basepairs to the next insertion/deletion was applied for SNPs and indels, respectively. This resulted in a total of 80,392 and 89,725 high quality SNPs available for imputation in L and D, respectively. Newly detected SNPs have been deposited to EVA with accession number PRJEB27233.

3.7 Imputation to Sequence

The WGS data SNPs from the 24 L and 23 D boars in the SSC7 QTL region were phased within breed using Beagle v.4.1 (Browning and Browning, 2016). Prior to imputation, the 660K array SNPs described in "high density genotypes" were compared with the WGS data SNPs using conform-gt (Browning and Browning, 2016) to remove array SNPs that were not present in the WGS data and to adjust corresponding SNPs to match chromosome strand and allele order. In L, 954 of the 4189 array SNPs in the QTL region were removed by conform-gt because they were not in the reference dataset or because the chromosome strand was unknown, whereas in D, 744 of 4539 SNPs were removed. The rest of the 660K array SNPs were included for imputation to sequence level using Beagle v.4.1 and default settings.

3.8 GWAS

A single-SNP GWAS was performed with the GENOTYPED dataset within each population using the following linear animal model in GCTA (Yang et al., 2011; Yang et al., 2014):

$$y*k = \mu + X\hat{\beta} + uk + ek \qquad (3)$$

where y*k was the pre-corrected phenotype of the k animal; μ the average of the pre-corrected phenotype; X was the genotype (0, 1, 2) of the k animal for the evaluated SNP; $\hat{\beta}$ was the unknown allele substitution effect of the evaluated SNP; uk was the random additive genetic effect, being that the vector of additive genetic effects was assumed to be distributed as ~N(0,Gσ_a^2), which accounted for the (co)variances between animals due to relationships by formation of an G matrix (genomic numerator relationship matrix build using the imputed 660K genotypes), σ_a^2 being the additive genetic variance; and ek was the random residual effect which was assumed to be distributed as ~N(0,Iσ_e^2).

The genetic variance explained by a SNP (σ_snp^2=2pqα2) was estimated based on the allele frequencies (p and q) and the estimated allele substitution effect (α). The proportion of phenotypic variance explained by the SNP was defined as σ_snp^2/σ_P^2, where σ_P^2 is total phenotypic variance (sum of the additive and residual variances) which was estimated based on model (3) without a SNP effect. Significant SNPs and QTL were detected using a p-value < 1.0 x 10-8. After the GWAS using the imputed 660K data, we extracted all SNPs from the imputed WGS data that are located 5 Mb upstream and downstream the most significant 660K SNP for the traits NVE and RIB. With this data, we performed WGS association analyses for NVE and RIB aiming to identify stronger association with these phenotypes. These analyses were also performed applying model (3) in GCTA (Yang et al., 2011; Yang et al., 2014). LD as measured by r2 was calculated between SNPs using Plink 1.9 (Purcell et al., 2007).

3.9 Identification of Functional Variants from Sequencing Data

The variants identified by WGS data analysis were used to find potentially functional mutations. All L and D animals with WGS and NVE data (34 animals) were grouped based on their genotype for the VRTN insertion g.20311_20312ins291. Three groups were created: Homozygous Wild Type (7 animals) (wt/wt), Heterozygous Insertion (18 animals) (wt/ins), and Homozygous Insertion (9 animals) (ins/ins). The filtered SNPs generated with sequence data was then used to search for variants that were overrepresented in the different groups.

Variant calls for all samples were annotated with the Ensembl variant effect predictor (McLaren et al., 2016), version 90.6, using Ensembl release 90 and Sus scrofa genome build 11.1. Known regulatory regions of the human LTBP2 gene (Davis et al., 2014) were mapped from GRCh37 to GRCh38, using the UCSC liftOver tool (Kent et al., 2002). These mapped coordinates were used as input for the Ensembl comparative genomics tool to identify the corresponding regions in build 11.1 of the Sus scrofa genome (Zerbino et al., 2018). The corresponding pig genome sequences were then used as input to EMBOSS Needle to generate local alignments and percentage identities for the promoters (Rice et al., 2000).

3.10 Genotyping of VRTN insertion variant

The 291-basepair insertion g.20311_20312ins291 (GenBank accession number AB554652, position 7:97615880), was genotyped using primers and PCR conditions reported by Yang et al. (2016). The PCR products were separated by 2% agarose gel electrophoresis and the genotypes were visually recorded by inspection of amplicon length. The insertion allele was represented by a 411-basepair amplicon whereas the wild type allele was 120 basepairs. One 96-well plate was filled with samples for genotyping, out of which 82 had phenotypes and the rest of the animals were parents and grandparents used to confirm observed genotypes.

Visualization of the genomic region containing the VRTN insertion was done in the WGS animals by the Integrative Genomics Viewer (IGV) software (Robinson et al., 2011; Thorvaldsdóttir et al., 2012). This allowed us to genotype the subset of 18/15 L/D animals with NVE phenotypes using their sequence data (see Supplementary Figure S1 for examples on how this was done). Only animals showing at least two forward and two reverse reads for the insertion and the wildtype allele were genotyped as heterozygous (wt/ins). For homozygous wt/wt or ins/ins genotypes, only animals with at least 5 reads covering the insertion point were included. The insertion is supported by both split-reads and discordantly mapped pairs (Supplementary Figure S1) as well as reduced coverage of aligned sequences. Genotypes of WGS animals were also derived by IGV for the SNP promoter mutation in VRTN (g.19034A>C, rs709317845, position 7:97614602) (Fan et al., 2013) and the missense mutation in LTBP2 (c.4481A>C, rs322260921, position 7:97751432) (Park et al., 2017).

4.0 Results

A schematic overview of the approach linking large-scale phenotypic, genetic and genomic data is given in Figure 2.

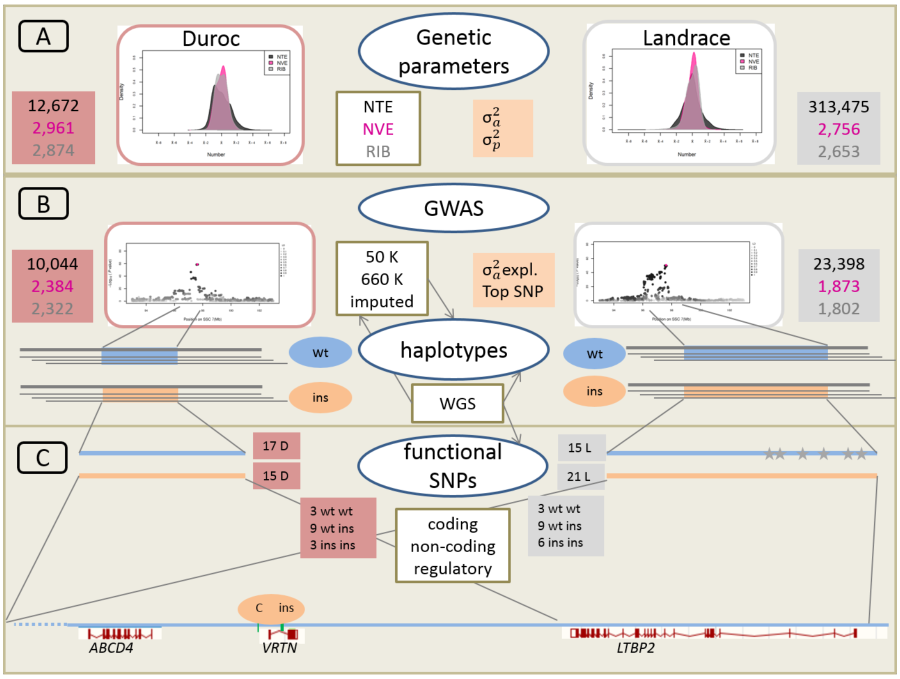


Figure 2: Schematic overveiw of the approach linking large-scale phenotypic, genetic, and genomic data. A. Phenotypic distribution and estimation of additive genetic variance (σ2a) and phenotypic variance (σ2p) for Duroc and Landrace breeds for number of teats (NTE), number of vertebrae (NVE), and number of ribs (RIB). Number of animals for each population and each trait is indicated in red and grey boxes for Duroc and Landrace, respectively. B. Genome wide association analyses (GWAS) with medium density (50K) and imputed high density (660K) SNP sets as well as imputed from whole genome sequence (WGS) data. Number of animals with 50K data in red and grey boxes. Estimation of haplotypes from 660K SNP data and search for identical core haplotype associated with two functional variants (ins) in the Vertnin (VRTN) gene compared to the wildtype (wt) haplotypes. C. Search for potentially functional SNPs and indels in coding (missense, nonsense) and non-coding (splice sites) regions in the core haplotypes of 16 D (3 wt wt, 9 wt ins, 3 ins ins) and 18 L (3 wt wt, 9 wt ins, 6 ins ins) animals for the wt and ins alleles. Mapping of regulatory enhancer sequences in LTBP2 from human genome. Identification of SNP and indel variants only present on each of the 15 wt haplotype in L breed indicated as grey stars. Underneath annotation of the relevant candidate genes from Sus scrofa reference genome build 11.1, and position of the two causative variants in the VRTN gene.

4.1 Heritabilities and GWAS

Table 1 shows the descriptive statistics for all traits and breeds. The mean value for NTE is around 3 teats higher in the LW and L breeds compared to D (16 vs. 13

for genotyped animals). The mean NVE and RIB is 1 unit higher in the L breed than in D (29.8 vs. 28.7 and 15.5 vs. 14.6, respectively).

The heritability (h^2), defined as the proportion of the total phenotypic variance explained by additive genetic variance, for NTE was 0.41 (LW), 0.39 (L), and slightly smaller for the D breed (0.28) (Table 2). For NVE, the h^2 increased considerably up to 0.59 in the L breed, and the increase was even more pronounced in the D breed (0.62). However, the h^2 for RIB only increased even further in D animals (0.78) but decreased again down to 0.24 in L animals (Table 2). Data on NVE and RIB were not available from the LW breed.

Table 2: Heritability h^2 and parameters for the most significant SNP for each trait in each breed from the GWAS.

Breed	Trait [1]	h^2 [2]	SNP	SSC7 [3]	$-\text{Log}_{10}(P)$	Freq. [4]	Effect [5]	SD [6]	Var._{exp} [7]
LW	NTE	0.41	AX-116757987	97.57	74	0.23	0.38	0.02	0.05
L	NTE	0.39	AX-116329721	97.62	50	0.70	0.33	0.02	0.05
	NVE	0.59	AX-116329717	97.53	84	0.69	0.49	0.03	0.34
	RIB	0.24	AX-116329717	97.53	54	0.69	0.49	0.03	0.20
D	NTE	0.28	AX-116777212	97.61	59	0.33	0.38	0.02	0.06
	NVE	0.62	AX-116329719	97.59	184	0.31	0.68	0.02	0.52
	RIB	0.78	AX-116329688	97.60	260	0.31	0.83	0.02	0.69

[1] NTE: number of teats, NVE: total number of vertebrae, RIB: number of ribs on the right side of the animal. [2] h^2: heritability. [3] SSC7: position in Mb on chromosome 7. [4] Freq.: frequency of the allele increasing NTE, NVE or RIB. [5] Effect: allele substitution effect. [6] SD: standard deviation of the allele susbtitution effect. [7] Var._{exp}: percentage total phenotypic variance explained by the most significant SNP.

The GWAS results for NTE show a strong QTL on SSC7 in all three breeds in the same region (Figure 3). In addition, other QTL segregate on several other chromosomes, especially in the LW and L breeds. Some QTL regions overlap between the three populations but a few QTL are breed specific (Table 3 and Supplementary File S1). All these additional QTL disappear in the L and D breeds in the GWAS for the traits NVE and RIB (Figure 4, Figure 5 and Supplementary File S1). Only a strong significant peak remains for SSC7 in the same region for the L and D breeds (Table 2).

The position of the most significant SNP (top SNP) differs slightly between traits and populations around 97.6 Mb on SSC7 (build 11.1). Only in the L breed, the top SNP shifts around 90 kb from 97.62 Mb for NTE to 97.53 Mb for NVE and RIB. However, Figures 6 and 7 show that the extent of LD between the top SNP and all other SNPs in the QTL region is much larger in the L breed than in D for both NVE and RIB. For all GWAS, the most significant SNP from the imputed 660K SNP set is also the top SNP from the imputed WGS association analyses (Supplementary File S2). The frequency of the allele of the top SNP that is related to increased NTE is more than twice as high in the L breed compared to D (0.70 vs. 0.33) and the lowest in LW (0.23) (Table 2). For NVE and RIB the allele frequency values in the L and D breeds are only slightly lower than for NTE (0.69 vs. 0.31).

However, the significance levels differ remarkably between traits and breeds. The –log10 of the p-value increases from 50 to 84 in the L breed and from 59 to 184 in D for NTE and NVE, respectively. Furthermore, scoring RIB instead of NVE increases significance levels considerably further but only in D and decreases extremely in the L breed (260 vs. 54, respectively). This is in line with the explained variance increasing from 5% and 6% for NTE up to 34% (L) and 52 % (D) for NVE. However, for RIB, we observe a further increase of explained variance up to 69% in D whereas a severe reduction is seen again in the L breed down to 20%. The size of the effect for NTE is comparable across breeds between 0.33 (L) and 0.38 (LW and D). For NVE the effect is much higher in D than in L (0.68 vs. 0.49) and for RIB it only increases in D up to 0.83 but remains the same in the L breed (0.49).

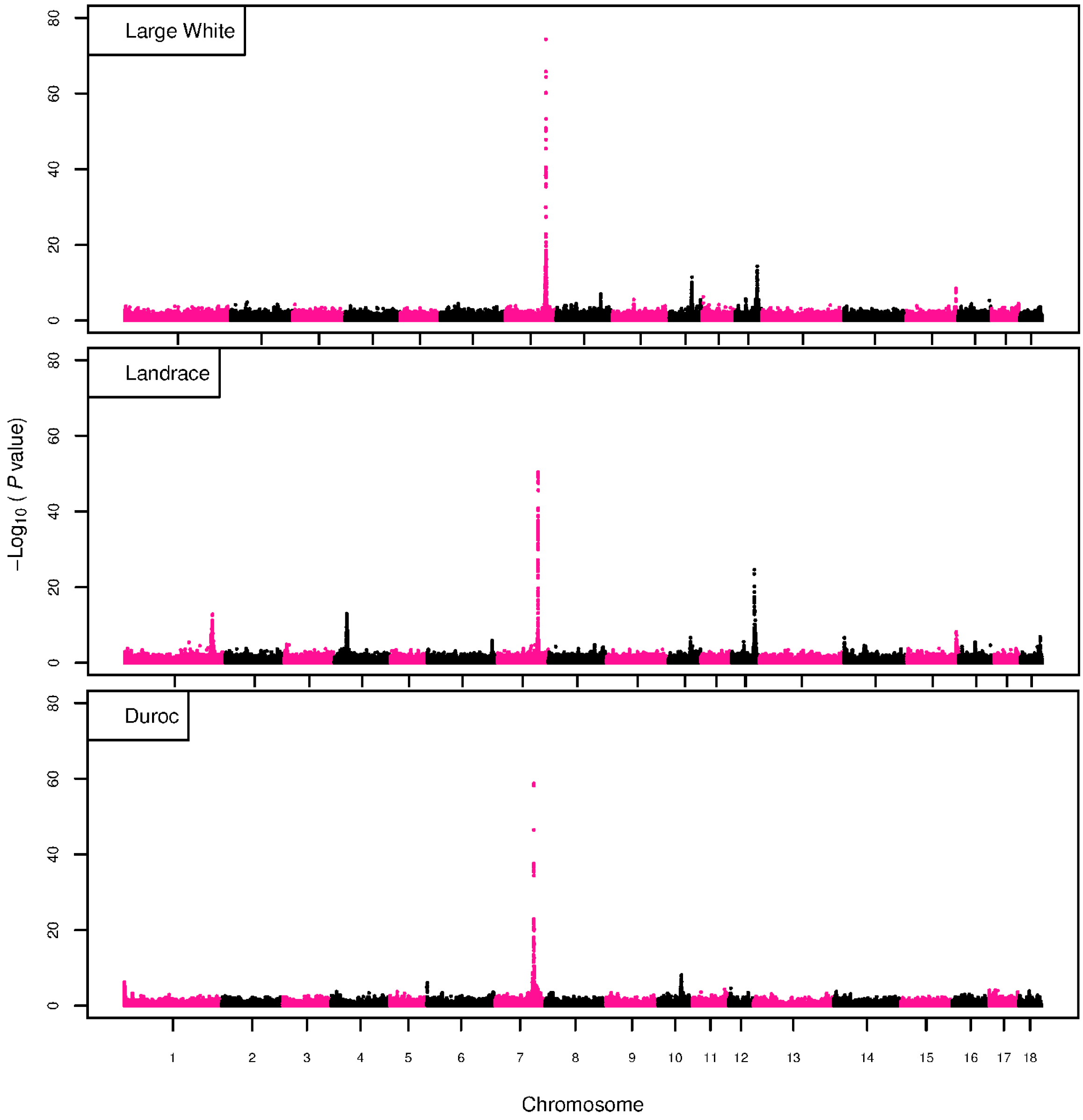


Figure 3: GWAS plot for NTE with imputed 660K SNP chip data. On the y-axis is the –log10(p-values) of single SNP association with NTE in LW, L and D breeds. On the x-axis is the physical position of the SNP across the 18 autosomes.

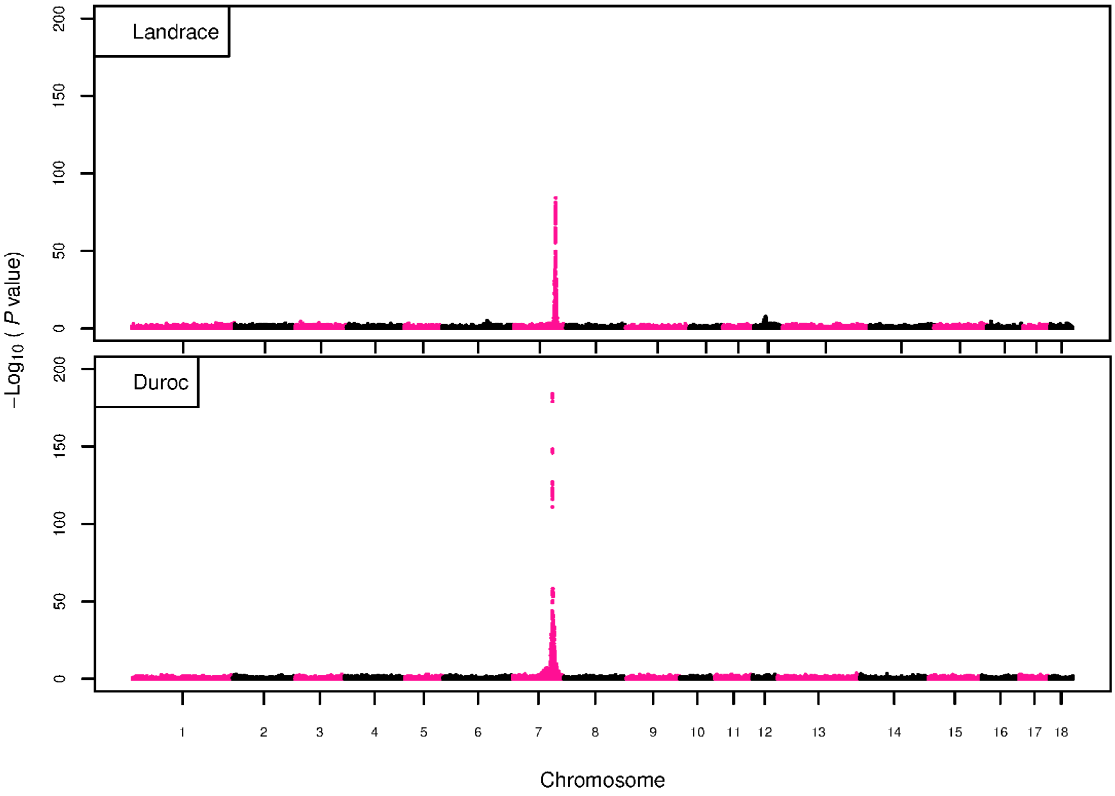


Figure 4: GWAS plot for total NVE with imputed 660K SNP chip data. On the y-axis is the –log10(p-values) of single SNP association with NVE in L and D breeds. On the x-axis is the physical position of the SNP across the 18 autosomes.

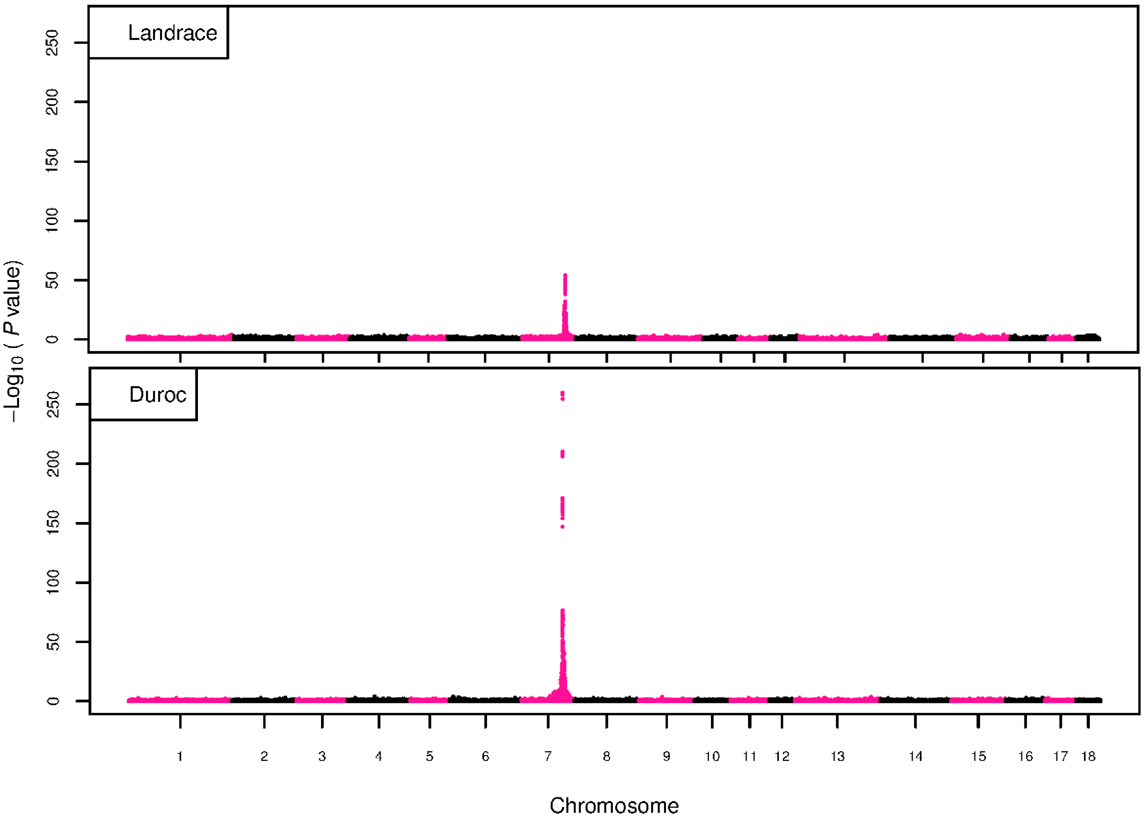


Figure 5: GWAS plot for RIB with imputed 660K SNP chip data. On the y-axis is the –log10(p-values) of single SNP association with RIB in L and D breeds. On the x-axis is the physical position of the SNP across the 18 autosomes.

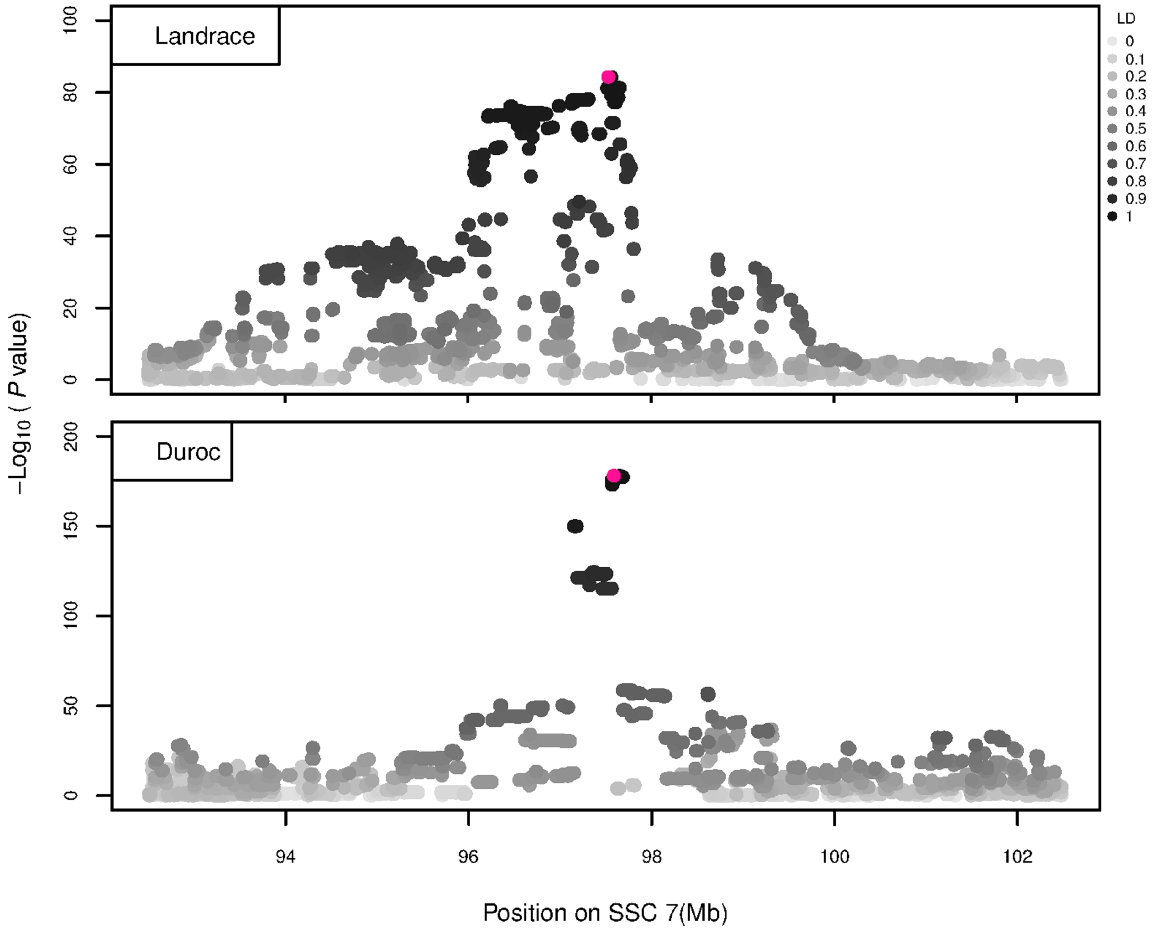


Figure 6: GWAS plot for NVE using imputed whole-sequence data. On the y-axis is the –log10(p-values) of single SNP association with NVE in L and D breeds. On the x-axis is the physical position of the SNP in the SSC7 QTL region. Linkage disequilibrium (LD) is given in a scale of 0 to 1 as a measure of the pairwise correlation between the most significant SNP (pink dot) and all other SNPs.

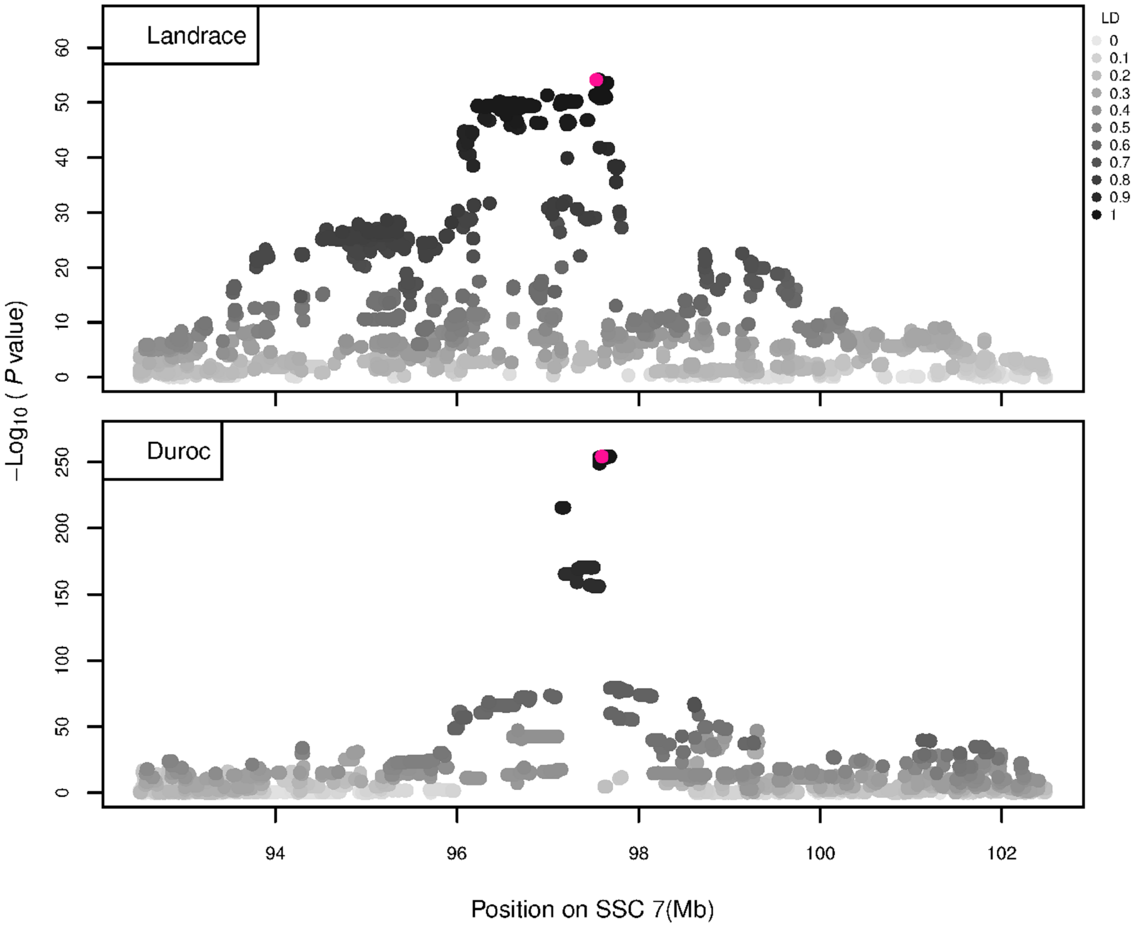


Figure 7: GWAS plot for RIB using imputed whole-sequence data. On the y-axis is the –log10(p-values) of single SNP association with RIB in L and D breeds. On the x-axis is the physical position of the SNP in the SSC7 QTL region. Linkage disequilibrium (LD) is given in a scale of 0 to 1 as a measure of the pairwise correlation between the most significant SNP (pink dot) and all other SNPs.

Table 3: Genomic regions associated with NTE in LW, L and D populations.

Breed	SSC	Position(Mb)	#significant SNPs	topSNP location(Mb)	topSNP *p*-value	gene/nearest gene
Large White	7	93.7-99.2	195	97.57	4.68e-75	*VRTN*
	10	47.2-47.9	13	47.88	3.30e-12	*FRMD4A*
	12	50.4-50.6	23	50.60	4.02e-15	*SMTNL2*
	15	134.8-134.9	11	134.80	3.27e-09	*ARL4C*
Landrace	1	23.9-24.1	25	24.05	1.55e-13	*ANKS6*
	4	29.0-30.9	57	29.21	1.03e-13	*EIF3E*
	7	95.9-97.8	147	97.62	4.13e-51	*VRTN*
	12	50.3-52.6	64	50.36	6.11e-10	*UBE2G1*
	15	134.62-134.64	2	134.64	6.59e-09	*ARL4C*
Duroc	7	82.8-99.3	261	97.61	1.69e-59	*VRTN*
	10	47.8-47.9	7	47.88	7.55e-09	*FRMD4A*

The QTL regions for NTE are presented with positions (build 11.1), number of significant SNPs (*p*-value<1.0e-08), topSNP location and *p*-value, and the gene/nearest gene of the topSNP.

4.2 Sequencing Data Analysis

The previously identified VRTN insertion (g.20311_20312ins291) (Mikawa et al., 2011) and promoter SNP (g.19034A>C, rs709317845) (Fan et al., 2013) were compared to NVE in the WGS animals and some relatives by PCR and IGV. Pigs with the insertion allele and the C allele of the promoter SNP were associated with increasing NVE in both L and D breeds, and the two variants were completely linked in the examined animals. The PCR genotyping of the insertion confirmed the findings by IGV.

When grouping animals by VRTN genotypes and analyzing the WGS data within the SSC7 QTL region for protein coding variants, no other compelling candidates for functional coding variants were found (Supplementary File S3, Table X7). Protein coding variants in this region occur either in far too many individuals or far too few individuals to be having an impact on the observed phenotypes (Supplementary File S3, Tables X3-X5 show variants unique to groups, and Table X6 shows variants shared between groups).

However, there were large differences between the three groups for some non-coding variants, affecting several genes (Supplementary File S3, Tables X2-X6). For example, multiple mutations were identified in ATP binding cassette subfamily D member 4 (ABCD4) that are far more common in the wt/wt and wt/ins samples, with ABCD4 having the most unique mutations of any gene in this locus. Several non-coding mutations in ABCD4 occur in 7/7 of the wt/wt samples and 18/18 wt/ins samples, and 0/9 of the ins/ins samples (Table 4 and Supplementary File S3, Table X6). These non-coding variants could have functional effects that impact NVE, but without transcriptomics data, and with poor functional annotation of the non-coding regions in this locus, it is not possible to determine which of these variants could have a functional impact.

Table 5: Identified variants on the same haplotype as VRTN insertion and promoter SNP.

Variant	Chromosome	Position	Ref/Alt	Consequence	Gene Symbol	rsID	LD D	LD L
7_97563673_T/G	7	97563673	T/G	Downstream	ABCD4	-	0.693193	0.970185
7_97568605_T/A	7	97568605	T/A	Intronic	ABCD4	rs711873120	0.989868	0.888563
7_97568606_A/C	7	97568606	A/C	Intronic	ABCD4	rs699009491	0.989868	0.888563
7_97568835_G/A	7	97568835	G/A	Intronic	ABCD4	rs692051374	0.989868	0.888563
7_97569136_C/T	7	97569136	C/T	Intronic	ABCD4	WU_10_2_7_103412699	0.993029	0.861442
7_97571221_C/T	7	97571221	C/T	Intronic	ABCD4	-	0.989868	0.888563
7_97571322_A/T	7	97571322	A/T	Intronic	ABCD4	-	0.989868	0.888563
7_97571384_CAGC/CGG	7	97571384	CAGC/CGG	Intronic	ABCD4	-	0.989868	0.888563
7_97572779_G/A	7	97572779	G/A	Intronic	ABCD4	-	0.989868	0.888563
7_97572788_C/T	7	97572788	C/T	Intronic	ABCD4	-	0.989868	0.888563
7_97574280_CG/CAG	7	97574280	CG/CAG	Intronic	ABCD4	-	0.989868	0.888563
7_97579254_T/C	7	97579254	T/C	Intronic	ABCD4	rs1112162366	0.994818	0.888563
7_97579520_G/T	7	97579520	G/T	Intronic	ABCD4	-	0.994818	0.888563
7_97606621_A/T	7	97606621	A/T	Intergenic	ABCD4-VRTN	rs331843703	1	1
7_97642098_C/G	7	97642098	C/G	Intergenic	VRTN-SYNDIG1L	rs337650751	0.996804	0.980266

At these identified variants, the animals (n=34) with *VRTN* homozygous ins/ins (ins/ins at the VRTN insertion *g.20311_20312ins291*, position 7:97615880, and CC at the VRTN promoter SNP *g.19034A>C*, position 97614602*)* and wt/wt have opposing homozygous genotypes and all the Qq animals have heterozygous genotypes. LD is calculated with respect to the two *VRTN* mutations in Duroc (D) and Landrace (L). More information can be found in Supplementary File S3, Table X6

4.3 Haplotype Analysis and Recombinant Identification

We determined the haplotypes that are associated with the two functional VRTN mutations and find a single relatively high frequency haplotype associated within each breed. However, several haplotypes at lower frequency are also associated with the two functional VRTN mutations; an example of the associated 80K and

660K haplotypes in the D breed is presented in Supplementary File S4. They all have a core haplotype of seven SNPs in common surrounding the two VRTN variants.

Duan et al. (2018) showed that the promoter variant and the PRE1 insertion increase VRTN expression in an additive way. To estimate the effect of both causal mutations separately we searched the genotype data for recombinant animals. Interestingly, we identified two sequenced recombinant animals, but only within the LW breed. One animal is wt/wt for the promoter SNP, while being heterozygous for the VRTN insertion. The other animal is heterozygous for the promoter SNP, while being homozygous for the VRTN insertion (Supplementary Figure S2). Haplotype analysis (on medium density SNP chip) confirmed that both animals carry a separate recombinant haplotype. The haplotypes are segregating with a combined frequency of about 1.8% in LW. Unfortunately, no phenotypic data on NVE and RIB are available for this breed.

Haplotype analysis also identified a haplotype that is specific for the L breed animals not carrying the VRTN insertion allele (Supplementary File S5, Table S1). The haplotype was not present in D or in L ins/ins animals and might explain the different effect on RIB found in the L breed. There are two missense mutations within this haplotype, both located in ABCD4, with SIFT values of 0.03 and 0.1. Moreover, two splice region variants, located in ABCD4 and VSX2 (visual system homeobox 2), are putative causal candidates within this wildtype haplotype observed only in L. The SNPs located on the 660K chip that are segregating with this haplotype are presented in Supplementary File S5, Table S2.

4.4 LTBP2 Regulatory Region Analysis

Park et al. (2017) describe an effect on number of thoracic vertebrae further downstream of VRTN and pinpoint LTBP2 as a candidate gene. Here, known human LTBP2 regulatory regions (Davis et al., 2014) were mapped to the corresponding regions of the pig genome. The majority of these regions were found to have a high sequence identity between the species, indicating that they may perform similar regulatory functions in the pig genome (Supplementary File S3).

Additionally, the previously reported LTBP2 variant (c.4481A>C, rs322260921, position 7:97751432) (Park et al., 2017) was investigated using sequenced animals. The LTBP2 variant showed no effect when sorting animals by the VRTN genotype (Supplementary File S3, Table S6-S7). Moreover, the SNP variant in the LTBP2 gene is not in complete LD with VRTN variants in our pig breeds (r2 of 0.70 and 0.44 in L and D, respectively) and it is outside the core region identified in haplotype analyses (Supplementary File S4). This makes it unlikely that the insertion is directly affecting LTBP2 regulation, but it is possible that pig LTBP2 has additional regulatory regions compared to the human version, and these could be affected by the insertion.

5.0 Discussion

5.1 Parameters of Genetic Variation at the Population Level

In a previous study, we identified a QTL for NTE on SSC7 in a population of 936 LW animals (Duijvesteijn et al., 2014). Expanding the data to 2,620 individuals of the same LW population and adding 6,090 and 3,798 animals of the two other purebred breeds also evaluated in this study (L and D, respectively), we identified the same QTL segregating in all three populations (Lopes et al., 2017b). In the current study, we expanded the data even further to more than 20,000 LW and L animals and more than 10,000 D animals, reconfirming the QTL region for NTE on SSC7.

In this study, we show that the size of the effect is comparable in all three breeds, but the frequency of the allele related to increased NTE at the top SNP is more than twice as high in L as in LW and D breeds (Table 2). Assuming that the underlying mutation was affecting NVE (Duijvesteijn et al., 2014; Rohrer et al., 2015), we used detailed phenotypes for the vertebral column available for the L and D breeds from CT scan data. Indeed, examining the phenotype closer to the causative variation increased heritability, that is explaining an additional 20% (L) – 34% (D) of the phenotypic variation by additive genetic effects from just one QTL.

Noise from other QTL segregating for NTE is not relevant for NVE, as can be seen from the GWAS results (Figure 4). With only 24% of the animals available for NVE in D and 8% in L, compared to the number of animals available for NTE in these breeds, a much higher significance of the effect is obtained. Moreover, the size of the effect is much higher because additional variation of loci further downstream in the developmental cascade of teat development (Veltmaat, 2017) is not diluting the genetic effect on NVE. A further reduction in noise by counting RIB was expected because this QTL has been reported to affect thoracic vertebrae only and domestic pig populations also show variability in number of lumbar vertebrae (Rohrer et al., 2015), which was included in the NVE count. However, a further increase in effect size and h² for RIB compared to NVE was only observed in D, whereas a strong reduction of h² was estimated in L.

The size of the effect on RIB remained the same (0.49) for NVE in L and in the expected range as reported for thoracic vertebrae (ribs) by Duan et al. (2018), with the homozygous QQ animals having 1 vertebra more. Apparently, the effect of this QTL is disturbed by other genetic variation affecting rib development in L, which is not present in the D breed. Breed-specific effects have been reported earlier for other traits in pigs (Lopes et al., 2017a; Sevillano et al., 2018). However, it is also important to highlight that the allele related to increased RIB in the L breed is going towards fixation, decreasing the genetic variability in this population and therefore could be responsible for the lower h2 and total phenotypic variance for RIB in L compared to D.

The total phenotypic variance for RIB in the L breed is 0.51±0.02 rib2 while in the D breed it is 53% higher (0.78±0.02 rib2, data not shown). Furthermore, the size of the allelic effect in D is much larger (0.68 NVE and 0.83 RIB) generating 1.3 vertebrae more in the homozygous state compared to homozygous wildtype animals and even developing 1.6 more ribs phenotypically. The sum of these effects together (allele frequency and h^2) is mirrored by the phenotypic variance explained by the top SNP which accumulates to 69% in D for RIB and is the main indicator for the expected breeding progress in this trait.

To examine whether other loci close to VRTN had an effect on NTE, NVE or RIB, as previously reported by Rohrer et al. (2015), the most significant SNP was included as a fixed effect in the model. Correction for the top SNP showed that no other genetic variation is present in our populations for either of the traits. Although our GWAS results indicate that the traits NVE and RIB are controlled by one QTL of large effect, the top SNP does not seem to be the causal mutation. The variance explained by the top SNP was only up to 88% of h2 for these traits in these breeds. Therefore, there is still genetic variance that is not captured by this SNP.

Trying to get closer to the causal mutation, we also performed a GWAS using imputed WGS data. However, no additional information was obtained from increasing the resolution to imputed WGS variants as the same top SNP was identified using both SNP chip and imputed WGS data. Although the added benefit of WGS data is including the functional mutation, the WGS dataset mostly contains haplotypes that are in complete LD with the 660K SNPs. The analysis of extremely large data set from a granddaughter design in dairy cattle shows clearly that with statistical evaluation of SNP chip and WGS information only, the identification of the causative quantitative trait nucleotide just based on concordance is not achieved (Weller et al., 2018). We therefore analyzed separately potentially functional SNPs in coding regions. In any case, because of high LD with the functional VRTN variants and extensive LD especially in L, it is impossible to disentangle the effect from these different alleles. Additional laboratory functional experiments would be needed.

5.2 Phenotypic Differences Between Breeds

The primary task of the method used to sample NVE and RIB (Gangsei and Kongsro, 2016) is 3D segmentation of bones. Thus, the counts (NVE and RIB) utilized in the present study are just favorable by-products. The rib counting is challenging due to so-called half-ribs (Fredeen and Newman, 1962), underdeveloped ribs that are barely visible in the CT images. The rib count (RIB) might be viewed as a proxy variable for number of thoracic vertebrae. A higher proportion of underdeveloped ribs in L compared to D populations could be an additional explanation for the different effects observed for RIB between these two pig breeds.

Figure 8 shows the differences in allele frequencies at the VRTN locus in relation to the distribution of NVE and RIB for each breed. Homozygous wt/wt L animals have one NVE and RIB more than wt/wt D animals. In D, a large proportion of the animals are wt/ins and have 29 NVE with 15 RIB. The estimated effect comes mainly from this contrast between wt/wt and wt/ins animals. Very few animals are homozygous ins/ins and have 29 NVE with 15 RIB or 30 NVE with 16 RIB. However, nearly all L ins/ins animals show 30 NVE and the estimate of the allelic effect comes from the contrast between wt/ins and ins/ins animals. More than 60% of heterozygous L animals appear to already have 30 NVE and they show more variation in RIB. This indicates that in L animals other genetic factors are present causing a higher background level of NVE and RIB and causing more variability in RIB independent of VRTN genotype. We cannot disentangle whether additional mutations in LD with the long haplotype in the L breed cause the reduced effect on rib formation or whether other variants in the L breed blur the effect.

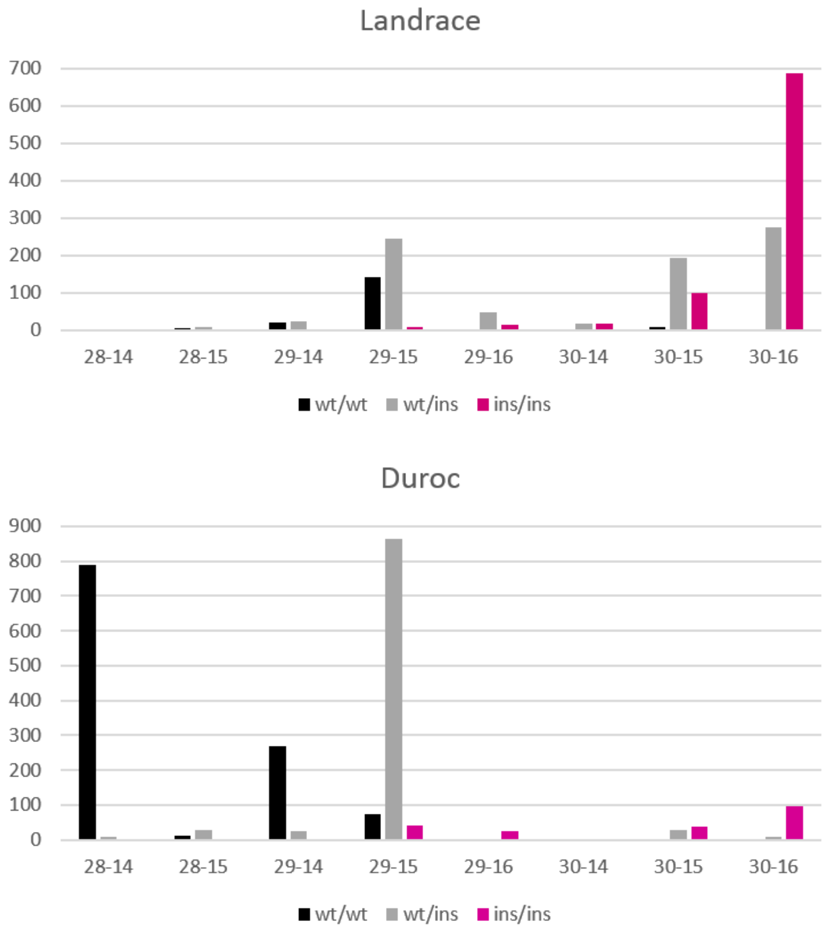


Figure 8: Distribution of NVE and RIB for Landrace and Duroc animals according to their genotypes wt/wt, wt/ins and ins/ins at the VRTN gene. A few animals with extreme phenotypes were discarded for better visibility.

5.3 Molecular Background of Life Development

In mammals, mammary gland complexes develop along a mammary line on each flank along the spine. The mammary line extends from the axilla to the inguin (Veltmaat, 2017). At designated points of the mammary line, mammary glands will develop in pairs. These points are determined by the underlying development of a vertebra. Vertebrae develop from the somites and mammary gland formation is initiated by factors in the dermal mesenchyme, which is also derived from the somites (Veltmaat et al., 2006). Further, each mammary gland is thought to be determined by specific genetic components, which determine whether its development will be initiated and continued.

Figure 9 shows schematically the somites as progenitor cells of vertebrae, ribs, and mammary glands. Segmental identity of each somite is maintained by the Hox code which controls the positional specification of each segment that later forms e.g. thoracic or lumbar vertebra (Wellik, 2007; Myers, 2008). In other words, the expression of Hox genes gradually changes along the axis from head to tail. For example, members of the paralog group Hox10 block rib formation whereas Hox6 proteins show rib promoting activity (Guerreiro et al., 2013).

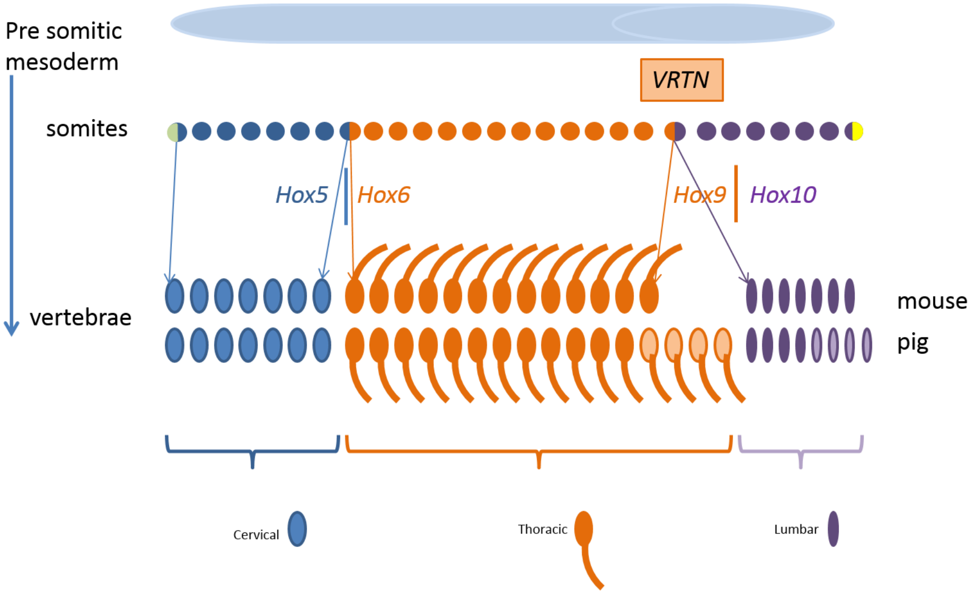


Figure 9: Development of the mouse and pig vertebral pattern from somites. Genes involved in the regulation and their spatial expression in italics. Variation in number of thoracic and lumbar vertebrae in pigs is indicated in light color. Parts of the figure was adopted from Gilbert (2000).

In our QTL region on SSC7, the direct effect of the two VRTN variants initiate the development of an additional vertebra in ins/ins animals. VRTN has been identified as a DNA transcription factor increasing the expression of NOTCH2 and HES1 (hes family bHLH transcription factor 1) and is therefore involved in the regulation of the synchronized oscillation of the segmentation clock (Duan et al., 2018). This causes a change in the embryonic development rate increasing the number of somites but we observe also breed differences in the segmental identity.

At the experimental level, increasing the number of cervical and thoracic vertebrae has been reported in transgenic mice by accelerating mRNA expression through reduction of number of introns of HES7 (hes family bHLH transcription factor 7) gene (Harima et al., 2013). Their results also indicate that the link between the segmentation clock and the hox gene activation can be dissociated which causes a partial transformation of lumbar vertebrae into a thoracic vertebrae. They observe a shift in the anterior border of HOXB6 (homeobox B6) and HOXB9 (homeobox B9) expression by one or two somites in mice mutants with an accelerated expression of HES7.

A partial transformation of lumbar vertebrae into a thoracic vertebra was also observed by knockout of HOXC8 (homeobox C8) gene in mice (Le Mouellic et al., 1992). Even environmental influences such as reduced temperature during embryonic development have been reported to affect the segmentation clock in zebrafish under experimental conditions (Jiang et al., 2000). So both number of somites and type of vertebrae can be affected by gene variants in the NOTCH pathway. Depending on the genetic capacity of the individual at the other loci that regulate rib development the animal will also develop additional ribs.

Other QTL for NTE are detected in the GWAS that are most likely due to variation further downstream in the developmental cascade for the formation of the mammary gland. Different genetic factors have been described regulating pairs of mammary glands at different locations and even between the left and right counterpart (Veltmaat et al., 2006; Rohrer and Nonneman, 2017). All pairs of mammary glands in mice have been shown to be different in terms of their timing of appearance, their molecular requirements, and their morphogenetic program (Veltmaat, 2017).

5.4 Identification of Other Functional Mutations

Analysis of the sequence variation in the SSC7 QTL region suggested that protein coding variants are unlikely to be having a large impact on the observed phenotypes. However, we identified two missense mutations in the ABCD4 gene located just upstream of VRTN. One missense variant has a SIFT code of 0.03

and is therefore expected to be deleterious for function of the protein. These two variants together with a large list of non-coding variants are present in all L wt/wt animals. Obviously, these mutations accumulated in the L breed and are not present in the D breed.

These mutations could be altering ABCD4, possibly changing its expression or function. Intriguingly there is seemingly a link between ABCD4 and the development of the spine. ABCD4 is believed to play a role in the intracellular processing of vitamin B12, and mutations affecting the ATPase domain of this protein have been shown to alter intracellular vitamin B12 trafficking (Coelho et al., 2012; Fettelschoss et al., 2017). Vitamin B12 is required as a cofactor in methionine synthase, and low levels of vitamin B12 during development are associated with higher levels of neural tube defects in humans (Ray and Blom, 2003; Groenen et al., 2004; Turner, 2018). With the addition of expression data, in future analyses it may be possible to better characterize variants in this region and identify other important functional mutations. The obvious sequence differences between the two breeds on the wildtype haplotypes could also explain why the size of the effect on NVE is larger in D than in the L breed.

5.5 Origin of VRTN Promoter SNP and Insertion

The insertion allele characterized by the PRE1 insertion element and the SNP in the promoter of VRTN are only 1.2 kb apart. Fan et al. (2013) describe both VRTN variants to be in complete LD in three experimental populations of Western and Chinese origin. The insertion allele segregates in some Chinese breeds but Chinese wild boar and most Chinese indigenous breeds are wt/wt. Also in our data set, both VRTN mutations are in strong LD, but Zhang et al. (2016) describe an experimental cross where only the promoter SNP is segregating. To the contrary, among our sequenced animals, we do not find the promoter SNP variant without the insertion. However, we do find the insertion without the promoter SNP in LW, showing that recombinant animals are segregating in the LW breed. These recombinant animals (1.8% frequency) could be used to estimate the effect of the insertion separately in the future to test whether the increase in VRTN expression caused by the insertion alone is sufficient to generate an additional vertebra. Zhang et al. (2016) do not report an estimate of the allelic effect for the promoter

SNP and describe that the effect in their data set is only due to a mutation in the LTBP2 gene further distal on SSC7.

6.0 Conclusions

In this study, a clear relationship between formation of the vertebral column and development of teats is observed in two populations of commercial pigs, L and D, differing largely in NTE. In both breeds, this difference in NTE is partly due to genetic variation in a region on SSC7, which has earlier been reported in several studies. By refining phenotype and examining NVE, noise from other loci is omitted, increasing overall heritability and significance of the SSC7 QTL region. However, the effect of two previously reported VRTN variants on thoracic vertebrae was found to be dependent on the genetic background. Allele frequencies at the QTL and accuracy of the phenotype differ between breeds and thereby influence genetic and phenotypic variance. Also, the overall population mean for RIB differs between L and D breeds.

At the molecular lever, the large number of non-coding and coding variants observed to be present only in L on the wildtype haplotype show that the genetic background in the 3 Mb region encompassing VRTN differs between the two breeds. Moreover, other more subtle differences, such as variants affecting ABCD4, can also be expected in the remainder of the genome although they were not detectable in the GWAS results. The relationship between quantitative genetic parameters and the underlying factors of the developmental cascade of skeletal and mammary gland development gives a good example how biological factors influence our population parameters used for practical breeding value estimation. Identification and proof of causative mutations for oligogenic or polygenic traits without functional laboratory studies remains nearly impossible, especially for non-coding variants.

7.0 List of Abbreviations

ABCD4	ATP binding cassette subfamily D member 4 gene
CT	Computer tomography
D	Duroc
HES1	Hes family bHLH transcription factor 1 gene
HES7	Hes family bHLH transcription factor 7 gene
HOXB6	Homeobox B6 gene
HOXB9	Homeobox B9 gene
HOXC8	Homeobox C8 gene
IGV	Integrative Genomics Viewer software
L	Landrace
LD	Linkage disequilibrium
LTBP2	Latent transforming growth factor binding protein 2 gene
LW	Large White
NTE	Number of teats
NVE	Number of vertebrae
RIB	Number of ribs
SSC7	Sus scrofa chromosome 7
VRTN	Vertnin gene
VSX2	Visual system homeobox 2
WGS	Whole genome sequence

8.0 Acknowledgements

The authors thank Cigene (NMBU, Norway), Hanne Hamland at Norsvin and the Technical Administration department at Topigs Norsvin Research Center for sample handling and genotyping. We are grateful to Chad Harland (University of Liege, Belgium) for help in genotyping the insertion variant from sequence data.

9.0 Funding

This work was financed by the Research Council of Norway, the STW-Breed4Food Partnership, project number 14283: From sequence to phenotype: detecting deleterious variation by prediction of functionality, and a BBSRC CASE studentship (award reference: 1523312).

10.0 Data Availability Statement

The datasets analyzed for this study are included in the manuscript and the supplementary material.

The supplementary data can be found on the USB stick provided with this thesis.

11.0 Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

12.0 Author Contributions Statement

MS performed SNP detection in WGS data, was involved in imputation and haplotype work, and contributed to writing the paper. ML calculated genetic parameters, performed GWAS analyses and was involved in writing the paper. HM performed functional SNP analyses, performed analyses of the LTBP2 and contributed to writing the paper. MD performed haplotype analyses and contributed to writing the paper. LG and JK performed phenotyping using CT images and was involved in writing the paper. MW and EG was involved in planning the project and contributed to writing the paper. BH was involved in planning the project, coordinated the studies and drafted the paper. All authors have read and approved the final manuscript.

13.0 References

Arakawa, A., Okumura, N., Taniguchi, M., Hayashi, T., Hirose, K., Fukawa, K., et al. (2015). Genome-wide association QTL mapping for teat number in a purebred population of Duroc pigs. Animal Genetics 46(5), 571-575.

Browning, B.L., and Browning, S.R. (2016). Genotype Imputation with Millions of Reference Samples. The American Journal of Human Genetics 98, 116-126.

Casbon, J. (2012). PyVCF - A Variant Call Format Parser for Python [Online]. Available: https://pyvcf.readthedocs.io/en/latest/INTRO.html [Accessed].

Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., Rose, D.B., Garrison, E.P., et al. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. Nature Methods 12, 966-968.

Coelho, D., Kim, J.C., Miousse, I.R., Fung, S., du Moulin, M., Buers, I., et al. (2012). Mutations in ABCD4 cause a new inborn error of vitamin B12 metabolism. Nature Genetics 44(10), 1152-1155.

Dall'Olio, S., Ribani, A., Moscatelli, G., Zambonelli, P., Gallo, M., Costa, L.N., et al. (2018). Teat number parameters in Italian Large White pigs: Phenotypic analysis and association with vertnin (VRTN) gene allele variants. Livestock Science 210, 68-72.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., et al. (2011). The variant call format and VCFtools. Bioinformatics 27(15), 2156-2158.

Davis, M.R., Andersson, R., Severin, J., de Hoon, M., Bertin, N., Baillie, J.K., et al. (2014). Transcriptional profiling of the human fibrillin/LTBP gene family, key regulators of mesenchymal cell functions. Molecular Genetics and Metabolism 112, 73-83.

Duan, Y., Zhang, H., Zhang, Z., Gao, J., Yang, J., Wu, Z., et al. (2018). VRTN is Required for the Development of Thoracic Vertebrae in Mammals. International Journal of Biological Sciences 14(6), 667-681.

Duijvesteijn, N., Veltmaat, J.M., Knol, E.F., and Harlizius, B. (2014). High-resolution association mapping of number of teats in pigs reveals regions controlling vertebral development. BMC Genomics 15:542.

Fan, Y., Xing, Y., Zhang, Z., Ai, H., Ouyang, Z., Ouyang, J., et al. (2013). A Further Look at Porcine Chromosome 7 Reveals VRTN Variants Associated with Vertebral Number in Chinese and Western Pigs. PLOS ONE 8(4), e62534.

Fettelschoss, V., Burda, P., Sagné, C., Coelho, D., De Laet, C., Lutz, S., et al. (2017). Clinical or ATPase domain mutations in ABCD4 disrupt the interaction between the vitamin B12-trafficking proteins ABCD4 and LMBD1. The Journal of Biological Chemistry 292(28), 11980-11991.

Fredeen, H.T., and Newman, J.A. (1962). Rib and vertebral numbers in swine. I. Variation observed in a large population. Canadian Journal of Animal Sciene 42, 232-239.

Gangsei, L.E., and Kongsro, J. (2016). Automatic segmentation of Computed Tomography (CT) images of domestic pig skeleton using a 3D expansion of Dijkstra's algorithm. Computers and Electronics in Agriculture 121, 191-194.

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907v2.

Gilbert, S.F. (2000). "Early Mammalian Development," in Developmental Biology, 6th edition. (Sunderland (MA): Sinauer Associates).

Gilmour, A.R., Cullis, B.R., Gogel, B.J., Welham, S.J., and Thompson, R. (2009). ASReml User Guide Release 3.0. VSN International Ltd, UK.

Groenen, P.M., van Rooij, I.A., Peer, P.G., Gooskens, R.H., Zielhuis, G.A., and Steegers-Theunissen, R.P. (2004). Marginal maternal vitamin B12 status increases the risk of offspring with spina bifida. American Journal of Obstetrics and Gynecology 191(1), 11-17.

Guerreiro, I., Nunes, A., Woltering, J.M., Casaca, A., Nóvoa, A., Vinagre, T., et al. (2013). Role of a polymorphism in a Hox/Pax-responsive enhancer in the evolution of the vertebrate spine. PNAS 110(26), 10682-10686.

Harima, Y., Takashima, Y., Ueda, Y., Ohtsuka, T., and Kageyama, R. (2013). Accelerating the tempo of the segmentation clock by reducing the number of introns in the Hes7 gene. Cell Reports 3(1), 1-7.

Jiang, Y.I., Aerne, B.L., Smithers, L., Haddon, C., Ish-Horowicz, D., and Lewis, J. (2000). Notch signalling and the synchronization of the somite segmentation clock. Nature 408(6811), 475-479.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., et al. (2002). The Human Genome Browser at UCSC. Journal of medical chemistry 12, 996-1006.

King, J.W.B., and Roberts, R.C. (1960). Carcass length in the bacon pig; its association with vertebrae numbers and prediction from radiographs of the young pig. Animal Science 2(1), 59-65.

Le Mouellic, H., Lallemand, Y., and Brûlet, P. (1992). Homeosis in the mouse induced by a null mutation in the Hox-3.1 gene. Cell 69(2), 251-264.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16), 2078-2079.

Lopes, M.S., Bastiaansen, J.W.M., Harlizius, B., Knol, E.F., and Bovenhuis, H. (2014). A Genome-Wide Association Study Reveals Dominance Effects on Number of Teats in Pigs. PLOS ONE 9(8), e105867.

Lopes, M.S., Bovenhuis, H., Hidalgo, A.M., Arendonk, J.A., Knol, E.F., and Bastiaansen, J.W. (2017a). Genomic selection for crossbred performance accounting for breed-specific effects. Genetic Selection Evolution 49(1), 51.

Lopes, M.S., Bovenhuis, H., van Son, M., Nordbø, Ø., Grindflek, E.H., Knol, E.F., et al. (2017b). Using markers with large effect in genetic and genomic predictions. Journal of Animal Science 95(1), 59-71.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., et al. (2016). The Ensembl Variant Effect Predictor. Genome Biology 17:122.

Mikawa, S., Sato, S., Nii, M., Morozumi, T., Yoshioka, G., Imaeda, N., et al. (2011). Identification of a second gene associated with variation in vertebral number in domestic pigs. BMC Genetics 12:5.

Myers, P.Z. (2008). Hox genes in development: The Hox code. Nature Education 1(1):2.

Park, H.-B., Han, S.-H., Lee, J.-B., and Cho, I.-C. (2017). High-resolution quantitative trait loci analysis identifies LTBP2 encoding latent transforming growth factor beta binding protein 2 associated with thoracic vertebrae number in a large F2 intercross between Landrace and Korean native pigs. Journal of Animal Science 95, 1957-1962.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., et al. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics 81(3), 559-575.

Ray, J.G., and Blom, H.J. (2003). Vitamin B12 insufficiency and the risk of fetal neural tube defects. QJM: Monthly Journal of the Association of Physicians 96(4), 289-295.

Ren, D.R., Ren, J., Ruan, G.F., Guo, Y.M., Wu, L.H., Yang, G.C., et al. (2012). Mapping and fine mapping of quantitative trait loci for the number of vertebrae in a White Duroc x Chinese Erhualian intercross resource population. Animal Genetics 43, 545-551.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16(1), 276-277.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., et al. (2011). Integrative Genomics Viewer. Nature Biotechnology 29(1), 24-26.

Rohrer, G.A., and Nonneman, D.J. (2017). Genetic analysis of teat number in pigs reveals some developmental pathways independent of vertebra number and several loci which only affect a specific side. Genetic Selection Evolution 49(1), 4.

Rohrer, G.A., Nonneman, D.J., Wiedmann, R.T., and Schneider, J.F. (2015). A study of vertebra number in pigs confirms the association of vertnin and reveals additional QTL. BMC Genetics 16:129.

Sargolzaei, M., Chesnais, J.P., and Schenkel, F.S. (2014). A new approach for efficient genotype imputation using information from relatives. BMC genomics 15(1), 478.

Sevillano, C.A., Guimarães, S.E.F., Silva, F.F., and Calus, M.P.L. (Year). "Breed-specific genome-wide association study for purebred and crossbred performance", in: Proceedings of the World Congress on Genetics Applied to Livestock Production).

Tang, J., Zhang, Z., Yang, B., Guo, Y., Ai, H., Long, Y., et al. (2017). Identification of loci affecting teat number by genome-wide association studies on three pig populations. Asian-Australasian Journal of Animal Sciences 30(1), 1-7.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2012). Integrative GenomicsViewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14(2), 178-192.

Turner, M. (2018). Folic acid and vitamin B12 fortification of food for preventing neural tube defects in Europe. BMJ 361:k1572.

Uzzaman, M.R., Park, J.E., Lee, K.T., Cho, E.S., Choi, B.H., and Kim, T.H. (2018). Whole-genome association and genome partitioning revealed variants and explained heritability for total number of teats in a Yorkshire pig population. Asian-Australasian Journal of Animal Sciences 31(4), 473-479.

van Son, M., Enger, E.G., Grove, H., Ros-Freixedes, R., Kent, M.P., Lien, S., et al. (2017). Genome-wide association study confirm major QTL for backfat fatty acid composition on SSC14 in Duroc pigs. BMC Genomics 18:369.

Veltmaat, J.M. (2017). "Prenatal Mammary Gland Development in the Mouse: Research Models and Techniques for Its Study from Past to Present," in Mammary Gland Development, eds. F. Martin, T. Stein & J. Howlin. (Humana Press, New York, NY: Springer Link), 21-76.

Veltmaat, J.M., Relaix, F., Le, L.T., Kratochwil, K., Sala, F.G., van Veelen, W., et al. (2006). Gli3-mediated somitic Fgf10 expression gradients are required for the induction and patterning of mammary epithelium along the embryonic axes. Development 133(12), 2325-2335.

Verardo, L.L., Silva, F.F., Lopes, M.S., Madsen, O., Bastiaansen, J.W., Knol, E.F., et al. (2016). Revealing new candidate genes for reproductive traits in pigs: combining Bayesian GWAS and functional pathways. Genetic Selection Evolution 48:9.

Weller, J.I., Bickhart, D.M., Wiggans, G.R., Tooker, M.E., O'Connell, J.R., Jiang, J., et al. (2018). Determination of quantitative trait nucleotides by concordance analysis between quantitative trait loci and marker genotypes of US Holsteins. Journal of Dairy Science 101, 1-19.

Wellik, D. (2007). Hox patterning of the vertebrate axial skeleton. Developmental Dynamics 236(9), 2454-2463.

Yang, J., Huang, L., Yang, M., Fan, Y., Li, L., Fang, S., et al. (2016). Possible introgression of the VRTN mutation increasing vertebral number, carcass length and teat number from Chinese pigs into European pigs. Scientific Reports 6:19240.

Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics 88(1), 76-82.

Yang, J., Zaitlen, N., Goddard, M., Visscher, P., and Price, A. (2014). Mixed model association methods: advantages and pitfalls. Nat Genet 46(2), 100-106.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., et al. (2018). Ensembl 2018. Nucleic Acids Research 46, D754-D761.

Zhang, L.-C., Yue, J.-W., Pu, L., Wang, L.-G., Liu, X., Liang, J., et al. (2016). Genome-wide study refines the quantitative trait locus for number of ribs in a Large White x Minzhu intercross pig population and reveals a new candidate gene. Molecular Genetics and Genomics 291, 1885-1890.