Understanding the evolutionary origins of genome structural novelty in mouse

By
Frances Burden

A thesis submitted for the degree of Doctor of Philosophy in Genetics
March 2024

Supervisors: Dr Peter Ellis and Dr Marta Farré

This thesis was supported by a grant from the Leverhulme Trust

Declaration

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

F Burden

Frances Burden

March 2024

Acknowledgements

I would like to thank my supervisors Dr Peter Ellis and Dr Marta Farré for their support in completion of this thesis. Thank you, Marta, for such patience in guiding me through the HPC inner workings during a global pandemic. Thank you, Peter, for all the scientific discussions on slack.

Thank you, Ellis, Farré and Griffin lab members it was a pleasure working with you all.

Thank you to Dr Valdone Maciulyte and Dr James Turner at the Crick institute for help with flow sorting.

I would like to thank my husband for listening to my presentations and scrutinising my figures.

Thank you, mum, for your continued support and offers of proof reading.

Contents

Publications

Articles 1 and 2 form and integral part of the results chapter 4.

1. Burden F, Ellis PJI, Farré M. A shared 'vulnerability code' underpins varying sources of DNA damage throughout paternal germline transmission in mouse. Nucleic Acids Res. 2023 Mar 21;51(5):2319-2332. doi: 10.1093/nar/gkad089.

2. Álvarez-González, L., Burden, F., Doddamani, D. *et al.* 3D chromatin remodelling in the germ line modulates genome evolutionary plasticity. *Nat Commun* **13**, 2608 (2022). https://doi.org/10.1038/s41467-022-30296-6.
   (Joint first author with L.Álvarez-González and, D.Doddamani)

Conferences

- Gene Regulation and Epigenetics Conference for Early Career Scientists 18th January 2022- held Virtually. Presentation of a talk about DNA damage in mouse spermatogenesis.

- SMBE Evolution of Reproduction 27-29 July 2022 Oeiras Portugal- Presentation of a talk entitled “Where does DNA damage occur during mouse spermatogenesis?”

Abbreviations

3D - Three dimensional
5hmC - Five hydroxymethyl cytosine
AC - Adenine-Cytosine
AGSC - Adult germinal stem cells or spermatogonia
ASE – Allele-specific expression
AT - Adenine-Thymine
ATR - Ataxia telangiectasia and Rad3 related.
bp - Base pair
sBLISS - Suspension break labelling in situ and sequencing
BRobS - Barcelona Rob system
BRD4 - Bromodomain containing protein 4.
BSA - Bovine serum albumin
CA-Cytosine-Adenine
ChIP-seq - Chromatin Immunoprecipitation sequencing.
ChromHMM - Tool used to learn chromatin-state signatures using a multivariate hidden Markov model (HMM)
CNV - Copy number variation
CRs - Chromosomal rearrangements
CSK - Cytoskeletal
Cut&Tag - Cleavage under targets and tagmentation.
DAPI - 4',6-Diamidino-2-Phenylindole, Dihydrochloride
DNA-PKcs - DNA-dependent protein kinase catalytic subunit
db - Database
DBrIC- DNA break Immunocapture
DDR- DNA damage response
dds - DESeq2 data set
DMC1- The meiotic recombination protein DMC1
DMEM- Dulbecco's Modified Eagle Medium
DMSO - Dimethyl sulfoxide
DNA - Deoxyribonucleic acid
dNTP - Deoxyribose nucleotide triphosphate
DSB – Double-strand break
EBR - Evolutionary breakpoint region
FACS – Fluorescent-activated cell sorting
FBS - Foetal bovine serum
FISH - Fluorescence in situ hybridisation
FSC - Forward scattered light
GO - Gene Ontology
GC - Guanine-Cytosine
gH2AX - phosphorylation of histone H2AX on serine 139
HBSS - Hank’s Balanced Salt Solution
HR - Homologous recombination
HSB - Homologous synteny blocks
H3K9ac - Histone H3 acetylated at lysine 9.
H3K27ac - Histone H3 acetylated at lysine 27.
H3K4me3 - Histone H3 methylated at lysine 4.
H3K9me3 - Histone H3 methylated at lysine 9.
H3K27me3 - Histone H3 methylated at lysine 27.
H4Kac - Histone H4 acetylation
H4K/5ac/8ac/16ac - Histone H4 acetylated at lysine 5/8/16
H2AZ- A variant of histone H2A
IF - Immunofluorescence

Kb - Kilo base pair
LINE - Long interspersed nuclear element.
MACS2 - Model-based analysis of ChIP-seq version 2
Mb – Megabase
mESCs- Mouse embryonic stem cells
$MgCl_2$ - Magnesium chloride
msHSBs – multi-species homologous syntenic blocks
MSUC- Meiotic silencing of unsynapsed chromatin
MSCI - Meiotic sex chromosome inactivation
NAHR - Non-allelic homologous recombination
NGS - Next generation sequencing
NHEJ - Non-homologous end joining
OD - Oxidative damage
PCR - Polymerase chain reaction
PBS - Phosphate buffered saline
PI - Propidium iodide
PMSC - Post meiotic sex chromatin
PRDM9 -PR domain containing protein 9
RACFs -Reconstructed ancestral chromosome fragments
ROI- Region of interest
RPA - Replication protein A
Rob - Robertsonian
RNA - Ribonucleic acid
RNA-seq - RNA sequencing
SC- Spermatocytes
SINE - Short interspersed nuclear elements
SNP - Single nucleotide polymorphism
SSC - Side scattered light
SSC - Spermatogonial stem cell
SPO11 - Meiotic recombination protein SPO11
ST - Spematids
STR - short tandem repeat
SVs – structural variants
TAD - Topologically associated domain
TE - Transposable Element
TG-Thymine-Guanine
TOP2B - Topoisomerase II beta
TSS - Transcription start site.
UCSC - University of California, Santa Cruz
WGS - Whole genome sequencing
ZGA - Zygotic genome activation
Z-scan4- A gene encoding a SCAN domain and four zinc finger domains.

List of tables

List of figures

Abstract

During evolution, each lineage follows its own independent path of accumulating genome variation. Variation exists at both sequence and structural level - the latter representing dramatic changes in the organisation of the genome. This structural variation is encompassed by the term chromosomal rearrangements (CRs). It can take many forms such as inversions, translocations, fissions or fusions.

Chromosomal rearrangements can have profound consequences both for fertility at the individual level and for inter-individual reproductive compatibility. Understanding how CRs arise and spread within the population is thus pivotal to understanding multiple areas of reproductive and evolutionary biology, from individual fertility through to speciation.

Formation of CRs can be described by the Integrative Breakage Model - generation of CRs requires the formation of double-strand breaks (DSBs) during gamete production, followed by rejoining of loci that are physically adjacent within the nucleus. In this thesis, I address this element by studying the genetic, and epigenetic contexts of DSBs occurring in spermatogenesis, in combination with the 3D organisation of chromatin in male germ cells and show that this explains the locations of evolutionary breakpoint regions throughout rodent evolution.

Once CRs are formed, selective dynamics will subsequently determine whether they go to fixation or not. An understudied aspect of this is the potential for "drive", in which genetic or epigenetic factors bias the meiotic process and lead to non-Mendelian inheritance of CRs from heterozygous carriers. In this thesis, I address this element by investigating the genetic and epigenetic effects at play in male mice heterozygous for a Robertsonian chromosome fusion reported to show non-Mendelian inheritance.

Key findings:

- EBRs are associated with DSBs formed during spermiogenesis and not with meiotic DSBs.
- Spermatid DSBs are associated with specific chromatin state changes during spermatogenesis, and with predicted non-B DNA structures that may regulate DNA tension during sperm head compaction.
- In the Rb6.16 fusion model I identified a range of non-synonymous gene variants linked to the fusion breakpoint, which may explain the reported transmission skewing.
- Unexpectedly, I found no evidence for chromatin silencing in the vicinity of the Robertsonian breakpoint(s), indicating that Robertsonian fusions may be regulated differently to other structural variants during spermatogenesis.

1. General Introduction

Chromosomes are the largest-scale organisational unit of DNA within the cell, with each chromosome being a single long DNA molecule containing multiple genes. A typical diploid organism has two copies of each chromosome, one inherited from each parent.

Chromosomal rearrangements (CRs) are structural changes to chromosomes that alter the ordering of genes along the various chromosomes in the cell, and/or the number of chromosomes present in the cell. These CRs can be divided into different classes: inversions reverse a segment of the chromosome and thus the ordering of the genes in the inverted segment; translocations move genes from one location to another; while fissions and fusions respectively break or join chromosomes and thus alter the chromosome number for the organism bearing the rearrangement. CRs may be polymorphic within a species or may be fixed differences between species.

In diploid organisms, chromosomal rearrangements affect pairing and recombination during meiosis, as these processes require strict linear identity between the paired chromosomal rearrangements. Mis-paired chromosomes may lead to mis-segregation of chromosomes and thus the production of non-viable gametes, while a lack of recombination restricts gene flow between different chromosomal lineages. CRs thus not only have the potential to affect fertility on an individual level, but also to act as a source of reproductive incompatibilities between related species. For example, mules (a hybrid between a donkey and a horse) have 63 chromosomes whereas donkey has 62 and horse 64. Gamete formation in mules is impaired due to the odd number of chromosomes and they are sterile. This in turn demonstrates that donkeys and horse are reproductively incompatible species.

Therefore, understanding how structural novelty is generated within a genome and the evolutionary origin of this novelty is an important question in biology because 1) Rearrangements may cause reproductive incompatibilities that can lead to speciation. 2) Recombination may be suppressed in the vicinity of the rearrangement. 3) Chromosomal rearrangements may alter expression of genes in the vicinity of the chromosomal breakpoints, due to disruption of the 3-dimensional (3D) genome architecture such as promoter/gene/enhancer interactions or topological associated domains (TADs). As structural variation is key in evolution this raises the pivotal question of where this variation comes from and how it is generated. This thesis will attempt to address some of these questions.

1.1 The structure of chromatin

In the context of the nucleus, DNA exists in the form of chromatin (Figure 1-1) – a large-scale complex between the DNA strand and multiple different proteins that regulate its behaviour. Chromatin-associated proteins fall into two classes: histones (replaced by protamines in sperm) that directly wrap and package DNA, and other proteins associated with core activities of replication, transcription and DNA repair. Histones are an integral core structural component of chromatin, whereas other components such chromatin remodelers are only facultatively associated with the chromatin dependent on the cellular context.

In general terms, the purpose of histones and protamines is to condense multiple metres of DNA into the size of a cell nucleus, while simultaneously controlling which regions of DNA are made accessible to other factors involved with replication, transcription, repair etc.

A nucleosome is the smallest unit of chromatin structure, consisting of 147 bp of DNA wrapped around a histone core (1). The core is an octamer, that contains two copies each of the core histones H2A, H2B, H3, and H4 (1) (Figure 1-1). The histone N-terminal tails stick out from their own nucleosome, contacting adjoining nucleosomes, affecting inter-nucleosomal interactions (2). The histone proteins can be post translationally modified (Figure 1-1) on the histone tails (reviewed in (3)). These post-translational modifications (also known as "histone marks") are used as signals to control multiple aspects of DNA activity within the nucleus. These may include altering DNA conformation, regulating the level of transcriptional activity from specific DNA regions, and coordinating DNA repair processes. In particular, chromatin can be broadly divided into two states: compact and often transcriptionally inactive, termed heterochromatin, or a more relaxed open, likely transcriptionally active state, termed euchromatin. These chromatin variations were first discovered by Heitz in 1928 (4).

Two distinct types of heterochromatin have been identified, facultative and constitutive. Facultative heterochromatin is used to describe "genomic regions containing genes that are differentially expressed through development and/or differentiation and which then become silenced." The term constitutive heterochromatin is used to describe "permanently silenced genes in genomic regions such as centromeres and telomeres" (2).

Euchromatin is used to describe regions of the genome that contain active genes, however the distribution of histone modifications across these regions can vary greatly. Regions containing a high density of marks often termed 'islands' tend to be active sites of transcription (5).

Histone marks are named according to the nature of the modification and which residues of which histone monomer are affected. Different histone marks are associated with different consequences for chromatin activity. For example, transcriptionally active genes have a high enrichment of H3K4me3 (trimethylation at lysine 4 of histone H3), which marks the transcriptional start site (TSS) (5, 6). H3K27me3 (trimethylation at lysine 27 of histone H3) is a repressive chromatin mark, that leads to gene silencing. H3K9me3 (trimethylation at lysine 9 of histone H3), also leads to the transient formation of repressive chromatin (7). H3K27ac (lysine 27 acetylation of histone H3) is associated with higher activation of transcription but is not exclusively associated with the gene itself and is therefore deemed to mark enhancers (8). In addition to transcriptional regulation, histone proteins may be modified in response to DNA damage. In particular histone H2AX (a variant of H2A) is phosphorylated at serine 139 in response to DNA double-strand breaks – this is known as gammaH2AX or γH2AX. Multiple post translational modification may occur on the same nucleosome (a bivalent modification), affecting different histone tails within the octameric structure (9).

The location of specific histone proteins within the genome can be studied using the technique of chromatin immunoprecipitation (ChIP-seq). This uses an antibody against the modification of interest to immunoprecipitate regions containing the mark from sonicated DNA, followed by sequencing to identify which regions of DNA contain the modification of interest. More recently alternative techniques such as cleavage under targets and tagmentation (Cut&Tag) have been developed. Cut&Tag uses a specific antibody to bind a chromatin protein in situ in live cells. A secondary antibody is used to bind the primary antibody to amplify the signal. A protein A-Tn5 transposase fusion protein then binds to the secondary antibody and cleaves the DNA, which is followed by library preparation and sequencing. These and other techniques are discussed in more detail in section 1.10.

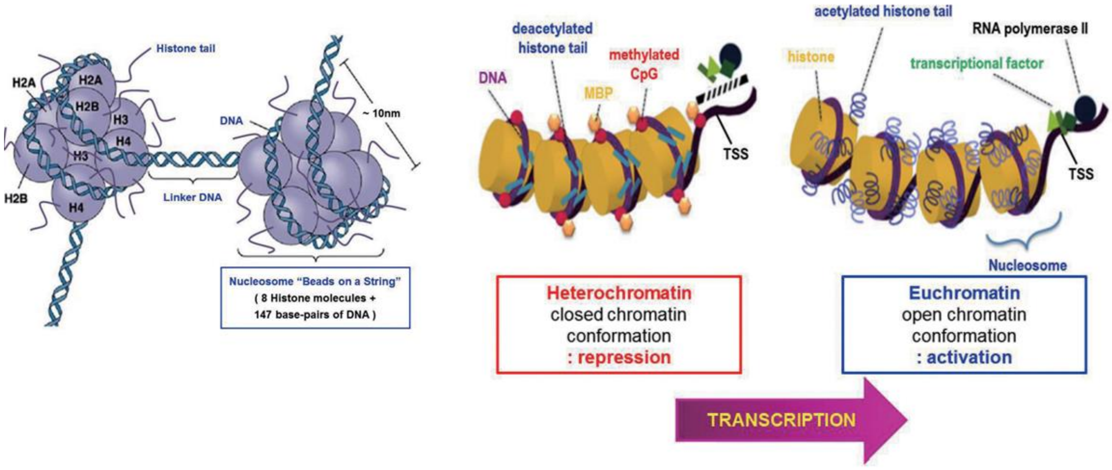


Figure 1-1: The structure of a nucleosome showing the histone octamer and a representation of heterochromatin and euchromatin.
The histone tails (shown in blue) can be covalently modified. Reproduced from: (10) under the creative commons licence (https://creativecommons.org/licenses/by-nc/4.0/).

1.2 3D genome organisation

Above the scale of individual nucleosomes, chromatin and the position of chromatin elements within the nucleus is highly regulated by several superimposed layers of organisation, which includes chromosome territories, within which chromatin is organised into compartments (open/closed), which in turn consist of topologically associated domains (TADs) and DNA loops (Figure 1-2).

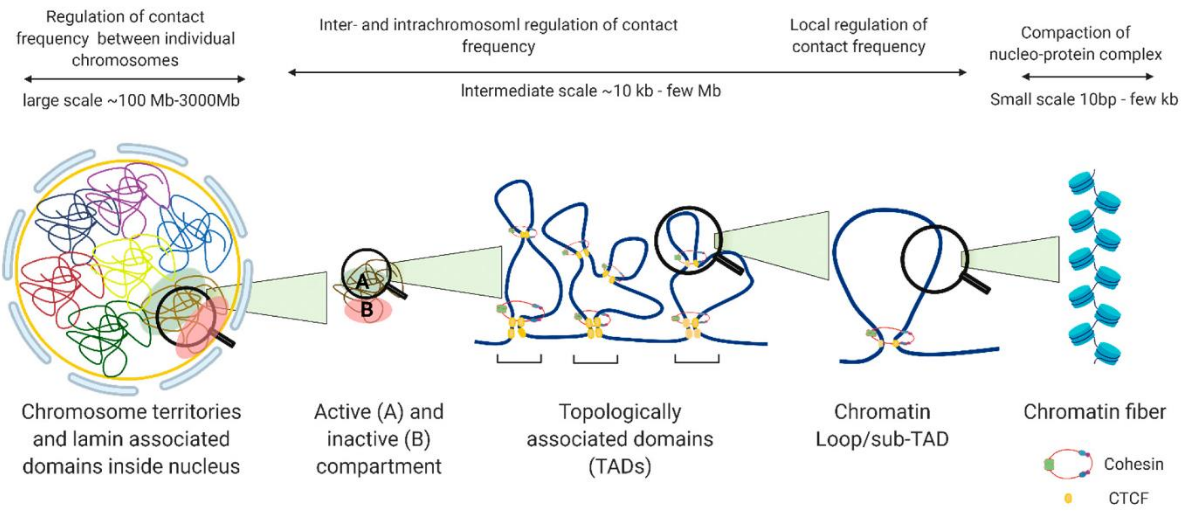

Regulation of contact frequency between individual chromosomes
large scale ~100 Mb-3000Mb
Inter- and intrachromosoml regulation of contact frequency
Local regulation of contact frequency
Intermediate scale ~10 kb - few Mb
Compaction of nucleo-protein complex
Small scale 10bp - few kb
A
B
Chromosome territories and lamin associated domains inside nucleus
Active (A) and inactive (B) compartment
Topologically associated domains (TADs)
Chromatin Loop/sub-TAD
Chromatin fiber
Cohesin
CTCF

Figure 1-2: Genome organisation in interphase cells.
This figure outlines the complexity of genome organisation. Chromosomal rearrangements can disrupt loop formation and TADs. Reproduced from: (11).

1.2.1 Chromosome territories

The arrangement of chromosomes in the interphase nucleus is not random and chromosomes occupy specific regions termed chromosome territories, this model was first proposed in 1885 by Carl Rabl and later confirmed in 1982 by Cremer *et al* (12). In interphase nuclei, the chromosomes only overlap with their immediate neighbours. Further experimental work with fluorescent in situ hybridisation (FISH) (13) has confirmed that chromosome territories are not random and that some chromosomes are located towards the nuclear periphery and some at the centre. Lieberman-Aiden *et al* in 2009 (14) identified the presence of chromosome territories in the human genome, using Hi-C (see section 1.2.2, for a description of the technique) at 1 megabase resolution. More recent work by Tavares-Cadete *et al* 2020 (15) has shown using multi-contact 3C, that in interphase nuclei, the human genome is largely not entangled. Large areas of chromosomal identity between different species have been maintained throughout evolution and these areas of identity maintain their nuclear positions in different species, irrespective of karyotypic rearrangements in the different phylogenetic lineages (16).

1.2.2 Compartments

Genome wide interactions of chromatin within the nucleus can be determined with techniques such as Hi-C, this enables the spatial interactions within the nucleus to be determined. Hi-C involves fixing a cell sample with formaldehyde and then using restriction enzymes to digest the DNA. The restriction fragments are then ligated using biotinylated nucleotides to label the ligation junctions. Ligation is carried out under very dilute conditions to favor intra-molecular ligation over inter-molecular ligation. The crosslinking is then reversed, proteins are degraded and the biotinylated DNA is purified. This results in the formation of chimeric DNA products, which represent loci that were interacting within the nucleus.

Chromosomes are organized into two distinct types of compartments, active compartments which are termed A compartments and inactive compartments which are termed B compartments. The compartments can vary in size but have a median size of ~3Mb in mouse (17). The compartments are identified through Hi-C experiments and the creation of interaction matrices (18). The distribution of the genome into compartments is linked to the transcriptional state of the chromatin, with A compartments composed of euchromatin (open chromatin) and B compartments of heterochromatin (Figure 1-1). Euchromatin is marked with active histone modifications such as H3K27ac, whereas heterochromatin is marked with repressive chromatin marks such as H3K27me3.

1.2.3 Topologically associated domains (TADs).

Topological associated domains or TADs have been identified by Hi-C analysis and are defined based on their interaction patterns, they are self-interacting sub domains within the A and B compartments and have a median size of 800kb in mouse (19). Genomic loci within the same TAD contact each other more frequently than genomic loci in different TADs. TAD boundaries composed of the CTCF protein, insulate interactions of loci in different TADs (19). The boundaries represent loci where there is a sharp break from preferential upstream interactions to preferential downstream interactions (19). The activity of promoters and enhancers within the same TAD appears to be weakly coordinated and genes within the same TAD may have similar expression patterns (20). Disruption of TAD boundaries can cause ectopic chromosomal contacts and long-range transcriptional mis-regulation (21).

TAD boundaries are maintained by CTCF and cohesins which also play a role in the formation of chromatin loops (22). TADs boundaries may be conserved between the same cell types in different species, such as mouse and human (19), but not all boundaries are conserved across evolution (19).

1.2.3.1 Chromatin loops

Chromatin loops are a substructure of TADs which also self-associate and can have insulative properties. Chromatin loops are not conserved between cell types but are thought to be related to specific regulatory events within individual cells (23). For a chromatin loop to occur CTCF must bind to specific CTCF binding sites that are orientated in opposite directions (forward and reverse) (24) on the same chromosome (Figure 1-3). The binding of CTCF together with cohesins stabilizes the chromatin loop.

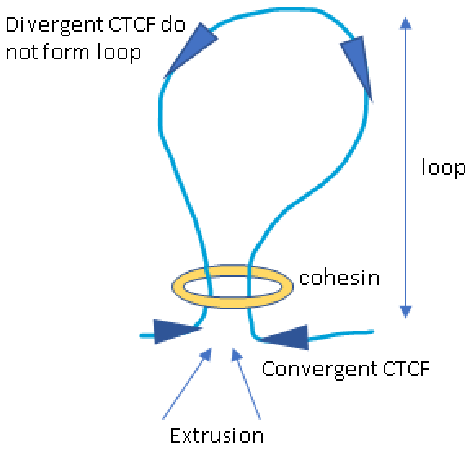


Figure 1-3: Diagram of a chromatin loop forming at convergent CTCF sites.

1.3 Structural variation within genomes and its origins

The investigation of structural variation within genomes of the same species and between different species is a key area of biological research. Structural variation is a broad term, which can be used to describe changes in genome structure (Figure 1-4). It is important to note that sequence and structural variation exists on a size spectrum, which ranges from single nucleotide polymorphisms (SNPs) up to large structural variants many megabases in size (reviewed in ((25))). Structural variants (SVs) can be divided into large SVs > 100kb called chromosomal rearrangements (CRs), which include inversions, fissions, fusions and translocations, or smaller SVs which include insertions or deletions (INDELs), copy number variants (CNVs) and transposable elements.

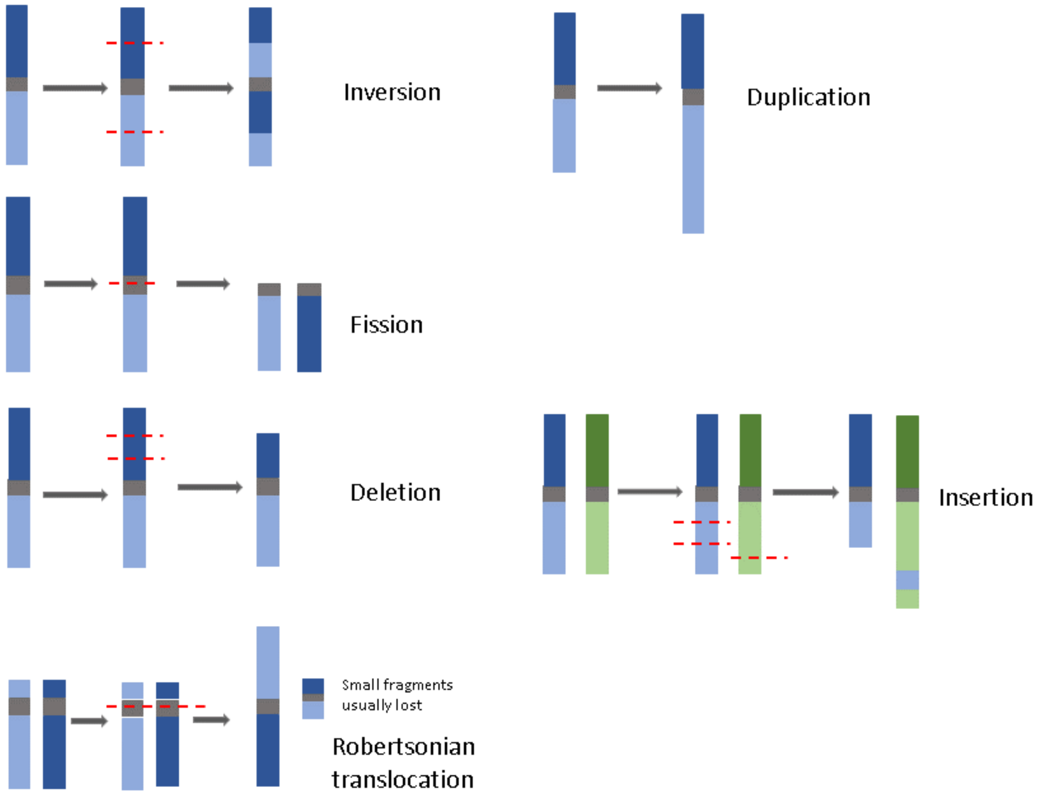


Figure 1-4: Schematic showing different types of chromosomal structural variants.

1.3.1 Chromosome classifications based on morphology.

Structurally, chromosomes are classified according to the location of the centromere, the constricted region of a chromosome that separates it into a short (p) and a long arm (q) (Figure 1-5). The centromere can be located in the middle of the chromosome with the p and q arm of equal length. This type of chromosome is termed metacentric. If the centromere is slightly off centre and the 2 arms are not of equal length, then the chromosome is termed sub-metacentric. If the centromere is shifted near to the end of the chromosome, with one arm significantly longer than the other, this is termed an acrocentric chromosome. If the centromere is located at the very end of the chromosome, then it is called telocentric (26). In a typical mouse karyotype, (I.e., one without any chromosomal fusions) all chromosomes except the Y chromosome are telocentric. The Y chromosome has a very small, short arm that cannot be visualised microscopically and is therefore technically classed as acrocentric.

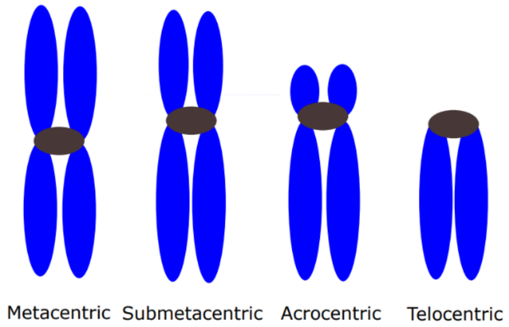


Figure 1-5: Chromosome classification according to arm type.
The black oval represents the centromere. Mouse chromosomes are all telocentric (unless any chromosome fusions have occurred).

1.3.2 Types of chromosome rearrangements

1.3.2.1 Inversions and translocations

Double-strand breaks (DSBs) can lead to the formation of chromosomal rearrangements. Different rearrangements may form depending on whether a break occurred in one or two chromosomes. An inversion occurs when a double-strand break forms and the chromosomal segment is inverted before the DSB is repaired. Translocations occur when a DSB occurs in more than one chromosome. Translocations are termed reciprocal when a break occurs in two chromosomes and then the fragments are exchanged between the non-homologous chromosomes with no loss of genetic material. A non-reciprocal translocation occurs when breaks occur in two chromosomes and one chromosome fragment is transferred to a non-homologous chromosome, some genetic material may be lost (Figure 1-6).

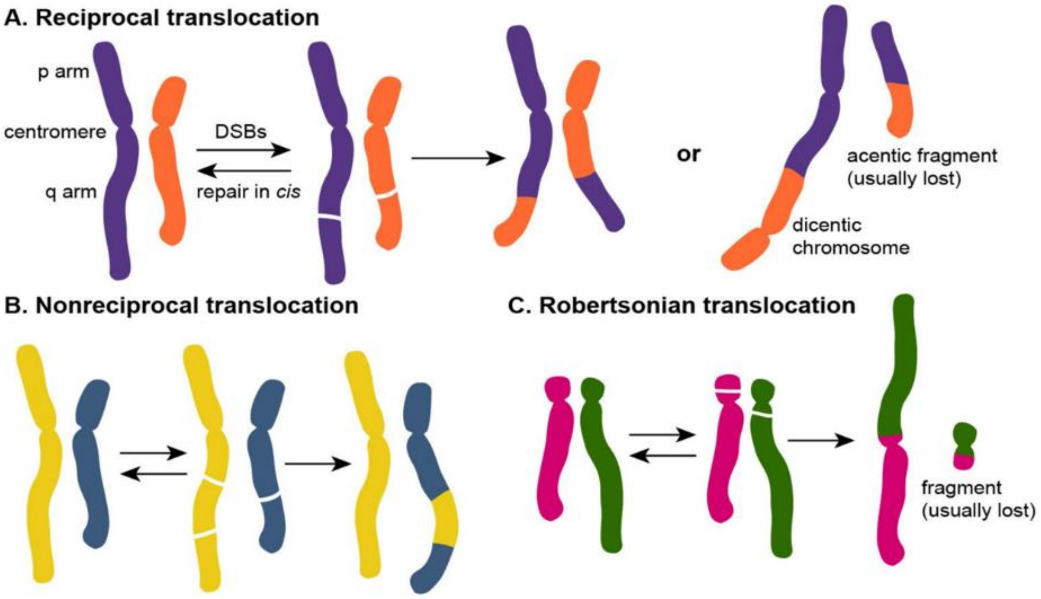


Figure 1-6: Schematic showing different types of translocations.
Reproduced from: (27) under the creative commons licence (https://creativecommons.org/licenses/by/4.0/).

1.3.2.2 Robertsonian translocations

Robertsonian (Rob) translocations occur when two chromosomes fuse (Figure 1-6). Fusions can occur between homologous or non-homologous chromosomes, with fusion of non-homologous chromosomes being more common (28). Fusion can occur between telocentric, acrocentric or metacentric chromosomes. If fusion occurs between acrocentric chromosomes, then the short arms may be lost, but this is not always the case. Rob fusions can alter the normal segregation pattern of the chromosomes during meiosis, producing trivalents (Figure 1-29), leading to disomic or nullisomic gametes (29). If upon generation of the Robertsonian fusion some DNA was lost, then this will result in unpaired chromatin when the chromosomes align at meiosis (Figure 1-7). Centromeric regions may still remain unpaired (without the loss of DNA) due to other structural rearrangements such as an inversion proximal to the breakpoint or simply because the homologous chromosomes fail to fully synapse.

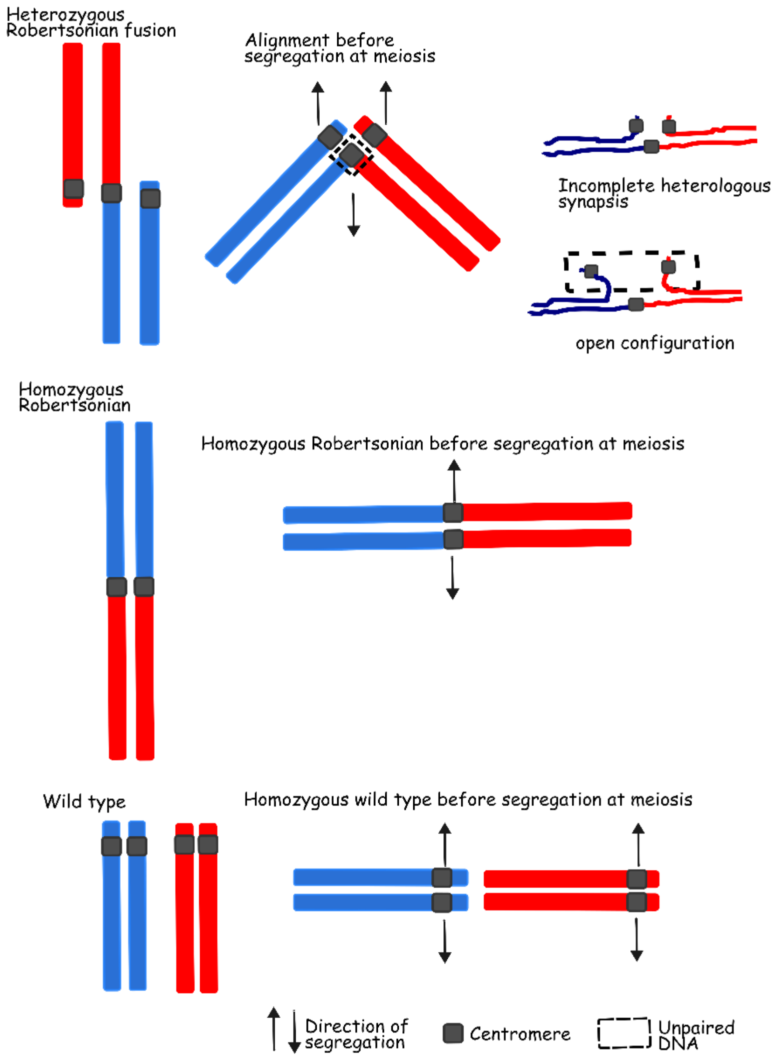

Heterozygous Robertsonian fusion
Alignment before segregation at meiosis
Incomplete heterologous synapsis
open configuration
Homozygous Robertsonian
Homozygous Robertsonian before segregation at meiosis
Wild type
Homozygous wild type before segregation at meiosis
Direction of segregation
Centromere
Unpaired DNA

Figure 1-7: An example of the alignment of heterozygous Robertsonian chromosomes (when DNA has been lost from both chromosomes), homozygous Robertsonian chromosomes and wild type chromosomes at meiosis.
DNA loss does not always occur. The dotted box outlines a region of unpaired chromatin due to the loss of the short arms of the acrocentric chromosomes.

Several examples of Rob fusions occurring in nature have been discovered, such as the Barcelona Rob system (BRobS) in which wild European house mice (*Mus musculus domesticus)* from the northeast of the Iberian Peninsula have chromosomal fusions of recent evolutionary origin (30–32).

More than one Rob fusion can be present, and the epigenetic state of the Rob chromosomes can depend on whether the fusion is in the homozygous or the heterozygous state. Heterozygous Rob fusions are when the homologous chromosomes of the Rob fusion are the non-fused chromosomes (Figure 1-8). Homozygous Rob fusions occur when the homolog also has the same Rob translocation (33).

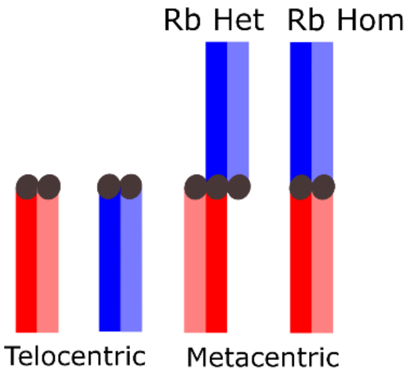


Figure 1-8: Schematic of standard telocentric chromosomes without Rob (Rob) fusions and Rob fusions in the Heterozygous (Het) and Homozygous (Hom) state. Adapted from (Vara et al 2021 (217)).

1.3.3 Evolutionary breakpoint regions (EBRs) and Homologous synteny blocks (HSBs)

When comparing genome structural variation between species, two key concepts are evolutionary breakpoint regions (EBRs) and homologous synteny blocks (HSBs). Evolutionary breakpoint regions (EBRs) are specific genomic locations where breaks occur during karyotype evolution (34). When comparing two genomes EBRs can be identified as those regions where the order of orthologous sequences differs among species. Instead, HSBs define the syntenic regions, i.e., those regions of the genome where the gene order has been conserved among species (Figure 1-9). Research indicates that EBRs and HSBs differ in their genomic content and context. In particular, the gene content around EBRs is increased relative to the genome wide average (35). EBRs are clustered in regions rich in repetitive elements and segmental duplications (36) and genes related to lineage specific biology (37). These results may appear counterintuitive, as repetitive regions do not have a high gene content. However, the exact location of EBRs is difficult to determine, so the analysis is often carried out in genomic windows of around 10kbp in size (38). These genomic windows can be both rich in repetitive elements and gene rich. Repetitive sequences may provide templates for non-allelic homologous recombination (NAHR), or non-homologous end joining (NHEJ), increasing the chances for chromosome rearrangements to occur.

Conversely, HSBs are less likely to be associated with repetitive elements, and more likely to contain constitutively expressed housekeeping genes with conserved functions across species (39).

For example, in humans, EBRs are not uniformly distributed across chromosomes, they correspond well to the location of tandem repeats (40).

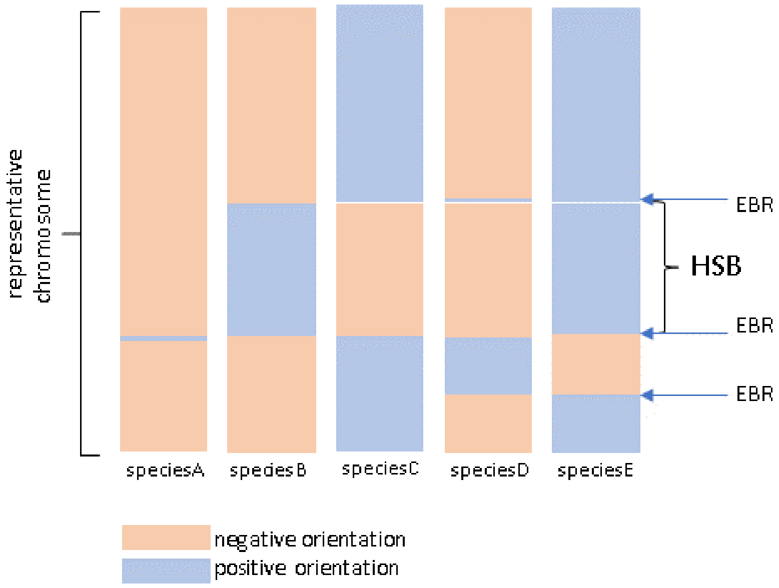


Figure 1-9: Schematic representation of an evolutionary highway plot showing evolutionary breakpoint regions (EBRs) and Homologous synteny blocks.
Shown is a representation of a chromosome from species A, with the other species columns B-E representing ancestors of species A. Syntenic blocks are shown in either pink or blue, with pink representing the negative orientation and blue representing the positive orientation with respect to species A chromosome. The small white gaps represent unaligned or unassembled regions.

These associations give clues to the nature of the mutational and selective forces generating chromosomal rearrangements. In particular, the association between repeats and EBRs (35) suggests that these repeats may destabilise DNA and predispose to the generation of rearrangements. Subsequently, a given rearrangement may alter gene expression if a breakpoint separates genes from their regulatory elements, and/or new regulatory regions may be brought into proximity to alternative genes. Thus, genes within the vicinity of EBRs may be predisposed to acquire new functions as a consequence of the rearrangement. Conversely, rearrangements that disrupt the expression of core housekeeping genes will be selectively disfavoured – thus EBRs will not be present in the vicinity of such genes.

1.3.4 Models of genome evolution

1.3.4.1 The Random breakage, Fragile breakage and Intergenic breakage models

The random breakage model developed by Nadeau and Taylor in 1984 (41) using mouse and human linkage maps was the first model to attempt to explain genome evolution. The hypothesis had two main assumptions. Firstly, between related species many large chromosomal blocks are conserved, HSBs, which is presumptive evidence for linkage conservation. Secondly, genomic rearrangements on autosomes may become fixed during evolution and are randomly distributed throughout the genome (41). When genomic sequences from the human and mouse genomes became available the second assumption was challenged by Pevzner and Tesler in 2003 (42). They found that genomic rearrangements on autosomes are not randomly distributed throughout the genome, but rather are concentrated within fragile regions where breakpoints can be reused i.e., between syntenic blocks on the human and mouse genomes some regions contain breakpoints for multiple rearrangement events. Fluorescent in situ hybridisation (FISH) and cross-species FISH experiments then followed (43) along with whole genome comparisons, which confirmed that the pattern of breakage was not random, but that breakage occurred in hotspots. However, the exact location of these EBRs was only discussed in the Intergenic Breakage Model (44), suggesting that selection prevents breaks occurring within genes and regulatory regions upstream from genes. The model holds that DSBs (the origin of EBRs) are not located at ‘’preferred’’ sites in the genome, instead they appear to be random but only in the sense that those that do not disrupt essential genes and/or gene expression actually become fixed.

1.3.4.2 Integrative Breakage Model

The Integrative Breakage Model is a multidisciplinary hypothesis for the study of genome evolution proposed by Farré *et al* 2015 (44). The model proposed that it is both the chromatin conformation combined with the DNA sequence that are important in understanding how and at which stage of the cell cycle chromosomal rearrangements are formed and consequently passed onto the next generation. If a DSB occurs in a region that forms secondary structure and this secondary structure is subsequently disrupted, or if the DSB modifies the expression of key genes related to development or basic cellular functions (such as housekeeping genes), then it will not become fixed and will not be of evolutionary consequence (44). The Integrative Breakage Model also explains the genomic content of multi species homologous syntenic blocks (msHSBs). These are genomic blocks that are conserved in several species that share a common order of homologous genes derived from a common ancestor. The msHSBs are enriched for gene networks that control embryonic and tissue development (39). Genes that aid adaptive responses tend to be located within EBRs (37, 39). Following on from this, the Integrative Breakage Model suggests that EBRs occur in regions of fragility, whether due to genomic sequence or epigenetic state, and that only those not breaking regulatory blocks would become fixed.

Techniques such as Hi-C (see section 1.2.2) enable researchers to determine which regions interact in the nucleus and therefore answer the question of whether genomic regions that tend to break and reorganize are interacting inside the nucleus. Zhang *et al* (2012) (45) have shown in mouse interphase nuclei that genomic regions involved in translocations are found in close proximity.

As structural variation is so important in evolution this raises the pivotal question of where this variation comes from and how it is generated. Novel structural variation arises when DNA is broken and re-joined incorrectly (otherwise variation will not be induced). This must happen in a germline cell to be passed on to the next generation. Then, to be fixed in the population, the novel variant must spread to fixation. This can occur either through genetic drift (the change in frequency of an existing gene variant in the population due to random chance), or via selective dynamics.

The evolutionary origin of structural novelty is therefore linked to the Integrative Breakage Model, (44) which states that rearrangements arise due to inaccurate repair of DSBs occurring in the germline, leading to rearrangements between loci that are in close physical proximity. The Integrative Breakage Model also proposes that chromatin conformation is a key aspect which must be considered in order to understand how and at what point in the cell cycle novel chromosomal rearrangements may be generated. There is also the possibility of transmission ratio distortion, which could lead to non-Mendelian inheritance of structural variants. For example, a Robertsonian chromosome could form when a DSB occurs in two chromosomes that are in close proximity. Meiotic drive refers to a class of mechanisms that cause deviation from a 1:1 Mendelian segregation ratio. In females this would be preferential segregation of a chromosome to the oocyte or the polar body rather than a 50:50 segregation ratio. In females, a Robertsonian chromosome may show meiotic drive as a result of a stronger centromeric signal. This may bias its segregation to the oocyte and not the polar body, as posited by Henikoff *et al* 2001 (46) and later shown by Chmatal *et al*, 2014 (47).

Using the Integrative Breakage Model to study the origin of structural novelty generates another important question-does variation arise predominantly in the male germline during spermatogenesis or in the female germline during oogenesis?

It is still unknown whether structural novelty arises equally in males or females or if it is biased to one of the sexes. Male and female gametogenesis have significant biological differences, which may affect both the formation of structural variants and subsequently how they are transmitted. In particular, the formation of novel variation will be influenced by differences in the frequency of

DNA breakage, by differences in the DNA repair mechanisms involved, and by differences in the 3D chromatin structure of male vs female germ cells. Similarly, the transmission of novel variants will be influenced by the checkpoints that act to remove cells with DNA rearrangements, and the unique selective dynamics that permit or prevent the opportunity for selection of haploid gametes.

1.3.4.3 Genome structure affects spatial organisation within the nucleus.

Robertsonian fusions that alter the standard karyotype of a species provide a convenient model with which to interrogate links between genome structure and the physical organisation of the DNA in the nucleus. Rodents are a particularly useful example, since even closely related rodent species can have different chromosome numbers, for example the house mouse (*Mus musculus*) has a diploid number of 40 (48), whereas the red viscacha rat (*Tympanoctomys barrerae*) has a diploid number of 102 (49). This thesis will focus on the house mouse *Mus musculus*. The standard somatic-cell karyotype of the laboratory mouse consists of 20 pairs of acrocentric chromosomes, but this can vary due to Rob fusions. Mice from the BRobS system have a diploid number ranging from 2n-27 to 2n=40 (31). The decrease in the diploid number results from the formation of metacentric chromosomes as a result of Rob fusions.

In mice with a standard karyotype of 2n=40 where all chromosomes are telocentric, the pericentric regions are closely associated with the nuclear envelope (50). During leptotene of meiosis the chromosome ends come together in specific regions of the nucleus, forming a polarized arrangement termed a bouquet (51). In normal meiosis, therefore the pericentromeric regions of different chromosomes are in close proximity and may interact, which may predispose to the formation of novel Rob fusions.

In animals carrying such fusions, the presence of metacentric (Rob fusions) alters the nuclear architecture. The pericentric regions of the metacentric chromosomes (I.e., the fusion partners) are not associated with the nuclear envelope but located in the nuclear interior. This may alter their expression, their regulation, their ability to pair during meiosis and DNA exchange during meiotic recombination. The pericentromeric regions of the remaining telocentric chromosomes (I.e., those not involved in the fusion) remain within the nuclear periphery (50). This illustrates the presence of an intimate bidirectional causal links between genome structure and function during meiosis: the requirement for specific types of pairing behaviour will influence the probability that specific structural rearrangements are able to occur, and these rearrangements in turn will alter the regulation of events during gametogenesis.

1.4 Gametogenesis in mouse: understanding germ line development.

Given the intimate connection between molecular genetic events during gametogenesis and the potential to generate structural rearrangements, I will now review the broad principles of gamete production in mammals. In this thesis I will concentrate primarily on spermatogenesis, with a specific focus on male-specific aspects of spermatogenesis and how these differ from oogenesis. In section 1.5 of the Introduction, I justify this focus with reference to the male-specific wave of DNA damage formation in spermatids, which I argue leads to a male-specific potential to induce chromosomal rearrangements during gametogenesis.

1.4.1 Key differences between oogenesis and spermatogenesis

Gametogenesis is the process by which cells undergo meiosis to produce gametes (eggs and sperm). This process differs in males and females (Figure 1-10), but both processes start with a mitotic phase, which is succeeded by a meiotic phase, which in females is asymmetric but in males is symmetric. In males there is a subsequent post-meiotic differentiation step in which the sperm nucleus condenses, and the cells change shape. The male gametes then have a free-living haploid stage where there are limited resources to repair DNA damage and they are exposed to oxidative stress. Following fertilisation, male specific de-condensation of the chromatin must occur. Finally, the fertilised zygote will repair DNA damage sustained by the sperm using resources derived from the oocyte.

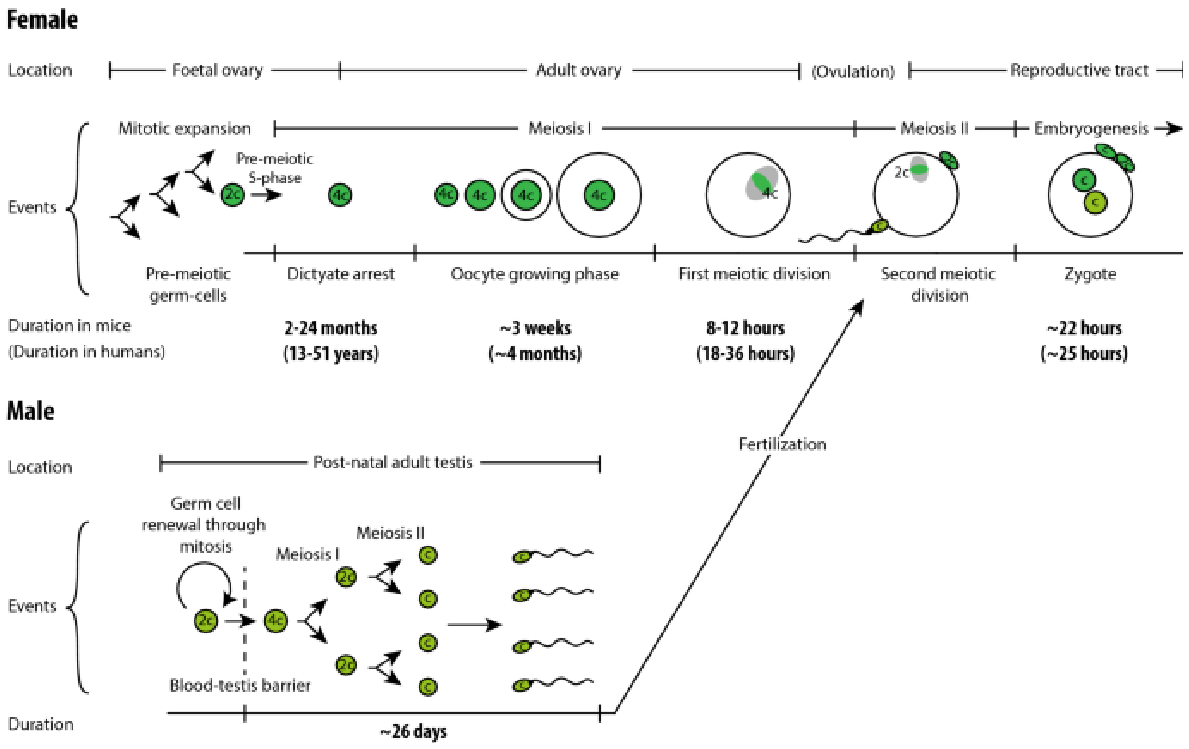


Figure 1-10: A comparison of female versus male meiosis.

The duration of female meiosis is substantially longer than that of male meiosis, mainly due to the dictyate arrest that can last for years. Reproduced (with no changes) from: (52)
with permission from: http://creativecommons.org/licenses/by/4.0/

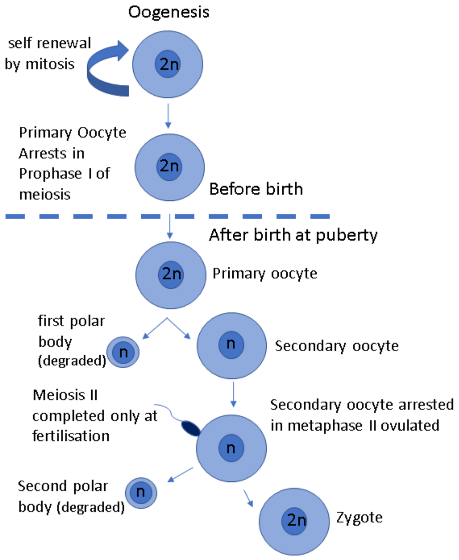

Oogenesis
self renewal by mitosis
2n
Primary Oocyte Arrests in Prophase I of meiosis
2n
Before birth
After birth at puberty
2n
Primary oocyte
first polar body (degraded)
n
n
Secondary oocyte
Meiosis II completed only at fertilisation
Secondary oocyte arrested in metaphase II ovulated
n
Second polar body (degraded)
n
2n
Zygote

Figure 1-11: Schematic of mouse oogenesis showing the main stages.
Oocytes only become haploid upon fertilisation.

In contrast to spermatogenesis, oogenesis is initiated in foetal development and is not a continuous and cyclic synchronised process. Oogonia differentiate from primordial germ cells in female mice shortly after birth and a finite number of oocytes are arrested at the first meiotic prophase, which mature to metaphase II oocytes upon hormonal stimulation. In contrast to males, there is no chromatin compaction and haploid selection acting on individual oocytes does not occur. In males, haploid selection acting on individual sperm can lead to transmission ratio distortion and thus affect the evolutionary fate of rearrangements once they have occurred. For example, in males heterozygous for the Robertsonian fusion Rob6.16, the fusion chromosome is under-transmitted to the next generation relative to its normal counterparts by 2.7-fold (53). In oocytes, in contrast to the haploid sperm, templates are available for homology directed repair. Meiotic division in oocytes completes on fertilization and so there is never a truly haploid stage (reviewed in (54)), in which the genome is vulnerable to alterations. In spermatids the first wave occurs during crossing over at prophase 1 and the second wave occurs during genome remodelling in spermatids. Whereas in oocytes only one wave of programmed DSBs occurs during prophase 1, which is required for meiotic recombination (reviewed in (55)). In homogametic females (e.g., XX mice) the sex chromosomes can undergo proper pairing and recombination. In heterogametic males (e.g., XY mice) this cannot occur except for the small region of homology at the pseudoautosomal region (PAR). This can lead to transcriptional repression through meiotic sex chromatin inactivation (MSCI). In contrast to spermatogenesis in which cytogenesis occurs evenly, in oogenesis unequal cytokinesis occurs producing a large ovum and a small polar body which is degraded (reviewed in (56)).

1.4.2 Mouse spermatogenesis

Spermatogenesis is the continuous and cyclical process occurring in the seminiferous tubules of the testis in which diploid spermatogonia form haploid spermatozoa (mature sperm). It can be broadly split into three phases.

The proliferative phase, in which the spermatogonia are rapidly dividing, the meiotic phase, involving recombination and segregation of genetic material (spermatocytes) and a post meiotic spermiogenic phase (57).

Spermatogenesis starts with adult germline stems cells (AGSC or spermatogonia) which have differentiated from primordial germ cells, they are the first intermediate cell type of the process of spermatogenesis. AGSC then undergo meiosis to produce primary spermatocytes (SC). Post meiosis, haploid round spermatids (ST) are generated.

In the post-meiotic phase, the spermatids transform through successive developmental stages, which involves remodelling of the cell shape and the sperm head, from a round to an elongating, to a condensed state. Finally mature sperm capable of fertilizing an oocyte are produced.

Morphological criteria can be used to subdivide spermatogenesis in the mouse into 16 steps (57) (Figure 1-12). In mice at steps 11-12 the replacement of histone proteins with transition proteins and then finally protamines is initiated (58). Widespread DNA strand breaks occur during steps 9–11 in elongating spermatids (59).

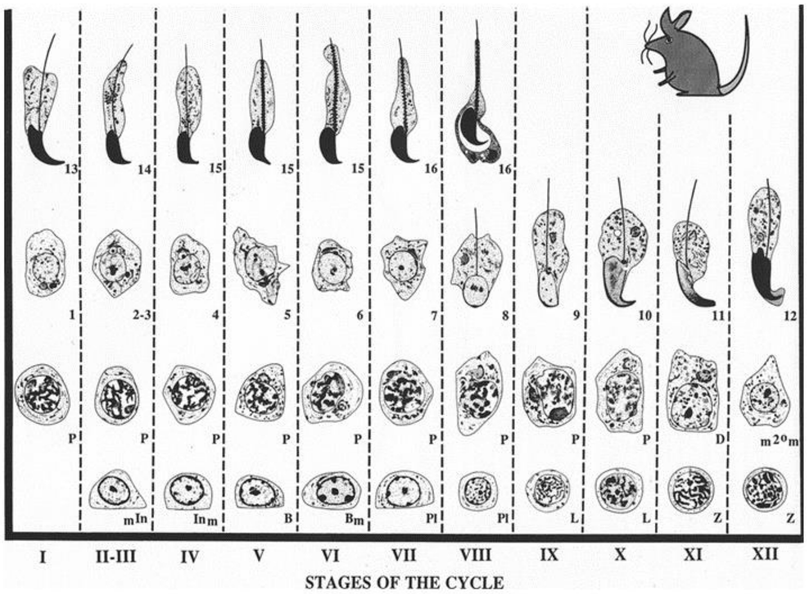


Figure 1-12: The division of mouse spermatogenesis into the different stages.

The Roman numerals represent the different stages and the subscript numbers below the cell images represent the different steps of spermatogenesis. Reproduced from (57).

During germ cell division in the testis, sister cells originating from the same stem cell remain connected to many others of the same type by intracytoplasmic bridges (Figure 1-26). These bridges are thought to have two functions: firstly, they promote synchronous development of germ cell clones as transcripts are shared across the bridges (60), and secondly, they allow for sharing of X and Y-linked transcripts between cells after the chromosomes have segregated during meiosis, thus enabling all cells to access essential sex-linked genes irrespective of which chromosome they carry. The bridges are formed during telophase of mitosis when cells fail to divide completely, i.e., cytokinesis is incomplete.

1.4.3 Mitotic phase of spermatogenesis

Mitosis, the first stage of gametogenesis is the process of cell division without reduction of DNA content, in which a cell replicates its chromosomes and then divides to produces two daughter cells.

Spermatogonial stem cells (SSC) replicate via mitosis. The cells undergo numerous mitoses to produce a large population of cells that will subsequently undergo meiosis and differentiate into sperm. This stage is also known as the proliferative stage (57) as it increases the cell numbers early on in the process of spermatogenesis. Mitotic division of SSC occurs to produce type A or type B spermatogonia (61). Type A spermatogonia replenish the stem cell population and type B spermatogonia develop into spermatocytes (SC) (57).

A type A spermatogonia (A_{single} or A_s) undergoes a self-renewing division producing two new A_s cells (Figure 1-13). These then divide to produce a pair of spermatogonial cells (A_{paired} or A_{pr}) which then continue to divide into 4–16 and even 32 spermatogonial cells ($A_{aligned}$ or A_{al}) via a sequence of mitotic cell divisions (62, 63). A_{pr} and A_{al} cells are connected by intercellular bridges as incomplete cytokinesis (cell division occurs) (63, 64).

When the A1 spermatogonia (that differentiate from the A_{al} cells) undergo mitosis they move to the seminiferous tubules (65), where five more mitotic divisions occur forming A2, A3, A4, Intermediate and B spermatogonia.

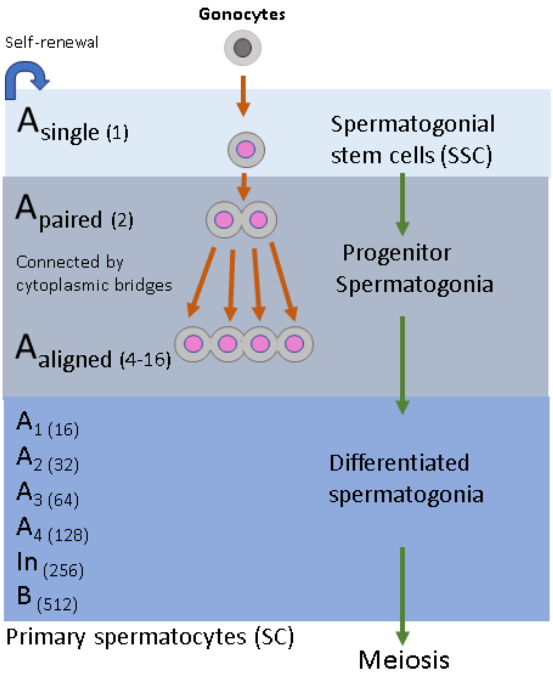


Figure 1-13: Spermatogonial stem cell development, showing the different mitotic divisions. The paired cells remain connected by cytoplasmic bridges. Once primary spermatocytes have been generated meiosis then occurs.

1.4.4 Spermatogonia

Spermatogonia are relatively immature cells which undergo multiple rounds of mitosis to produce a large population of cells that will undergo meiosis to form mature sperm (57). Multiple rounds of mitosis (differentiation) are required to help build up a large population of cells so that sexually mature mammals can produce millions of sperm per day. Not all cells are committed to mitosis, a population of undifferentiating cells must be maintained to allow spermatogenesis to commence regularly from a stem cell pool within each seminiferous tubule (66).

There are three main types of spermatogonia, stem cell spermatogonia, proliferative spermatogonia and differentiating spermatogonia. The stem cell population is required for spermatogonia maintenance. At the end of the differentiating process the most mature spermatogonia divide to form spermatocytes.

Spermatogonia can be divided into different classes: A spermatogonia, in which the nucleus does not contain heterochromatin, B spermatogonia with profuse heterochromatin, and intermediate types in mouse. In the spermatogonial compartment, three further types can be distinguished A_{single} (A_s or stem cell spermatogonia) (67), A_{paired} (A_{pr}), and $A_{aligned}$ (A_{al}). If the A_s cells divide fully, two stem cells are formed. If cytokinesis (cytoplasmic division) is incomplete, then the cells are termed A_{paired} and are connected by intracellular cytoplasmic bridges. The paired cells can divide further to form aligned spermatogonia with up to 16 cells connected (66).

The A_{al} spermatogonia differentiate into A1 differentiating spermatogonia. The A1 cells then undergo five divisions finally forming B spermatogonia. After a mitotic division the type B spermatogonia form primary spermatocytes.

1.4.5 Spermatocytes

The primary spermatocytes formed by division of the type B spermatogonia divide to form preleptotene spermatocytes (57). These are the last cells of spermatogenesis to undergo the s-phase of the cell cycle. The DNA is replicated to give 2n cells. Spermatocytes then pass through two meiotic divisions. The first meiotic prophase is a prolonged stage in which recombination occurs. This is followed by two rapid meiotic divisions to produce haploid spermatids.

1.4.6 Meiotic phase

This occurs after the differentiation phase and starts with young primary spermatocytes in the preleptotene stage. Recombination of the chromosomes occurs, and genetic material is halved in each cell during meiosis I and meiosis II (Figure 1-15). The meiotic prophase stage of meiosis I is typically long (1.5-2 weeks), which is followed by two more rapid divisions forming the haploid spermatids (57). The transition of cells through prophase is a continuum, without stepwise changes. Morphologically the stages can be differentiated by nuclear changes. The stages are preleptotene, leptotene, zygotene, pachytene and diplotene. In zygotene homologous chromosomes are paired along their length via the synaptonemal complex (SC), genetic recombination via crossing over occurs, through the generation of double-strand breaks initiated by the type II topoisomerase Spo11 (68). The nucleus increases in size and the sex vesicle forms.

During the brief diplotene stage the synaptonemal complex breaks down, which allow the homologous chromosomes to separate. They cannot separate at regions of crossing over (chiasmata). Diplotene cells are the largest of all germ cell types (57). The remaining stages of metaphase, anaphase and telophase are termed meiosis I, resulting in the formation of secondary spermatocytes. The second meiotic division, meiosis II forms spermatids. Cells at all stages of meiosis II are smaller than meiosis I and all stages are brief with no extended prophase as in meiosis I.

1.4.7 Marks of meiotic recombination hotspots PRDM9 and DMC1

PRDM9 is a zinc finger domain containing protein located on chromosome 17 that determines the positions of meiotic recombination hotspots in most mammals (69). Not all regions bound by PRDM9 will become a recombination hotspot (70). PRDM9 binds to DNA sequences at the centre of hotspots (71). Therefore, ChIP-sequencing data of PRDM9 can be used to map the location of meiotic recombination hotspots. Meiotic DSBs are not randomly distributed along the chromosomes but occur in specific regions of the genome and are clustered into hotspots (72). PRDM9 tri-methylates nearby histone H3 proteins at lysine 4 and 36 forming H3K4me3 and H3K36me3 respectively (73, 74). PRDM9 recruits SPO11 at a fraction of binding sites to form double-strand breaks (DSBs) (68). End resection of the DSBs occurs with the ssDNA binding DMC1 (reviewed in (75)) Figure 1-14.

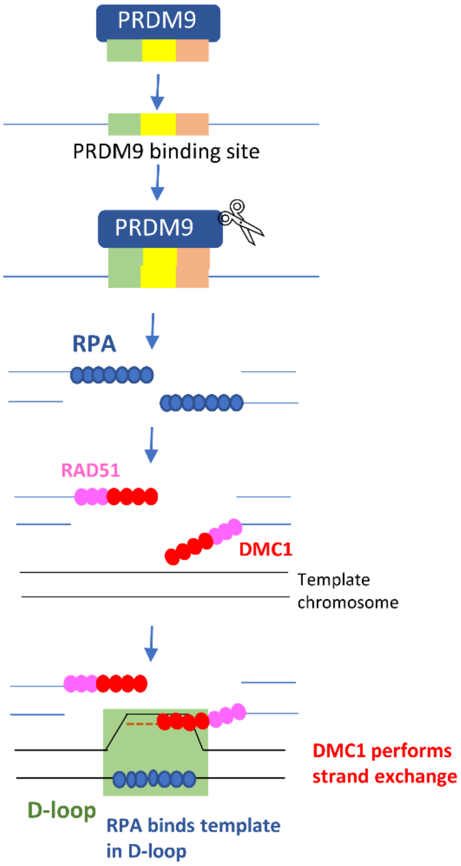


Figure 1-14: Schematic representation of the role of PRDM9 and DMC1 in meiotic DSB formation.

DMCI is a meiotic recombination protein, that mediates homologous chromosome pairing during homologous recombination (76). Therefore, like PRDM9, DMCI ChIP-seq data can be used to indirectly determine the location of meiotic DSBs.

Around 200-300 programmed meiotic DSBs occur per cell during the leptotene/zygotene stage of meiosis (77), most will be resolved (repaired) without the generation of a cross-over.

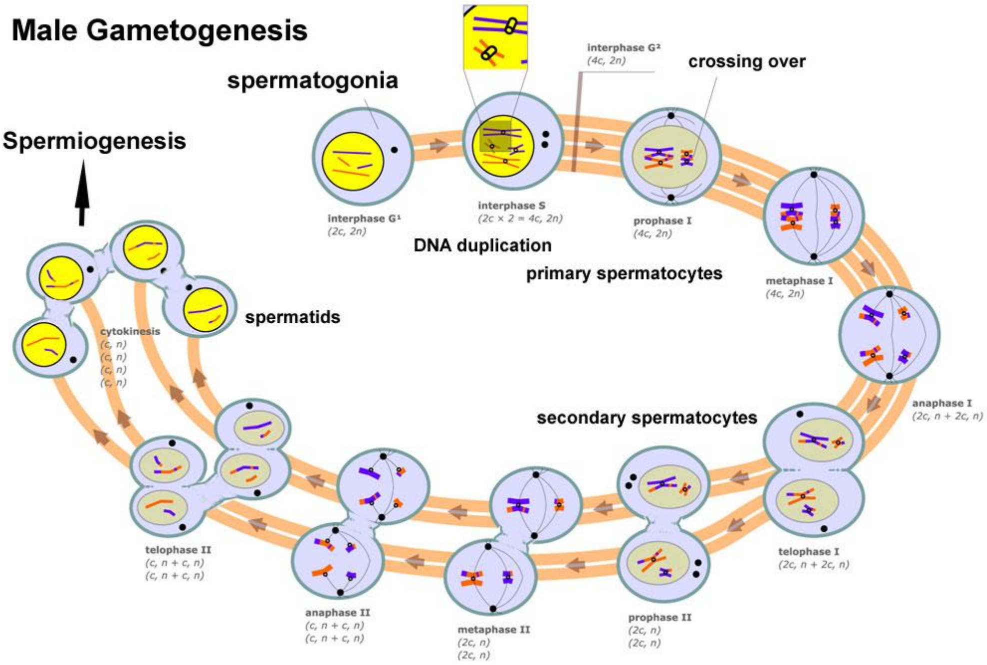


Figure 1-15: Male gametogenesis.
The three main cell types of spermatogenesis: spermatogonia, spermatocytes and spermatids are shown. The spermatogonia are immature cells that can differentiate into primary spermatocytes, these cells then undergo 2 rounds of meiosis resulting in the production of haploid spermatids. The spermatocytes do not divide fully during meiosis but remain attached by cytoplasmic bridges. Reproduced from: Hill, M.A. (2023, June 30) Embryology Male gametogenesis.jpg. Retrieved from https://embryology.med.unsw.edu.au/embryology/index.php/File:Male_gametogenesis.jpg

1.4.8 Post Meiotic phase.

Spermatids must develop into mature spermatozoa and this process takes around 14 days in mice (78). It occurs without cell division. Extensive remodelling of the sperm head must occur, with the shape being distinct between species. In mouse it forms a sickle like shape (Figure 1-12). As the nucleus decreases in size the DNA must be more tightly packaged to be accommodated. A 75% reduction in cell volume occurs (79).

To achieve this compaction the 3D genome is dramatically re-structured during gametogenesis, by global changes to chromatin structure, most notably most histone proteins are replaced by protamines. Compaction helps to protect the DNA from damage as well as allowing packaging in the condensed sperm head.

Not all histone proteins are replaced by protamines and approximately 1% remain in mature sperm in mouse (80). Histones may remain in regions that are important for embryonic development (81), which suggests that they may mark regions that are epigenetically regulated in early embryo development. The transition from less densely packaged chromatin, packaged with nucleosomes, to a highly condensed protamine-based chromatin structure takes place during spermiogenesis, the post-meiotic haploid phase of spermatogenesis. In round spermatids there is increased acetylation of histone proteins (hyperacetylation). This may weaken the interaction of DNA with the histone proteins, enabling testis specific histone variants to be inserted. These histone specific variants are then replaced by transition proteins (TP1 and TP2), which finally are replaced with the basic proteins termed protamines (Figure 1-16).

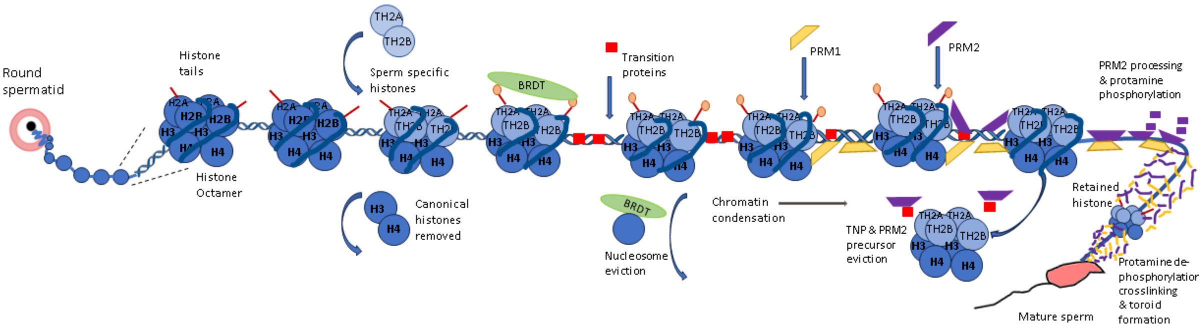


Figure 1-16: Male gametogenesis showing the role of the transition proteins.
BRDT (green oval) binds to acetylated residues. PRM1 (protamine 1) is shown in yellow and PRM2 (protamine 2) in purple. Not all histone proteins are removed in the mature sperm and some (~1%) remain, shown as ‘retained histone’.

The switch from nucleosomes to protamines changes the DNA structure from a supercoiled state to a toroid conformation (Figure 1-17) (82, 83). During this switch double-strand breaks (DSBs) occur (84). Chromatin compaction is required to accommodate the sperm chromatin in the small sickle shaped sperm head, the chromatin is compacted to an almost crystal-like structure (83). The compaction helps to prevent DNA damage as well as enabling the chromatin to be contained in the hydrodynamically shaped sperm head.

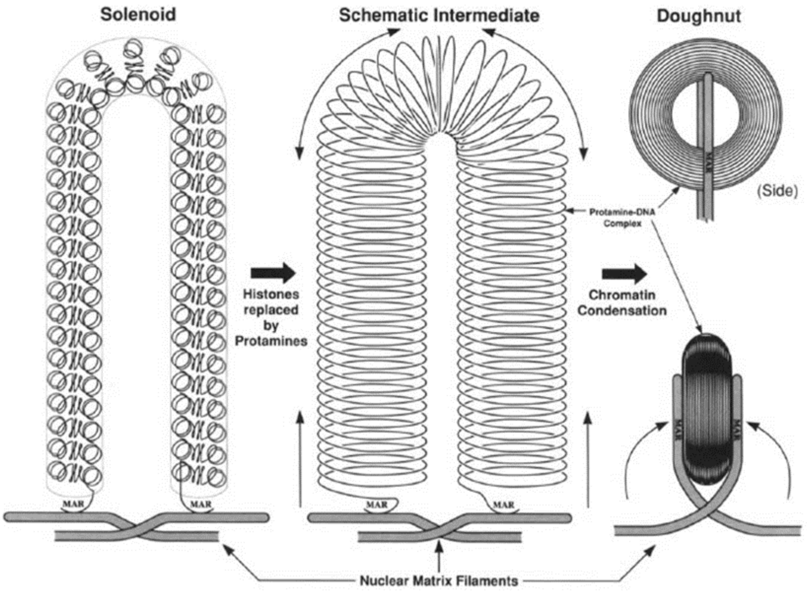


Figure 1-17: Model of solenoid equivalent in one sperm DNA loop.

The model is diagrammed for one loop of DNA that is 47 kbp in length, and the comparative structures are drawn to scale, viewed from above (top) and from the side (bottom). In the parent cell, the DNA is in the solenoid configuration (left). As the histones are replaced by the protamines, each turn of the solenoid becomes two concentric circles (centre). In the doughnut structure, the protamine-bound DNA circles are collapsed into a toroid-shaped structure made up of 72 circles of DNA with an average diameter of 65 nm (right). The schematic intermediate (centre) is drawn only as an instructive diagram, and is not predicted as a real, functional intermediate that occurs during spermiogenesis. The actual transition must be much more complex, involving transitional proteins that are not considered in this model. Reproduced with permission from Oxford University Press (licence number 5581851136048) from: (85).

1.4.9 Round spermatids.

The round spermatid is a male gamete that has just completed the second meiotic division in the testis and therefore has a haploid gene content. As the name implies, they are round in appearance and are approximately 10µm in size (86). Spermatids can be morphologically classified into different steps, steps 1-8 are round spermatids and from step 9-16 onwards they transition through the elongating and condensing steps (57) (Figure 1-12 and Figure 1-18). The term spermiogenesis is used to describe the transition of round spermatids to motile mature sperm. As round spermatids are haploid cells any DBSs cannot be repaired by homologous recombination but must rely on the non-homologous end joining (NHEJ) pathway (see section 1.7.1).

1.4.10 Why do DSBs occur in round spermatids?

The 3D genome is dramatically re-structured during gametogenesis, which involves global changes to chromatin structure, most notably most histone proteins are replaced by protamines which helps the DNA to compact. DNA compaction helps to protect the DNA from damage as well as allowing packaging in the remodelled condensed sperm head. Not all histone proteins are replaced by protamines and approximately 1% remain in mature sperm in mouse (80). Histones remain in regions that are important for embryonic development (81), which suggests that they may mark regions that are epigenetically regulated in early embryo development. The transition from less densely packaged chromatin, packaged with nucleosomes, to a highly condensed protamine-based chromatin structure takes place during spermiogenesis, the post-meiotic haploid phase of spermatogenesis. In round spermatids there is increased acetylation of histone proteins (hyperacetylation). This may weaken the interaction of DNA with the histone proteins, enabling testis specific histone variants to be inserted. These histone specific variants are then replaced by transition proteins (TP1 and TP2), which finally are replaced with the basic proteins termed protamines.

Protamine bound DNA is less supercoiled than histone bound DNA, forming protamine toroid loops, which help the DNA to compact to 1/10th the size of a somatic cell (87). The switch from nucleosomes to protamines therefore substantially alters the winding number of the DNA, as it transitions from a supercoiled state to a toroid conformation. During this switch double-strand breaks (DSBs) occur in order to relieve the helical tension caused by the rewinding process. It is thought that the induction of DSBs occurs as a matter of necessity, and that post meiosis around 5-10 million DSBs may occur per cell (88). These DSBs are likely to be induced by topoisomerase II beta (TOP2B), as is the case for strand breaks associated with other remodelling phenomena such as transcription and DNA replication (89). The transient DSBs it creates help to unwind and detangle the DNA to reduce supercoiling.

1.4.11 Elongating, condensing spermatids and mature sperm

In elongating spermatids at step 9 (Figure 1-12), the tail has started to develop. Bilateral flattening of the nucleus occurs, which continues in step 10. In step 11 spermatids, further elongation of the sperm head occurs, but pronounced condensation of the sperm head has not yet begun (Figure 1-12). Step 12 spermatids have the longest nucleus of any stage (57) of spermatogenesis and chromatin condensation has occurred. In step 13 spermatids, the spermatid head has shortened and starts to take on a sickle shaped appearance, further chromatin compaction occurs. In step 14 the sperm head further shortens and in step 15 it narrows (57). In step 16 (the final stage) the spermatid head forms a prominent hook. Excess cytoplasm and organelles are removed from the mature sperm, the mature sperm are then released from the Sertoli cells into the lumen of the seminiferous tubules. These non-motile spermatozoa then enter the epididymis where they develop into mature motile sperm. Here the spermatids are exposed to oxidative stress, which leads to further chromatin compaction but can also cause oxidative damage to the DNA.

1.4.12 Retained histones in mature sperm.

As discussed previously, not all histones are replaced with protamines in mature sperm. Estimates for retention range from between 1-10% of histones, depending on the species. It was initially thought that the location of retained histones was random and was a result of the incomplete exchange with protamines. However, retained histones have been found to be important in embryo development. Disruption of sperm histone methylation during either spermiogenesis or at fertilization alters embryonic gene expression and development, suggesting modified histones in sperm chromatin are required for embryonic development (90, 91).

In mammals, zebrafish and frogs, developmentally important genes are marked by modified histones in sperm, a feature that correlates with their expression in the early embryos (80, 92).

Different histone subunits can be retained at different genomic locations. Histone H4 is retained at distal intergenic regions. Modified histones show enrichment in specific genomic elements, with the modification type determining enrichment location, for example H3K4me3 is found at CpG islands (as shown in the spermatid ChromHMM data Figure 4-2 chapter 4 and H3K9me3 in satellite repeats (81).

1.4.13 Post fertilization chromatin changes that occur in the embryo.

Mature sperm are transcriptionally inert, due to genome compaction (93, 94). Upon fertilisation the compaction of the sperm genome will have to be reversed, with the removal of protamines and their replacement with histone proteins. Initially maternal factors from the egg cytoplasm control development while the zygote genome is silent. Then development switches from being controlled by maternal factors to zygotic control, this is called the maternal-to-zygotic transition (MZT). The first wave of transcription is termed zygotic genome activation (ZGA) and begins during the 2-cell stage in a mouse implantation embryo, it is associated with large changes in chromatin structure (95). Falco *et al* 2007 (96) identified that the gene *Zscan4* is expressed during ZGA in the late 2-cell embryo stage. Srinivasan *et al* in 2020 (97) found that ZSCAN4 binds to nucleosomal microsatellite DNA and protects mouse two-cell embryos from DNA damage. This would suggest that genome remodelling in the embryo represents a period of instability in which the genome is vulnerable to damage.

1.5 The Integrative Breakage Model in the context of male and female gametogenesis

Having discussed the principles of evolutionary genomic rearrangement formations and the overall processes of gametogenesis, I will now consider the Integrative Breakage Model of genome evolution in the context of the germline events involved in reproduction. Recall that this model (44) states that chromosomal rearrangements arise due to inaccurate repair of DSBs occurring in the germline, which can then cause translocations between loci that are in close physical proximity. Male and female gametogenesis have distinct biological differences which will impact DSB formation, the repair processes available at the different stages of gametogenesis and the 3D organisation of the genome. Here I will outline these differences. In this, it is useful to divide gametogenesis into three stages – premeiotic, meiotic and post meiotic events. Premeiotic events encompass those occurring in primordial germ cells, oogonia and spermatogonia. Meiotic events occur in oocytes and spermatocytes, while post meiotic events are effectively restricted to the male germline since the egg only completes meiosis at the instant of fertilisation.

1.5.1 DSB formation during gametogenesis

There are two major waves of DSB formation that occur during gametogenesis. Meiotic DSBs occur in both male and female gametogenesis catalysed by the SPO11 protein. These double-strand breaks are required for homologous recombination between the homologous chromosomes and are vital for proper meiotic disjunction. Importantly the location of DSB hotspots is conserved across male and female meiosis, although the frequency with which any given DSB hotspot incurs a break may vary between the sexes (98). Thus, either male or female data can be used to study the potential link between meiosis recombination hotspots and evolutionary chromosome rearrangements.

The second wave of DSBs is specific to spermiogenesis (see section 1.4.10), occurring at the transition from round to elongating spermatids as a result of chromatin remodelling. A further male-specific vulnerability arises in the mature sperm themselves, which are subject to oxidative attack during their lifespan as free-living cells in the epididymis. Breakages at this point cannot be repaired by the sperm as the chromatin is fully condensed with protamines and inaccessible to repair enzymes, and so these breaks are repaired post-fertilisation, in the zygote.

In contrast to these programmed waves of DSBs during and after meiosis, breakages prior to meiosis are relatively rare. Since much of the developmental programme prior to meiosis is shared, it is likely that the location of premeiotic DSB hotspots is also shared between the male and female germlines, though this remains to be established.

1.5.2 DSB repair during gametogenesis

In addition to the differential vulnerability to DSB formation at different stages of gametogenesis, DNA repair fidelity will differ depending on the chromosomal content of each germ cell stage, as this will impact on which repair processes can be used. Prior to meiosis, in primordial germ cells, oogonia and spermatogonia, the cells are diploid with the full chromosomal content. Repair fidelity is therefore relatively high and similar to that seen in somatic cells.

In meiotic cells (oocytes and spermatocytes), repair fidelity is again likely to be high due to the availability of template DNA. Repair in these stages proceeds exclusively via homologous recombination. DSBs in meiotic cells are catalysed by SPO11, nucleolytic resection of 5′ DNA at the break site then occurs to produce a 3′ ssDNA overhang. The Recombinase proteins DMC1 and RAD51 bind the exposed 3′ ssDNA (Figure 1-14). This creates a nucleoprotein filament which performs a homology search to identify the allelic locus on the other parental chromosome. Subsequent strand invasion displaces a loop of DNA (D-loop) to begin the process of DSB repair via homologous recombination (for a review see Lam and Keeney, 2015 (99)).

However, in haploid spermatids, homologous recombination cannot occur as there is no template that can be used. Therefore, more error prone process such as microhomology mediated end joining (MHMEJ) and non-homologous end joining (NHEJ) will have to be used. The repair of DSBs in round spermatids has been investigated by Ahmed *et al* in 2010 (100). They show that both the classical-NHEJ pathway and the Parp1/XRCC1 dependent NHEJ pathway are active in round spermatids, although the classical pathway only has a small contribution. The zygotic repair pathways that repair damage on the incoming sperm are less well characterised, however it is likely that this too uses relatively error prone pathways since it occurs prior to DNA synthesis and fusion of the male and female pronuclei (101). Any DNA breaks remaining in the mature sperm have the potential to affect the next generation if the breaks cannot be repaired by the oocyte. In humans, sperm DNA fragmentation has been used as a biomarker to detect early pregnancy loss (102), highlighting the importance of investigating the context in which DSBs occur for both spermatids and mature sperm.

1.5.3 Differences in the 3D organisation of the male and female gamete genome

As already discussed, the paternal genome is extensively remodelled during spermiogenesis with the replacement of histone proteins with transition proteins and finally protamines, this achieves the almost crystal-like compaction of the genome in mature sperm. This inevitably will put the genome under torsional strain, leading to the formation of DSBs to relieve this strain. Torsional strain can also be absorbed by changes in the DNA state from a B-DNA conformation to a non-B DNA conformation such as the transition to Z-DNA. Genome remodelling in this manner does not occur in oogenesis, therefore the female genome is not exposed to such torsional strain (and likely DNA state changes into Non-B DNA) and there is no wave of post-meiotic DSBs. Therefore, the differences in 3D genome structure between male and female gametogenesis may impact both the occurrence of DSBs and the context in which they are repaired.

In addition to the changes associated with protamination and de-protamination, the profound transcriptional changes associated with male germ cell progression through meiosis and post-meiosis, necessitate large scale changes in chromatin structure to activate and de-activate the relevant genetic programmes. To study these organisational changes, TAD structure has been investigated at different developmental stages of spermatogenesis through Hi-C experiments (103). This shows that TAD structure changes throughout spermatogenesis. Vara *et al* in 2019 (103) have shown that inter and intra-chromosomal interaction ratios decreased 2-fold for all chromosomes in spermatogonia (early precursor cells to mature sperm) compared to fibroblasts.

This suggests that when cells commit to meiosis initiation, by differentiation into spermatogonia, a drastic remodelling of chromosomal territories occurs within the nucleus. They also observed that the decrease in inter and intra chromosomal interactions occurred simultaneously with changes to the A and B compartments. As meiosis progressed most compartments were lost in spermatocytes at prophase I (the meiotic stage where homologous chromosomes, condense, align, pair synapse and then recombine) (103). Post-meiotic cells then again developed a higher order chromatin structure with the reappearance of A and B compartments, but this pattern was less distinct than in spermatogonia and the compartments were of larger size (103).

Oocyte chromatin has been less well studied due to the difficulty in isolating sufficient material, however single-cell Hi-C studies indicate an attenuation of A/B compartments in the mature oocyte, similar to that seen in spermatocytes. Moreover, in the fertilized zygote the male and female pronuclei retain differential 3D organisation (104).

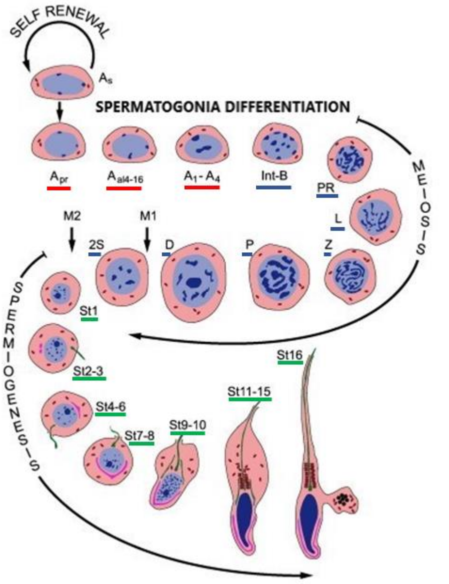


Figure 1-18: Schematic of mouse gametogenesis showing the 3 main cell types of spermatogenesis. The spermatogonia are underlined in red, the spermatocytes which undergo meiosis are underlined in blue and the spermatids are underlined in green, adapted from:(105). Reproduced with permission from Elsevier under licence number: 5581860247998.

1.6 Understanding male-specific contributions to genome rearrangements

Given the factors discussed above, there are thus two separate male-specific life stages where there is an additional vulnerability to DNA strand breakage and rejoining - i.e., the prerequisites for chromosomal rearrangement. Moreover, in both of these stages, DNA strand breaks are repaired via error-prone rather than high-fidelity mechanisms. To understand how these may contribute to the formation of chromosomal rearrangements, it is important to review in more detail how these DSBs occur and are repaired. In this section I will therefore review the general factors that predispose DNA to breakage. In general, DSBs may be associated with specific primary or secondary DNA structures that render the DNA more fragile and labile to damage. Primary sequences known to be associated with DNA damage include repetitive elements / short tandem repeats and transposable elements (84), while secondary structures associated with DNA damage include alternative DNA structures such as G-quadruplexes and R-loops. These are not mutually exclusive categories as specific repetitive elements may trigger DNA fragility precisely due to a propensity to fold into alternative secondary structures.

DNA damage may also be triggered by specific environmental insults occurring in specific cell stages. In this context a key contributor is oxidative damage occurring in the mature sperm head, since the sperm cell has few resources available to detoxify reactive oxygen species (ROS) and no ability to directly repair ROS-mediated damage. Another contributing factor (as mentioned above) is the torsional changes involved in sperm head compaction.

1.6.1 Primary sequence contributions to fragility: Repetitive elements

Transposable elements (TE) are repetitive DNA sequences that constitute around 40% of the mouse genome (106), they are mobile and can move from one location to another (107). The major TE families in the human and mouse genomes are LINEs and SINEs (long/short interspersed nuclear elements) and endogenous retroviruses (ERVs). Long interspersed elements LINE-1 (L1) are the dominant category of transposable elements in placental mammals. LINE elements are very numerous in the mouse genome (106) and contribute about 20% of the genome size (106, 108). They can be both beneficial acting as source of evolutionary novelty or detrimental if they are inserted into genes. Genomic re-arrangements can also occur through non-allelic homologous recombination between copies of repetitive elements. If a LINE insertion directly disrupts an exon of a gene, then this may disrupt the function of the protein produced. For example, a de novo LINE insertion has been shown to be the cause of a small number of cases of Haemophilia A in humans (109). LINE insertions within introns may also lead to the disruption of gene expression through transcriptional elongation inhibition (110).The insertion of L1 elements may lead to structural variation through non-allelic homologous recombination between the LINE elements (111). The L1 antisense promoter can also can generate tissue specific transcripts that may affect gene expression (112). This tissue specific expression could be beneficial if it leads to the expression of genes with an adaptive role. These definitions are not iron clad and initially detrimental effects can become coopted over evolutionary time to generate novel beneficial effects.

Short tandem repeats (STR) are regions where two or more nucleotides are repeated, and the repeats are next to each other. Some trinucleotide repeats may impair replication fork progression, which could lead to chromosomal fragility and double-strand breaks, for example in humans, CGG repeats in the X chromosome can cause fragile X syndrome (113).

1.6.2 Secondary structure contributions to fragility: Non-B DNA

Non-B DNA is DNA which does not form the typical right-handed helix and there are many different forms such as, Z-DNA, short tandem repeats (STRs), G-quadruplexes and R-loops to name a few. Non-B DNA is an important consideration in the context of the Integrative Breakage Model, as the 3D structure of the DNA in the non-B DNA form may influence the generation or repair of DSBs, which in turn may have evolutionary consequences. The formation of non-B DNA can relieve torsional strain during genome remodelling in spermiogenesis with the caveat that non-B DNA may be more prone to DSBs or be recognised by components of the DNA damage response as damage. The optimal substrates for the mismatch repair (MMR) proteins share similar features with some non-B DNA structures, such as the junction of B to Z-DNA (114). In oogenesis genome remodelling as occurs in spermiogenesis with the replacement of histones with protamines does not occur, therefore the genome of oocytes does not undergo such torsional strain, so non-B DNA structures in oogenesis may play less of a role in DSB formation and repair and therefore in genome evolution.

1.6.2.1 Z-DNA

Z-DNA is DNA in which the double helix has a left-handed conformation with two anti-parallel chains held together by Watson–Crick base pairs (115). This differs from B-DNA in which the double helix adopts a right-hand conformation. The Z-DNA backbone adopts a zigzag conformation (Figure 1-19). The Z-DNA helix does not have a major and minor grove as with B-DNA but instead the base pairs are offset to the side away from the axis forming one grove (115).

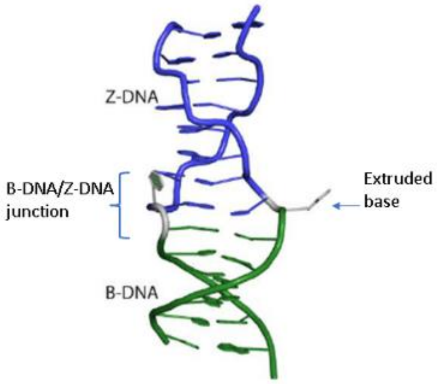


Figure 1-19: Schematic showing the junction between B-DNA and the non-B Z-DNA.
(PDB code 2ACJ). Adapted from (116) (and reproduced with permission from the creative commons licence (http://creativecommons.org/licenses/by/3.0/)).

B-DNA can be converted to Z-DNA by “flipping over of the base pairs so that they have an upside-down orientation relative to that of B-DNA" (115). In Z-DNA the phosphate groups of the sugar phosphate backbone are closer together than in B-DNA. Electrostatic repulsion between the phosphate groups under standard cellular conditions (e.g., standard salt concentration) favour B-DNA (115). At high salt concentration the repulsion between the phosphate groups is reduced and Z-DNA is the stable conformation. Z-DNA has a higher energy state than B-DNA (117).

The DNA sequence can be used to predict regions in which Z-DNA is likely to form. Switching between B-DNA to Z-DNA will use energy, therefore sequences in which this energy penalty is reduced are more readily converted to Z-DNA (117).

The sequences most readily converted to Z-DNA have alternated purines and pyrimidines, especially of C and G (118, 119). Z-DNA also occurs in regions with CA on one strand and TG on the other. Negative supercoiling will stabilize Z-DNA (120).

During spermatogenesis the removal of histones and the replacement of these proteins with protamines, could cause extensive negative supercoiling of the DNA. This may stabilize DNA in the Z-conformation.

1.6.2.2 G-quadruplexes

G-quadruplexes (or G4) are a type of non-B DNA structure formed in regions rich in guanine (Figure 1-20). They are formed in regions containing 3 runs of 4 or more guanines. They were first discovered by Sen and Gilbert in 1988 (121). They can be predicted bioinformatically, and several databases exist containing these predicted structures such as the Non-B database (https://nonb-abcc.ncifcrf.gov/apps/Query-GFF/feature/). G-quadruplexes can cause DNA instability and DNA damage and they can be found at sites of active transcription (122). G-quadruplexes can be unwound by helicases *in vitro* (123) and they have been shown to prevent genetic instability in vivo. Without helicases, G-quadruplexes can persist leading to replication fork stalling and collapse, which in turn leads to DSB formation (124)

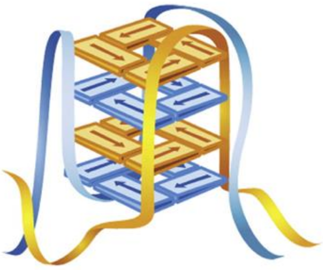


Figure 1-20: A schematic showing a G4 DNA structure.
Reproduced from: (114) with permission from Elsevier under licence number 5581881324701.

1.6.2.3 R-loops

R-loops are a type of non-B DNA, formed from a DNA:RNA hybrid and a displaced strand of ssDNA (Figure 1-21). There are opposing views in the literature that R-loops may cause DSBs if they stall replication at a site of active transcription or that they may stabilise the DSB to aid the repair process, as reviewed in Gan *et al*, 2011 (125) and Bader and Bushell 2020 (126).

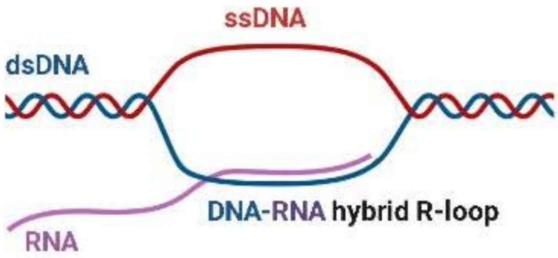


Figure 1-21: Schematic showing the key components of an R-loop i.e. the DNA:RNA hybrid and the displaced strand of ssDNA.

R-loops can be detected by ChIP-seq by using antibodies raised against the DNA-RNA hybrid, with the most common antibody used being clone S9.6. DNA:RNA hybrids are widespread in the genome with approximately 5% coverage in the human genome (127). R-loops can range in size from 200bp-500bp in humans (128) to several kilobases (129) with the resolution often limited by the choice of detection technique, with techniques such as ChIP-seq the lower limit of resolution is determined by the fragment size of the input DNA for the immunoprecipitation.

1.6.2.3.1 R-loop formation during transcription

If R-loops form during transcription, then they are likely to be found within genes. R-loops are thought to occur when a recently transcribed RNA strand invades the duplex DNA behind the polymerase and then hybridises to the template DNA which forms the DNA:RNA hybrid. The non-template DNA is displaced as a single strand of DNA. R-loops can form in regions of high GC content (127). R-loops are also associated with the formation of G-quadruplexes in transcribed regions (130). If an R-loop forms in front of the polymerase, then it may affect replication fork progression and lead to activation of the DNA damage response pathway.

DNA:RNA hybrids can be degraded by RNase H enzymes (131). Eukaryotes have two RNase H enzymes (H1/H2) both of which can degrade RNA in DNA:RNA hybrids (for a review see (132)). To avoid the generation of R-loops, eukaryotic cells co-transcriptionally package the newly formed RNA transcripts into ribonucleoprotein particles (RNPs) and then export them to the cytoplasm (133).

1.6.2.3.2 R-loop formation to stabilise a DSB and aid repair

In contrast to the belief that R-loops are the cause of DSBs Ohle *et al* in 2016 (134) showed in *Schizosaccharomyces pombe* that R-loops form as part of the homologous recombination DSB repair mechanism and that the RNase H enzyme is an essential component for complete DSB repair. Ohle *et al* 2016 (134) also showed that there was a strong increase in Polymerase II levels around the sites of DSBs, from which they could not detect increased RNA levels, therefore assuming that the RNA hybridises with its template forming DNA:RNA hybrids.

1.6.3 Environmental contributions to fragility: Oxidative DNA damage

Oxidative DNA damage can result from the action of Reactive Oxygen species (ROS) on DNA. ROS are formed as a by-product of aerobic metabolism. ROS are partially reduced forms of atmospheric O_2, often resulting from the excitation of O_2 to form singlet oxygen (O•) or from the transfer of 1, 2, or 3 electrons to O_2 to form, a superoxide anion ($•O_2–$), H_2O_2, or a hydroxyl radical (•OH–), respectively (reviewed in (135)).

Exposure of spermatozoa to oxidative damage occurs under physiological conditions. The epididymis is an oxidizing environment and oxidation of spermatozoa is required to finish the DNA packaging process (Figure 1-22) and generate a more compact sperm head. Spermatozoa also produce ROS themselves. More compact DNA is less susceptible to DNA damage, so oxidation of spermatozoa is required for sperm to reach their full fertilizing potential (136, 137), but too much oxidation can be harmful, so it is a double edge sword.

Mechanisms exist to mop up excessive ROS such as the multi-substrate enzyme glutathione peroxidase. This can transform H_2O_2 into water (reviewed in (138)). Knockout (KO) of GPX in mouse models results in subfertility (139).

The mammalian glutathione peroxidase family is split into 8 classes, with class 5 expression being highly restricted to the caput epididymis (140). If GPX5 is mutated, then this can lead to increased oxidative damage of sperm within the caput epididymis (141).

Sperm nucleus glutathione peroxidase 4 (*SnGPx4*) is predominantly expressed in late spermatids and spermatozoa (142). This is an enzyme which uses H_2O_2 to create inter and intra-protamine disulfide bounds on thiol groups of the cysteine-rich protamines, further condensing the sperm nucleus (143) (Figure 1-22). Therefore, like with *Gpx5* disruption, *SnGPx4* knockout will also result in a less compact sperm head more susceptible to OD.

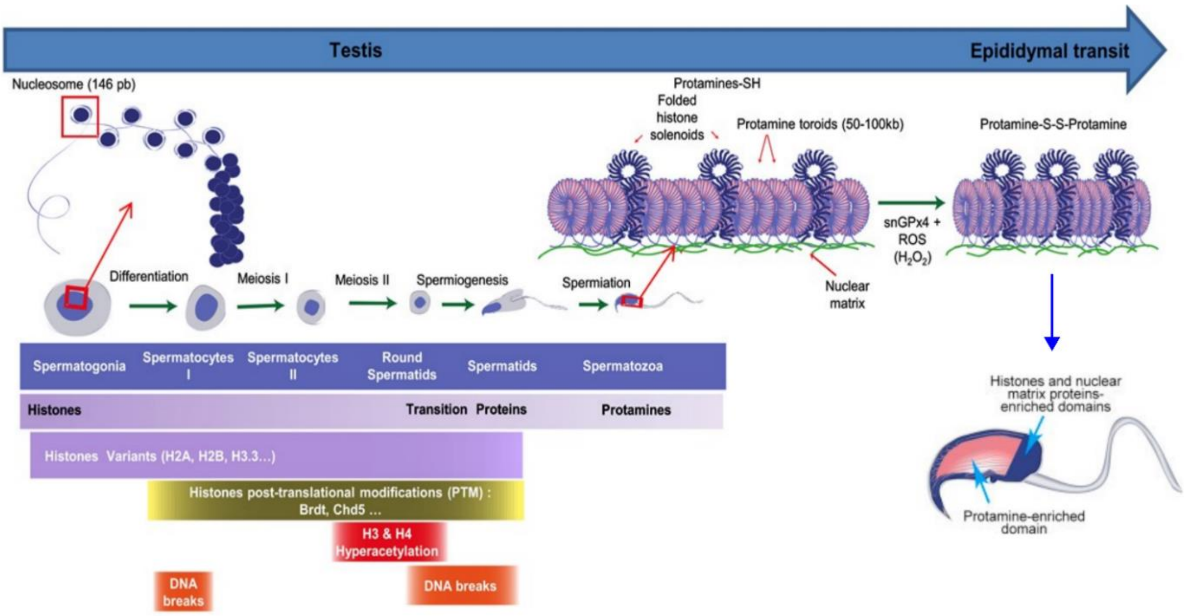


Figure 1-22: Changes in sperm chromatin structure as spermatogenesis progresses.
Reactive oxygen species (ROS) such as hydrogen peroxide allow further compaction of the chromatin. The enzyme *SnGPx4* uses hydrogen peroxide to create protamine-protamine disulfide bonds, leading to further chromatin compaction. Chromatin compaction is required to reduce the chromatin volume so that the chromatin can be contained in the compact hydrodynamic sperm head. Compact chromatin is also less accessible to DNA damaging agents. In mature sperm (lower right image) histone and nuclear matrix attached regions are found preferentially within the peripheral and basal regions (shown in dark blue in the picture of the sperm head), whereas protamine bound regions occupy a more central location (shown in pink). Adapted from: (143).

Sperm oxidative damage preferentially affects the basal and peripheral regions of the sperm nucleus (144). These are regions which are enriched in retained histones, so the DNA is not as compact and therefore more sensitive to OD. Oxidative damage in mature sperm is also high in the histone bound regions that are attached to the nuclear matrix (144).

Chromosome position within the sperm head had also been shown to impact the level of oxidative damage, with a basal location being more sensitive to oxidative attack (144). Hammoud *et al*, 2009 have shown in humans that these regions of the paternal genome are enriched for genes involved in the regulation of post fertilisation DNA replication events and the onset of the embryonic developmental program (92). Mature sperm lack a functional DNA repair mechanism, so OD sustained in the sperm will have to be repaired by the oocyte.

1.6.4 Environmental contributions to fragility: Chromatin compaction

This factor has been discussed already (see section 1.6.3). Mechanistically, DNA torsion is relieved by topoisomerase II. It does so by transiently inducing a covalently-bound DNA double-strand break, enabling the release of torsional strain. If the cleave site is not re-ligated, then this results in a DSB (145). In spermatids which are haploid, the TOP2 covalently bound DNA must be repaired by NHEJ as there is no template available for the homologous recombination repair pathway.

1.7 Pathways of DNA damage repair

Having discussed the mechanisms of DNA damage induction in spermatogenesis, I turn here to mechanisms of DNA repair, together with some of the molecular markers used to study this process. DNA repair mechanisms necessarily depend upon the cell type in which the break occurred, the stage of the cell cycle and the type of DNA ends at the break site. In spermatogonia and spermatocytes DNA repair fidelity is high, mostly occurring through homologous recombination as the cells are diploid. In contrast, spermatids are haploid cells and so lack a template for homology directed repair. DSBs occurring in 1n spermatids must therefore be repaired by error prone mechanisms such as non-homologous end joining (NHEJ), (Figure 1-23 and section 1.7.1), rather than the more accurate homologous recombination.

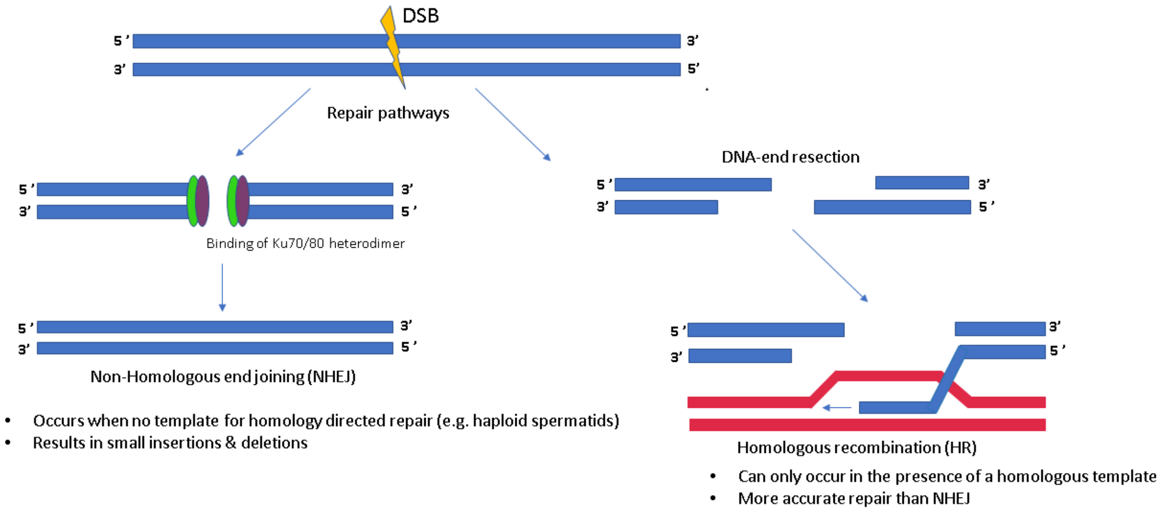

DSB
5'
3'
3'
5'
Repair pathways
DNA-end resection
5'
3'
3'
5'
5'
3'
3'
5'
Binding of Ku70/80 heterodimer
5'
3'
3'
5'
Non-Homologous end joining (NHEJ)
• Occurs when no template for homology directed repair (e.g. haploid spermatids)
• Results in small insertions & deletions
5'
3'
3'
5'
Homologous recombination (HR)
• Can only occur in the presence of a homologous template
• More accurate repair than NHEJ

Figure 1-23: Summary schematic of non-homologous end joining and homologous recombination.

1.7.1 Non-Homologous end joining

NHEJ involves several steps: DSB recognition, processing, and ligation. The ligation process will differ depending on the type of DSB, for example whether the DNA ends have been covalently modified and there are many different end processing factors.

Once a DSB has occurred, various protein complexes are recruited to the site of the DSB. The first is Ku, a heterodimer that consists of two subunits of Ku70 and Ku80, which can form a ring that can surround the DSB (146), helping to stabilise it (Figure 1-24), this then leads to the recruitment of DNA-dependent protein kinase (DNA-PK) and its activation by autophosphorylation (147). Autophosphorylation of the DNA-dependent serine/threonine kinase occurs on its large catalytic subunit (DNA-PKcs) and only takes place after collocation of the DNA ends. Autophosphorylation is required for correct DNA end accessibility for other NHEJ proteins (148) facilitating end processing but not end joining. Artemis is another key protein involved in the NHEJ process. It is recruited by Ku and DNA-dependent protein kinase catalytic subunit (DNA-PKcs) phosphorylates it (149). This induces its endonuclease activity, and it is recruited to the DSB. The DNA-PKcs-Artemis complex is key to bringing DNA ligase IV to the DSB (150) to repair it.

The ligase complex IV/XRCC4 then joins compatible DNA ends, aided by the XLF/Cernunnos protein interacting with XRCC4 (151, 152). If the DNA ends are not initially compatible, then NHEJ can occur with other nucleases and ligases to ensure that the ends can be ligated by the IV/XRCC4 complex.

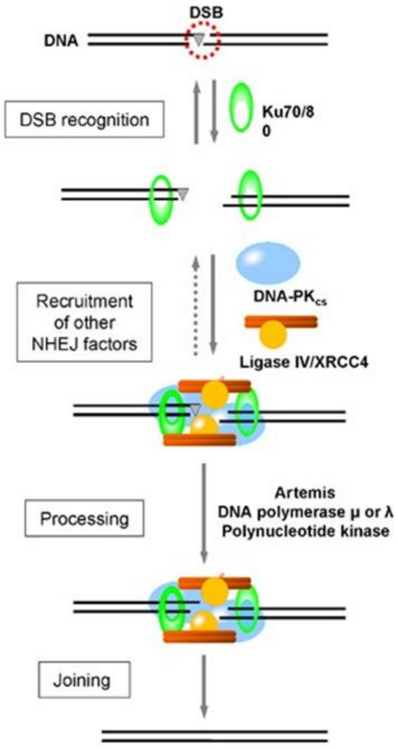


Figure 1-24: Schematic of a model of non-homologous end joining.
At a double-strand break Ku70/80 heterodimer (Green) binds to the DNA ends recruiting DNA-PKCS (shown in blue). This activates the DNA-PK kinase activity, leading to autophosphorylation which enables the subsequent processing and ligation steps. The small triangle symbolizes a DNA end that needs processing before ligation. Reproduced from: (153) under the creative commons licence (https://creativecommons.org/licenses/by/4.0/).

As previously discussed NHEJ is error prone, as no template is used in the repair process, consequently this can lead to small insertions or deletions during the repair process. Deletions can occur when the DNA ends are processed if they are damaged or not complementary. NHEJ of blunt ends in mammalian fibroblasts is usually precise (154).

As a backup to NHEJ, an alternative form of joining called microhomology-mediated end-joining (MHMEJ) may occur. This involves a few nucleotides of homology, (in yeast between 7-22bp (155)) that may be used to join the ends. The process occurs independently of Ku (156). The process starts with end resection- the exposed micro homologous sequences are then annealed forming an intermediate with 3′-flap and gaps on both sides of the DSB. The non-homologous 3′ flap must then be removed by nucleases (157) to allow DNA polymerase to fill the gap. The final step is ligation of the DNA ends by ligase. MHMEJ is an error prone process causing various sized deletions, and it has been shown that translocated chromosomes can occur in cells where components of the NHEJ pathway have been mutated (158). The chromosomal translocations often contain microhomologies at the junctions (159).

1.7.2 Homologous Recombination

Homologous recombination (HR) is an essential pathway for DSB repair in diploid cells in the S and G2 phase of the cell cycle, unlike NHEJ and MHMEJ, it is much more accurate, as a template is available to direct the repair (Figure 1-23). Briefly it starts with nuclease 5'–3' broad resection of break ends, generating 3' ssDNA overhangs, which are coated with replication protein A (160). The recombinase RAD51 is loaded onto the 3'ssDNA via the BRCA2 protein (161), replacing replication protein A. This forms a nucleoprotein filament which is used to search for complimentary sequence. Once found a D-loop is formed (Figure 1-14), with the broken strand acting as primer and the complete strand as a template. Once enough DNA has been synthesised to fill the break, HR proceeds via displacement of the newly synthesised DNA strand from the D-loop which then anneals to the complementary sequence at the non-invading end. Alternatively, a structure called a Holliday junction may be formed which consists of a 4-way junction between the recombining strands (162). It may form through the annealing of the non-invading end to the displaced strand of the D-loop in another step, or by invasion of the two resected ends and their concurrent extension. The Holliday junctions are resolved without crossing over and DNA exchange by helicases and Topoisomerases. Alternatively specific nucleases resolve the junctions with exchange of genetic material (cross overs) as occurs in prophase 1 of meiosis.

As spermatids are haploid cells, repair cannot occur by homologous recombination and must occur via one of the other pathways such as NHEJ. Conversely, HR is the sole pathway of DNA repair operating during meiosis, and thus meiotic DSBs are repaired exclusively by this highly accurate repair pathway.

1.7.3 Non-allelic Homologous recombination

Non-allelic homologous recombination is a form of DNA repair which involves homologous recombination between two regions of DNA that have highly similar sequence but are not alleles. The *Mus musculus* genome contains blocks of repetitive DNA. If a meiotic DSB is formed within a repeat, then it has the potential to induce genomic rearrangement through repair with non-allelic sequences. This can generate genome rearrangements such as inversions, duplications and translocations (reviewed in (163)). If the homologous repair pathway is ever active in a haploid cell (i.e. a spermatid) then by definition it thus must result in non-allelic homologous recombination and consequently the potential for genome rearrangement.

1.7.4 DNA structures and chromatin marks associated with DNA damage

DNA repair mechanisms can be studied using chromatin immunoprecipitation (and related techniques such as Cut&Tag – see sections 1.10.4 and 1.10.6) to interrogate chromatin marks associated with DSBs. The most well-known of these is gamma-H2AX, a histone mark specifically involved with recruitment of DNA repair proteins. In this thesis, given the availability of a suitable dataset, I also examine the distribution of BRD4 – a chromatin-binding protein with a specific role in NHEJ.

1.7.4.1 BRD4 and DSBs

BRD4 is a member of the bromodomain and extraterminal (BET) family of proteins. BRD4 is characterized by two tandem bromodomains (BD1, BD2). BDs bind acetylated lysine residues on target proteins, including histones (164–166). BRD4 is involved in the DNA damage response, specifically the non-homologous end joining (NHEJ) pathway (167) – which is of particular importance in round spermatids as discussed previously. DSBs result in increased H4 acetylation (H4Ac) and phosphorylation of H2AX (γH2AX) at both ends of the breaks, which induces BRD4 recruitment. BRD4 then recruits other proteins of the DNA repair complex (168).

1.7.4.2 Gamma H2AX as a marker of DNA damage

GammaH2AX is the serine 139 phosphorylated form of histone H2AX. H2AX is a histone H2A variant. Phosphorylation of serine 139 on histone H2AX is induced upon DNA damage and DSB induction (169). Consequently, GammaH2AX is used as a marker for DSBs in molecular biology assays such as ChIP-seq and Cut&Tag. GammaH2AX regions are induced early in the cellular response to DSBs (169), the DNA damage response (DDR) and can form large foci, this may help to recruit DSB repair factors to the site of DNA damage (Figure 1-25). GammaH2AX knockout cells have defects in the DSB-induced cell cycle checkpoint response (170).

GammaH2AX is thought to help aid DNA repair by anchoring the broken ends together via the recruitment of cohesins (171, 172). This can occur because of nucleosome repositioning at damaged sites and reduced chromatin density (173–175).

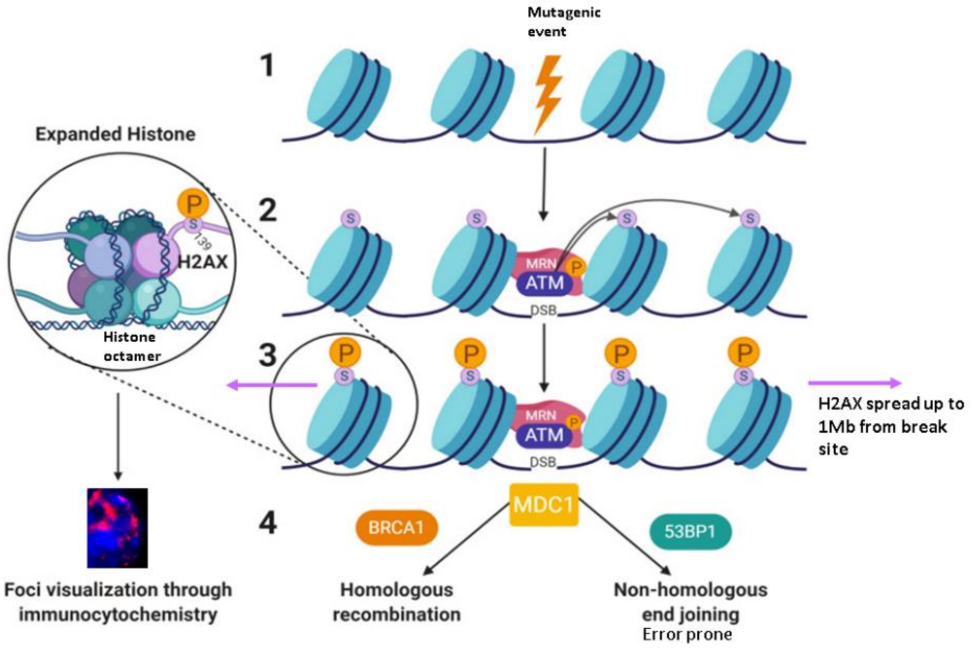


Figure 1-25: Schematic showing a DNA damage and repair pathway involving H2AX.
(1) Mutagens (e.g., ionizing radiation) induce double-strand breaks (DSBs). **(2)** The MRN complex composed of MRe11, Rad50, and Nbs1 proteins, is then recruited to the DSB. The MRN complex then recruits and activates ATM kinase. **(3)** ATM kinase phosphorylates the H2AX histone protein on the serine 139 residue (expanded histone) creating phosphorylated foci that can be visualized through immunofluorescence.
(4) The mediator of DNA damage checkpoint protein 1 (MDC1) is recruited to the DSB. After modification via ATM, MDC1 recruit proteins, such as BRCA1 and 53BP1, to direct the DNA damage repair pathway through homologous recombination or nonhomologous end-joining. Image adapted from: (176) with permission from the creative commons licence: https://creativecommons.org/licenses/by/4.0/.

1.8 Selective dynamics that influence how chromosomal variants spread through populations

I have introduced where and when chromosomal rearrangements might happen, this section will discuss the different models of how chromosomal rearrangements might spread and become fixed within individuals. Chromosomal rearrangements must initially arise in the heterozygous form before they can spread through a population to generate homozygous individuals. As a given rearrangement spreads within the population, the differing haplotypes compete with each other and suffer one of three fates: either the new variant dies out, it spreads to fixation, or it persists. Long term persistence as a stable polymorphism is however rare as the restricted gene flow between haplotypes leads to progressive functional divergence, and eventual hybrid breakdown and/or hybrid sterility. Chromosomal rearrangements are thus a key factor in initiating and/or reinforcing reproductive barriers during speciation. In particular, the hybrid dysfunction model of speciation (first proposed by Dobzhansky (177)) predicts that speciation occurs as a result of structural chromosomal changes such as inversions, translocations or deletions becoming fixed within a population. In the heterokaryotypic hybrid, recombination among the rearranged chromosomes generates gametes which are unbalanced. This can cause reduced fertility or complete sterility (underdominance).

There is a vast literature on the selective dynamics associated with speciation which is too extensive to recapitulate here (however see references (38) and (178) for review). In this thesis, I focus on a less studied aspect of this process: namely the potential for transmission ratio distortion (non-Mendelian inheritance) of chromosomal rearrangements among the progeny of heterozygous carriers. In particular, I have investigated a Robertsonian fusion of chr6 and chr16 in mice that has previously been shown to undergo male drive - i.e., male parents pass on the fused and unfused versions of the chromosome at non-Mendelian frequencies.

1.8.1 Overview of transmission ratio distortion

Transmission ratio distortion (TRD) refers to the non-Mendelian inheritance of alleles (or linked haplotypes) due to processes operating during gametogenesis and/or fertilisation. Conceptually, instances of TRD can be grouped into two categories:

- Processes operating during meiosis that lead to a skewed ratio of haploid gametes being produced.
- Processes operating after meiosis, that bias either the survival or fertilising capacity of gametes dependent upon their haploid gene content.

In the first category, the gamete production bias subsequently leads to transmission ratio distortion among the offspring. Such cases are referred to as “true meiotic drive”, and examples include knob loci in maize, that can promote their preferential segregation during meiosis to the position that will become the egg (179). Typically, true meiotic drive operates during female meiosis, since in this case only one of the meiotic products forms a viable gamete and the “drive” consists of an intragenomic competition to avoid segregation into the polar body and thus be included in the egg.

In the second category, an initially equal gamete ratio nevertheless produces biased transmission ratios due to functional differences between gametes. This is described as haploid selection and was first proposed by Haldane in 1924 (180). In contrast to true meiotic drive, haploid selection is predominantly seen in males (or isogamous organisms), since all meiotic products have the potential to become gametes and compete with each other. Haploid selection during oogenesis is likely to be minimal, as meiosis II is only triggered upon sperm fertilisation.

In this thesis, I will refer to all forms of transmission ratio distortion operating during spermatogenesis as “male drive” (see also section 1.8.5) since the net effect is to favour the transmission of specific alleles or haplotypes when passing through the male line during reproduction.

1.8.2 Male drive in the context of spermatid development

Given that sperm are all genetically unique, a naïve view would suggest that these differences must inevitably lead to haploid selection on their viability or fertilizing potential. Against this, Burgos and Fawcett in 1955 (181) discovered that spermatids were linked by cytoplasmic bridges formed by incomplete cytokinesis (Figure 1-26). Willison *et al* in 1988 (182) discovered biochemical evidence that transcripts could be passed between these bridges. This will allow the spermatids linked by cytoplasmic bridges to effectively become a homogeneous population. If a transcript is not shared between the spermatids, then this may cause functional differences between them. Protein products and whole organelles can also be shared by adjacent cells (183). Thus, the prevailing view since the mid-1950s has been that sperm are effectively diploid, and that there is little or no possibility for haploid selection to occur in mammals.

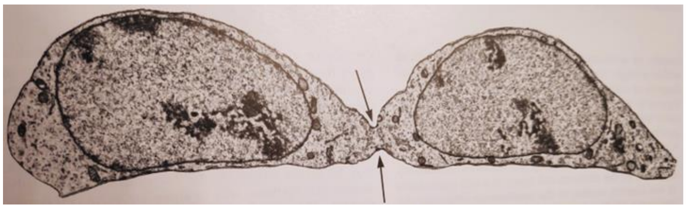


Figure 1-26: Two spermatogonial cells connected by a cytoplasmic bridge.
(arrows). Bridges range in diameter from 1-3µm. Reproduced from: (57) (page 51).

However, in recent years results have begun to accumulate that challenge this dogma. In particular (see following section), understanding the mechanisms of certain rare examples of male drive has shown that some genes escape sharing across the cytoplasmic bridges between spermatids and therefore become subject to haploid selection (184). Most recently, Bhutani *et al* (185) claimed based on single cell sequencing that a large fraction of spermatid-expressed genes escape sharing to at least some extent and that haploid selection is more prevalent than previously appreciated. Additional experimental evidence shows that selection on sperm longevity can affect offspring fitness in fish (186).

1.8.3 Known examples of male drive

Only three cases of male drive have been unambiguously identified in mammals, largely because the driving variant was associated with chromosome-scale differences, so inheritance of the chromosomal differences was used as a proxy or marker for the drive gene. Given the nature of male meiosis, it is generally assumed that these are due to haploid selection rather than meiotic drive, but this is not formally proven in all cases. The first on mouse chromosome 17, termed the t-complex, was first discovered in 1932 in wild mouse populations by Dobrovoloskaia-Zavadskaia and Kobozieff (187). The t-complex contains a series of closely linked inversions that contain a single responder gene named *Smok*^{TCR} (that evades sharing through the intracytoplasmic bridges), together with other genes termed distorters that act on *Smok*^{TCR} to promote drive (184).
The second is on mouse chromosome 6. Chromosome fusions involving mouse chromosome 6, such as in Robertsonian 6.16 mice, show an under transmission of the fusion chromosome in heterozygous animals (53). This was first thought to occur due to disruption of the *Spam1* gene, which encodes a hyaluronidase enzyme, which was thought to escapes transcript sharing through the cytoplasmic bridges. Sperm without functional SPAM1 were thought to show a reduced fertilisation potential due to absence of the *Spam1* encoded hyaluronidase enzyme which aids in sperm penetration of the oocyte. However other hyaluronidase enzymes or other genes may also be involved as *Spam1* or *HyaL5* knockout heterozygous animals that were crossed to generate double knockouts, generated the expected mendelian rations of the progeny (188).

The third case occurs when a deletion on the Y chromosome leads to sex ratio skewing in the offspring of affected XYR^{III}qdel males (189). Driving variants undergoing haploid selection will eventually spread to fixation, so there is likely to be limited variation in any given population, hampering the ability to identify heterozygous animals and investigate drive through traditional pedigree analysis.

1.8.3.1 Haploid selection

Haploid selection is distinct from meiotic drive and describes selection acting on haploid biased or haploid limited genes, such as during the haploid stage of oogenesis or spermatogenesis, this concept was first proposed by Haldane in 1924 (180).
Genes expressed in the haploid state are directly exposed to selection, whereas in the diploid state selection may be partially or fully masked by a homologous allele. Meiotic drive is a type of intragenomic conflict, whereby the meiotic process is manipulated by one or more loci to bias the transmission of one or more alleles over another.

Other Instances of male drive- in this case known to operate by haploid selection have been investigated without identifying the exact genes involved. For example, Immler *et al* 2014 (186) investigated sperm variation within a single ejaculate and the effects on offspring development in Atlantic salmon. Alavioon *et al* 2017 (190) used zebrafish as a model to show that that selection on phenotypic variation among intact fertile sperm within an ejaculate affects offspring fitness. The study also showed that there was genetic variation among sperm selected according to phenotype.

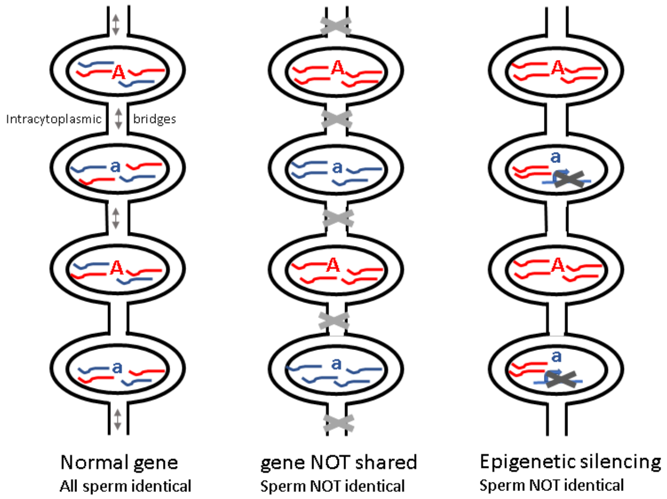


Figure 1-27: Schematic indicating different transcript sharing scenarios between the cytoplasmic bridges. For a normal gene, transcripts of allele A and allele a (represented by the coloured lines) are shared between the cytoplasmic bridges, so all sperm are functionally equivalent. The middle panel shows what happens when the product of an allele of a gene is not shared between the spermatids through the cytoplasmic bridges. The spermatids are not functionally equivalent. Epigenetic silencing of one allele of a gene (right panel) can also lead to functionally different spermatids.

1.8.4 Drive as a force for genome structural evolution

Male drive can potentially contribute to genome structural evolution via the non-Mendelian inheritance of variant chromosomes from heterozygous carriers. Driving genes can be organised into supergene clusters which are kept together by genome rearrangements, often inversions that prevent recombination during meiosis (reviewed in (191)). The clusters usually contain one gene that evades sharing between spermatids (a responder) and a range of other genes known as distorters (that may or may not be shared), but which act on the responder to trigger the drive, reviewed in (192). Tight genetic linkage is maintained between the responder and distorter genes as the inversion is selected for during evolution.

The typical mammalian example of a male drive complex associated with chromosomal inversions is the t complex in mice on chromosome 17 (reviewed in (193)). The evolution of this gene cluster has been extensively studied (184), interpretation of the evolutionary history is complex since t haplotypes also carry fatal genes and are not likely representative of the general case. Without fatal genes, a "driving" cluster will rapidly sweep to fixation, taking any associated genome inversion(s) with it. Numerous distorter genes may act on a single responder (194), so genes that escaping transcript sharing (i.e., potential responder genes) will become the focus for recurrent genomic instability during evolution. These genes may trigger repeated episodes of genome rearrangements as new distorter genes arise and become linked to the responder via successive chromosomal inversions.

1.8.5 Mechanisms of male drive

Chromosomes that are rearranged, such as Robertsonian fusions are often transmitted at non-Mendelian ratios and undergo drive. Mechanistically, any given chromosomal variant could be subject to drive due to either genetic variants linked to the fusion breakpoint, or epigenetic difference triggered by pairing abnormalities during meiosis. In either case, identifying the proximate mechanism of the transmission ratio skewing will rely on identifying either genetic variants or epigenetic differences that affect genes which escape transcript sharing between spermatids and thus provide the necessary substrate on which haploid selection can act.

1.8.5.1 Drive via genetic differences linked to chromosome rearrangements

The presence of a chromosomal rearrangement such as a Robertsonian fusion can lead to the suppression of recombination of genes closely linked to the fusion. This can lead to the accumulation of SNPs surrounding the fusion. If these SNPs occur in genes which impact fitness, then this could impact the spread of the chromosomal fusion within a population. In particular, the suppression of recombination in the vicinity of the rearrangement breakpoint has been modelled and shown to lead to accumulation of deleterious mutations, akin to the "Muller's ratchet" model of Y chromosome degeneration (195). Different types of chromosome rearrangements might be more or less effective at supressing recombination. For example, Inversions which change the gene order might be more effective at supressing recombination than a rearrangement which does not change the gene order such as a fusion. However, studies have shown that Robertsonian translocations can restrict gene flow in mouse (196),(197). We therefore predict that there may be an accumulation of deleterious mutations tightly linked to the centromere of any given Rob fusion chromosome, that these mutations in turn may potentially affect the fertility or viability of animals bearing the Rob chromosome, and that if the mutations affect genes that escape transcript sharing, haploid selection and male drive will result.

1.8.5.2 Drive via epigenetic regulation of rearranged regions during meiosis

Normal meiotic chromosome pairing (and by proxy recombination) cannot occur in the region of a chromosomal rearrangement. Therefore, there may be epigenetic changes in these chromosomal regions to transcriptionally silence the unpaired regions at the pachytene stage of cell division. In particular, regions that do not pair during meiosis are subject to profound transcriptional silencing during pachytene – a process is known as meiotic silencing of unsynapsed chromatin (MSUC), (see section 1.9.5 of this introduction) (198). Turner *et al* in 2005 (199) showed that all unsynapsed regions are transcriptionally repressed relative to synapsed regions. It is believed that MSUC may help protect the genome from invasion by parasitic sequences such as retroviruses by silencing of any unpaired sequences.

For the unsynapsed axes of the sex chromosomes, gene silencing by MSUC is sustained through to post-meiotic stages of germ cell development, and thus the majority of sex-linked genes remain inactive in spermatids (200). The limited evidence available to date indicated that post meiotic silencing also applies to unsynapsed autosomal regions. Therefore, the regions adjacent to chromosomal rearrangement breakpoints are predicted to undergo transcriptional silencing in most meiotic and post meiotic germ cells. If the silenced regions contain genes that escape transcript sharing in spermatids, then this may in turn create functional differences between cells that will be subject to haploid selection.

In Robertsonian fusion mice, the fused vs unfused chromosomal copies can potentially be differentially affected by MSUC (201). This therefore provides a mechanism in which rearranged chromosomes may become subjected to male drive through an epigenetic mechanism.

1.9 Effect of chromosomal rearrangements in meiosis

Having reviewed the possibility for chromosomal rearrangements to undergo drive during gametogenesis, I will now outline the meiotic checkpoints and related processes that monitor synapsis and recombination and thus typically guard against the production of chromosomally abnormal gametes.

1.9.1 Meiosis and the meiotic checkpoints

Meiosis involves a single round of DNA replication followed by two segregation events, one which separates the homologous chromosomes and one which separates the homologous chromatids. This needs to be tightly controlled to prevent mutations being passed on to the next generation. This is achieved via several meiotic checkpoints, which will eliminate the majority of rearrangements occurring before or during meiosis. However, rearrangements occurring in haploid post-meiotic cells will not be eliminated by meiotic checkpoints and will therefore have a much higher chance of being passed on to the next generation.

There are three main meiotic checkpoints, one which detects failure of the chromosomes to properly synapse (MSUC), another which detects unrepaired DSBs and one which controls bipolar attachment to the spindle at metaphase 1 (the spindle assembly checkpoint). The meiotic control network uses numerous components of the DNA damage response (DDR) pathway, including conserved checkpoint sensor kinases ATM and ATR (202). Sister chromatids are held together by cohesins established during the pre-meiotic s-phase (203), but there is no such linkage for the homologous chromosomes, so during meiosis homologous chromosome pairs are identified and connected, this occurs during prophase I via homologous recombination between the homologous chromosomes. Homologous recombination ensures correct chromosome segregation during meiosis as it physically connects the homologous chromosomes (204). After premeiotic chromosome replication, programmed DSBs are induced by the SPO11 enzyme (68) SPO11 is then removed and 5' resection of the DSB end produces a 3' ssDNA. RAD51 and DMC1 (disrupted meiotic cDNA1) are strand invasion proteins (reviewed in (205)) which use the ssDNA ends to search for a homologous sequence. Homologous recombination is biased towards the homologous chromosome (reviewed in (206)) and the distribution of the cross overs along the chromosomes is non-random (as shown in various papers such as Latos-Bielenska and Vogel in 1990 (207)). Only stable strand invasions are processed into crossovers, while the rest are repaired as non-crossovers (208).

An intricate signalling network known as the meiotic checkpoints controls the processes described above and creates dependencies between the different processes. This is essential to ensure that the events occur at the correct time, so that the different process do not interact. If meiotic checkpoints are activated, then meiosis can be delayed, to allow time to repair damaged DNA, or if this cannot occur then the cell can be removed by apoptosis, this will prevent the segregation of broken chromosomes and the generation of aneuploid offspring.

Most meiotic control dependencies are linked with the formation of DSBs, which although an intrinsic essential part of meiosis have the potential to cause chromosomal breaks, other checkpoints are associated with DNA replication or with proper pairing of the chromosomes via the synaptonemal complex. As already mentioned, the serine/threonine kinases ATM and ATR are key. Blunt (209) or protein conjugated ends will activate ATM. ATR is activated by single stranded DNA (210) (formed by DNA processing) that is coated with replication protein A (RPA) (211) or by ssDNA/dsDNA junctions. Cofactors are also required for damage recognition by the kinases. ATM requires the complex of MRE11-RAD50-NBS1, termed MRN (212). The regulatory protein ATRIP (ATR interacting protein) detects ssDNA and activates ATR (reviewed in (213)). While ATR detects ssDNA/dsDNA junctions via RAD9-RAD1-HUS1, also known as the PCNA-like 9-1-1 complex. Unsynapsed meiotic chromatin activates BRAC1 and TOPBP1 (214) which in turn activates ATR. ATM/ATR also activate two other checkpoint serine/threonine kinases termed CHK1 and CHK2, which relay the signals of ATM/ATR (reviewed in (215)).

There is a temporal separation between chromosomal replication and DSB formation, so that recombination via crossovers only occurs once the chromosomes have been replicated (216).

DSB formation will activate the meiotic control network (MCN), end-resection will occur, initiated by MRN/CtIP (217). End resection will mean that the more error prone repair pathways such as end joining are less likely to occur, resected ends are poor substrates for key components of the NHEJ pathway such as Ku (218). MRE11-dependent endonucleolytic incisions near DSBs starts resection (219).

Asynapsis of homologous chromosomes will activate the meiotic control network, via the accumulation of gammaH2AX on the unsynapsed chromosomes (reviewed in (220)), If this asynapsis persists, then this can lead to the formation of heterochromatin and transcriptional silencing (meiotic silencing of unsynapsed chromatin (MSUC)). This can result in the loss of the spermatocyte at metaphase if genes essential for survival are silenced (221). MSUC is similar to meiotic sex chromosome inactivation (MSCI) in the unpaired region of the X/Y chromosomes, although this does not result in cell death (198).

1.9.2 Meiotic spindle assembly checkpoint

The meiotic spindle assembly checkpoint (SAC) (Figure 1-28) ensures accurate segregation of the chromosomes in meiosis, preventing gains or losses. The SAC pauses the cell cycle until correct segregation of the chromosomes has occurred. It senses kinetochore microtubule attachments and the tension that is generated when the chromosomes are assembled in a bipolar manner

(222). Kinetochores are multi protein structures which bind to centromeric chromatin and to microtubules (223). The spindle assembly checkpoint prevents anaphase until all the chromosomes are stably attached to the spindle via kinetochore-microtubule attachments, only then is the block on progression to anaphase lifted.

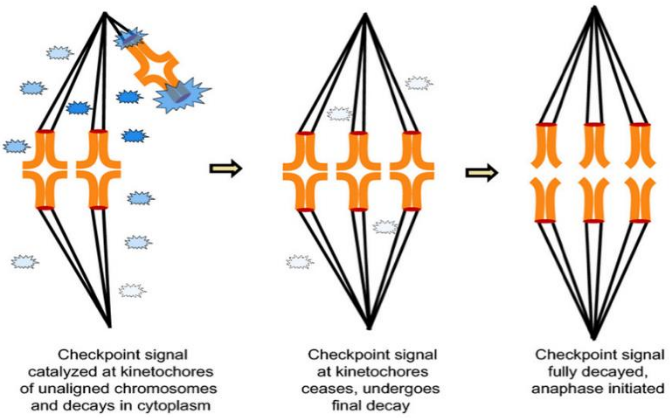

Checkpoint signal catalyzed at kinetochores of unaligned chromosomes and decays in cytoplasm
Checkpoint signal at kinetochores ceases, undergoes final decay
Checkpoint signal fully decayed, anaphase initiated

Figure 1-28: The spindle assembly checkpoint.
The kinetochores of unaligned chromosomes generate a spindle assembly checkpoint signal in the cytoplasm which will decay over time. (Left panel blue stars). The checkpoint can be regulated by both mechanical tension and microtubule attachment.
The left diagram shows paired homologous chromosomes at meiosis I. The signal from the unaligned kinetochore (larger blue star) is strong as it is not attached to microtubules and lacks tension.
The checkpoint signal diffuses into the cytoplasm and decays (faded blue stars). In the middle diagram, all the chromosomes have achieved bipolar attachment. The time required for complete decay of the checkpoint signal allows sufficient time for the final chromosome to attach and align at the metaphase plate. In the diagram on the right, the checkpoint signal has fully decayed which allows anaphase onset.
Reproduced from: (224) with permission of John Wiley and Sons under licence 5581910014995.

1.9.3 The effect of Inversions, fusions and fissions on meiosis

The presence of chromosomal rearrangements such inversions, fusions and fissions can have an impact on the normal progression of meiosis as homologous recombination can be disrupted as well as the normal 3D architecture of the nucleus, which can include an increased rate of heterologous interactions in primary spermatocytes, and alterations in both chromosome synapsis and axis length (225).

1.9.4 Heterozygous metacentric chromosomes and meiosis

A metacentric chromosome in a heterozygous state will pair with a homologous acrocentric or the homologous arm of another metacentric. This type of pairing will give rise to trivalent structures or more complex chains or rings, which can lead to meiotic defects (50).

Different types of synapsis around the centromere can occur in heterozygous metacentrics, synapsed, open and asynapsed (Figure 1-29). Full synapsis (Figure 1-29b) occurs when the centromeres of acrocentric chromosome synapses with the centromere of the metacentric. Trivalents remaining in the open configuration (Figure 1-29d) are not synapsed around the telocentric end containing the centromere (221, 225–227). These un-synapsed regions can interact with other chromosomes or with the sex chromosomes. This may disrupt the normal progression of meiosis.

Asynapsed chromosomes can be subject to meiotic silencing of unsynapsed chromatin (MSUC). This acts as a type of cell cycle checkpoint. In males without Rob fusions, it is only usually the non-complimentary region of the X and Y chromosome that are unpaired and therefore subject to meiotic sex chromatin inactivation (MSCI). This silencing occurs due to epigenetic modification of histone proteins. In mice with Rob fusion, chromosomes can be partially asynapsed (open) or fully asynapsed. These asynapsed regions can be marked with repressive histone modifications such as H3K9me3 (225). Heterologous interactions differ depending on the synaptic state of the Rob fusions. If the Rob fusion is in a synapsed configuration, then more centromeric associations of acrocentric chromosomes occur and the sex body is separated from the autosomes. However, if the Rob fusions fail to completely synapse, then intrachromosomal associations are disrupted and the sex body can be associated with the fused chromosomes (autosomes) (228). This can result in large regions of heterochromatin representing MSUC within a cell.

Structural variations such as a reciprocal translocation or Robertsonian fusion, without significant loss of genetic material can affect gene expression, if post meiotic sex chromatin (PMSC) repression occurs and is maintained during spermiogenesis. This in turn could have evolutionary consequences if the genes involved alter the fertilising ability of the sperm.

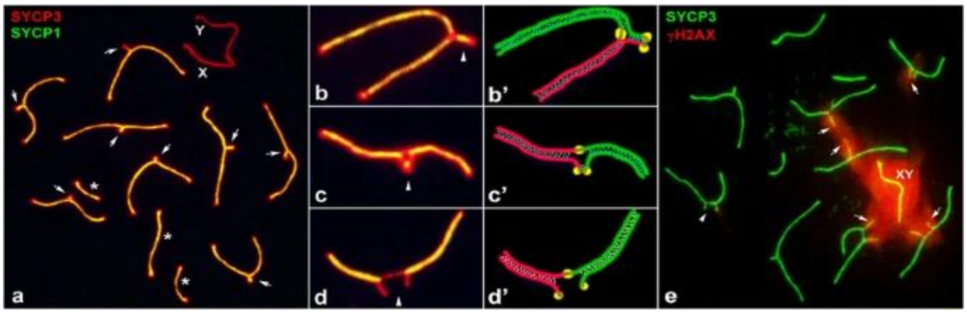


Figure 1-29: Pachytene spermatocytes of a 2n=32 simple heterozygous mouse and the constituent trivalents. **a)** Immunolabelling with the synaptonemal complex protein SYCP3 (red) and SYCP1 (green). Overlap of both proteins indicates synapsed regions. All eight trivalents (arrows) appear as closed configurations. Only the sex chromosomes (XY) have a large unsynapsed region. Telocentric bivalents are indicated by asterisks. b–d) Different degrees of synapsis found on trivalents. **b)** Complete heterologous synapsis of the telocentric chromosomes. **c)** Incomplete heterologous synapsis. **d)** Open configuration. b'–d') Schematic representations of the trivalent configurations. Green and red colours represent the homologous regions within the trivalent, blue lines represent synapsis, yellow circles represent centromeres. **e)** Immunolabelling with the synaptonemal complex protein SYCP3 (green) and the MSUC marker γH2AX (red). Many trivalents appear in an open configuration. Most of them show an intense labelling with γH2AX in the unsynapsed region (arrows). Some of them appear associated to the sex chromosomes (XY), which also show a conspicuous γH2AX mark. Some trivalents in open configuration do not show γH2AX labelling (arrowhead). Reproduced From: (50) with permission from Springer Nature under licence number 5581890493533.

1.9.5 Meiotic silencing of unsynapsed chromatin (MSUC)

During meiosis unsynapsed (unpaired) chromatin on autosomes is marked by a kinase of the DNA damage response (DDR), Ataxia telangiectasia and Rad3 related (ATR). This kinase phosphorylates serine 139 of H2AX forming gammaH2AX (229) and this is termed meiotic silencing of un-synapsed chromatin (MSUC) (230). Meiotic silencing may have evolved as a genome defence mechanism or to aid in the detection and elimination of cells with synaptic errors (199) (231).

A key component of MSUC is a variant of the H2A histone protein, H2AX. H2AX is enriched in testis relative to other cell types (232) and is a key component of the nucleosome during meiosis (233). H2AX also has a central role in the DNA damage response (see section on GammaH2AX as a marker of DNA damage). H2AX phosphorylation by ATR (a kinase involved in DNA repair) (234) occurs at the zygotene to pachytene transition where it marks the X and Y chromatin, it is this histone modification that leads to transcriptional repression and Meiotic sex chromatin inactivation and MSUC. Mice lacking H2AX fail to undergo MSCI (233). The BRCA1 (breast cancer 1, early onset) protein is required to target ATR to the X and Y chromosomes (234).

Fayer *et al* 2016 (201) have investigated how Robertsonian translocations modify the genomic distribution of γH2AFX and H3.3 in mouse germ cells. They found that the proximal 6–15 Mb portions of the chromosomes involved in Robertsonian fusions in spermatocytes are prone to meiotic silencing and the non-translocated homologs may have a slightly increased risk of being silenced. They did not investigate meiotic silencing in Rob translocations in spermatids, therefore it is possible that spermatocyte silencing is not maintained into round spermatids and this has not been investigated previously. In this thesis I will attempt to investigate this further, by looking at allele-specific expression in spermatids from mice with a Robertsonian fusion chromosome, to determine if either the fused or non-fused allele is preferentially expressed or whether there is equal expression of both alleles.

1.9.6 Meiotic checkpoints in female gametogenesis

The susceptibility of the male and female genomes to chromosomal defects differs, as does the response to chromosome mis-segregation, likely due to differences in the meiotic checkpoints between male and female gametogenesis. As in males, there are two checkpoints that operate during meiotic prophase 1 in oocytes, one that monitors DSB repair and another that monitors correct chromosome synapsis. However, in males there is a prompt cell death response to meiotic arrest due to expression of the Y-linked “meiotic executioner” gene *Zfy,* which triggers mid-pachytene apoptosis in cells arrested at either of these checkpoints. In females, there is no mid-pachytene apoptotic response, likely due to the absence of *Zfy*. Elimination of arrested cells is therefore delayed until late pachytene or diplotene (235, 236). Similarly, the meiotic spindle assembly checkpoint in females is thought to not be as robust as the spindle assembly checkpoint in males. In female mice with the Mlh1−/− mutation where almost no meiotic crossovers form, univalents are produced in meiosis, but their oocytes occasionally progress through meiosis I to extrude a polar body (237), in contrast in males with the same mutation metaphase arrest occurs (238). Robertsonian chromosomes in males and females are also processed differently, male mice with a Robertsonian chromosome have a greater incidence of unpaired and non-aligned chromosomes in meiosis I and increased metaphase I arrest and apoptosis in spermatocytes (239), whereas in females metaphase arrest may not occur (47).

Thus, in females, oocytes are prone to meiotic errors that lead to mis-segregation of entire chromosomes and production of aneuploid offspring. While the clinical consequences of this are profound, including maternal age effects on fertility and the risk of chromosomal disorders (240) there are few implications for genome structure evolution, since the majority of rearrangements originating in the female germline are non-viable and do not enter the evolutionary record.

1.10 Overview of genomic, transcriptomic and epigenomic techniques

1.10.1 Fluorescence-activated cell sorting (FACS)

A fluorescent-activated cell sorter (FACS) is a machine that can be used to isolate different cell populations. The machine generates a narrow stream from a cell suspension by passing the suspension through a small nozzle. The stream (containing droplets of single cells) is passed through an excitation light source (laser beam) which refracts the light. The forward scatter (FSC) can be used to indicate the cells size and the side scatter (SSC) the granularity or complexity of the cell, but this can vary between samples and flow cytometers. The cells can be labelled with additional dyes or antibody correlated fluorochromes. The cells can be excited by lasers of different wavelengths, the lasers required will depend on the stains used. The emission spectra, after excitation of the sample by different lasers is collected and analysed by the cytometer. The cell droplet can be differentially charged depending on the sort gating criteria to either be deflected into a collection tube or into the waste.

1.10.2 Whole genome sequencing

Whole genome sequencing (WGS) of DNA can now be carried out at relatively low cost and has an almost limitless range of uses from evolutionary biology research through to medical research. The importance of determining the conservation or variation of nucleotides in a species cannot be underestimated, both for genome structural analysis such as determining the presence of homologous synteny blocks and structural rearrangements and for small scale genomic rearrangements such as single nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELs). This has been made possible by the unprecedented advancement of sequencing technologies in the last few decades.

Next generation sequencing (NGS) technology uses massively parallel sequencing, (sequencing the genome many times in small and random fragments) to produce thousands to millions of sequences in parallel. This improves the speed and accuracy and reduces the cost.

In NGS (Figure 1-30) the DNA is immobilised to a solid support; each nucleotide is then washed through the system separately. The enzyme ATP sulfurylase is used to convert pyrophosphate into ATP, which is then used as the substrate for luciferase. Light is produced in proportion to the amount of pyrophosphate (241) and there is a unique fluorescent signal for each nucleotide.

Sequencing techniques then evolved to generate paired end data. This improves sequencing accuracy when mapping to the reference genome especially in repetitive DNA.

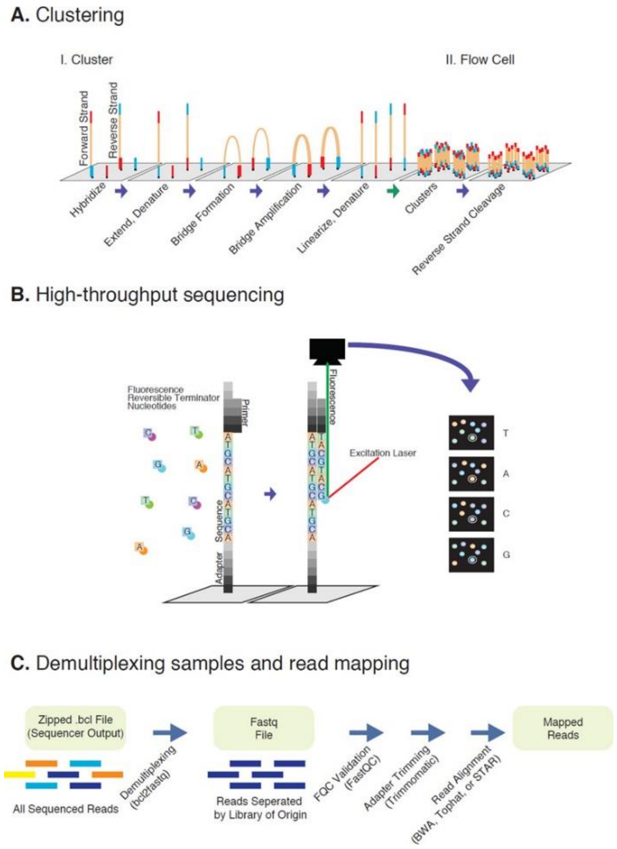


Figure 1-30: Schematic illustrating the NGS workflow for whole genome sequencing.

A) Denatured NGS library fragments are flowed across a flow cell and hybridize to the complementary Illumina adapter oligos. Complementary fragments are extended, amplified via bridge amplification PCR, and denatured; this results in identical single-stranded library clusters. **B)** Fragments are primed and sequenced utilizing reversible terminator nucleotides. laser excitation and fluorescence detection are used to identify the base pairs. **C)** The raw data is demultiplexed into individual libraries and assessed for quality using tools such as FastQC. Adapter reads can be removed to reduce technical noise (with tools such as Trimmomatic). Finally reads are aligned to the reference genome assembly. Reproduced from: (242).

WGS can be used to determine nucleotide differences between two strains of mice, for example C57BL/6 (wild type) and a Robertsonian fusion line. This is done bioinformatically using the whole genome sequencing data. The wild type strain is used as the reference to which the other strain is mapped. Using a file of a known set of SNPs within the reference genome, SNPs in the other strain can be determined using bioinformatics pipelines such as those within the Genome Analysis Tool Kit (GATK) (243). The SNPs can be filtered for read depth and quality scores to obtain high confidence calls. WGS data can also be used to detect insertions and deletions (INDELs) as well as copy number variants.

1.10.3 RNA-seq

RNA sequencing is a next generation sequencing technique that can be used to interrogate either the whole transcriptome or the mRNA produced within a cell. It involves RNA extraction from a cell sample and then the removal of ribosomal RNA or the selection of mRNA. The RNA is then fragmented, and cDNA is produced. The libraries produced can either be stranded or unstranded. The RNA is reverse transcribed into cDNA, followed by second strand cDNA synthesis. After end repair, A-tailing is carried out to add dAMP to the 3'-ends of the double stranded cDNA. Adapter ligation is then carried out, where dsDNA adapters with 3'-dTMP overhangs are ligated to A-tailed library insert fragments. The additions of adapter sequences allow many libraries to be sequenced on the same lane. Dual indexed library fragments are then amplified by PCR and the resulting fragments are then purified and sequenced.

1.10.4 Chromatin immunoprecipitation

Chromatin immunoprecipitation or ChIP-seq is a technique that can be used to immunoprecipitate regions of DNA that are bound by an antibody of interest. It can be used to investigate histone marks, DNA repair proteins, double-strand breaks or any protein-DNA interaction. Briefly the DNA/proteins of interest are fixed with formaldehyde, the nuclei are isolated, and the DNA broken into fragments of approximately 300bp either by enzymatic digestion or sonication. Immunoprecipitation is then carried out to precipitate the DNA of interest, which is then amplified and sequenced. A variant of ChIP-seq termed DNA Break Immunocapture (DBrIC) has been used to detect double-strand breaks. In DBrIC nicks and gaps in the DNA are repaired using T4 ligase and polymerase. Then DSBs that remain are labelled with TdT and biotin-14-dATP. The DNA is fragmented using Shearase and immunoprecipitated. Like ChIP-seq the resolution is governed by the size of the sheared DNA fragments. ChIP-seq techniques have evolved into other methods such as Cut&Tag.

1.10.5 Chromatin states

Genomic regions with different combinations of chromatin marks (or states) can be bioinformatically identified from ChIP-seq data using tools such as ChromHMM. This enables chromatin stage changes across developmental trajectories to be investigated, or different chromatin states can be compared against another condition, such as the location of spermatid DSBs. Chromatin states can be classified into active, repressed, or poised states based on the intensity of different histone marks within the state (see section 1.1). For example, H3K27ac and H3K4me3 are active histone mark, H3K9me3 and H3K27me3 are repressive chromatin marks. States with both active and repressive chromatin marks are termed poised.

1.10.6 Cut&Tag

Cut&Tag stands for cleavage under targets and tagmentation and is a method that can be used to investigate the location of histone modifications or transcription factors like chromatin immunoprecipitation-seq (ChIP-seq). In contrast to ChIP-seq, Cut&Tag does not require sonication but instead uses a modified Tn5 transposase that is bound to protein-A (PA). The protein A recognises the antibody bound DNA and simultaneously cuts the DNA and adds sequencing adapters at sites where the protein of interest was bound. Cut&Tag is carried out on fresh (or cryopreserved) unfixed permeabilised cells which are bound to concanavalin A (con-A) coated magnetic beads to immobilise them. Antibody incubation is performed with cells in their native state. Once the primary antibody has bound its targets, a secondary antibody is added to amplify the signal. The PA-Tn5 transposase is then added which will bind to the secondary antibody. In the presence of magnesium, the Tn5 transposase will Cut&Tag the DNA (cut the DNA and insert sequencing adapters). The cells are then immobilised on a magnet via the con-A beads and extensive wash steps are then carried out to remove background signal with the antibody bound DNA remaining within the cells that are attached to the magnetic beads. This results in lower background signal than standard ChIP-seq, therefore fewer cells are required as input and less sequencing depth is required.

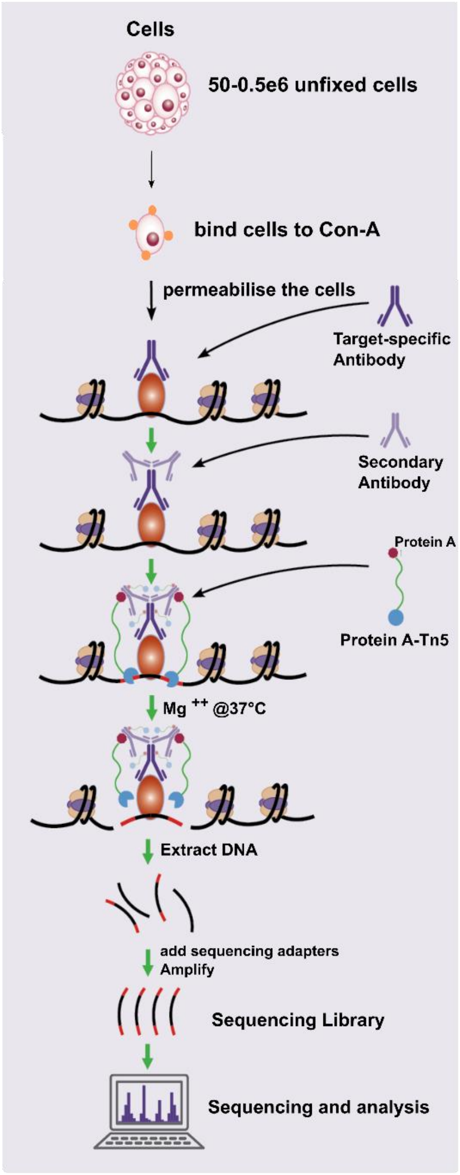


Figure 1-31: Schematic of Cut&Tag showing the different stages. adapted from www.activemotif.com.

1.11 Specific predictions and aims

We predict that post-meiotic spermatids share a number of specific vulnerabilities that will predispose them to the generation of novel structural variation, including: a high rate of DNA damage, dependence on error prone DNA repair processes, and the potential for non-Mendelian inheritance of variants due to competition between haploid gametes.

Therefore, in this thesis we will investigate:

(a) The evolutionary breakpoint regions (EBRs) where genome rearrangements have occurred during rodent evolution.

(b) The three-dimensional organisation of chromatin in the nuclei of germ cells throughout spermatogenesis.

(c) The locations where DNA damage occurs in the male germline, with a particular focus on the round spermatid stages, to determine DNA motifs and chromatin states that predispose the genome to breakage.

(d) The patterns of gene expression in spermatids from animals that are wild type, heterozygous or homozygous for a specific chromosomal rearrangement (a Robertsonian fusion of chromosome 6 and 16) that is claimed to undergo non-Mendelian inheritance in heterozygotes.

Collectively, aims (a) to (c) test the various aspects of the Integrative Breakage Model of structural novelty formation, initiating via DNA strand breakage, proceeding via re-joining of DNA strands in close proximity, and ultimately generating structural variation. Further analysis of (a) will give clues to the selective dynamics that retain or eliminate specific types of rearrangements, while (d) will investigate whether structural variation can itself trigger non-Mendelian inheritance in the male germline and thus influence its own transmission.

2. Methods

2.1 The breeding and characterisation of Rob6.16 mice

2.1.1 Rob6.16 mouse breeding

Mice with a Robertsonian translocation of chromosome 6 and chromosome 16 were obtained from the Jackson laboratory (JAX stock:000885, strain Rb(6.16)24Lub) and transferred to Charles River Laboratories (Manston, UK). The original line was derived from mice that came from Alfred Gropp. The original breeding strategy that was used to generate the homozygous line is unknown. However, it is likely that the original wild caught mouse with the Robertsonian fusion was bred to a laboratory mouse strain. It is likely that a filial mating between heterozygous mice from the first few generations was then carried out to generate mice homozygous for the Rob6.16 fusion, which were then inbred for numerous generations before being transferred to the Jackson laboratory (see Figure 2-1).

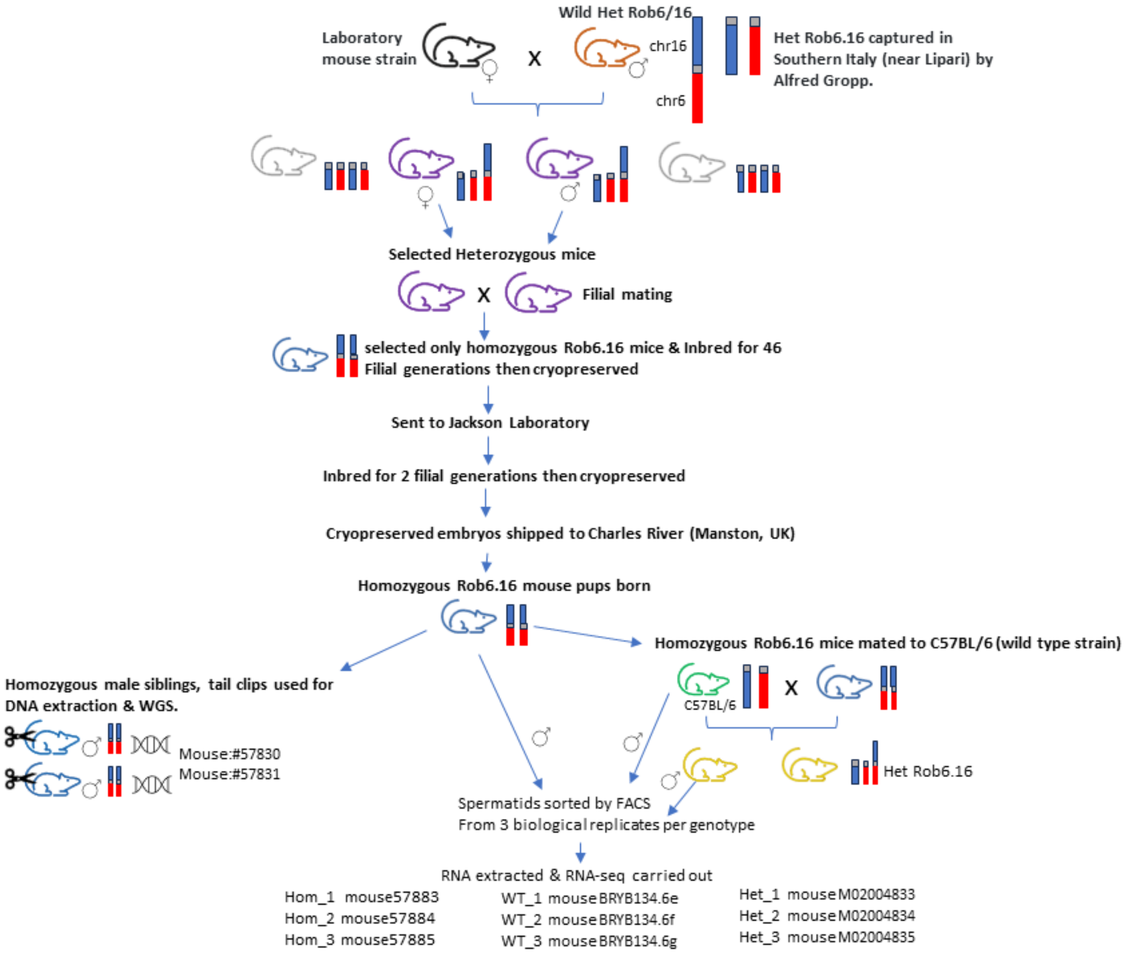

Laboratory mouse strain
♀
X
Wild Het Rob6/16
♂
chr16
chr6
Het Rob6.16 captured in Southern Italy (near Lipari) by Alfred Gropp.
♀
♂
Selected Heterozygous mice
X
Filial mating
selected only homozygous Rob6.16 mice & Inbred for 46 Filial generations then cryopreserved
Sent to Jackson Laboratory
Inbred for 2 filial generations then cryopreserved
Cryopreserved embryos shipped to Charles River (Manston, UK)
Homozygous Rob6.16 mouse pups born
Homozygous male siblings, tail clips used for DNA extraction & WGS.
♂
Mouse:#57830
♂
Mouse:#57831
Homozygous Rob6.16 mice mated to C57BL/6 (wild type strain)
C57BL/6
X
♂
♂
♂
Het Rob6.16
Spermatids sorted by FACS
From 3 biological replicates per genotype
RNA extracted & RNA-seq carried out
Hom_1 mouse57883
Hom_2 mouse57884
Hom_3 mouse57885
WT_1 mouse BRYB134.6e
WT_2 mouse BRYB134.6f
WT_3 mouse BRYB134.6g
Het_1 mouse M02004833
Het_2 mouse M02004834
Het_3 mouse M02004835

Figure 2-1: The Rob6.16 breeding strategy.
The breeding strategy shown at the top of the figure (from the capture of wild mice with the Robertsonian fusion, up until the cryopreserved embryos were shipped to Charles River) is the likely strategy, as the actual steps taken are unknown. Homozygous Rob6.16 mice (blue) were mated to wild type C57BL/6 mice (green) to produce Het Rob6.16 mice (yellow).

C57BL/6 mice were obtained from Charles River laboratories (Manston, UK) to act as a wild type control strain. Homozygous Rob6.16 mice were bred to wild type C57BL/6 mice to generate heterozygous Rob6.16 mice. For the details of the mice used for experiments see Supplementary Table 8-1.

When comparing between the three genotypes (wild type, homozygous Rob6.16 and heterozygous Rob6.16), age matched adult mice between 22-24 weeks were used for experiments. The mice were older than the typical 8-12 weeks normally used (as the age range for adult mice), as we were limited by the availability of the Aria flow cytometers which we used at the Crick Institute in London.

2.1.2 Karyotyping the Rob6.16 heterozygous and homozygous mice

A ~1cm^3 piece of mouse thigh muscle or both lungs was added to a small petri dish with approximately 1ml of RPMI media (containing 10% foetal bovine serum (FBS) and 1% Penicillin-Streptomycin). The muscle or lung was fragmented into very small pieces using sterile scissors. This suspension was then transferred into an adherent T25 tissue culture flask and the volume made up to 5mls, this was incubated at 37°C 5% CO2. The aim was to culture primary fibroblasts from these tissues. Every 2 days the culture was topped up with an additional 1ml of media. After around 7 days, fibroblast cells started to attach to the bottom of the flask and 90% of the media was removed and discarded and replaced with fresh media. Once the cells were 80% confluent the primary culture was split 1:3 into one T75 flask (passage 1). The spent media was removed from the flask and discarded. The flask was washed with 2mls of phosphate buffered saline (PBS), which was then discarded. 1ml of 0.25% tryspin was added and the cells incubated until they just started to detach. The trypsin was stopped with 9mls of complete media and the cell suspension transferred to one T75 flask. The cells were monitored every 1-2 days and split 1:2 or 1:3 as required. Once the cells were growing in log phase, a T75 flask of cells was used for karyotyping.

The cells were treated with 10µl/ml of a 10µg/ml stock of colcemid for 45 minutes at 37°C 5% CO_2 to arrest the cells in metaphase. They were then harvested as described above, keeping the time in trypsin to a minimum. Once the cells had detached the trypsin was inactivated by the addition of complete cell culture media. The cells were then centrifuged at 1900rpm for 5 minutes and the supernatant was discarded. The cell pellet was then resuspended in 10ml of 0.075M KCl pre-warmed at 37°C and incubated for 10 minutes at 37°C. The sample was then centrifuged at 1900rpm for 5 minutes and the supernatant was discarded (leaving about 0.5ml remaining above the pellet). The cells were then fixed in 5mls of fresh Carnoy's fixative (a 3:1 ratio of methanol:glacial acetic acid), while vortexing. 5mls more fixative was added, and the cells were centrifuged at 1900rpm for 5 minutes. The supernatant was discarded, and the cell pellet was resuspended in 5mls more fixative, the sample was centrifuged again, and the supernatant was discarded. The cell pellet was resuspended in ~100µl of fixative (volume was adjusted depending on the cell density) and 2x5µl drops were spotted onto a humidified glass slide. Once dry the slide was washed in 70% acetic acid for 5 seconds. The slides were air dried and then two drops of vectashield containing DAPI was added, a coverslip was then placed on top. Excess DAPI was blotted from the slides, and they were viewed at 100X magnification under a fluorescence microscope. Images of optimally spread metaphase spreads at 100x were taken and the chromosomes were counted to confirm the genotype (chapter 3, Figure 3-11).

2.1.3 Rob6.16 mice whole genome sequencing

DNA was extracted from recently euthanised homozygous Rob6.16 mice from a tail sample which was harvested approximately 0.5 cm from the tail tip. The tail samples from two homozygous Rob6.16 mice (called Rob1 and Rob2) were obtained and stored separately at –20°C until ready to proceed to DNA extraction. Before DNA extraction each tail sample was cut into several small pieces using a scalpel, and placed into a separate PCR clean 1.5ml Eppendorf tube. The samples were lysed overnight in lysis buffer (100mM Tris-HCL, 5mM EDTA pH8, 200nM NaCl [Fisher scientific S/3160/60], 0.2% SDS. Per sample 20µl of 20mg/ml proteinase K was added to the lysis buffer) and then DNA was extracted using a phenol chloroform-based approach. The DNeasy blood and tissue kit (Qiagen) was tested, however, RNA was carried over with this kit (as determined by Nanodrop readings), so a phenol-chloroform extraction was used instead.

Briefly, 500µl of lysis buffer was added to the tail samples (which included proteinase K). The samples were vortexed at low speed to aid lysis and then they were incubated overnight at 56°C. After the overnight incubation any undigested bone/cartilage/hair was removed from the sample by centrifugation at 2000g for 3 minutes. The supernatant was transferred to a fresh 1.5ml tube. Phenol:Chloroform:Isoamyl Alcohol (25:24:1, v/v) (Fisher scientific 15593031 was warmed to room temperature before use and swirled gently to mix, 520µl was added to the lysed tail sample. The tubes were shaken vigorously to obtain a homogenous suspension. Samples were then allowed to separate into an upper aqueous and lower organic layer for ~15 minutes at room temperature. To separate the sample into a tight lower organic phase and an upper aqueous phase containing DNA the samples were spun at 13,000 rpm for 5 minutes at room temperature. The upper aqueous layer (containing DNA) was pipetted into a fresh tube (approx. volume 420µl). An equal volume of isopropanol ~420µl was added to the aqueous layer, the tubes were capped and inverted well to mix. Once small strands of precipitated DNA were visible the tubes were spun at 13,000rpm at 4°C for 30 minutes to pellet the DNA. The supernatant was discarded, and the pellet was washed twice with 70% ethanol. The pellets were allowed to air dry at room temperature for 5-10 minutes until all traces of ethanol had evaporated. The DNA pellets were then resuspended by adding 80µl of TE buffer (100mM Tris-Cl, 1mM EDTA) to the pellets without pipetting. The tubes were placed at 4°C overnight to allow the pellets to fully hydrate. The next day DNA concentration was measured in duplicate using the Nanodrop. To obtain an accurate reading for DNA concentration alone, without RNA contamination skewing the absorbance readings, the DNA was also measured on a high sensitivity Qubit kit [Invitrogen] (first diluting the DNA 1:5). The DNA concentrations obtained ranged from 400-437ng/µl (Qubit measurement). The homozygous Rob6.16 DNA was diluted in TE buffer to 100ng/µl in 50µl of TE and sent to Novogene (Cambridge, UK) for whole genome sequencing at 27x coverage.

2.1.4 Genotyping wild type C57BL/6 mice, Robertsonian Rob6.16 mice and heterozygotes Rob6.16 mice by PCR

Using the SnpEff INDEL and SNP data (see section 2.12.1), 16 pairs of primers (four pairs for chr6 and four for chr16) that spanned INDELS (see Table 2-1 and alignment maps Supplementary Figure 8-1 to Supplementary Figure 8-8), were designed to genotype the Rob6.16 mice. Therefore, different sized PCR products were obtained depending on the genotype of the sample (Table 3-6).

NCBI genome data viewer was used to obtain FASTA files from the mm39 genome of a region ~100bp upstream and ~ 100bp downstream of the INDEL. This sequence was then loaded into primer3, which was used to design primers using the default settings. If primer3 could not find a primer pair, then the sequence range was extended slightly. If a primer pair was still not found then primers were designed manually and checked in the Thermo Fisher Multiple Primer Analyzer tool: (https://www.Thermo Fisher.com/uk/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/multiple-primer-analyzer.html) for the presence of cross primer dimer and self-primer dimer.

Robertsonian FASTA sequences were obtained from the SnpEff VCF files, which had been split into a VCF file of INDELS and a VCF file of SNPs. gatk *FastaAlternateReferenceMaker* with the default settings was used with the FASTA file of the mm39 genome to produce the Robertsonian FASTA files (one FASTA of the SNPs and one FASTA of the INDELS). These files were then used in the MUSCLE aligner (244) to create alignments between the WT C57BL/6 reference and the Robertsonian samples. The position of any insertions and SNPs was marked, as was the location of the forward and reverse primers (see Supplementary Figure 8-1 to Supplementary Figure 8-8).

The UCSC (University of California, Santa Cruz) In *silico* PCR tool (http://genome.ucsc.edu/cgi-bin/hgPcr?db=mm39) was used to check for the presence of off target PCR products. The regions amplified by these PCR primers are repetitive regions, so off target products are possible. We minimised this where possible or made sure that any off-target bands were of a different size to the target product/products.

The primer pairs were tested on the wild type C57BL/6 DNA first to determine that a product of the correct size was obtained. Primers that successfully amplified a wild type band were tested on the Robertsonian sample. To obtain heterozygous DNA, wild type and Robertsonian DNA of equal concentration were mixed in a 1:1 ratio and then run in the PCR reaction. This was because tissue samples from heterozygous mice were not available to extract DNA from, at the time of the PCR genotyping optimisation.

Table 2-1: PCR primers designed for Rob6.16 genotyping reactions.

chr6_fwd_1	ACACACCAAGCCAGAGAAAA
chr6_rev_1	TTGGTTACTGCGCCAACA
chr6_fwd_2	GCAGATATCACAGCAGCACC
chr6_rev_2	ATGGAGGGGTGAGGTTTCTG
chr6_fwd_3	AGGAGACCAGACTGAACACT
chr6_rev_3	CCCACTCCTCTCTTGCACC
chr6_fwd_4	AGTTCAAATCCCAGCAACCAC
chr6_rev_4	TGATCATCATGGCAGGACGT
chr16_fwd_1	TGGAGAAAGAGGGGAGAGGG
chr16_rev_1	GAATGAGTTCCAGGACAGCC
chr16_fwd_2	ACCCTCCTCTCCAGTCTCTT
chr16_rev_2	AAATGGGAGGGGCAGGAAAA
chr16_fwd_3	GGCCTCCCATTACCTGTG
chr16_rev_3	AGTGCTTAAGGTTGGAGGCT
chr16_fwd_4	CAGGCACCTCATATCCTCCA
chr16_rev_4	TGTCTCCTCACAGCAAACCA

PCR primers at 25nM scale and standard purification were obtained from Integrated DNA Technologies (IDT). The primers were diluted to a stock concentration of 100µM in nuclease free water. Working stocks at 10µM were used to set up the PCR reactions. (2.5µl of 10µM working stock in a 50µl PCR reaction gave a final concentration of 0.5µM).

15ng of input DNA was used per PCR reaction, with 1µl of 10mM dNTP mix, 0.5µM each of the forward and reverse primer, 0.25units of DNA polymerase, 10µl of 5X Green GoTaq polymerase buffer, 2mM $MgCl_2$ and nuclease free water up to a final volume of 50µl.

The DNA polymerase used was from Promega (GoTaq® G2 Flexi DNA Polymerase, catalogue number M7801). The GoTaq polymerase buffer contains loading dye, so post PCR the reactions were loaded directly onto a gel.

The melting temperatures of the primers were designed to be as close as possible with a maximum difference ideally being 5°C. The melting temperature was used as a guide to inform the annealing temperature of the PCR reactions.

The IDT website was used to calculate the annealing temperature of the primers with 0.5µM forward and reverse primer per reaction with 2mM $MgCl_{2.}$ The primers spanning the INDELs had annealing temperatures of around 58°C.
Two rounds of primer testing were carried out (one on wild type C57BL/6 DNA and one on Robertsonian 6.16 DNA) before a blind genotyping exercise was run to determine if the primers could be used to identify, a wild type, a Robertsonian or a mixed wild type/Robertsonian sample (the equivalent of a het sample).
The cycling conditions used for the PCR were as follows: 95°C for 2 minutes, then 30 cycles of 95°C for 15 seconds, 58°C for 30 seconds, 72°C for 30 seconds. Then a final incubation of 72°C for 5 minutes. Post PCR 20µl of each PCR reaction was run on a 2% agarose gel (containing SYBR safe) at ~100V for ~1 hour.

2.2 Testis digestion for FACS and spermatid sorting on the FACS Aria and FACS JAZZ

The tubule digestion protocol differed depending on if the sample was sorted on the Aria or the Jazz FACS. We used the Aria flow cytometers at the Crick Institute in London and the Jazz flow cytometer at the University of Kent.

In summary (Figure 2-2), both testes were dissected out from a mouse. The outer coat (Tunica albuginea) was snipped and then the tubules were gently squeezed out from the testis into dissociation buffer by running forceps along its length. The coat was discarded.

2.2.1 Testis digestion and staining for the FACS Aria

For sorting on the Aria dissociation buffer was composed of 5mg/ml Collagenase, 2.5mg/ml Dispase II, 2.5mg/ml DNase I in HBSS. The tubules were allowed to digest at 37°C for 30 minutes shaking at 800rpm, tapping the tube every 10 minutes. Digestion was then blocked by transferring the digested tubule cell suspension from one testis into 20 ml of DMEM + 10% FBS. The dissociated testis cell suspensions from the same mouse were then combined into one tube and filtered through a 30µm smart cell strainer to remove any cell clumps. The cell suspension was then aliquoted into tubes for:

a. Unstained control (500µl taken from 40ml dissociated sample)
b. PI control (500µl from 40ml)
c. Hoechst control (500µl from 40ml)

d. Hoechst + PI sample to sort (remaining 38.5ml)

Cells were then spun down at 300g for 5 minutes at 14°C, using a swing-out rotor at this step and all other spinning steps. The supernatant was removed and discarded. The cells were resuspended in DMEM + 10% FBS (1ml for the control samples and 10ml for the combined Hoechst and propidium iodide (PI) bulk sample). The samples not containing Hoechst were then spun down to wash, to be used as controls for FACS set up, before the bulk and control Hoechst samples were stained.

Hoechst 33342 staining:

a. 5µl of 1mg/ml Hoechst 33342 solution to 1ml of cell suspension in control samples
b. 50µl of 1mg/ml Hoechst 33342 solution to 10ml sorting sample for a final concentration of 5µg/ml.

The samples were stained at 37°C for 45 minutes in dark, then spun down at 500g for 3 minutes. The supernatant was discarded. The cells were then resuspended in 1ml of PBS 1% FBS 2mM EDTA to wash and spun down again at 500g for 3 minutes. The supernatant was discarded, and the cell pellets were resuspended in 1ml of PBS 1% FBS 2mM EDTA and filtered through a 70µm strainer into 5ml round-bottom FACS polystyrene tubes. Just before sorting 20µl of 50µg/ml PI was added to 1ml of cell suspension for a final concentration of 1µg/ml. The cells were kept on ice until they were sorted. Cells were sorted into chilled 1.5ml Eppendorf tubes that contained a small amount of PBS 1% FBS, 2mM EDTA.

Round spermatids were sorted on a FACS Aria equipped with a UV laser. For sorting of round spermatids, the following sorting strategy was performed: A forward scatter (FSC) and side scatter (SSC) plot was created to exclude dead cells and debris (which has low FSC) from the analysis. Dead cells were excluded based on high propidium iodide (PI) staining and a gate of viable cells was created (Figure 2-2B). As elongating spermatozoa are very small, exclusion based on low FSC will also eliminate much of their contribution from the analysis and sorting.

The viable cells were plotted as Hoechst Blue Height (x-axis) vs Hoechst Blue Area (Y-axis) and a gate of single cells were selected. The single cells were then gated on Hoechst blue 450/50 (Y-axis) against Hoechst red 660/20 LP635 530/30. This staining pattern was used to determine the 1n spermatid population (Figure 2-2B).

Three wild type, three Rob6.16 heterozygous and three Robertsonian homozygous mice were sorted. 300,000 cells were sorted per 1.5ml tube, with an approximate sort volume of 1.2mls. Depending on the sorting speed, four to five 1.5ml tubes each containing 300,000 cells were collected per mouse. 30,000-200,000 cells were collected into one 1.5ml tube from each mouse to stain with PNA-lectin and DAPI to use as a purity check. The tubes containing 300,000 cells were spun at 300g for 5 minutes and the supernatant was discarded. The cell pellets were resuspended in 250µl of lysis solution (RNAqueous micro total RNA isolation kit, AM1931) and snap frozen on dry ice and then transferred to -80°C storage.

Immunofluorescence purity check:
The 30,000-200000 sorted cells were fixed with an equal volume of 4% PFA solution, giving 2% final and incubated for 5 minutes at room temperature. Approximately half of this cell suspension was then cytospinned per slide at 1000rpm for 5 minutes. The slides were then stained with a 1:500 dilution (in PBS 0.1% triton) of a 1mg/ml PNA lectin Alexa Fluor 488 stock for 30 minutes at room temperature in the dark. The slides were then washed twice gently in PBS before being allowed to air dry. One drop of VECTASHIELD anti-fade mounting media containing DAPI was then added per slide and a coverslip placed on top. Excess DAPI was blotted from the slides, and they were visualised using a fluorescent microscope with DAPI and FITC filters. 200 cells were counted per slide (or as many as possible if cell density was low) and cells were classified as either round spermatids, elongating spermatids or neither of these categories (for example spermatogonia). The purity obtained with this approach was 97.5%-100% round and elongating spermatids per sample, with approximately 20% elongating spermatids per sample.

2.2.2 Testis digestion and staining for the FACS Jazz

Digestion buffer of a similar composition to that used for testis digestion for the Aria was used. The digestion protocol was based on the method used by Bhutani *et al* 2021 (185).

Two digestion solutions were used and heated to 37°C before use, solution 1 containing 1mg/ml Collagenase type I and 6 units/ml DNase I and solution 2 containing 1mg/ml Collagenase type I, 6 units/ml DNase I and 0.05% trypsin. The tubules from both testes were squeezed out into 12mls of digestion solution 1. The tubules were incubated at 37°C for 10 minutes with gentle agitation. The tubule fragments were then left to settle and the supernatant containing somatic cells was discarded. The tubule fragments were then transferred into 12mls of lysis solution 2. Incubation at 37°C for 12.5 minutes was performed with gentle agitation. The solution was gently pipetted every 5 minutes. Digestion solution two was then spiked with an extra 0.025% trypsin to a final concentration of 0.075%. The tubules were incubated for a further 12.5 minutes at 37°C. A sample of the digested tubules was taken and visualised under a light microscope to confirm that digestion was complete. If not, the sample was incubated for a few more minutes. The dissociated testis cell suspension was then filtered through a pre-wetted 70µm filter to remove cell clumps. 0.5ml of cell suspension was taken for an unstained control, 0.5ml for a PI only sample and 0.5ml for a Hoechst only sample. The remaining cell suspension was used for a bulk Hoechst/PI-stained sample. The bulk and Hoechst only sample were stained with Hoechst at 5µg/ml final for 45 minutes in the dark and the samples were processed as described above. PI at 1µg/ml final was added just before the samples were processed.

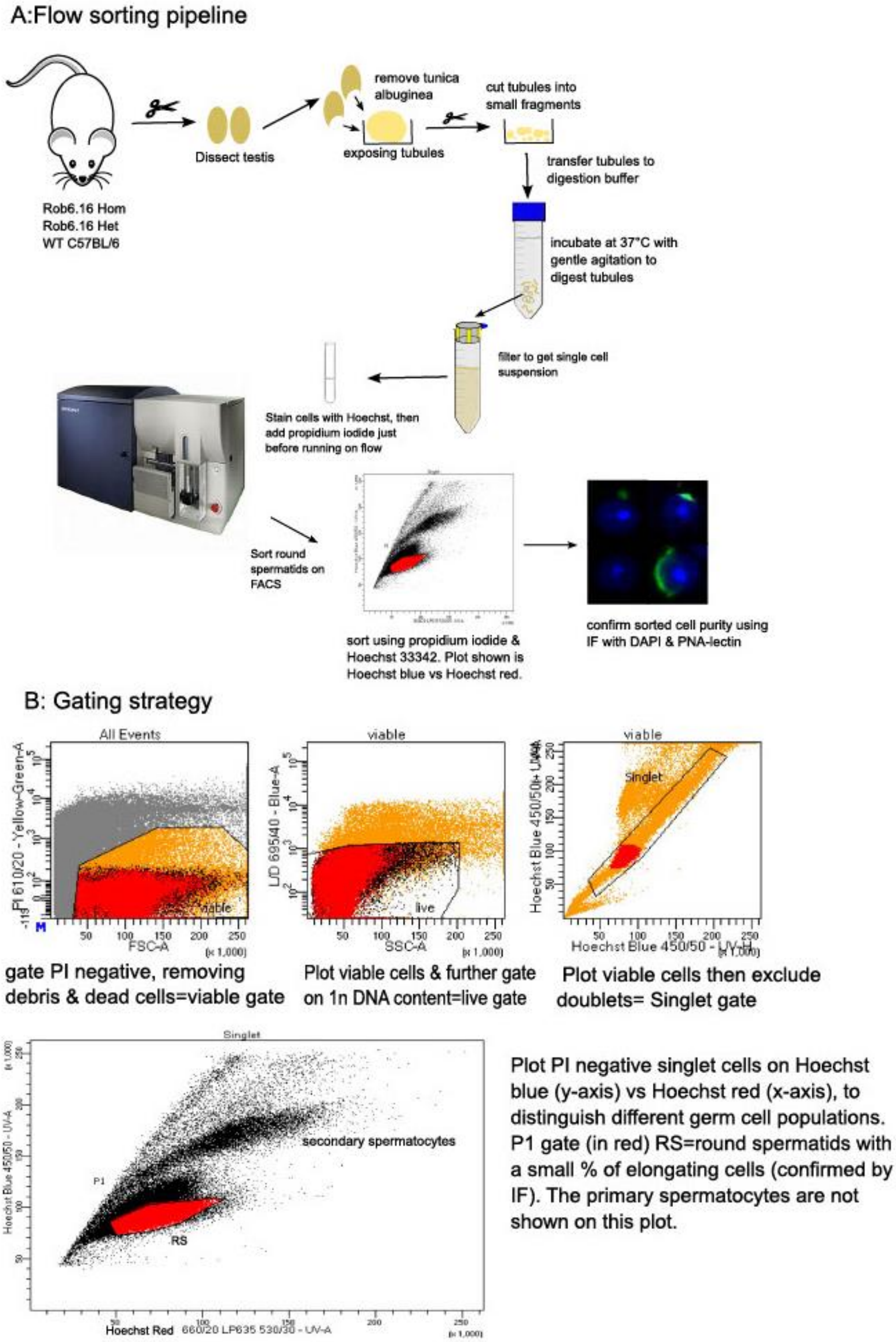

A:Flow sorting pipeline
Rob6.16 Hom
Rob6.16 Het
WT C57BL/6
Dissect testis
remove tunica albuginea
exposing tubules
cut tubules into small fragments
transfer tubules to digestion buffer
incubate at 37°C with gentle agitation to digest tubules
filter to get single cell suspension
Stain cells with Hoechst, then add propidium iodide just before running on flow
Sort round spermatids on FACS
sort using propidium iodide & Hoechst 33342. Plot shown is Hoechst blue vs Hoechst red.
confirm sorted cell purity using IF with DAPI & PNA-lectin
B: Gating strategy
All Events
PI 610/20 - Yellow-Green-A
FSC-A
viable
gate PI negative, removing debris & dead cells=viable gate
viable
L/D 695/40 - Blue-A
SSC-A
live
Plot viable cells & further gate on 1n DNA content=live gate
viable
Singlet
Hoechst Blue 450/50 - UV-A
Hoechst Blue 450/50 - UV-H
Plot viable cells then exclude doublets= Singlet gate
Singlet
Hoechst Blue 450/50 - UV-A
P1
RS
secondary spermatocytes
Hoechst Red 660/20 LP635 530/30 - UV-A
Plot PI negative singlet cells on Hoechst blue (y-axis) vs Hoechst red (x-axis), to distinguish different germ cell populations. P1 gate (in red) RS=round spermatids with a small % of elongating cells (confirmed by IF). The primary spermatocytes are not shown on this plot.

Figure 2-2: Schematic showing the steps of spermatid sorting.

A) Testis dissection method. Briefly, the testes are dissected from the mouse and the tunica albuginea is removed and discarded. The tubules are chopped into small fragments and transferred to dissociation buffer containing collagenase type I, DNase I and trypsin. Digestion is carried out until a single cell suspension is obtained. The cell suspension is filtered through a 70µm filter to remove cell clumps. The sample is stained with PI and Hoechst 33342 and sorted on the FACS. Cell purity is confirmed by immunofluorescence (IF) staining with DAPI and PNA-lectin. **B)** Flow sorting gating strategy. RS=round spermatids.

2.3 RNA extraction from sorted spermatids and library preparation

RNA was isolated from spermatids of three wild type mice, three homozygous Rob6.16 mice and three heterozygous Rob6.16 mice. From each mouse RNA from two pools of 300,000 spermatids was extracted with the RNAqueous micro kit (Thermo Fisher) on two columns with DNase I treatment as per the manufacturer's instructions. Once the RNA was eluted from the columns, RNA from the same animal was pooled. This gave three biological replicates per genotype, with each RNA sample originating from 600,000 spermatids. RNA was quantified on the Nanodrop. The RNA was eluted from the columns in the smallest volume possible to obtain the highest RNA concentration per µl possible (approx. 17µl).

~1.2µg of RNA per sample was sent to Novogene (Cambridge, UK) for paired end 150bp library preparation and sequencing using the Illumina NovaSeq platform. RNA quality was checked on a bioanalyzer before library preparation. RIN values ranged from 8.7-9.7. Messenger RNA was purified from total RNA using poly-T oligo-attached magnetic beads. After fragmentation, the first strand cDNA was synthesized using random hexamer primers followed by second strand cDNA synthesis. The libraries were then end repaired, A-tailing was carried out and adapters were ligated. Size selection was then carried out, libraries were amplified by PCR, and then finally purified before sequencing. Only replicate three of the Rob6.16 heterozygous sample was split over more than one lane of the sequencer; the remaining samples were not split over different lanes, but were sequenced on different flow cells. For RNA-seq statistics please refer to Table 5-3.

2.4 Cut&Tag

Cleavage under targets and tagmentation (Cut&Tag), a variant of ChIP-seq was used to investigate R-loop signals and gammaH2AX in a DSB inducible cell line (DIvA U20S). DSBs can be induced at specific sites in the DIvA cells by the addition of 4-Hydroxytamoxifen (4OHT) to the cell culture media.

2.4.1 DIvA cell culture

DIvA cells are a human U20S cell line developed by Lacovoni *et al* and Massip *et al* (245, 246), which allow the generation of DSBs at defined positions in the genome. The DIvA cell line expresses the AsiSI restriction enzyme fused to a modified oestrogen receptor ligand binding domain and to an auxin-inducible degron (AID-AsiSI-ER) (both under puromycin selection). Treatment of the DIvA cells with 4-hydroxy tamoxifen (4OHT) causes the nuclear localization of the AsiSI restriction enzyme and the rapid induction (<1 hour) of multiple (approximately 150) sequence-specific DSBs, widespread across the genome (245).

DIvA cells are an adherent cell line, which were grown in T75 flasks at 37°C, 5% $C0_2$. The media was composed of DMEM Glutamax + glucose 4.5g/L –pyruvate (Invitrogen 61965), 1mM sodium pyruvate (Fisher 11548876), 10% FBS, 1% Pen/Strep (Fisher 11548876), and Puromycin at 1 µg/ml (Fisher 15490717). The cells were split 1:2 to 1:6 depending on their growth rate, and they were not allowed to get more than 80 % confluent. Media was changed approximately every 3 days. Splitting was carried out as follows: The spent media was removed from the flask and discarded, the adherent cells were rinsed in PBS (no calcium, no magnesium) and this was discarded. 3mls of 0.25% trypsin was added and the cells were incubated at 37°C, 5% $C0_2$ until the cells detached. Complete cell culture media up to a total volume of 10mls was then added to inactivate the trypsin and a portion of the cell suspension was carried forward for splitting.

To induce DSBs at AsiSI restriction sites throughout the genome, the DIvA cells were treated with 300nM (Z)-4-Hydroxytamoxifen (4OHT Sigma-Aldrich H7904) [diluted in100% ethanol] for 4 hours at 37°C 5% $C0_2$. 4OHT was not added to some flasks to act as a control. Post DSB induction, the induced DIvA cells or control cells were harvested using Enzyme Free Cell Dissociation Solution Hank's Based (Merk S-004-C) with gentle scraping. This was to maintain the cell surface receptors that are required for the cells to bind to the con-A beads at the start of the Cut&Tag protocol. The cells were then counted using trypan blue to assess viability. Aliquots of 1e5 cells or multiples of 1e5 cells were then frozen down in 1.5ml PCR clean tubes in freeze media (10% FBS, 40% complete media, 10% DMSO).

2.4.2 Checking DSB induction efficiency

To assess the efficiency of DSB induction both a sample of control cells and induced cells were fixed in 4% PFA (diluted in PBS) for 30 minutes at 4°C. The cells were then washed twice in ice cold PBS and stored in PBS at 4°C. 15,000 fixed cells were cytospinned per slide, with 3 slides per control cells and 3 per induced cells. The cells were then permeabilized as follows: 1 minute in cytoskeletal (CSK) buffer 0.1% triton. CSK buffer is composed of 10mM PIPES pH to 6.8, 300mM sucrose, 100mM NaCl, 3mM $MgCl_2$, 1mM EDTA, made up in deionized water, filter-sterilized and stored in aliquots at -20°C. The cells were then rinsed twice in PBS and incubated for 20 minutes on ice in CSK buffer 0.5% triton. The slides were then washed twice in TBS 0.025% triton and blocked with TBS 10% FBS, 1% BSA for 2 hours at room temperature.

Immunofluorescence staining was carried out with one antibody per slide of either GammaH2AX, S9.6 or a secondary antibody only as a control. Anti-phospho-Histone H2AX (ser139) Millipore 05-636 was used at 1:400. Anti-DNA-RNA Hybrid Antibody, clone S9.6 R-loop antibody MABE109 was used at 1:50. Staining was carried out overnight at 4°C in a humidified chamber. The secondary antibodies were washed from the slides with 3x 5-minute washes in TBS 1% BSA.

Goat Anti-Mouse IgG H&L (Texas Red) Abcam ab6787 was used as the secondary at 1:500 and incubated for 2 hours at room temperature in the dark. Slides were then washed 3x 5 minutes with TBS 1% BSA and mounted using VECTASGIELD anti-fade mounting media containing DAPI.

2.4.3 DIvA Cut&Tag

To determine if the Cut&Tag protocol would work on the DIvA cell line, an initial pilot experiment was carried out using 80,000 fresh cells per reaction, with a control and an induced library prepared with gammaH2AX. Libraries of the correct size and concentration were obtained and sent for sequencing at the Earlham Institute (Norwich, UK).

The Cut&Tag protocol outlined in Figure 1-31 was run using the Active Motif kit 53160 as per the kit instructions. The cryopreserved DIvA cells were prepared as follows: 1e5 cryopreserved DIvA cells from either Control (not induced for DSBs) or induced samples were used per Cut&Tag reaction. The cells were thawed quickly in a 37°C water bath and the DMSO in the freeze media was diluted out by the addition of PBS. Cells were spun at 200g for 5 minutes. The supernatant was removed and discarded, and the cell pellet was resuspended in 1ml of PBS. The cells were spun again at 200g for 5 minutes and the supernatant was discarded. The protocol as described in the Active Motif kit catalogue number 53160 was then followed. Briefly, Cut&Tag uses an antibody-based enzyme tethering strategy to target specific histone modifications or proteins of interest. Instead of sonication of fixed chromatin and immunoprecipitation as performed in ChIP-seq, in Cut&Tag, unfixed cells were bound to concanavalin A beads (to facilitate the washing steps), and the antibody incubations were performed with cells in their native state. The cells are first permeabilized then the primary and secondary antibodies are added, the chromatin was digested and next generation sequencing (NGS) libraries were prepared in a single step by tagmentation using the protein A-Tn5 (pA-Tn5) transposase enzyme that has been pre-loaded with sequencing adapters (Figure 1-31). Tn5 was activated by the addition of a buffer containing magnesium. This results in the chromatin being cut close to the antibody bound site, with the addition of sequencing adapters to the cut ends. Library preparation is then carried out by PCR with the addition of specific sequencing indexed primers, so that the libraries can be multiplexed. The libraries were purified using solid phase reversible immobilization (SPRI) magnetic beads and quantified on a high sensitivity Qubit kit (Invitrogen).

Cut&Tag requires lower cell inputs than ChIP-seq and can be used with between 50-500,000 cells per reaction. The primary antibody is specific to the protein of interest in this instance gamma H2AX or DNA:RNA hybrids. The addition of the secondary antibody helps to amplify the signal. As the DNA is cut close to the antibody binding site, robust results can be achieved with a lower sequencing depth of 3-5 million reads per sample.

The following antibodies were used: Anti-phospho-Histone H2A.X (Ser139) antibody, clone JBW301, Merk 05-636 (1µg per reaction). Anti-DNA-RNA hybrid antibody, clone S9.6 MABE1095 Merk Sigma-Aldrich (2µg per reaction).
To degrade R-loops (DNA:RNA hybrids), RNase H (20U per reaction of NEB M0297S) was added for the duration of the primary antibody incubation). Primary antibody incubations were carried out overnight at 4°C with shaking. 1µg of Rabbit Anti-mouse IgG H&L ab46540 secondary antibody was used per reaction and incubated for one hour at room temperature.

A unique combination of i7 and i5 index was added to each PCR reaction, for the separate libraries, so that the libraries could be pooled (multiplexed) and sequenced on one lane.

Three technical replicates of the DIvA cells were carried out. Replicate 1 was produced with 14 PCR cycles as per the cycling conditions of the Active Motif protocol. To increase the library yield to meet library sequencing requirements, replicates 2 and 3 were produced with 16 PCR cycles. Library QC and sequencing was carried out by Novogene UK, with 10 million reads requested per library at 150bp PE (paired-end), equivalent to 3Gb of data. See Table 3-2 for library statistics.

2.4.4 Dissociated testis Cut&Tag

To determine if the Cut&Tag protocol would work on a dissociated testis sample, (without further optimization) a pilot experiment was carried out using a freshly dissociated testis (prepared with 0.25% trypsin digestion in DMEM (no glucose, no glutamine, no phenol red (A14430-01, Thermo Fisher Scientific)) and 100,000 cells per reaction. A frozen formaldehyde fixed testis sample was also tested, but library preparation was not successful (see Figure 3-10), so only cryopreserved unfixed samples were used. A library was prepared with gammaH2AX (Millipore 05-636, 1/100 dilution) and one with H3K27me3 (ab6002 Abcam, also 1/100 dilution). Libraries of suitable concentration and size were obtained and sent to the Earlham Institute (Norwich, UK). The aim of the data analysis was to determine if the libraries had the correct read length periodicity and whether enrichment of gammaH2AX and depletion of H3K27me3 reads were observed on the sex chromosomes compared to the autosomes.

2.4.5 Cut&Tag data analysis

The Cut&Tag data analysis pipeline developed by the Henikoff lab was followed: (https://yezhengstat.github.io/CUTTag_tutorial/). Briefly FASTQ files of raw 150bp PE data were obtained from Novogene and analysed on the high-performance computer at the University of Kent. The libraries were all sequenced on one lane and between 0.9-3Gb of data was obtained per library. Read quality was checked using FastQC (v0.11.9, Andrews 2020). Cut&Tag reads will often fail FastQC, on the per base sequence content as there is discordant sequence content at the beginning of the reads, but this does not mean that the data is of poor quality (see Figure 3-1).

Reads were trimmed using Trimmomatic (v0.39) (247) using the PE setting and AVGQUAL:20 to remove reads with an average phred score below 20. This removed between 0.01-0.03% of all reads. If required post-trimming, FastQC can be re-run to determine if the trimming settings were stringent enough. Since the percentage of trimmed reads was so low and that trimming of reads for Cut&Tag analysis is not recommended, untrimmed reads were aligned to the hg38 genome (from which random and unplaced chromosomes had been removed) using bwa-mem (v0.7.17-r1188, Li, 2013). *Samtools flagstat* (v1.7-22-gf9e5e35) was used to obtain alignment statistics. *Samtools index* was used to index the bam files.

The periodicity of the libraries was calculated by converting the aligned bam files to sam format using *samtools view -h* and extracting data from the 9th column, which contains information on the fragment size. Library periodicity plots were generated in R using ggplot2 with geom_line. The level of duplicate reads was calculated using Picard. Sam files were sorted by coordinate using *PicardCMD SortSam* and the duplicates were marked using *PicardCMD MarkDuplicates.* If duplicate read levels were below 1% then duplicate reads were not removed. For the DIvA library prep, duplicate reads levels typically ranged from 0.65-0.85%.

For data visualization of the read density over the AsiSI sites, further processing of the bam files was required. From sorted bam files Deeptools *bamCoverage* (248) was used to generate a BigWig file, which was then used as input to Deeptools *computeMatrix* (v3.5.1) (248). The computeMatrix output was used in Deeptools *plotHeatmap* to plot the read density centred over the top 80 most cut AsiSI sites. Plots extending 1kb upstream and downstream of the AsiSI sites were produced. The data for the 80 most cut AsiSI read sites was obtained from Clouaire *et al* 2018 (249).

Cut&Tag blacklisted regions were not removed from the files before analysis. I tested whether the presence of blacklisted regions made a difference to the deeptools plots at the 1kb scale, and it did not - because we were looking specifically at the AsiSI sites and there were no blacklisted regions ~+/- 1kb from in the vicinity of these sites. However, for larger scale analysis blacklisted regions would likely be present and so should be removed before plotting.

2.5 Data set mining

To investigate the correlation of spermatid DSBs to different repeat sequences, histone marks and Non-B DNA sequences data was gathered from the NCBI repository and the Non-B database (Table 2-2).

Table 2-2: Publicly available data files used for analysis.

Accession	study	Ref	files downloaded	Mark	Cell type
ERX1946916	PRJEB20038	Grégoire *et al.,* 2018	ERR1886418.1	DSBs	steps 1-9 spermatids
ERX1946917			ERR1886419.1	DSBs	steps 1-9 spermatids
ERX1946918			ERR1886420.1	DSBs	steps 15-16 spermatids
ERX1236388	PRJEB11644	Kocer A *et al.*, 2015	ERR1162981	WT sample	WT Cauda epididymal spermatozoa
ERX1236389			ERR1162982	Oxidative damage	*Gpx5-/-* cauda epididymal spermatozoa
ERX1236390			ERR1162983	Oxidative damage	*Gpx5:SnGPx4-/-* cauda epididymal spermatozoa
SRX207541	GSM1046836	Erkek *et al.,* 2013	SRR625509	input for DSB data	sonicated sperm DNA
SRX207545	GSM1046840		SRR625513	H3K4me3 rep1	ST
SRX207546	GSM1046841		SRR625514	H3K4me3 rep2	ST
SRX207547	GSM1046842		SRR625515	H3K27me3 rep1	ST
SRX207548	GSM1046843		SRR625516	H3K27me3 rep2	ST
SRX332345	GSM1202707	Hammoud *et al.,* 2014	SRR948800.2	H3K4me3	ST
SRX332348	GSM1202710		SRR948805.2	H27me3	ST
SRX332350	GSM1202712		SRR948807	H3K4me1	ST
SRX332353	GSM1202715		SRR948811.2	H3K27ac	ST
SRX332363	GSM1202725		SRR948825.2	input	ST
SRX332356	GSM1202718		SRR948814.1	5hmC	ST
SRX332360	GSM1202722		SRR948819/SRR948820	H2AZ	ST
SRX099896	GSM810677	Tan M *et al.,* 2011	SRR350906.2	Kac	ST
SRX099897	GSM810678		SRR350907.2	Kcr	ST
SRX719833	GSM1519002		SRR1596612.1	BRD4	ST

Accession	study	Ref	files downloaded	Mark	Cell type
SRX719834	GSM1519003	Bryant JM et al., 2015	SRR1596613.1	H3K9me3	ST
SRX719835	GSM1519004		SRR1596614.1	H3K9ac	ST
SRX719836	GSM1519005		SRR1596615.1	H4K5ac	ST
SRX719837	GSM1519006		SRR1596616.1	H4K8ac	ST
SRX719838	GSM1519007		SRR1596617.1	H4K12ac	ST
SRX719839	GSM1519008		SRR1596618.1	H4K16ac	ST
SRX719840	GSM1519009		SRR1596619.1	H4Kac	ST
DRX117176		Yamaguchi K *et al.*, 2018	DRR124323.1	H3-C method 1	sperm
DRX117177			DRR124324.1	H3-C method 2	sperm
DRX117176			DRR124330.1	H3-N method 1	sperm
DRX117177			DRR124335.1	H3-N method 2	sperm
DRX117176			DRR124328.1	H3K4me3 method 1	sperm
DRX117177			DRR124333.1	H3K4me3 method 2	sperm
DRX117176			DRR124329.1	H3K9me3 method1	sperm
DRX117177			DRR124334.1	H3K9me3 method 2	sperm
DRX117176			DRR124331.1	H4 method 1	sperm
DRX117177			DRR124336.1	H4 method 2	sperm
SRX3697622	GSE110582	Marsico G *et al.,* 2019	GSM3003548_Mouse_all_w15_th-1_minus.hits.max.PDS.w50.35.bed	G-Quadruplex experimental	Mouse skin C57BL/6J strain
N/A	N/A	https://nonb-abcc.ncifcrf.gov/apps/Query-GFF/feature/	Non-B DNA DB mm38	Z-DNA	predicted from mm38 genome
N/A	N/A	https://nonb-abcc.ncifcrf.gov/apps/Query-GFF/feature/	Non-B DNA DB mm38	G-quadruplex	predicted from mm38 genome
N/A	N/A	https://nonb-abcc.ncifcrf.gov/apps/Query-GFF/feature/	Non-B DNA DB mm38	short tandem repeat	predicted from mm38 genome

Accession	study	Ref	files downloaded	Mark	Cell type
SRX7176634	GSM4175885	Srinivasan *et al.*, 2020	GSM4175885_GFP-Zscan_GFP_ChIP.bw	ZCAN4	Embryonic stem cells (ESC) cell line: E14Tg2a
SRX7756509	GSE145598	Bouwman B *et al.*, 2020	GSM4322063_Enterocyte_High_M1.bed	sBLISS	Enterocyte (Intestine)
SRX7756512			GSM4322066_Enterocyte_High_M2.bed	sBLISS	Enterocyte (Intestine)
SRX332343	GSE49621	Hammoud *et al* 2014	SRR948796.1/SRR948797.1/SRR948798.1	H3K4me3	Spermatogonia
SRX332344		Hammoud *et al* 2014	SRR948799.2	H3K4me3	Spermatocytes
SRX332345		Hammoud *et al* 2014	SRR948800.2	H3K4me3	Spermatids
SRX332346		Hammoud *et al* 2014	SRR948801.1/SRR948802.1/SRR948803.1	H27me3	Spermatogonia
SRX332347		Hammoud *et al* 2014	SRR948804.2	H27me3	Spermatocytes
SRX332348		Hammoud *et al* 2014	SRR948805.2	H27me3	Spermatids
SRX332351		Hammoud *et al* 2014	SRR948808.1/SRR948809.1	H3K27ac	Spermatogonia
SRX332352		Hammoud *et al* 2014	SRR948810.2	H3K27ac	Spermatocytes
SRX332353		Hammoud *et al* 2014	SRR948811.2	H3K27ac	Spermatids
SRX332361		Hammoud *et al* 2014	SRR948821.1/SRR948822.1/SRR948823.1	input	Spermatogonia
SRX332362		Hammoud *et al* 2014	SRR948824.2	input	Spermatocytes
SRX332363		Hammoud *et al* 2014	SRR948825.2	input	Spermatids
SRX3114879	PRJNA399533	Maezawa *et al* 2018	SRR5956512	ATAC-seq	Spermatocytes
SRX3114880		Maezawa *et al* 2018	SRR5956513	ATAC-seq	Spermatocytes
SRX3114883		Maezawa *et al* 2018	SRR5956516	ATAC-seq	Spermatids
SRX3114884		Maezawa *et al* 2018	SRR5956517	ATAC-seq	Spermatids

2.5.1 Data types

Various data types were downloaded from NCBI and a brief description of each is outlined below.

2.5.1.1 Spermatid DSB data (DBrIC)

To study spermatid DSB locations, data obtained by DNA break Immunocapture (DBrIC) was analysed using the same parameters described in the original study (84). DBrIC involves nick and gap repair using T4 ligase and polymerase. The DSBs that remain are labelled with TdT and biotin-14-dATP. The DNA is fragmented using Shearase and immunoprecipitated. This technique does not show precise DSB locations, but indicates the near vicinity of each labelled DSB, with the resolution governed by the size of the sheared DNA fragments.

Three fastq files were downloaded from NCBI, ERR1886418 and ERR1886419 (spermatids steps 1-9), and ERR1886420 (spermatids steps 15-16) (84) (Table 2-2). These data files were chosen to avoid methodological bias, as the data was obtained from the same publication and generated by the same technique.

2.5.1.2 DSB detection by sBLISS (in enterocytes)

To study enterocyte DSB locations (to use this data as an alternative source of DSB data to compare to the spermatid DBrIC DSB locations data), sBLISS (in-suspension break labelling in situ and sequencing) data was downloaded from NCBI (GSE145598) (250). This cell type was chosen solely on the basis that a suitable data set of DSBs was available within the same mouse strain (C57BL/6) as the spermatid DSB data. In sBLISS the cells are harvested, fixed and crosslinked. DSB ends are then blunted, and adapters are ligated, followed by next generation library preparation and sequencing. This method does not require an immunoprecipitation step. sBLISS also indicates the precise location of each detected DSB. Both sBLISS files used were replicates from enterocytes with high levels of the cell surface markers CD73 representing cells from the tip of the villus. CD73 is a cell surface glycosylphosphatidylinositol-anchored glycoprotein. It is essential for the generation of extracellular adenosine from 5’-adenosine monophosphate (5’-AMP) (251).

2.5.1.3 Histone marks

ChIP-seq FASTQ files from either spermatogonia, spermatocytes, round spermatids or epididymal sperm (80, 81, 252–255) were downloaded from NCBI (Table 2-2), reads were pre-processed as described below. *Samtools merge* was used to combine BAM files of the same histone mark into one file.

2.5.1.4 Oxidative damage data

Paired end cauda epididymal sperm oxidative damage (OD) data was obtained from PRJEB11644 (141) see Table 2-2. Three data files were downloaded, one control and two with oxidative

damage. Data from file ERR1162982 with a *Gpx5* mutation will be referred to as moderate OD and data from file ERR1162983 with both a *Gpx5* and *SnGPx4* mutation will be referred to as severe OD. The oxidative damage data was obtained by Oxidized DNA Immuno-Precipitation (OxiDIP). A method which uses an anti-8-OHdG antibody that specifically recognizes
oxidized guanine residues, (a fingerprint of oxidative DNA damage) to immunoprecipitate oxidised regions (256).

2.6 Pre-processing of raw data files obtained from NCBI

All raw data downloaded from NCBI was passed through a pre-processing quality control pipeline. This involved checking read quality and trimming the reads where necessary. The raw reads downloaded from NCBI ranged from 54.5 to 66.5 million for the DBrIC samples, 3.1-31.7 million for the oxidative damage samples, 13.4-305.1 million for the spermatid marks and 4.9-40.9 million for the retained histone samples from epididymal sperm. Read quality was checked using FastQC (v0.11.9, Andrews 2020). This tool enables visual representation of the quality of raw sequencing reads and enables the user to determine what further trimming of the reads is required, before proceeding to read alignment. FastQC can identify whether the sequencing quality at the end of the read falls below an average quality threshold and it can be used to determine if adapter sequences are still present. Raw reads were further processed with the tool Trimmomatic (v0.39) (247) see Table 2-3. Adapter sequences (as determined by FastQC were removed within Trimmomatic using the ILLUMINACLIP option. Trimmed FASTQ files were then aligned to the mouse genome mm10 using bwa-mem with default parameters (v0.7.17-r1188, Li, 2013). The percentage of mapped reads ranged from 87.6-92.0% for the spermatid DBrIC samples, 96.9-97.0% for the oxidative damage samples, 80.7-99.2% for the spermatid marks and 95.7-99.5% for the retained histone data.

Table 2-3: Preprocessing pipeline for the different data files downloaded from NCBI.

Files	**FastQC**	**Trimmomatic settings**	**Bwa-mem** v0.7.17-r1188, Li, 2013).
Spermatid DSB files	Run before & after Trimmomatic	AVGQUAL:30 (to remove reads with an average Phred score of <30) MINLEN:30 (to remove reads with an average length <30)	Default settings
Histone marks & ATAC-seq	Run before & after Trimmomatic	SE, ILLUMINACLIP (to remove any adapter sequences). AVGQUAL:20, (to remove reads with an average length <20).	Default settings

2.7 Peak calling

2.7.1 Spermatid DSBs, histone marks and ATAC-seq data

The tool MACS2 (Model based analysis of ChIP-seq) (MACS2 v.2.2.7.1) (257) was used to identify DSB, histone and ATAC-seq peaks. MACS2 can be used to identify either broad or narrow peaks within the data and can be used with a control (or input sample) to increase the specificity of peak calling. The –broad setting will link nearby peaks. For the DSB data, peaks were called using the same parameters described in the original study (84). Briefly, MACS2 with the default settings and *--bw600 -q0.01 --broad --broad-cutoff 0.1* was used (see Supplementary Table 8-5 for peak statistics). The broad setting is suitable for the DSB peak detection as we were not interested in defining the exact location of each DSB peak, but broader regions where DSBs occur. *--broad-cutoff* 0.1 means that the q value cutoff for the broad peak setting will be 0.1. -q 0.01=Minimum FDR (q-value) cutoff for peak detection. –bw600=band width for picking regions to compute fragment size. This value is only used when building the shifting model. The band width of the peaks is half of the estimated sonication fragment size with a default of 300bp. The band width was kept the same as the original paper as it relates to the fragment sizes of the peaks obtained in the ChIP-seq experiment. DSB peaks called with these settings ranged from 146bp to 6662bp with a mean of 267bp (see Supplementary Table 8-5). For the histone data either the default settings of *callpeak* to produce narrow peaks or with *--broad --broad-cut-off 0.05* to produce broad peaks was used. Histone marks were defined as narrow or broad based on the ENCODE project (258). For marks not defined on ENCODE, the broad settings were used. (See Supplementary Table 8-5 for peak statistics and Figure 4-15 and Supplementary Table 8-3 for coverage of histone marks per chromosome). Single end ATAC-seq data was analysed with the default settings of *callpeak* to produce narrow peaks with the same parameters described in the original study (255) (MACS2 with the default setting of *callpeak* and *-q 0.2*).

2.7.2 Oxidative damage data

Paired end cauda epididymal sperm oxidative damage data was obtained from PRJEB11644 (141). Data from file ERR1162982 with a *Gpx5* mutation was termed moderate OD and data from file ERR1162983 with both a *Gpx5* and *SnGPx4* mutation was termed severe OD. A windowed data analysis approach was used to calculate the amount of oxidative damage in 50kb windows (259). Briefly read 1 and read 2 FASTQ files were independently aligned to the mouse genome mm10 using *bwa-mem* (v0.7.17-r1188). The bam files were sorted and read 1 read 2 files were merged using *samtools merge* (260) to treat the files as SE as per the original paper. 50kb windows of the mm10 genome were created using bedtools *makewindows* (261). *Bedtools coverage* (261) with

the –*mean* was used to calculate the mean coverage at each 50kb window. A global scaling factor was calculated as the mean read coverage in a low-damage subset (10%) of the 50kb bins, these bins had the lowest 10% read coverage in the 50kb windows over the mean of all three files downloaded. This global scaling factor was used in a custom Perl script to calculate scaled OD values per 50kb window of the mm10 genome. Relative enrichment of DNA damage was assessed through the fold change of the enriched OD samples over input WT sample.
Any 50kb window of the genome which overlapped with GSAT_MM regions from the RepeatMasker track was removed from both OD files. The top 1% of the highest damaged windows were used for further analysis (see chapter 4 Table 4-5).

2.7.3 Non-B DNA sequences

The non-B DNA data was not peaks per se, but regions of the genome predicted to form non-B DNA sequences. Predicted non-B DNA sequences (Z-DNA, short tandem repeats (STR) and G-quadruplex) were downloaded from the Non-B DNA Database (https://nonb-abcc.ncifcrf.gov/apps/Query-GFF/feature/) from the mm10 genome. Per Non-B DNA type, data for each chromosome was merged into a master file. For the predicted Z-DNA and STR samples *bedtools intersect* (261) was used with the –v option to obtain a file of predicted Z-DNA regions that did not overlap STR and a file of STR regions that did not overlap predicted Z-DNA. While experimentally validated G-quadruplex regions were obtained from GSE110582 (262).

2.8 Further processing of the peak data

2.8.1 Spermatid DSBs

To assess whether peak files of the different round spermatid stages (stage 1–9 and stage 15–16) could be merged to facilitate cross-comparison with datasets not stratified into different developmental stages in spermatids. We first examined the overlap between files by using an UpSetR plot (263). To generate the UpsetR plot, first we generated 1kb windows of the mm10 genome and then intersected each DSB peak file with these windows. Per sample we then summed the number of DSB peaks within each window, these values were then input into UpSetR. A 1kb windowed UpSetR analysis showed a good correlation between the different DSB files (Figure 4-9). A 5kb windowed UpSetR analysis showed more overlap, as expected (Supplementary Figure 8-9). We then performed a correlation analysis with the 1kb windows using corrplot (Supplementary Figure 8-10), which also showed a good correlation. Therefore, peaks files of the different round spermatid stages (stage 1-9 and stage 15-16) were merged for further analysis.

2.8.2 sBLISS enterocyte data and comparison to spermatid DBrIC data

For an initial comparison of the degree of concordance between spermatid DBrIC DSB locations and enterocyte DSB locations, the enterocyte files (GSM4322063 termed sBLISS 63 and GSM4322066 termed sBLISS 66) were analysed separately and the DSB locations analysed at maximum (1bp) resolution (Figure 4-13). A high degree of overlap would indicate that the sites of DSB in spermatids and enterocytes were similar, despite differences in the physiological environments. A low degree of overlap would indicate that the DSBs were occurring at distinct sites within spermatids and enterocytes. For further analysis both the sBLISS 63 and sBLISS 66 files were extended +/-133bp to simulate peaks the same size as the mean of the peaks obtained in the spermatid DBrIC data. Any peaks overlapping regions of centromeric satellite repeats (GSAT_MM from the RepeatMasker track) were removed. *Bedtools intersect* (version bedtools2-2.29.2) (19) with the -u option was used to obtain a unique file of the extended sBLISS 63 peaks that overlapped the extended sBLISS 66 peaks. This file was then used for motif analysis and permutation testing. We finally used this file with *bedtools intersect* with the **-**v parameter to identify spermatid DSB peaks not overlapping enterocyte DSBs.

2.9 ChromHMM

ChromHMM (v1.22) (264) was used for chromatin state analysis of the spermatid ChIP-seq data. ChromHMM is a tool which can be used to divide the genome into different states based on the intensity of histone marks. “ChromHMM learns chromatin-state signatures using a multivariate hidden Markov model (HMM) that explicitly models the combinatorial presence or absence of each mark. ChromHMM uses these signatures to generate a genome-wide annotation for each cell type by calculating the most probable state for each genomic segment” (264).

The first step of the analysis involves producing a tab separated table (the cellmark file), which contains all the information relating to the data files to use in the analysis, such as the cell type, the histone mark, the name of the bam file and the name of the control file. The next step is binarization of the data, for which bam files of the histone marks are required, the cellmark file, a file of the size of the chromosomes of the genome used for alignment to produce the bam files, as well as the desired binsize. The corresponding cell type specific input was used as a control to adjust the binarization threshold locally. The default binsize of 200bp was used with the concatenated strategy. Once binarization was completed the model was learned with varying numbers of states. To run the *learnmodel* script the binarized data is supplied to the *learnmodel* script, with the desired number of output states and the genome of the files used. (Figure 4-2). See Table 4-1 for coverage of the states per chromosome. The number of states should be adjusted to obtain maximum resolution between the different states but without obtaining states with extremely similar profiles. For the spermatid data, as 16 chromatin marks were used 16, 20, 30 and 50 states were tested, with 16 states chosen as the optimum model.

ChromHMM (v1.22) (264) was also used for chromatin state analysis across different developmental stages of spermatogenesis (Table 2-2) by applying the default binsize of 200 bp with the concatenated strategy. The corresponding cell type specific input was used as a control to adjust the binarization threshold locally (as for the 16 histone mark spermatid data). Once binarization was completed the model was learned with varying numbers of states (4-9), with 8 states (from E1 to E8) chosen as the optimum model.

To identify how chromatin states change during spermatogenesis, genomic locations of states in each cell type (spermatogonia, spermatocytes and spermatids) were compared using *bedtools intersect* (261). Regions of the genome missing a ChromHMM state in any cell type were removed. The dominant states in any of the three cell types (E7, E3 and E1) (Figure 4-2) were merged into a joint state named E0. Genomic locations were then labelled according to the states in each cell type, and the transitions from one chromatin sate to another were plotted using ggalluvial with ggplot2 in R. Consecutive 200 bp regions with the same three-cell type state combination were merged, and 34 combinations with more than 0.1% genomic coverage were identified. This cut off was chosen because the total coverage of all these regions represented >98% of the mm10 genome.

2.10 Statistical analysis

2.10.1 Permutation tests

RegioneR (1.22.0) (265) run with R 4.0.3 was used for permutation testing between the merged spermatid DBrIC DSB locations file and the histone ChIP-seq data or files of DNA structure (Supplementary Table 8-6 and Figure 4-10). Randomization was carried out per standard chromosome with an unmasked mm10 genome, 1,000 permutations and *nonoverlapping=FALSE*. Genomic association tester (GAT 1.3.4) (266) was used to compute the fold change between the merged spermatid DBrIC DSB locations file, the top 1% of moderate and severe OD damage regions, the histone ChIP-seq data and files of DNA structure Figure 4-18. The spermatid DBrIC DSB locations file, OD damage, non-B DNA files or histone ChIP-seq files were used as the segment file, the workspace was the mm10 genome and the annotation file was either the DBrIC DSB locations file, OD damage, a histone ChIP-seq file, or a file of DNA structure. All iterations were run with --*num-samples=10,000.*

2.10.2 DSB motif analysis

MEME 5.1.1 (267), (the most commonly used tool to search for nucleotide sequence motifs), was used to search for the top three motifs in the merged spermatid DSB sample with the option –*revcomp*, using both the given strand and the reverse complement strand when searching for motifs. The same settings were used to search for motifs in the spermatid specific breaks, enterocyte specific breaks, shared spermatid/enterocyte breaks and all enterocyte breaks (Table 4-3). From the input fasta files, MEME will output position-dependent letter probability matrices that describe the probability of each possible letter at each position in the pattern. The motifs do not contain gaps. MEME uses statistical modelling to automatically choose the best width and the number of occurrences of the motif. Each motif is given an E-value to describe the statistical significance of the motif. The MEME tool describes the E-value as "an estimate of the expected number of motifs with the given log likelihood ratio (or higher), and with the same width and site count, that one would find in a similarly sized set of random sequences (sequences where each position is independent and letters are chosen according to the background letter frequencies" (267).

Bedtools getfasta was used to extract the sequence contained within the extended enterocyte DSB peak file described above.

From the total number of peaks in the input file and the number of peaks with a given motif, the motif occurrence as a percentage of the total peaks can be calculated. This will allow comparison of motif abundance between different data sets, such as the spermatid and enterocyte DSBs.

2.10.3 Gene ontology enrichment

Using BioMart v2.46.1 in R v4.0.3, unique protein coding gene IDs in mm10 or mm39 for each classification were identified. These were input into the PANTHER db (268) for the gene ontology (GO) enrichment. Statistical overrepresentation test was selected with either GO biological process complete or PANTHER (v17.0) pathways. Where statistically significant results were obtained only GO terms with ≥1.5-fold enrichment and FDR < 0.05 were considered statistically significant. Plots were created with ggplot2 v3.3.5 in R.

2.11 Data visualization

Circular plots were created using the R tool circlize_0.4.15 (269) within R 3.6.1 with the outer track as the mm10 Ideogram. The UpSetR plots were created using the R tool UpSetR_1.4.0 (263) within R 3.6.1 using either 1kb or 5kb binarized DSB data as input. PyGenomeTracks-3.6 (270, 271) was used to visualize the same genomic regions from different browser tracks, to illustrate regions of overlap between samples. The PyGenome scripts make_tracks_file and pyGenomeTrack were used. Heatmaps were generated in R 4.0.3 using the program pheatmap (v1.0.12, Kolde, 2019). The 1kb DSB corrplot was created using the R tool corrplot (v0.92, Wei and Simko 2021) within R 4.0.3. SNP density per 1Mb window per chromosome was plotted using ggplot2. The region of interest (ROI) of chr6 and chr16 was plotted using gggenes (version 0.5.0) in R, to show the different categories of genes within the ROI.

2.12 SNP identification methods

Raw FASTQ files of whole genome sequencing data (from mouse tail) were obtained from Novogene for two homozygous Rob6.16 samples, Rob1 and Rob2. The data was processed using the High-Performance Computer (HPC) at the University of Kent, through a GATK bioinformatics pipeline. The aim was to obtain filtered homozygous SNPs that were concordant between the two Rob6.16 samples sent for sequencing. Concordant SNPs were selected to remove SNPs that were due to the variation between individuals. Heterozygous SNPs were excluded from analysis, as the Rob6.16 animals are inbred. Only high confidence SNPs supported by multiple reads (>15) were carried forward, to ensure that the SNPs identified were real and not artefactual.

The mouse genome (mm39) was downloaded from NCBI. GATK *CreateSequenceDictionary* was used to convert the mm39 genome into FASTA format. *Samtools index* was used to create an index of this genome. Fastp was used for read trimming with the following settings, --n_base_limit 15, --qualified_quality_phred 5, --unqualified_percent_limit 50. Paired end reads were aligned to the mm39 reference genome with bwa-mem 0.7.17-r1188, using the -M option to mark shorter split hits as secondary for Picard compatibility in later steps. The coverage of the bam files was checked using *samtools idxstats* with the default parameters. Samtools was used to check the read groups of the bam files and GATK *AddOrReplaceReadGroups* was used to replace any missing parameters, with read group parameters required to run Picards mark duplicates at later steps. GATK *MarkDuplicatesSpark* (4.3.0.0) was used with the default parameters to mark duplicate reads, so that they could be ignored in downstream processes. GATK *SetNmMdAndUqTags* script was then run with the default parameters to calculate NM, MD, & UQ tags by comparing with the mm39 reference, samtools *index* (1.10) was then used to index the resulting output files. Known *Mus musculus* variants were downloaded from Ensembl (file date 20220807, ##source=ensembl; version=108; url=https://e108.ensembl.org/mus_musculus). This file is required for the subsequent BQSR step of the GATK pipeline, and it used to determine which SNPs are in fact known *Mus musculus* variants. GATK *IndexFeatureFile* with the default settings was used to index this VCF file. The GATK *BQSRPipelineSpark* script was used for base quality score recalibration (BQSR) with the default settings. GATK *HaplotypeCaller* was used to call raw unfiltered SNPs and indels simultaneously, with the options --native-pair-hmm-threads 50 (50 threads used per native pairHMM implementation) and -ERC GVCF mode for emitting reference confidence scores. GATK *GenotypeGVCFs* was then used to convert the output of HaplotypeCaller into VCF format (which is required to run the select variants script). GATK *SelectVarian*ts script with the default settings was used to create separate files of SNPs and INDELS. GATK *VariantFiltration* was used to filter variants for high confidence calls. This script does not remove the variants it only marks them as failing, so failed variants were removed in a subsequent script. The GATK recommended filtration settings of, --filter-name "QD_filterLessThan2" -filter "QD < 2.00", --filter-name "QUAL30" -filter "QUAL < 30.00", --filter-name "FS_filterGreaterThan60" -filter "FS > 60.000", --filter-name "MQ_filterLessThan40" -filter "MQ < 40.00", --filter-name "SOR_filterGreaterThan3" -filter "SOR > 3.000, --filter-name "ReadPosRankSum-8" -filter "ReadPosRankSum < -8.00", --filter-name "MQRankSum-12.5" -filter "MQRankSum < -12.500" were used.

Filter QD is filtering on quality by depth, it is the variant confidence (from the QUAL field) divided by the unfiltered depth of non-hom-ref samples. This annotation is intended to normalize the variant quality to avoid inflation caused when there is deep coverage.

The QUAL filter is filtering on variant confidence quality score. QUAL is the Phred-scaled probability that the site has no variant, so by filtering on QUAL <30, only sites with >99.9% probability of a variant are kept.

Variants that failed the hard filtering were removed using VCFtools (0.1.16) using the option --remove-filtered-all. GATK *VariantFiltration* was used to mark the Het genotypes using the options -G-filter "isHet == 1" and -G-filter-name "isHetFilter". GATK *SelectVariats* was used to transform filtered variants to no call using --set-filtered-gt-to-nocall. These no call variants were then removed from the VCF file using grep. Heterozygous variants were removed from the analysis because we sequenced homozygous samples and we only wanted to investigate homozygous variants, as the mouse line Rob6.16 24lub is inbred and we were not interested in the variation between individual mice. The VCF files were then further filtered on read depth (DP) and genotype quality (GQ) using VCFtools with the parameters --minDP 15 --maxDP 1500 --minGQ 30. This marked the variants that failed these parameters as no call which were then subsequently removed.

GATK *IndexFeatureFile* was used to index the second VCF file, both files were then used in GATK *SelecVariants* script with the -conc option to filter for concordant variants. VCFtools was then used to produce VCF files of just chr6 and chr16 with the options --chr chr6 or --chr chr16. SnpEff (5.0e) (272) was used for variant annotation with the following settings used to obtain mostly protein coding annotations: -no-downstream -no-intergenic -no-intron -no-upstream -no-utr. SnpSift part of the SnpEff tool was used to obtain one variant type per line of the VCF file, so that further filtering could be carried out. Only protein coding annotations of HIGH, MODERATE or LOW impact were selected for further analysis. From this file, a list of genes containing HIGH and MODERATE impact SNPs was obtained for chr6 and chr16. This gene list was intersected with a list of genes likely not shared between spermatids (see section 2.14).

2.12.1 Identifying the size of the region of interest (ROI) on the Robertsonian 6.16 chromosome

GATK *selectvariants* was used to select all concordant filtered homozygous SNPs in the Robertsonian 6.16 data. *Bedtools makewindows* was used to make 50kb and 1Mb windows of the mm39 genome (with random and unplaced chromosomes removed). The selectvariants SNP data was then intersected using *bedtools intersect* with the 50kb and 1Mb windows of the mm39 genome. The number of SNPs within each genomic window was then summed using an awk script. This data was plotted using ggplot facet_grid in R to determine the size of the region at the start of chr6 and chr16 with increased SNP density. VCFtools with the option --window-pi with the hard filtered GATK variant file was also used to calculate nucleotide diversity (π). From the chromosomal start, to the point at which the π value drops is considered the region of interest (ROI). It is within this region that genes involved in transmission ratio distortion of the fusion chromosome may be located.

2.13 Spermatid RNA-seq data analysis

Data analysis was carried out on the high-performance computer at the University of Kent. The fastq files obtained from Novogene were run through FastQC (273) to determine if there were any overrepresented adapter sequences (there were none). Raw read depth ranged from 19.4-30.1 million reads per sample (Table 5-3). The RNA-seq reads were trimmed using Trimmomatic (v0.39) (247) with the PE option and MINLEN:100 AVGQUAL:30 (to remove reads with a minimum length of <100 and reads with an average quality score of <30). Trimmed reads were aligned to the mm39 genome using STAR (v 2.7.5c) (274). The mm39 genome was first indexed and then STAR was run with the following options to align the reads to the mm39 genome: *--outSAMtype BAM Unsorted SortedByCoordinate --readFilesCommand zcat --quantMode TranscriptomeSAM GeneCounts*. For samples split over more than one lane of the sequencer multiple 1P and 2P fastq files were specified within STAR.

2.14 Identifying genes within the ROI on chr 6 and chr16 that are not shared between spermatids

Genes whose transcripts are not shared between spermatids have been termed genoinformative markers or GIM by Bhutani *et al* 2021 (185). They carried out single cell RNA-sequencing to quantify allele-specific biases in spermatids and developed a new computational technique to jointly infer genotype and allelic expression biases (both technical and biological) in single haploid cells. The majority of GIM that they identified were autosomal although they did identify some GIM genes on the sex chromosomes.

The genoinformative marker (GIM) data from single cell RNA-sequencing of cells obtained from four mice (from F1 offspring from a cross of the distantly related inbred mouse strains C57BL/6 and PWK/PhJ) was downloaded from the Bhutani paper (185). Bhutani *et al* developed a Bayesian inference framework to jointly infer the haploid genotypes of each cell and the tendency of each gene to have genoinformative expression (i.e., incomplete sharing across cytoplasmic bridges). "The Bayesian model was fit to shuffled data and the posterior distributions for the parameters between real and shuffled data were compared. Cutoffs for confident GIMs were selected to achieve worst-case false discovery rates of 0.2 for each individual" (185). Using R the data was filtered to contain only the gene stable IDs of the confident GIMs on chr 6 and chr16. BioMart was then used to obtain the gene start and gene end coordinates of the gene stable IDs. This file was then further filtered to only contain the genes within the ROI of chr 6 and chr 16. The output file was saved in bed format, so that *bedtools intersect* could be used to intersect the high confidence GIM gene list with the output file from our SnpEff analysis.

A list of genes within the ROI obtained from SnpEff with high or moderate predicted SNP effects was compiled and this was then intersected with the confident GIM markers within the ROI (Figure 5-3). Genes containing SNPs with high or moderate predicted effects that overlapped the GIM data were of particular interest (Supplementary Table 8-11). A literature search was carried out to determine if any of the GIM genes were involved in the acrosome reaction or affected sperm motility or fertility. These are marked in Supplementary Table 8-11 In the comments column.

Genes with high and moderate effect SNPs that did not overlap the confident GIM genes from the Bhutani paper (Supplementary Table 8-12) were also identified. It is possible that these genes if they are not shared could also lead to TRD, if the gene products are involved in sperm motility or fertilisation.

2.15 Protein modelling of missense mutations within the ROI

UCSC genome browser (mm39) was used to obtain the AlphaFold (275, 276) models of the PLA210 and the Protamine 2 proteins (both proteins having been identified as containing missense mutations from the WGS data). The confidence scores (pLDDT values) of the models were low and so further structural prediction analysis was not carried out.

2.16 Calculating differential gene expression between WT-Het and Het-Hom Robertsonian samples

FeatureCounts (v2.0.6) (277) was used with the aligned.sortedByCoord.bam files output by STAR to obtain one file of all read counts per feature for all 9 samples (3 WT, 3 Het and 3 Hom). This file was then used to determine which genes with high or moderate effect SNPs within the ROI were expressed with a count ≥ 10 in all round spermatid samples. 15 out of 18 genes with high and moderate effect SNPs within the ROI of chr6 were expressed with a count ≥ 10 in all samples. 44 out of 61 genes with high and moderate effect SNPs within the ROI of chr16 were expressed with a count ≥ 10 in all samples.

The FeatureCounts output file was then used in the R package DESeq2 (v1.40.2) (278). Rows with a count <10 (across all the samples) were removed from the dds (DESeq2 data set). The dds was releveled to the wild type genotype. Then files of WT to Het, WT to Hom and Het to Hom were generated using the ‘contrast' option in DESeq2, specifying alpha=0.05 (to only include features (genes) with a p-value of <=0.05). The contrast output files form DESeq2 were then manipulated further in excel. Gene IDs, chromosome number, gene start position and gene end positions were downloaded from BioMart (279) and combined with the DESeq2 output file based on the gene IDs. Genes with p-adjusted values of ≤0.05 in both the WT-Het and Het-Hom comparisons were highlighted. Then only genes in the region of interest on chr6 and chr16 that were significantly (P ≤0.05) differentially expressed (in either direction) in both comparisons were taken forward for further analysis. This list of genes was then input into STRING-DB (280) and the PANTHER db (268) to look for enriched pathways or biological processes. Any gene that was differentially expressed (in either direction) in WT-Het and Het-Hom that contained a missense mutation was searched for in the Ensembl database to determine if there was a known fertility phenotype.
Genes that were significantly (P<=0.05) down regulated within the ROI in the WT-Het (a negative Log2 fold change) and in the Het-Hom (a positive Log2 fold change) DESeq2 comparisons were listed. These may represent genes that are silenced through post-meiotic sex chromatin repression (PMSCR).

2.17 Calculating allele-specific expression in the heterozygous Robertsonian samples

The bam files output by STAR (274) were sorted by coordinate and read group information was added using the picard script *AddOrReplaceReadGroups*. The high confidence concordant hard filtered homozygous Robertsonian SNPs were converted to heterozygous calls in the VCF file (0/1 or 0|1) using gsub. Only heterozygous biallelic SNPs should be used to calculate allelic expression, but since we only sent for sequencing the homozygous Robertsonian samples, this is the only data source available for use. The SNPs obtained from the homozygous sample were high confidence concordant homozygous SNPs, that must be heterozygous in the F1 hybrid males that were made by crossing the Rob6.16 homozygous mice with wild type C57BL/6 mice.

GATK (243) *ASEReadCounter* tool was used to calculate the raw read counts for each allele of each SNP within the heterozygous samples.

The allele-specific expression score (ASE) was then calculated per SNP as follows:

ASE score= (abs(Reference sum /Total count - 0.5) + 0.5)

This can give values ranging from 0.5 to 1.

The ASE score was then used to carry out a binomial test within R, using the function binom.test. The data was first rounded to give integers using the function Round within R. The ASE score * total count was used as x (the number of successes), the total count as n (the number of trials) and a median value of 0.5 as p (the hypothesized probability of success). The p-values obtained were then adjusted using the function p.adjust with the “fdr” method. The significance threshold was set to an adjusted p-value threshold of 0.05.

ASE was also calculated per gene (as opposed to per SNP). To do this, genome information was first obtained from BioMart GRCm39 (279) of chromosome, gene start, gene end, gene ID and gene name. For genes without any gene name information, awk was used to add the name ‘UNKNOWN’ so that all lines of the file contained the same number of columns. The ASEReadCounter output data was then intersected with this file of gene information using *bedtools intersect*. The data was then further processed in R. The reference and total allele counts per SNP for the three replicates were then summed. The file was then grouped by gene name and the sum of all the reference and total counts for SNPs within the same gene calculated, then averaged per SNP. For example, if the reference read sum (of het1, het2 and het3) of a gene was 1000 and the gene contained four SNPs, then the reference read average per gene per SNP would be 250. The same calculation was carried out for the total count per SNP per gene.

The ASE score per gene was then calculated in R using the following formula:

ASE score per gene= (abs(Average Reference sum of Het1/2/3 per SNP per gene / Average Total count Het1/2/3 per SNP per gene - 0.5) + 0.5).

The per gene ASE score was then used to carry out a binomial test within R, using the function binom.test. The data was first rounded to give integers using the function Round within R. The binomial test was carried out with the per gene ASE score * the average total count per SNP per gene, as x (the number of successes), the average total count per SNP per gene as n (the number of trials) and a median value of 0.5 (the expected ASE score if there is no allelic imbalance, as p the hypothesized probability of success).

The p-values obtained were then adjusted using the function p.adjust with the “fdr” method. The significance threshold was set to an adjusted p-value threshold of 0.05.

2.18 Code availability

No new data was generated for the spermatid DSB analysis study. Code used for plotting can be found at https://github.com/Farre-lab/Spermatid_DSB_paper

https://zenodo.org/record/7433522#.Y5iePXbP1PZ and https://github.com/Farre-lab/EBRs_HiC_Spermatogenesis_paper

3. Method development

3.1. Background

To investigate the chromatin architecture around double-strand breaks and the transmission dynamics of a Robertsonian fusion (chr6.16) claimed to undergo non-Mendelian inheritance, state of the art methods were required. These methods were not fully developed or established at the University of Kent and so the first step was method development and in vitro testing, to determine if they could be used to answer our experimental questions. I will now outline the methods that I tested or developed.

Cut&Tag stands for Cleavage Under Targets and Tagmentation (see methods). This protocol can be used to map chromatin features through immunoprecipitation of genomic regions bound by a specific antibody and is a successor to Cut and Run, a variant of ChIP-seq. It was developed by the lab of Dr S Henikoff and was first published in 2019 (281). The method differs from ChIP-seq: In Cut&Tag a primary antibody is used to bind to the regions of interest (as in ChIP-seq), but then a secondary antibody is used to amplify the signal. Once antibody is bound, a fusion of Protein A or Protein G, and transposase Tn5 (pAG-Tn5) is added. This selectively cuts the DNA and ligates sequencing adapters at the antibody-bound chromatin loci. The protocol can be carried out in intact nuclei (or cells), whereas nuclei are always used in ChIP-seq. Chromatin fragments in the vicinity of the target are amplified using primers that recognize the adapter-ligated DNA. The DNA is purified, and then sequenced.

Cut&Tag is faster than ChIP-seq, as steps such as end repair and adapter ligation are not required. It requires fewer cells (as few as 10,000), whereas for ChIP-seq 100s of thousands or millions of cells are required depending on the cell type. Cut&Tag has an improved signal to noise ratio, as the target bound chromatin is separated from the intact nuclei as part of the protocol. Consequently, the sequencing cost per sample is significantly reduced as only ~5 million reads are required per sample as opposed to > 20 million for ChIP-seq.

In this thesis, I used Cut&Tag as part of two different experiments: (i) I tested the resolution of Cut&Tag with different kinds of marks, including broad marks, gammaH2AX and R-loops; and (ii) I determined whether the protocol was amenable for use with primary cells either formaldehyde fixed frozen cells (to allow for easy shipping of material) or fresh cells.

A second set of methods were established to characterise the genomic landscape of the homozygous Robertsonian 6.16 mouse. Using whole genome sequencing we determined the SNPs and INDELs characterising the homozygous Robertsonian mice when compared to the mouse reference genome (mm39). Using these SNPs and INDELs we designed a rapid PCR genotyping assay to screen new animals from extracted DNA. This method speeds up our existing approach based on cell culture and karyotyping (see methods). Moreover, the identification of SNPs characterising Robertsonian mice is key to allowing us to detect allele-specific expression analysis through RNA-sequencing.

Finally, we improved upon published protocols, to separate round spermatids from a dissociated testis sample by FACS analysis. The protocol was not previously established at the University of Kent, and it required optimisation of the FACS setting to obtain high purity samples (>90%). Such high purity samples are needed for RNA-seq experiments, so that spurious reads from a contaminating cell population do not skew the resulting analysis.

3.1. Cut&Tag pilot experiments

In general, histone marks may be either well-localised to specific regions of DNA, or more broadly present across large swaths of the genome. Localised marks may be present as small signal peaks spanning a few hundred base pairs, up to wider peaks (also known as broad marks) spanning regions up to a megabase of DNA. In contrast, more diffuse marks may show enrichment over much larger areas of the genome, up to chromosome scale, in the context of marks associated with sex chromosome silencing. We wished to determine whether Cut&Tag methodology was suitable for the detection of peaks at different scales. To test this, we carried out Cut&Tag profiling of gamma-H2AX in two different systems (the DIvA cell line and dissociated testis, see methods sections 2.4.1 and 2.2.1). During the DNA damage response (DDR) gamma-H2AX accumulates in “foci” at the sites of DSBs. These are initially small but rapidly expand to encompass the whole TAD containing the DSB. Thus, in this context we expect to see localised peaks of ~1Mb at DSB sites. In contrast, during meiotic sex chromosome inactivation, gammaH2AX is highly enriched throughout the entirety of the sex chromosomes (232), and thus we expect to see a global enrichment of reads mapping to the X and/or Y chromosomes in testis samples.

3.1.1. Looking at punctate marks using Cut&Tag with DIvA cells

To assess whether Cut&Tag is a suitable technique to detect punctate marks, we used the human DIvA cell line (DSB inducible via AsiSI). This cell line is an experimental system in which DSBs can be induced at defined locations (~ 100 AsiSI restriction sites) throughout the genome upon treatment of the cell line with 4OHT. The cell line expresses an AsiSI restriction enzyme fused to an oestrogen receptor ligand-binding domain. This is activated by 4OHT addition to the cell culture media, which causes relocation of the enzyme from the cytoplasm into the nucleus, leading to DSBs at approximately 100 defined locations in the human genome. GammaH2AX was used in a Cut&Tag assay to tagment DNA in the vicinity of the AsiSI cut sites. In addition to the known cut sites, gammaH2AX will also be present at any other DSBs occurring within the cultured cells. These break sites will be quasi-randomly distributed across the genome, generating a diffuse background signal with the potential for peaks at fragile sites that are frequently broken. Analysis of the background signal outside the vicinity of AsiSI cut sites is beyond the scope of this thesis.

3.1.1.1. Pilot analysis of gammaH2AX localisation in DIvA cells

A pilot experiment was carried out to determine if Cut&Tag libraries could be produced from the DIvA cell line, when following the Cut&Tag kit manufacturer's instructions, for gammaH2AX, with both an induced (4OHT treated) and a control sample. FastQC was used to determine the read quality for each library, the reads were 25bp in length and the number of raw reads obtained is shown in Table 3-1. The samples failed the FastQC per base sequence content flag as shown in Figure 3-1. This is normal for Cut&Tag libraries see (https://yezhengstat.github.io/CUTTag_tutorial/) and does not mean that the library preparation step failed. The pattern observed for the per base sequence content may be caused by the cleavage preference of the Tn5 transposase, which is mainly "dictated by intrinsic parameters, including DNA motif and DNA shape" (282).

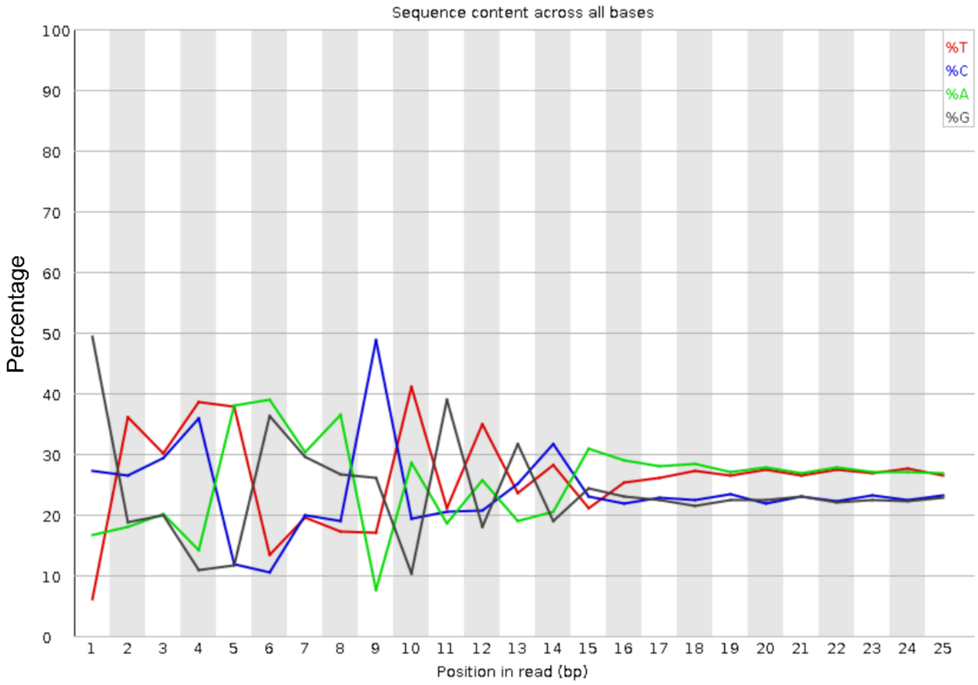


Figure 3-1*:* The typical per base sequence content observed for a Cut&Tag library from an induced DIvA sample immunoprecipitated with gammaH2AX.
The discordant sequence content at the beginning of the reads (which causes the library to fail the per base sequence content FastQC flag) is not common for other library types but is common in Cut&Tag.

Bowtie2 was used to obtain the alignment rate for each of the libraries shown in Table 3-1.

Table 3-1: Bowtie2 alignment summary statistics for the first pilot experiment using DIvA cells and a dissociated testis.

Sample	Number of raw reads	Overall alignment rate	Aligned concordantly exactly one time	aligned concordantly > 1 times	Sequencing depth (fold)
DIvA gammaH2AX control	2,312,344	94.77%	68.20%	26.58%	0.0196
DIvA gamma H2AX treated	2,428,024	94.38%	72.43%	21.95%	0.0206
Dissociated testis gammaH2AX (see section 3.1.2)	5,299,757	96.77%	71.07%	25.69%	0.0486
Dissociated testis H3K27me3 (see section 3.1.2)	6,884,132	93.34%	78.97%	14.37%	0.0631

The overall alignment rate was high (94.77% for the control DIvA sample and 94.38% for the induced sample), so further analysis was carried out. From the total number of raw reads and the read length, the sequencing depth was calculated (Table 3-1). The sequencing depth obtained from a Cut&Tag library will depend on the prevalence of the histone mark that is being immunoprecipitated. The depth obtained for the libraries was low as was expected. The fold coverage of gammaH2AX was slightly higher for the induced DIvA cells, 0.02 compared to 0.019 for the uninduced sample.

To determine the success of the library prep, the mapped fragment size distribution can be plotted. The transposase used in Cut&Tag will specifically tagment (insert sequencing adapters) into accessible chromatin. In a cellular context, this is governed by two factors. Firstly, linked DNA between adjacent nucleosomes is more accessible than nucleosome-bound DNA. This leads to a primary "laddering" pattern of fragment sizes with a periodicity of around 180bp. Secondly, for nucleosome-bound DNA, the DNA surface in contact with the nucleosome has lower accessibility than the surface facing outwards from the nucleosome. This leads to a secondary pattern with a periodicity of ~10bp, often termed 'sawtooth periodicity'. The combination of these two patterns leads to the characteristic library fragment distribution shown in Figure 3-2.
Figure 3-2 from the protocols.io website of the developers of the Cut&Tag protocol (https://www.protocols.io/view/bench-top-cut-amp-tag-kqdg34qdpl25/v3) (281) shows the fragment length distribution at single base pair resolution for histone modifications, which shows the typical 'saw-tooth' pattern of 10bp periodicity, which occurs in a successful Cut&Tag experiments, as well as the tri-modal distribution of peaks which represent peaks approximately the length of one, two and three nucleosomes. Higher order fragments representing 4+

nucleosomes are possible in theory, but are excluded during standard library preparations which aims for fragment sizes < 500bp.

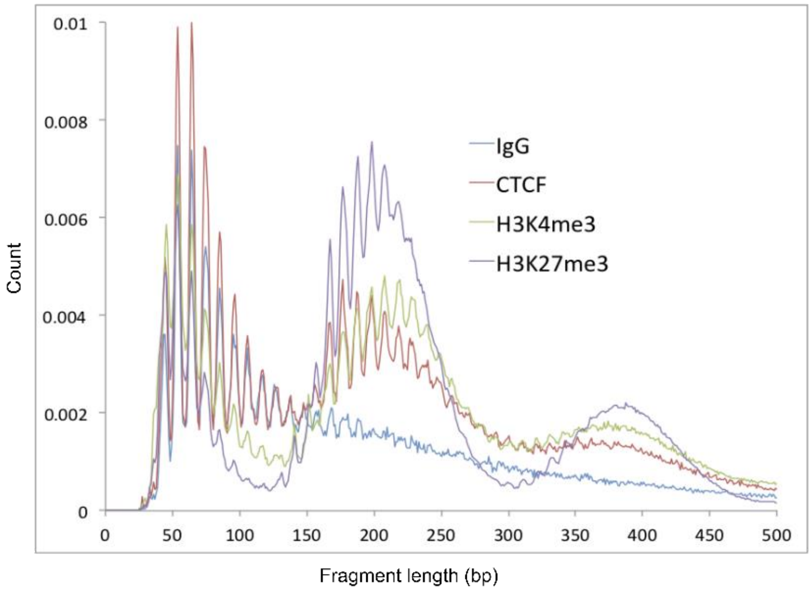


Figure 3-2: An example of the expected Cut&Tag periodicity.
Reproduced from the protocols.io website (https://www.protocols.io/view/bench-top-cut-amp-tag-kqdg34qdpl25/v3) of the developers of the Cut&Tag protocol (281).
CTCF, H3K4me3 and H3K27me3 all show a tri-modal pattern of peaks whereas the IgG control shown in blue does not.

Figure 3-3 shows the periodicity obtained from our test run on DIvA cells.

These library preps were successful as library fragments with the correct periodicity were obtained (the 10bp sawtooth pattern was evident, as were peaks representing the length of a nucleosome), with a similar distribution to that described in the analysis pipeline developed by the Henikoff lab (https://yezhengstat.github.io/CUTTag_tutorial/), see Figure 3-2.

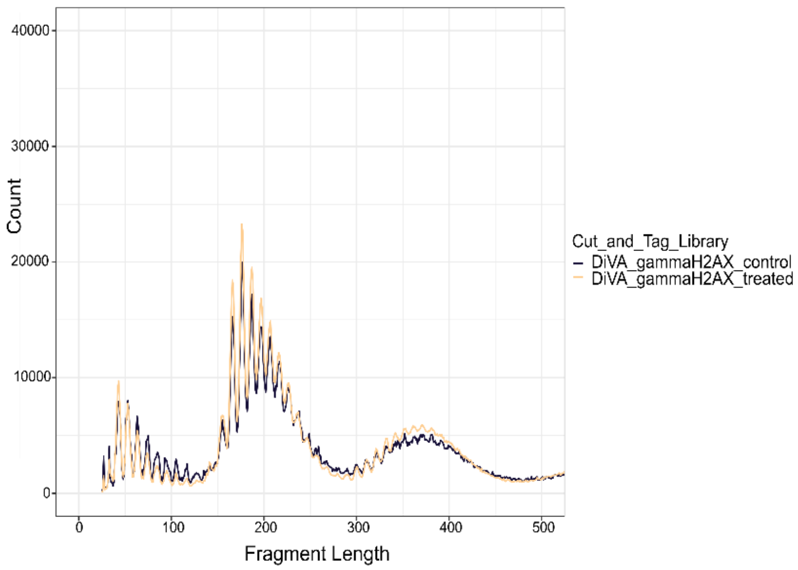


Figure 3-3: Fragment size distribution obtained from DIvA Cut&Tag libraries.
This was a first run 'proof of principle' experiment to determine if libraries with the correct size distribution could be produced from induced and control DIvA cells. The 10bp sawtooth pattern of periodicity was obtained as well as peaks representative of the length of one, two and three nucleosomes. Shown in blue is the uninduced or control DIvA sample and in orange the DSB induced (treated) DIvA sample.

As part of the analysis for the Cut&Tag DIvA cell line, we plotted the gammaH2AX signal around the 80 most cut AsiSI sites identified by Clouaire *et al* 2018 (249).

At a scale corresponding to the typical size of gammaH2AX foci surrounding DSBs, this showed the expected pattern, with ~1Mb regions of gammaH2AX enrichment surrounding each induced DSB.

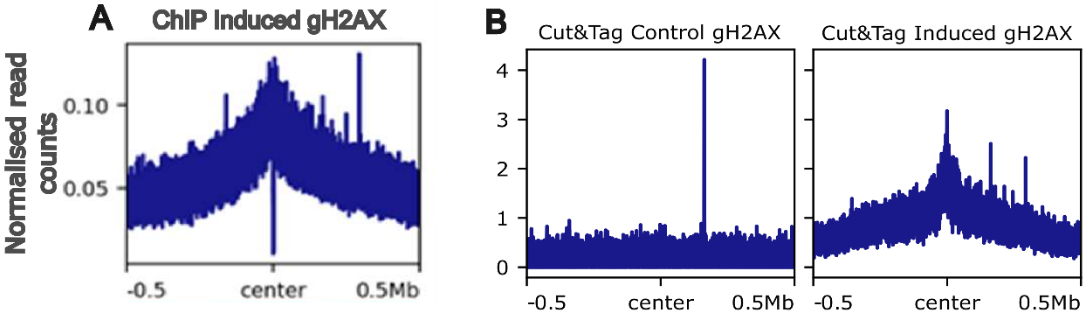


Figure 3-4: ChIP-seq and Cut&Tag plots of gammaH2AX signal around the 80 top AsiSI cut sites within the DIvA cell line.
A) ChIP-seq of gammaH2AXl in induced DIvA cells +/-0.5Mb from the top 80 AsiSI sites. Figure A reproduced with permission from the Thesis of Ane Stranger (raw data originally from Clouaire *et al 2018* (249)*.* **B**) Cut&Tag with gammaH2AX on control and induced DIvA cells +/- 0.5Mb from the top 80 AsiSI cut sites. The narrow peaks of largest signal represent blacklist regions as these were not removed from the input data.

However, at a finer scale there were intriguing deviations from the expected signal pattern. We were expecting a drop in gammaH2AX signal at the break site, as has been previously published by Lacovoni *et al* 2010 (245) for ChIP-seq of gammaH2AX in the DIvA cell line as shown in Figure 3-5A, but we observed the opposite pattern shown in Figure 3-5B. A drop in gammaH2AX at the

DSB sites is expected due to two factors – firstly histones are removed to allow DSB repair factors to access the DNA, and secondly the DNA strands opposing the break are resected to give single-stranded overhangs which are not a substrate for standard adaptor ligation protocols. These two factors mean that sequences immediately adjacent to the break are specifically lost from typical ChIP-Seq libraries (283).

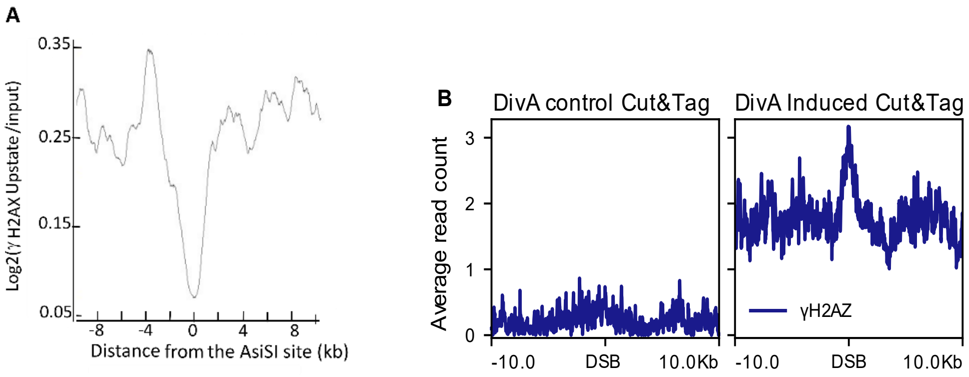


Figure 3-5: GammaH2AX profiles obtained by ChIP-seq and Cut&Tag in the DIvA cells lines.
A) Plot showing the previously published gammaH2AX signal +/- 8kb from the AsiSI site for ChIP-seq in DIvA cells (reproduced from (284)). It shows that γH2AX is depleted in the immediate vicinity (~+/- 2kb) of AsiSI sites.
B) Deeptools plot from my own data showing the gammaH2AX signal at the top 80 most cut AsiSI sites, labelled as 'DSB'. Notably, there is a peak of gammaH2AX signal at the AsiSI DSB sites when using Cut&Tag, in contrast to the dip observed in the published ChIP-Seq experiments.

Consequently, we wished to investigate the possible cause of this unexpected signal. We hypothesised that the increase in signal in the vicinity of the breaks could be due to the Tn5 transposase cleavage preference. It has recently been shown that following strand resection during DSB repair, RNA is synthesized at the break site – thus rather than forming ssDNA overhangs at the DSB site (as previously believed), these regions actually form DNA:RNA heteroduplexes adjacent to the break site (134). While heteroduplex DNA is not a substrate for DNA ligase (explaining the loss of these regions during conventional ChIP-Seq adapter ligation), heteroduplexes can be cleaved by Tn5 transposase (285). Thus, sequences immediately proximal to the DSBs might be detected by Cut&Tag despite being "invisible" to conventional ChIP-Seq. We therefore wished to test if the peak of gammaH2AX signal decreased or disappeared if the DIvA cells were treated with RNase H (that preferentially degrades DNA:RNA hybrids) before antibody addition and tagmentation. The experiments run to investigate this are described below.

3.1.1.2. Understanding the unexpected signal proximal to break sites in DIvA cells

Having observed the unexpected peak of Cut&Tag gammaH2AX signal at AsiSI DSB sites in the DIvA cell line we wished to investigate whether this was caused by the cleavage preferences of the Tn5 transposase, i.e. preferential cleavage of DNA:RNA hybrids or R-loops (285) which might be present at the DSB sites, as Ohle *et al* 2020 (134) have suggested that R-loops form at DSBs to help stabilise them.

We investigated the association of R-loops and DSBs in the human U20S DIvA cell line, in which DSBs can be induced throughout the genome by the addition of 4OHT to the cell culture media (see Cut&Tag methods section). We induced some DIvA cells for DSB and kept a population of DIvA cells as a control. We cryopreserved the cells using the optimised method published on the active motif website (https://www.activemotif.com/catalog/1320/cut-tag-service) and tested the efficacy in Cut&Tag. Cryopreserved cells are expected to perform well in Cut&Tag (if the post thaw viability is high). A caveat here, is that owing to limited sample amounts, we were unable to check the viability of the sample aliquots used for this experiment. Future Cut&Tag experiments will either require additional aliquots to be set aside for the purpose of viability testing, or be conducted on fresh non-cryopreserved cells.

Pre-Cut&Tag we then treated some of the samples with RNase H to try and degrade DNA:RNA hybrids (R-loops) which are the substrate for RNase H. By comparing the gammaH2AX signal pre and post RNase H treatment as well as R-loop signal from the S9.6 antibody in the final Cut&Tag libraries we hoped to be able to determine the contribution of DNA:RNA hybrids to the peak of signal at the DSB sites.

To determine if DSB induction in the DIvA cell line was successful, immunofluorescence (IF) was used to visualise the gammaH2AX foci as well as any signal obtained from the S9.6 R-loop antibody. Figure 3-6 panel F clearly shows an increase in gammaH2AX foci, relative to the control cells, so these samples were carried forward for library preparation.
Library preparation from control and induced DIvA samples with 1e5 cells as input was successful, as libraries with the expected periodicity were obtained (Figure 3-7).

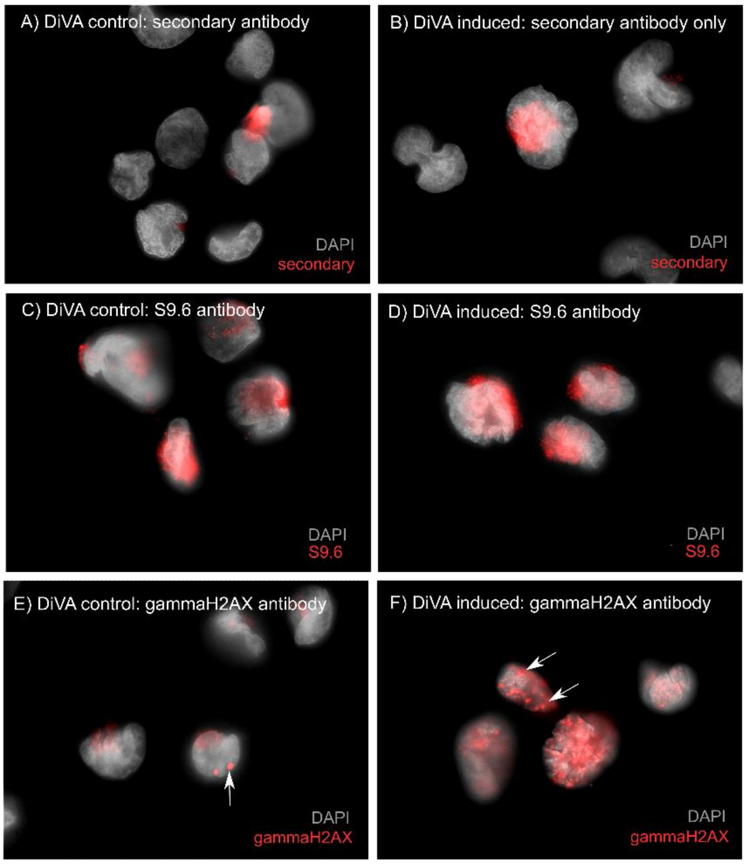

A) DiVA control: secondary antibody
DAPI
secondary
B) DiVA induced: secondary antibody only
DAPI
secondary
C) DiVA control: S9.6 antibody
DAPI
S9.6
D) DiVA induced: S9.6 antibody
DAPI
S9.6
E) DiVA control: gammaH2AX antibody
DAPI
gammaH2AX
F) DiVA induced: gammaH2AX antibody
DAPI
gammaH2AX

Figure 3-6: Immunofluorescence images of DIvA cells from control and induced samples.
Induced samples were treated with 300nM 4-OHT for 4 hours to induce DSBs. Gray represents DAPI staining of the nucleus and red signal from the secondary antibody Goat Anti-Mouse IgG H&L (Texas Red ®) Abcam ab6787 (1:500). Panels **A** & **B** show signal in the absence of a primary antibody. Panels **C** & **D** show control & induced cells with S9.6 antibody (for R-loop detection) as the primary. Panels **E** & **F** show control and induced cells with gammaH2AX antibody as the primary, the white arrows represent examples of gamma H2AX foci. Panel F clearly shows the gamma H2AX foci within the cells (white arrows). All images are x100 magnification.

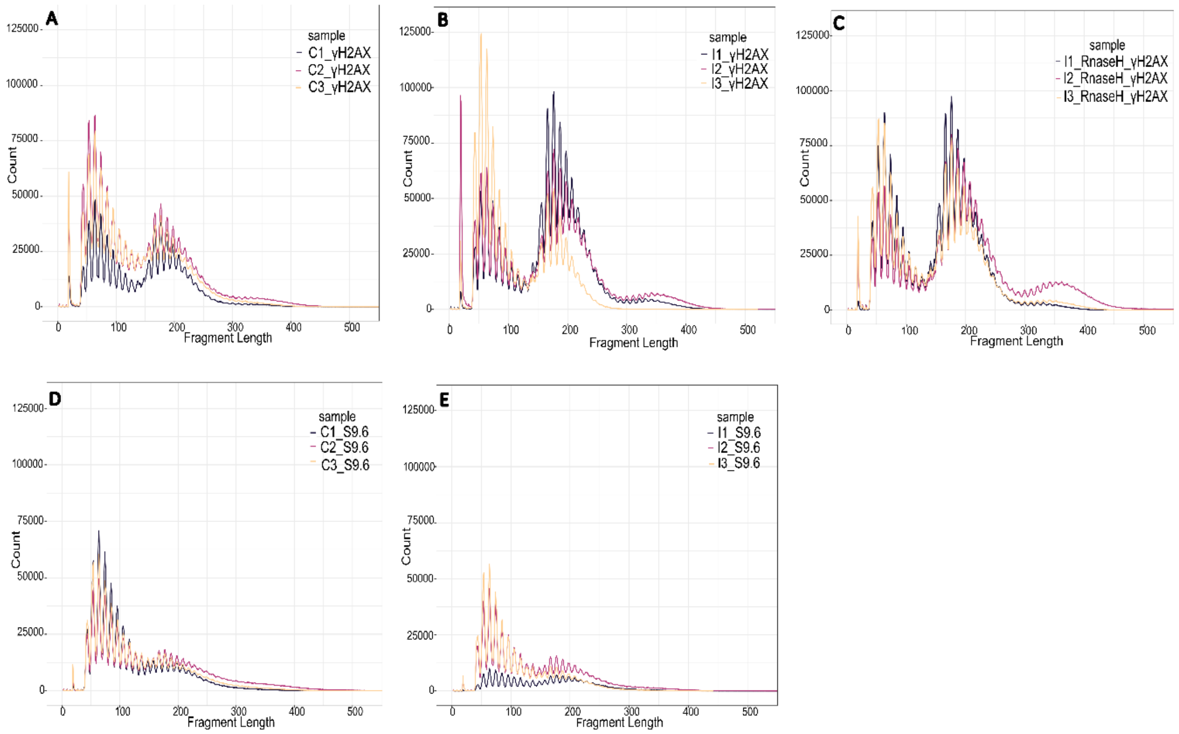


Figure 3-7: Cut&Tag library fragment length from all the technical replicates of control and DSB induced DIvA cells.
A) Periodicity of the three technical replicates of control (C) uninduced DIvA cells (not induced for DSBs), immunoprecipitated with gamma H2AX. **B**) Periodicity of the three technical replicates of induced (I) DIvA cells (induced for DSBs with 4OHT treatment), immunoprecipitated with gamma H2AX. **C**) Periodicity of the three technical replicates of induced (I) DiVA cells (induced for DSBs with 4OHT treatment), treated with RNase H, then immunoprecipitated with gammaH2AX antibody. **D**) Periodicity of the three technical replicates of control (C) uninduced DIvA cells, (not induced for DSBs) immunoprecipitation with the S9.6 antibody targeting R-loops. **E**) Periodicity of the three technical replicates of induced (I) DIvA cells (induced for DSBs with 4OHT treatment), treated with RNase H, then immunoprecipitated with gammaH2AX antibody.

The Cut&Tag data was analysed as described at https://yezhengstat.github.io/CUTTag_tutorial/ (see methods). The raw data obtained per library and the percentage duplication rate per library is shown in Table 3-2. Between 6.03 to 20.02 million reads were obtained per library with 100,000 cryopreserved DIvA cells as input. The % of duplicated reads was low and ranged from 0.64 to 0.85%, indicating that a large number of duplicates were not introduced during the PCR step of library amplification.

Table 3-2: Summary statistics from the DIvA Cut&Tag libraries.

Library	Library concentration (Qubit ng/µl)*	Raw Reads	Raw data (Gb)	Read duplication %
Control_DIvA_rep1_S9.6	0.840	18031754	2.7	0.705
Control_DIvA_rep2_S9.6	1.970	17584126	2.6	0.664
Control_DIvA_rep3_S9.6	1.150	17391962	2.6	0.708
Control_DIvA_rep1_gammaH2AX	0.508	11163946	1.7	0.682
Control_DIvA_rep2_gammaH2AX	0.644	18873428	2.8	0.670
Control_DIvA_rep3_gammaH2AX	0.508	16226214	2.4	0.789
Induced_DIvA_rep1_S9.6	1.160	6032614	0.9	0.638
Induced_DIvA_rep2_S9.6	1.570	19458916	2.9	0.802
Induced_DIvA_rep3_S9.6	0.970	18362292	2.8	0.840
Induced_DIvA_rep1_gammaH2AX	0.774	18092066	2.7	0.636
Induced_DIvA_rep2_gammaH2AX	1.200	18140534	2.7	0.652
Induced_DIvA_rep3_gammaH2AX	0.432	16375970	2.5	0.847
Induced_DIvA_rep1_RNase H_gammaH2AX	0.844	19291490	2.9	0.636
Induced_DIvA_rep2_RNase H_gammaH2AX	0.740	20219862	3.0	0.711
Induced_DIvA_rep3_RNase H_gammaH2AX	0.606	19444128	2.9	0.826

*All libraries were eluted in a volume of 20µl.

Deeptools (248) was used to produce a plot of the read density at the top 80 most cut AsiSI restriction sites as determine by Clouaire *et al* in 2018 (249) (Figure 3-8), with density 1kb upstream and 1kb downstream plotted. This clearly shows an increased relatively narrow peak of gamma H2AX signal in the induced cells as previously obtained in the pilot experiment (Figure 3-5B). Treatment with RNase H did not decrease the gammaH2AX signal and there was not an appreciable signal obtained with the R-loop antibody.

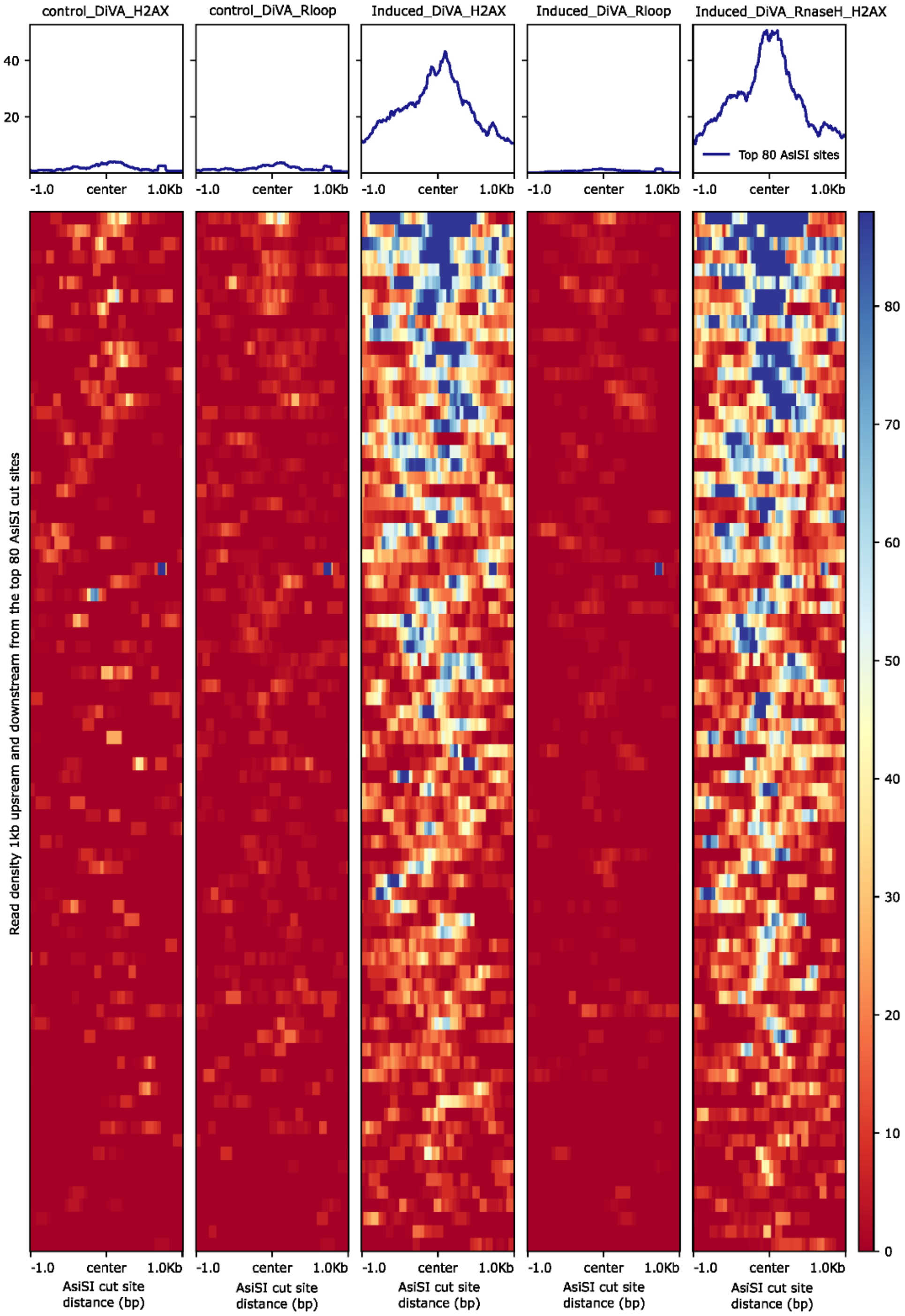


Figure 3-8: Deeptools plot showing the read density for the top 80 AsiSI sites, 1kb upstream and 1kb downstream from the AsiSI site for one technical replicate (replicate 1).

Control_DIvA represents cells NOT induced for DSB formation and DIvA_induced represents cells induced for DSBs formation by the addition of 4OHT to the culture media. For H2AX libraries, DNA was immunoprecipitated with an anti-gammaH2AX antibody. For R-loop libraries DNA was immunoprecipitated with the S9.6 antibody which recognises DNA:RNA hybrids.

From these experiments taken together, we therefore conclude that:

- Cut&Tag is indeed suitable for profiling of localised histone marks such as gammaH2AX and faithfully reproduces most of the known features of gammaH2AX localisation in the DIvA cell line.
- Cut&Tag appears to give increased signal in the immediate vicinity of DSBs undergoing repair. The reasons for this are unclear but do not appear to be related to DNA:RNA heteroduplex formation since RNase H treatment does not eliminate the signal. Moreover, the anti-R-loop antibody gave no signal in the vicinity of the AsiSI sites, indicating that the extent of RNA:DNA heteroduplex formation at these sites under the conditions tested is not detectable. It is likely that any such heteroduplexes are only transiently present during the repair process.

3.1.2. Looking at chromosome-scale marks via Cut&Tag in freshly dissociated versus fixed testis cells

To determine whether Cut&Tag worked on primary cells, a freshly dissociated testis was used to produce two libraries (one prepared with gammaH2AX and the other with H3K27me3). This also allowed us to test the utility of Cut&Tag for detecting extremely broad whole-chromosome enrichment or depletion of specific histone marks.

Specifically, in pachytene spermatocytes undergoing MSCI, gammaH2AX is highly enriched on the sex chromosomes. In other germ cell stages, and on the autosomes at all germ cell stages, this mark is very low abundance as it is specific for DNA damage. Therefore, in a dissociated testis with all germ cell stages present the vast majority of the gammaH2AX will be derived from pachytene cells and will be strongly enriched on the sex chromosomes, as observed by immunofluorescence (233). Conversely, the broad mark H3K27me3 is depleted on the sex chromosomes during MSCI, and this depletion is maintained into round spermatids, as has been shown by Moretti *et al* (2016) (286) using ChIP-seq. Therefore, in a dissociated testis with all germ cell stages present, the vast majority of cells present will show a depletion of this mark on the sex chromosomes. Therefore, by examining read counts for the autosomes and the sex chromosomes we can determine if we obtain the expected pattern.

As per the previous experiment, mapping statistics were obtained from bowtie2 (Table 3-1) and the periodicity of the mapped reads was plotted to determine if the libraries had the correct fragment size (Figure 3-9).

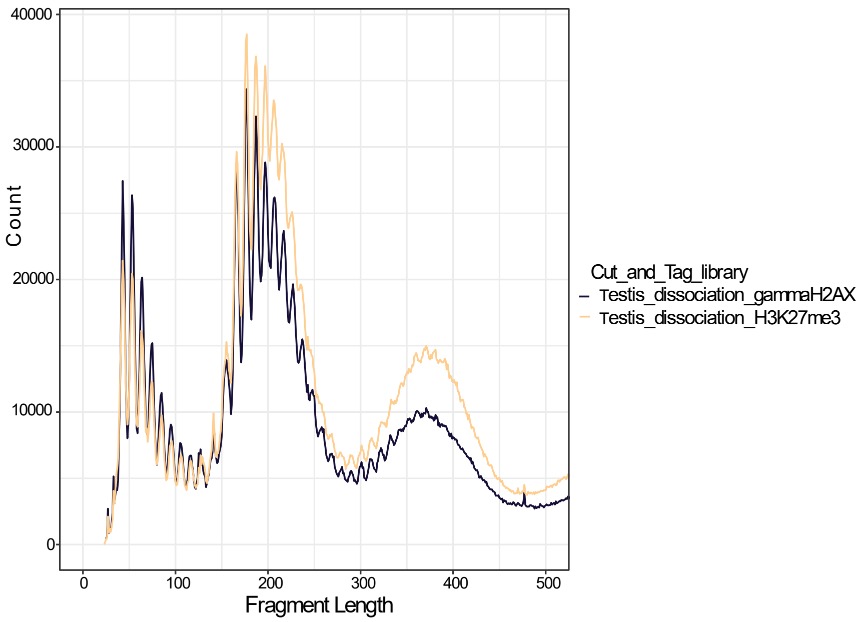


Figure 3-9: Fragment size distribution obtained from dissociated testis Cut&Tag libraries.
The gammaH2AX sample is shown in blue and the H3K27me3 is shown in orange. The expected 10bp sawtooth periodicity was obtained as well as peaks representative of the length of approximately one, two and three nucleosomes.

As per our expectation, we observed a strong chromosome-wide enrichment of gammaH2AX on the sex chromosomes and depletion on the autosomes (Table 3-3). Conversely, we observed H3K27me3 enrichment on the autosomes and depletion on the sex chromosomes (88).

Table 3-3: Cut&Tag reads per kb, on a dissociated testis sample.

Mark	chr X reads/kb	chr Y reads/ kb	autosomal reads/kb	chr X enrichment/depletion (fold) *	chr Y enrichment/ depletion (fold) *
gamma H2AX	6.4	5.2	3.55	3.6	2.9
H3K27me3	0.6	0.4	5.10	-4.2	-6.3

*Reads on the autosomes were divided by two, to count them as haploid when calculating fold enrichment/ fold depletion.

Following the success of the work on freshly dissociated cells, library preparation was attempted from a 1% formaldehyde fixed frozen dissociated testis sample (in duplicate) with H3K27me3. In this experiment, libraries with the correct periodicity were not obtained (Figure 3-10). The libraries had the periodicity pattern obtained for an IgG control like that shown in Figure 3-2. The sawtooth 10bp periodicity was obtained as per Figure 3-2, but the inter-nucleosome periodicity was lost. Since the reason for this change was unclear, we decided not to continue with experiments on fixed cells and that Cut&Tag should only be carried out with fresh or cryopreserved unfixed cells.

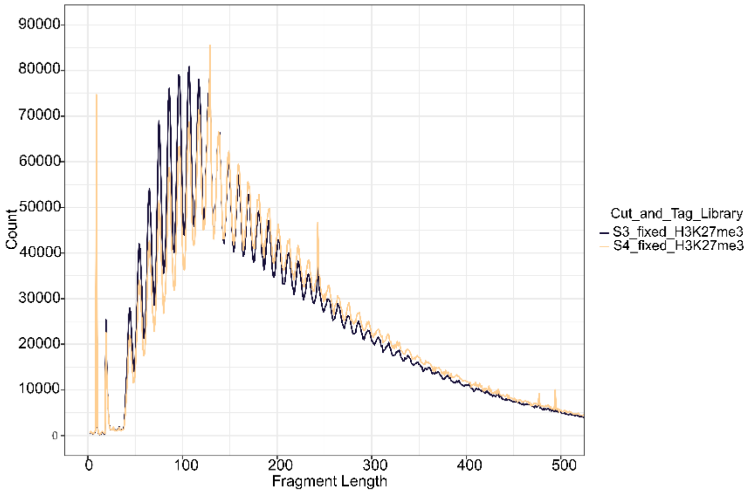


Figure 3-10: Fragment size distribution of formaldehyde fixed testis sample used for library preparation with H3K27me3 antibody.
These samples had been fixed in 1% formaldehyde and frozen at -80C before library prep. S=sample.

Table 3-1 shows the raw reads obtained for the Cut&Tag libraries, for the two DIvA samples tested in the pilot experiment, 2.31 and 2.42 million reads were obtained from ~80,000 cells as input, from which we could resolve inter-nucleosomal fragments.
From the dissociated testis samples, we obtained 5.30 and 6.88 million reads and from this relatively low read number we were able to obtain the expected pattern of enrichment and depletion of gammaH2AX and H3K27me2 on the sex chromosomes. This is in stark contrast to the number of reads required to obtain accurate results from ChIP-seq as the Encode website (https://www.encodeproject.org/chip-seq/histone/#restrictions) states that for narrow marks 20 million reads per sample are required and 45 million reads for broad peak experiments. However, these numbers will vary will depending on the cell type. Lannelli *et al* 2017 (287) carried out ChIP-seq on induced DIvA cells and obtained 37.5 million reads per sample. This was ~7 fold higher than the number of reads required for the gammaH2AX Cut&Tag (~5.3 million). This highlights the substantially reduced sequencing costs of data obtained by Cut&Tag due to the reduced sequencing depth required.

Overall, these pilot experiments trialling Cut&Tag in the DIvA cells line and in a dissociated mouse testis show that Cut&Tag can detect both gammaH2AX foci at DSBs and broad gammaH2AX domains as well as broad H3K27me3 marks in both fresh DIvA cells as well as in a suspension of freshly dissociated testis. However, Cut&Tag is not a viable option for cells fixed in formaldehyde and frozen (as library periodicity is lost (Figure 3-10)). Cryopreserved cells are expected to perform well in Cut&Tag (if the post thaw viability is high) and methods for successful cryopreservation have been published on the active motif website (https://www.activemotif.com/catalog/1320/cut-tag-service). Using this cryopreservation method, we obtained Cut&Tag libraries of the correct periodicity and size, so cryopreserved cells can be used as input for Cut&Tag library preparation.

3.2. Whole genome sequencing to identify haplotype-specific SNPs that distinguish the fused vs unfused chromosomal copies of chromosomes 6 and 16

We first had to establish a breeding colony of Rob6.16 mice before we could sacrifice animals to use for whole genome sequencing.

3.2.1 Breeding Robertsonian mice to use as the model to study non-Mendelian inheritance

We obtained the Rb(6.16)24Bnr strain (#000885) from the Jackson laboratory, which was used to produce an F1 generation of homozygous Rob6.16 pups (see methods). We karyotyped the line to determine it was homozygous for the Rob6.16 fusion (Figure 3-11A). Tail clips from two of these male pups (siblings) were used for whole genome sequencing to obtain the SNPs and INDELs present within this line. The remaining homozygous F1 mice, once sexually mature were mated to wild type C57BL/6 mice to produce heterozygous mice. The presence of one fusion chromosome was confirmed by karyotyping (Figure 3-11B). Once these male mice reached ~22 weeks they were used for spermatid sorting (section 3.3) for RNA-seq analysis.

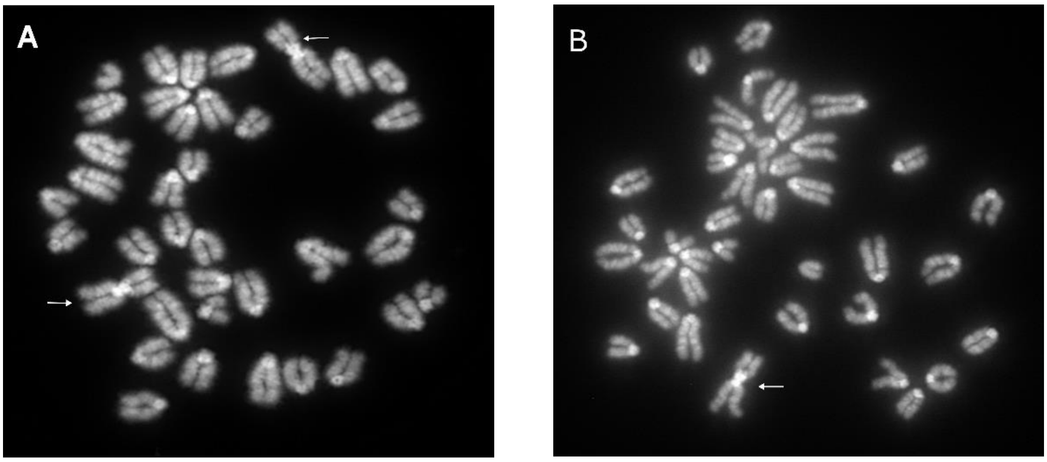


Figure 3-11: Metaphase spreads of homozygous and heterozygous Rob.16 mice.
Shown in A is a homozygous male- the metacentric Rob6.16 fusions are marked with white arrows. Shown in B is a heterozygous male, the metacentric fusion is marked with the white arrow.

3.2.2 Whole genome sequencing statistics for the Robertsonian samples

The whole genome sequencing data (from DNA extracted from two homozygous Rb6.16 24Lub mouse tails) was processed as described in the methods section. Fastp was used for read trimming (see methods) and to obtain the duplication rate. Fastp filtered out 0.04% of the reads for Rob1 and 0.02% for Rob2. The percentage duplication rate was 12.3% for sample 1 and 12.8% for sample 2. *Samtools idxstats* was used to calculate the fold coverage of the mapped reads within the bam files across the whole genome and per chromosome. The fold coverage for the whole genome for both the Robertsonian samples was 27X. Fold coverage per chromosome of the bam files Rob1 and Rob2 is shown in Table 3-4. This shows that the fold coverage for chr6 was slightly higher than the fold coverage for chr16. Chr6 had coverage of 27.7 and 27.5-fold (for Rob1 and Rob2) and chr16 had coverage of 25.8 and 25.7-fold (Rob1 and Rob2). *Samtools flagstat* was used to calculate the % of mapped reads per sample (Table 3-5), the % of properly paired reads was high at ~97%. *Samtools coverage* was used to calculate the number of reads obtained per chromosome for both the Rob1 and Rob2 samples (Supplementary Table *8-2*). For the Rob1 sample this was 28,076,914 for chr6 and 17,172,180 for chr16. The % coverage per chromosome was also calculated, for the Rob1 sample this was 97.3% for chr6 and 96.4% for chr16 (Supplementary Table 8-2).

Table 3-4: Fold coverage per chromosome for the whole genome sequencing data for the Robertsonian6.16 samples Rob1 and Rob2 calculated using *samtools idxstats.*

	Fold coverage	
Chr	**Rob1 bam file**	**Rob2 bam file**
chr1	26.7	26.6
chr2	46.2	45.2
chr3	26.0	25.8
chr4	26.6	26.6
chr5	26.6	26.6
chr6	27.7	27.5
chr7	26.6	26.7
chr8	26.3	26.3
chr9	48.0	46.9
chr10	26.4	26.3
chr11	26.4	26.5
chr12	27.8	27.6
chr13	26.5	26.4
chr14	26.8	26.7
chr15	25.9	25.8
chr16	25.8	25.7
chr17	26.3	26.3
chr18	25.7	25.6
chr19	25.4	25.5
chrX	14.1	14.2
chrY	9.2	9.7

Table 3-5: *Samtools flagstat* results from the WGS of Rob1 and Rob2 samples.

Samtools flagstat filters	Rob1	Rob2
Total read number (QC passed & QC failed)	520,124,555	516,596,458
% mapped	99.21	99.30
% properly paired	97.32	97.59
% singletons	0.46	0.38

Having processed the WGS data through the GATK pipeline and SnpEff (see methods) to obtain the SNPs and INDELs present within the homozygous Rob6.16 mouse, the INDELs can be used to differentiate between Rob6.16 homozygous, Rob6.16 heterozygous and wild type C57BL/6 mice. Concordant INDELs present within the homozygous Rob6.16 sample were identified as part of the SNP pipeline and were separated into a file of INDELs for chr6 and a file of INDELs for chr16. There were 33,088 concordant INDELs (annotated and unannotated) on chromosome 6 and 16,233 on chromosome 16 for the two homozygous Rob6.16 mice sent for WGS.

By designing PCR primers that span the INDELs present in the Rob6.16 homozygous mouse line, two PCR products will be produced for heterozygous mice and one band for homozygous and wild type mice. The wild type and homozygous band will differ in size, the size being dependant on whether the primers spanned an insertion or a deletion.

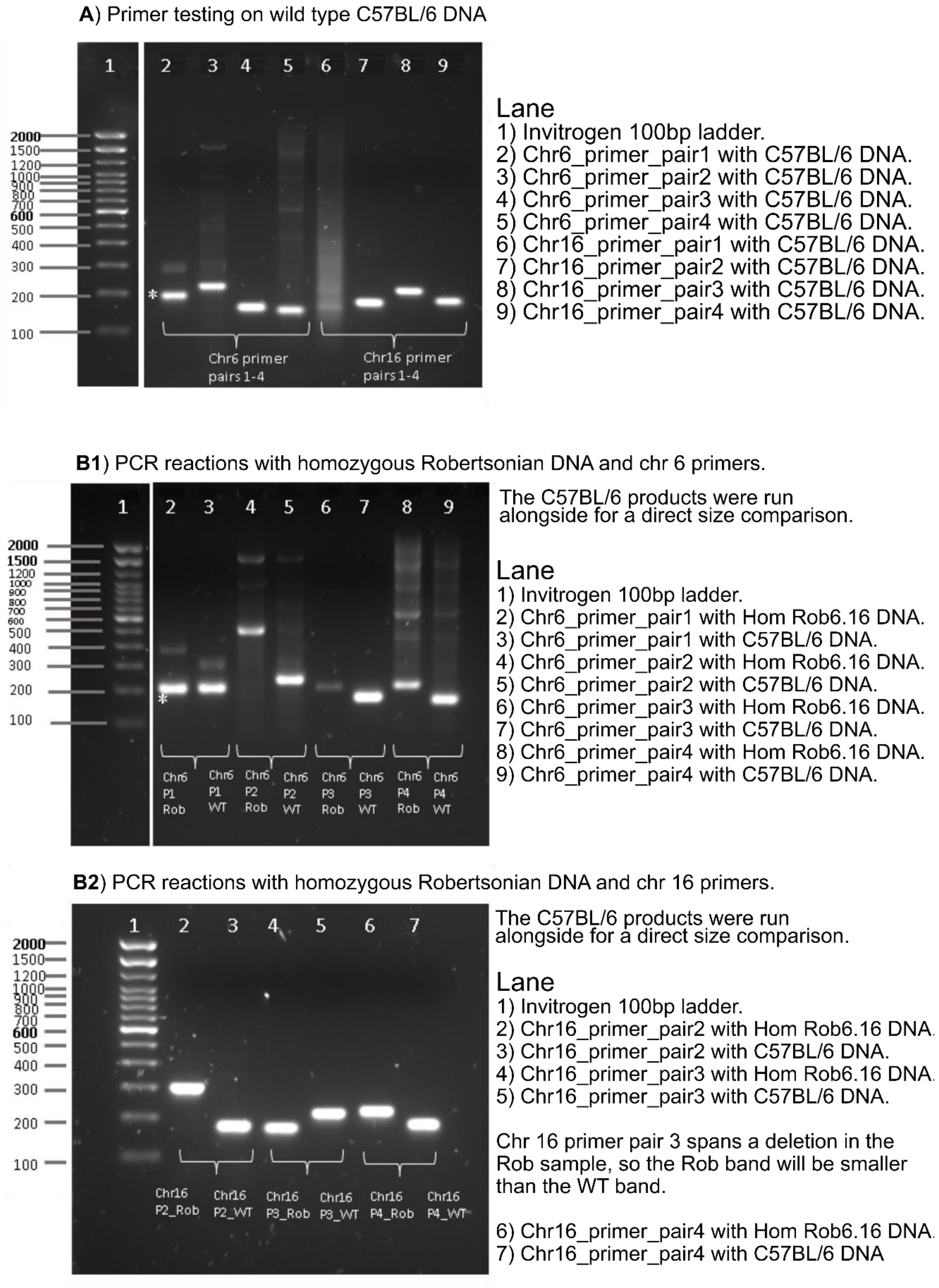
3.2.3 Genotyping PCR design for the Robertsonian mice

INDELs of approximately >20bp insertion or deletion were chosen to be able resolve bands of different sizes on a 2% agarose gel. The INDELs chosen were as close to the fusion point (the start of chr6 and chr16) as possible. PCR primers that spanned INDELs in the homozygous Rob6.16 mouse line were designed as described in section 2.1.4 and they were tested in several rounds of PCR.

The first round of testing was carried out on wild type DNA from C57BL/6 mice (Figure 3-12A), to determine the specificity of the primers. Table 3-6 shows the expected product sizes.

Table 3-6: Expected product sizes of the PCR reactions:

Primer	Product size WT (bp)	Product size Rob (bp)	Comments
chr6_1	286	377	Insertion in Rob (will also give a non-specific product of 195bp)
chr6_2	224	284	Insertion in Rob
chr6_3	160	185	Insertion in Rob
chr6_4	153	204	Insertion in Rob
chr16_1	163	198	Insertion in Rob
chr16_2	179	308	Insertion in Rob
chr16_3	220	175	Deletion in Rob
chr16_4	184	224	Insertion in Rob



A) Primer testing on wild type C57BL/6 DNA

Lane
1) Invitrogen 100bp ladder.
2) Chr6_primer_pair1 with C57BL/6 DNA.
3) Chr6_primer_pair2 with C57BL/6 DNA.
4) Chr6_primer_pair3 with C57BL/6 DNA.
5) Chr6_primer_pair4 with C57BL/6 DNA.
6) Chr16_primer_pair1 with C57BL/6 DNA.
7) Chr16_primer_pair2 with C57BL/6 DNA.
8) Chr16_primer_pair3 with C57BL/6 DNA.
9) Chr16_primer_pair4 with C57BL/6 DNA.

B1) PCR reactions with homozygous Robertsonian DNA and chr 6 primers.

The C57BL/6 products were run alongside for a direct size comparison.

Lane
1) Invitrogen 100bp ladder.
2) Chr6_primer_pair1 with Hom Rob6.16 DNA.
3) Chr6_primer_pair1 with C57BL/6 DNA.
4) Chr6_primer_pair2 with Hom Rob6.16 DNA.
5) Chr6_primer_pair2 with C57BL/6 DNA.
6) Chr6_primer_pair3 with Hom Rob6.16 DNA.
7) Chr6_primer_pair3 with C57BL/6 DNA.
8) Chr6_primer_pair4 with Hom Rob6.16 DNA.
9) Chr6_primer_pair4 with C57BL/6 DNA.

B2) PCR reactions with homozygous Robertsonian DNA and chr 16 primers.

The C57BL/6 products were run alongside for a direct size comparison.

Lane
1) Invitrogen 100bp ladder.
2) Chr16_primer_pair2 with Hom Rob6.16 DNA.
3) Chr16_primer_pair2 with C57BL/6 DNA.
4) Chr16_primer_pair3 with Hom Rob6.16 DNA.
5) Chr16_primer_pair3 with C57BL/6 DNA.

Chr 16 primer pair 3 spans a deletion in the Rob sample, so the Rob band will be smaller than the WT band.

6) Chr16_primer_pair4 with Hom Rob6.16 DNA.
7) Chr16_primer_pair4 with C57BL/6 DNA

Figure 3-12: Primer optimisation on Wild type C57BL/6 DNA (panel A) and Rob6.16 DNA, panels B1/B2. WT=C57BL/6 DNA and Rob=Rb6.16 24Lub DNA. (*= non-specific product of 195bp)

Figure 3-12A shows that Chr6 primer pairs 1-4 tested with WT C57BL/6 DNA all gave bands at the correct size. Chr6 primer pair 1 also gave a known non-specific band at 195bp (marked with the

*). Chr16 primer pair 1 on the C57BL/6 DNA gave a smear without a strong specific band, so it was not tested further.

The second round of testing was with Robertsonian DNA and all the primer pairs, except chr16 primer pair 1 which was not tested, for the reason outlined above. The results are shown in Figure 3-12 B1/B2.

Chr6 primer pair 1 was not further tested due to the high intensity of the non-specific band.

A sample of wild type, homozygous Robertsonian and mixed wild type/Robertsonian DNA (Het equivalent, as Het mice were not yet available at the time of PCR optimisation) was tested blind to determine whether from the pattern of bands in PCR reactions, chr6 primer pairs 2,3,4 and chr16 primer pair 2,3,4 could be used to determine the genotype (Figure 3-13).

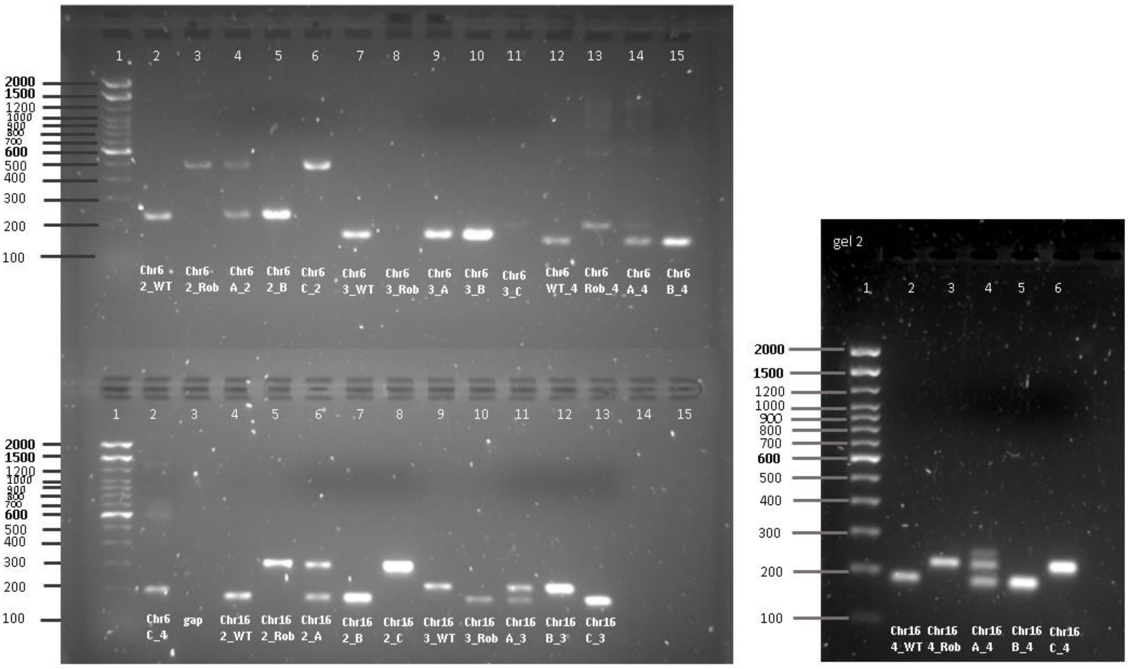


Figure 3-13: Blind genotyping of a wild type, homozygous Robertsonian chr6.16 sample and a mixed wild type/homozygous chr6.16 Robertsonian sample.
10µl of a 50µl PCR reaction was run on the gel for the wild type and homozygous chr6.16 Robertsonian PCR reactions to compare the product size to the unknown genotyped samples A-C. The figure shows that chr6 primer pair 2 can clearly identify A as the Het sample, B as the wild type and C as the homozygous chr6.16 Robertsonian sample, although the Robertsonian product is larger than expected at ~500bp. Lanes 7-11 contain the PCR reactions for chr6 primer pair3. When loading 10µl of a 50µl PCR reaction, the band of the homozygous chr6.16 Robertsonian sample is not visible in lane 8. A single feint band larger than the wild type band is observed in lane 11 (when loading 20µl of the PCR reaction), this also identified sample C as the homozygous chr6.16 Robertsonian sample. Lanes 12-15 and comb 2 lane 2 show chr6 primer pair 4. There are 2 band obtained for chr6 A_4, indicating that this is the heterozygous sample, although the band for the Robertsonian sample is quite feint. Lanes 4-15 gel 1 comb 2 and gel 2 show genotyping reactions for chr16. The Robertsonian bands for these chromosomes are more intense and therefore easier to identify, than the chr6 genotyping reaction.

For chr6 primer pairs 3 and 4, the Robertsonian band was hard to identify when loading 20µl of a 50µl PCR reaction and running on a 2% gel. The bands will be easier to identify if a larger volume of the PCR product is loaded.
Primer pair 2 for chr6 unexpectedly detected an insertion of ~300bp in the Robertsonian fusion, although only a 60bp insertion was expected. This was consistent across multiple technical repeats during primer validation, and also across multiple individual Rob fusion animals genotyped with these primers. Inspection of the raw data file showed no obvious reason for this discrepancy. We speculate that there may be a tandem repeat insertion in the Rob line which is collapsed in the WGS comparison to WT. Alternatively there may be secondary structures in the region that impairs accurate sequencing of the region. We did not further characterise this 300bp insertion by Sanger sequencing, as the purpose of the experiment was simply to identify differences between the haplotypes that could be genotyped.

Chr16 primer pairs 2-4 give more consistent bands than Chr6 primer pairs 2-4. However, bands are still visible for the Robertsonian genotype for the chr6 primer pairs, especially if more PCR product was loaded per well. Therefore, chr6 primer pairs should be used in conjunction with the chr16 primer pairs for a high confidence genotype call. Genotyping the SNPs within the ROI of the fused chr6/16 provides an alternative to karyotype analysis, which is slower and requires more input material, such as a whole spleen as opposed to an ear clip (live animals) or tail snip (following culling) for DNA extraction and genotyping by PCR. The primers can be used to confirm the genotype of the mice sent for RNA-sequencing.

3.3. Fluorescent-activated cell sorting of round spermatids

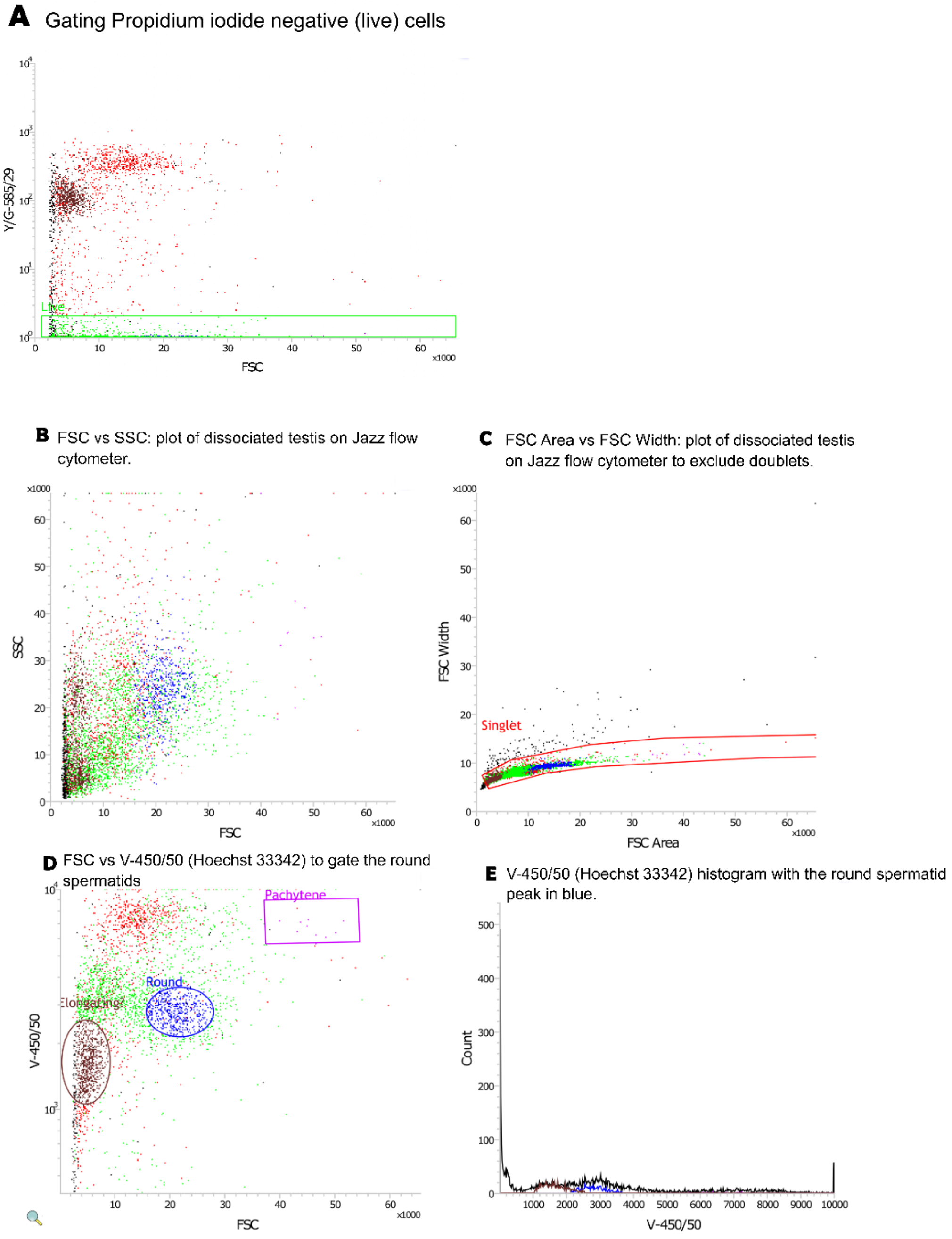
Round spermatids used for RNA-Seq analysis (chapter 5) were sorted at the Francis Crick Institute with the assistance of Dr Valdone Maciulyte from Dr Turner's research group, and the FCI flow sorting (FCI) core. BD Aria and Fusion FACS machines were used at the Crick Institute (London) to sort wild type, heterozygous and homozygous Rob6.16 mice. These machines are equipped with a UV laser and can simultaneously detect red and blue fluorescence from the Hoechst 33342 dye (see Supplementary Figure 8-11). This enables the use of standardised protocols (288) which give high purity sorts. The purity of the flow sorted Robertsonian 6.16 homozygous samples was confirmed by immunofluorescence staining with PNA-lectin and DAPI. Purity was 99.5% round and elongating spermatids, with approximately 20% elongating spermatids per sample (see Table 3-7). The viable (PI negative) cells post dissociation ranged from 44.6-51.7% for the three homozygous Rob6.16 samples sorted, 51-60.4% for the three wild type samples sorted and 19.9-55.2% for the three heterozygous samples sorted.

Table 3-7: Immunofluorescence classification of the cell counts obtained from the WT, Het and Hom Rob6.16 samples sorted at the Crick in London on Aria and Fusion BD FACS machines.

Sample	Number of Round spermatids	Number of Elongating spermatids	Other	% round & elongating spermatids
WT_1	185	15	-	100%
WT_2	188	11	1	99.5%
WT_3	176	22	2	99.0%
Het_1	181	14	5	97.5%
Het_2	176	26	3	98.5%
Het_3	157	20	3	98.5%
Hom_1*	43	7	-	100%
Hom_2*	156	40	4	98%
Hom_3*	32	12	-	100%

*Fewer cells were collected for the immunofluorescence purity check, so instead of 200 cells as many cells as possible were counted per slide. 'Other' refers to cells not identified as round or elongating spermatids often pre-leptotene or pachytene cells.

In addition, we wished to establish our own flow sorting protocol at Kent to facilitate follow-up experiments. This type of work requires round spermatids of high purity for RNA-seq and/or ChIP-Seq analysis, to prevent reads from contaminating cell types skewing the results. The FACS Jazz flow cytometer at the University of Kent is *not* equipped for simultaneous detection of red/blue signal from Hoechst dye staining, instead the separate detector channels are coupled to separate laser pinholes. This means that side population analysis with Hoechst blue and Hoechst red signal cannot be carried out as is common in other protocols (289). Flow sorting of wild type round spermatids at the University of Kent was therefore carried out using a modified sorting strategy. The dissociated testis was stained with Hoechst 33342 and propidium iodide (PI). PI staining was used to distinguish live cells from dead cells, while round spermatids were identified using a combination of Hoechst staining in the blue channel (V450/50, indicating ploidy) and forward scatter signal (indicating approximate cell size). The dissociated testis was prepared as described in the methods section 2.2. The Jazz settings were optimised daily to obtain the optimal sorting rate and cell recovery, such as flow rate and drop delay. See Figure 3-14 for the typical dot and density plots obtained. Per mouse aliquots of 200,000 round spermatids were sorted into 1.5ml tubes. A total yield of ~1.5e6 round spermatids was obtained per mouse. Post sorting an aliquot of the cell suspension was manually counted to check cell recovery. For the same sort, recovery ranged from 60-89% depending on the tube counted. The purity of the sorted samples was checked with PNA-AF488 and DAPI staining and then manual cell counting using an immunofluorescence microscope. Purity ranged from 90-93% round spermatids per tube for the same sort settings and animal (Table 3-7). See Figure 3-15 for the typical immunofluorescence cell classification categories. The high purity obtained means that this protocol is suitable for collecting cells to carry out RNA-seq or Cut&Tag experiments.



A Gating Propidium iodide negative (live) cells

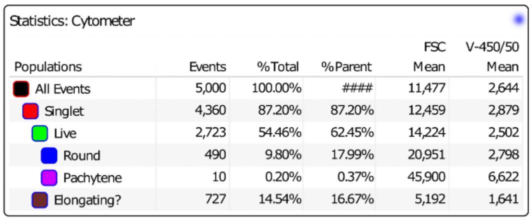
B FSC vs SSC: plot of dissociated testis on Jazz flow cytometer.

C FSC Area vs FSC Width: plot of dissociated testis on Jazz flow cytometer to exclude doublets.

D FSC vs V-450/50 (Hoechst 33342) to gate the round spermatids

E V-450/50 (Hoechst 33342) histogram with the round spermatid peak in blue.

Jazz cytometer: dissociated testis



Statistics: Cytometer

Populations	Events	%Total	%Parent	FSC Mean	V-450/50 Mean
■ All Events	5,000	100.00%	####	11,477	2,644
🟥 Singlet	4,360	87.20%	87.20%	12,459	2,879
🟩 Live	2,723	54.46%	62.45%	14,224	2,502
🟦 Round	490	9.80%	17.99%	20,951	2,798
🟪 Pachytene	10	0.20%	0.37%	45,900	6,622
🟫 Elongating?	727	14.54%	16.67%	5,192	1,641

Figure 3-14: Typical gating strategy on the FACS Jazz flow cytometer for a dissociated mouse testis.

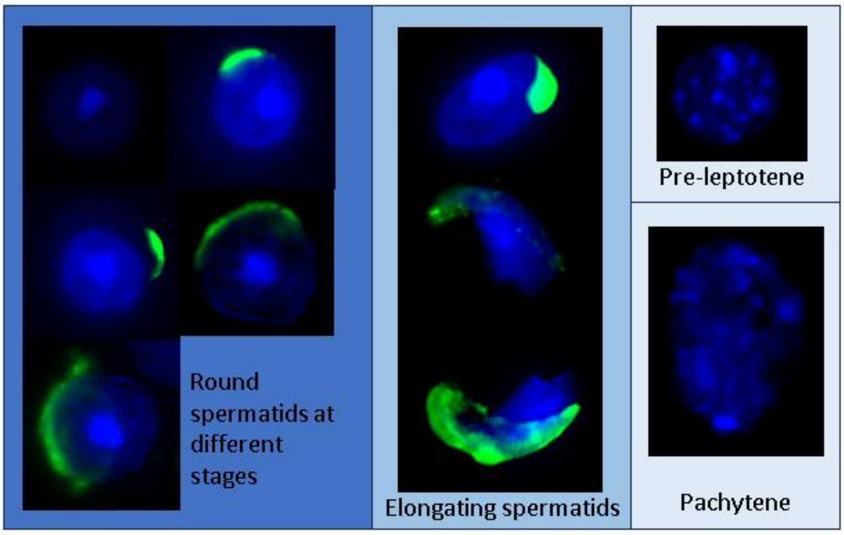


Figure 3-15: Representative images showing round, elongating, pre-leptotene and pachytene cells stained with PNA-AF-488 (2µg/ml) and DAPI (Vectashield 1.5µ/ml).
The nucleus is shown in blue (stained with DAPI) and PNA-AF488 staining is shown in green (representing the acrosome). Round spermatids are identified by the presence of an acrosome and a single DAPI-dense chromocenter, while elongating spermatids are additionally identifiable by their characteristic asymmetric morphology. Other cell types typically have symmetrical nuclei and lack acrosomes. Different germ cell stages are distinguished by size and chromatin distribution.

Table 3-8: Round spermatid immunofluorescence purity check from a cell sort on the BD Jazz flow cytometer.

Slide	Number of Round Spermatids	Number of Elongating spermatids	Number of Pachytene cells	Number of Pre-leptotene cells	other	Round spermatid purity
RS_1	154	26	1	14	5	90%
RS_2	158	26	0	13	3	92%
RS_3	156	23	0	4	17	90%
RS_4	156	27	1	15	6	92%
RS_5	93	38	1	4	1	96%
RS_6	166	20	0	10	4	93%

Where possible 200 cells were counted per slide. Slides RS_1 to RS_6 represent different aliquots of cells collected from the same sample. For representative cell images of the different categories see Figure 3-15.

3.4. Discussion

The overall goals for this part of the project were to (1) validate Cut&Tag for the study of both punctate and chromosome-wide epigenetic marks, (2) identify specific sequence variants present in the Rob fusion model to facilitate genotyping of animals and allow analysis of allele-specific gene expression, and (3) ensure we had working protocols to sort round spermatids for analysis. We were able to achieve all these goals.

In terms of Cut&Tag validation, we chose this method for epigenetic analysis because Cut&Tag will only solubilize the DNA that has been tagmented by the transposase, and this is what will be sequenced. This gives a lower background signal than ChIP-sequencing. The cell input requirements are also much lower with 50-500,000 cells required for Cut&Tag whereas between 300,000 to many millions of cells can be required for ChIP-seq depending on the cell type and mark immunoprecipitated. We showed that Cut&Tag detects punctate and whole-chromosome enrichment for gammaH2AX in DIvA cells and whole dissociated testis respectively, reflecting the presence of specific DSBs in DIvA cells and MSCI in pachytene spermatocytes. Therefore, the method will also be applicable to flow sorted spermatids, or pachytene cells. This will enable epigenetic studies to be carried out in conjunction with RNA-sequencing or whole genome sequencing on the same sample.

Unexpectedly, we detected increased signal in the vicinity of the DSBs in the DIvA cell work, and conducted a follow-up experiment to determine if this was due to Tn5 tagmentation of DSB-associated RNA/DNA heteroduplexes. We did not see a decrease in the gammaH2AX signal at the DSB site in the sample treated with RNase H, indicating that either RNase H is not effective at cleaving the R-loops in this system, (Active motif have investigated the combination of RNase H and RNase A instead, personal communication) or that the peak of gammaH2AX at the cut sites is not due to the presence of R-loops. The increased gamma H2AX signal at the cut site is however unlikely to be an artifact as it was observed in two biological replicates. Instead, this peak may be due to some other undefined factor, such as the cleavage preferences of the Tn5 transposase. Alternatively, it is possible that the peak of gammaH2AX signal at the break sites that we observed could be specific to the gammaH2AX antibody that we have used (Millipore 05-636, clone JBW301). A Cut&Tag library preparation with an alternative clone of gamma-H2AX antibody would determine if the increased signal at the DSB sites is still observed. We also did not observe a strong increase in read depth at the top 80 AsiSI sites in the control compared to the DSB induced DIvA cells with the R-loop antibody (S9.6) in the Cut&Tag experiments. This could be due to several factors: 1) The S9.6 antibody is not able to bind the R-loops within the DIvA cells, due to the DNA conformation. 2) R-loops are only present at relatively low percentage cut sites (e.g. if these are only transiently formed) - not enough to see a large increase in read density. 3) R-loops are not present at the cut sites. To further investigate the presence of R-loops in the DIvA cell line, additional experiments would have to be carried out to determine the optimal conditions for RNase activity within the constraints of the Cut&Tag buffer of the primary antibody incubation step. Or a separate RNase digestion could be carried out before starting the Cut&Tag protocol to enable optimal enzymatic conditions. Despite the experimental conditions tested not being optimal to detect R-loops this method has potential to be used to investigate other epigenetic marks surrounding the DSBs. Since carrying out the Cut&Tag experiments to try and detect R-loops, using the standard anti-mouse Active Motif kit, Active Motif have brought out a kit specifically designed to detect R-loops (catalogue number 53167). This kit contains optimized DNA Binding Buffer and modified DNA purification columns, suggesting that the binding conditions of the original kit for R-loop detection were sub-optimal.

In terms of genetic characterisation of the Rob fusion model, we successfully obtained and analysed almost 30x genome coverage for two separate homozygous individuals from the Rob6.16 line. This enabled robust identification of sequence variants present in this line, and the development of simplified genotyping protocols to trace inheritance of the fusion chromosome. Further characterisation of the genetic landscape of this model is presented in Chapter 5, section 5.1, and RNA-Seq work from sorted round spermatids in Chapter 5 section 5.3.

In terms of spermatid flow sorting from dissociated mouse testis, we were able to both successfully reproduce an existing flow sorting protocol using Hoechst 33342 side population analysis (red and blue signal) obtained with a UV laser on the FACS Aria and develop our own protocol using Hoechst 33342 signal on the violet laser of the FACS Jazz. Propidium iodide staining was used to exclude dead cells in both types of sorts. Both Hoechst 33342 and propidium iodide are relatively cheap reagents, amenable to staining other cell types. More than 1 million round spermatids could be obtained per mouse, enough for RNA-sequencing analysis, epigenetic studies (such as Cut&Tag) and a purity estimate. The purity obtained sorting with the UV laser on a FACS Aria was higher (98-100% round spermatids) compared to ~93% when using the violet laser alone (FACS Jazz). To optimise the recovery of the round spermatids we found that the collection tubes should contain some FBS (to prevent spermatids sticking to the sides of the tubes) and it was essential to optimise the drop delay of the FACS to achieve the optimal purity and recovery. Without optimal drop delay the purity of the sorted samples was below 75% (on the Jazz) and the yield was very low. When sorting the testis cell suspension from multiple animals, we staggered the dissociation steps. This ensured that the cells were as fresh as possible for the start of the sort, this reduced the time that the samples were sitting on ice and helped to improve the % viability of the sample.

Instead of flow sorting, round spermatids could be eluted by centrifugal elutriation, where separation by FACS is not possible. Centrifugal elutriation equipment was not available at the University of Kent. Centrifugal elutriation does not rely on cell staining but separates cells according to their sedimentation velocity. Prolonged staining with propidium iodide can lead to cell death, this is not required with centrifugal elutriation. Without the need for Hoechst staining, cells purified by centrifugal elutriation are amenable for use in immunofluorescence studies or other assays where it would be disadvantageous to use Hoechst-stained cells. Aliquots of round spermatids can be cryopreserved post sorting and used for epigenetic studies such as Cut&Tag, or other studies where intact viable cells are required. Ideally the cryopreservation media and cell density per vial would be optimised to obtain the maximum recovery of viable cells post thaw. This was beyond the scope of this thesis, however in pilot experiments I have shown that Cut&Tag can be carried out with cryopreserved DIvA cells.

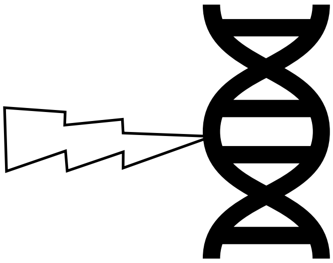
In summary: in this chapter I have validated an epigenetics profiling method for use in studying chromatin from separated germ cells, identified the sequence polymorphisms that will underpin the analysis, and validated the cell sorting methods required to prepare the samples.

4. Where does DNA damage occur during mouse spermatogenesis and what are the evolutionary consequences of this damage?

This chapter reports the results from the following papers:

Burden, F., Ellis, P.J.I., Farré, M., A shared 'vulnerability code' underpins varying sources of DNA damage throughout paternal germline transmission in mouse, *Nucleic Acids Research*, Volume 51, Issue 5, 21 March 2023, Pages 2319–2332, https://doi.org/10.1093/nar/gkad089

Álvarez-González, L., Burden, F., Doddamani, D., Malinverni, R., Leach, E., Marín-García, C., Marín-Gual, L., Gubern, A., Vara, C., Paytuví-Gallart, A., Buschbeck, M., Ellis, P. J. I., Farré, M., & Ruiz-Herrera, A. (2022). 3D chromatin remodelling in the germ line modulates genome evolutionary plasticity. Nature communications, 13(1), 2608. https://doi.org/10.1038/s41467-022-30296-6 (Joint first author with L.Álvarez-González and D.Doddamani)



Studying the structural changes that have occurred in related mammalian genomes is an important area of evolutionary biology which helps to unravel the genomic basis of speciation. Comparative genomics has shown that large genomic regions are maintained syntenic with the same order of loci in several species, while these regions are demarcated by breaks of synteny, so called evolutionary breakpoint regions (EBRs) (39, 40, 290, 291) (see Section 1.3.3). These EBRs are caused by chromosomal rearrangements (CRs) and are not randomly distributed in the genome, instead they tend to cluster in certain locations. Theoretical work suggests that CRs must originate in the germ line to be passed onto the next generation and occur in regions accessible in germ cells and/or early totipotent developmental stages (44). As such, heritable chromosomal reorganisations can occur before meiosis (in proliferating primordial germ cells, spermatogonia and oogonia), during meiotic division (in spermatocytes and oocytes) or in post-meiotic stages (i.e., round spermatids), highlighting the existence of a constraining role of EBRs in the germ line that needs further investigation.

For a chromosome rearrangement to take place, a break in DNA and an incorrect repair must occur. Possible mechanisms of DNA breakage and repair during spermatogenesis can occur through several different sources (i) formation and repair through homologous recombination (HR) and non-allelic homologous recombination (NAHR) of meiotically programmed double-strand breaks (DSBs) catalysed by SPO11 during early prophase I (i.e., primary spermatocytes in leptotene and pachytene stages) (292), (ii) formation and repair through non-homologous DNA end joining (NHEJ) or microhomology-mediated end joining (MMEJ) of DSBs generated in later stages of spermatogenesis (i.e., round spermatids) (84, 293), and (iii) zygotic repair of SSBs (single strand breaks) and DSBs generated by oxidative damage in mature sperm (294). However, the exact relation of EBRs and DSBs in the germline is still unknown. Moreover, there is a fine balance between chromatin remodelling, architectural proteins and cell-specific gene expression during spermatogenesis (103, 225, 295, 296) that can affect the potential outcomes of genetic damage in the germ line. It is not known, however, which of these sources contribute most to the formation of transmissible evolutionary chromosomal reorganisations.

In this chapter we study how DNA and chromatin change during male gametogenesis, and how this is related to evolutionary chromosome rearrangements. In mammals the production of fertilisation-competent sperm involves a drastic reduction in nuclear size—over 75% reduction in cell volume (79), to produce streamlined and hydrodynamic cells capable of fast independent motility. This dramatic loss of cytoplasmic content is accompanied by an equally drastic transformation of chromatin structure and organization (297). During the elongating stage of spermiogenesis, the genome is remodelled via the replacement of histones by protamines, resulting in a highly compact sperm nucleus and consequently the sperm head (Figure 1-16). While not all histone proteins are replaced—estimates for retention range from 1% to 10% in different species, fully protaminated chromatin packing is extremely space-efficient and approaches the theoretical crystal limit for DNA condensation (298). Protamine bound DNA is less supercoiled than histone bound DNA because wider supercoils are produced (82, 299). Consequently, the remodelling process requires significant changes in the DNA winding number to eliminate the free negative supercoils produced during histone replacement. This is believed to be enzymatically mediated by topoisomerase II beta (TOP2B) (or similar enzymes (88, 300, 301). These enzymes catalyse the scission of DNA strands to allow free rotation of the helix and/or unknotting of tangled strands. It is estimated that this chromatin remodelling requires between 5 and 10 million transient double-strand breaks (DSBs) per cell (88). If these transient breaks are not correctly re-ligated by the enzyme generating them, this leads to damage that must be repaired by one of several DNA DSB repair processes (145) (302) (See section 1.5.2). Any unrepaired breaks remaining in the mature sperm have the potential to affect the next generation, if not repaired by the oocyte. For example, increased sperm DNA fragmentation index (DFI) is associated with miscarriage (303). As such, determining the context in which DSBs occur within mouse spermatids (post meiotic cells) is a pivotal question both for understanding evolutionary transformations in genome structure and also for delineating the vulnerability of different regions of the sperm genome to clinically significant DNA damage.

Previous work in this field (84) has shown that spermatid DSBs are not randomly distributed, but rather are associated with particular categories of genomic repeats (LINE, satellite and simple repeats). Further work looking at mature sperm has shown that genomic repeats, in particular SINEs, are also enriched for oxidative damage occurring during epididymal maturation, and that this form of damage is also correlated with histone retention in sperm, and with 3D localisation in the sperm nucleus (141). It is as yet unclear whether there is a common ‘vulnerability code’ in which particular sequence or topological features underlie the susceptibility of sperm chromatin to multiple different types of damage at successive developmental stages.

The aims of this chapter are:

1) To investigate the epigenetic context of EBRs and their association to DSBs in the male germline.

2) To understand the genetic and epigenetic factors associated with the distribution of DSBs in the male germline, and their relation to gametogenesis and early embryonic stages.

4.1 Identifying EBRs in the mouse lineage

The EBR data that I have used in this thesis was produced by D. Doddamani as part of the paper “3D chromatin remodelling in the germ line modulates genome evolutionary plasticity”. Briefly, DESCHRAMBLER (304) was used to determine evolutionary rearrangements of rodent genomes by using 14 Rodentia chromosome-level genome assemblies and two outgroups (human and rabbit) (Figure 4-1). First, pairwise alignments using mouse as a reference genome were run with LASTZ and reconstructed ancestral chromosome fragments (RACFs) were generated by DESCHRAMBLER.

The EBRs in each lineage were identified within the mouse genome. EBR locations were determined as the regions between RACFs, with the smallest size shared across all lineages considered as the breakpoint. EBRs were then phylogenetically classified depending on the lineage in which they occurred as: 1) ancestral if they occurred before the Myodonta species; 2) recent, if they occurred between the Myodonta ancestor and Muridae ancestors; or 3) mouse-specific, if they happened after the split of mouse and rat. We identified 134 recent EBRs, 44 ancestral EBRs and 54 mouse specific EBRs. This gave a total of 232 EBRs with a median size of 21,502bp (305).

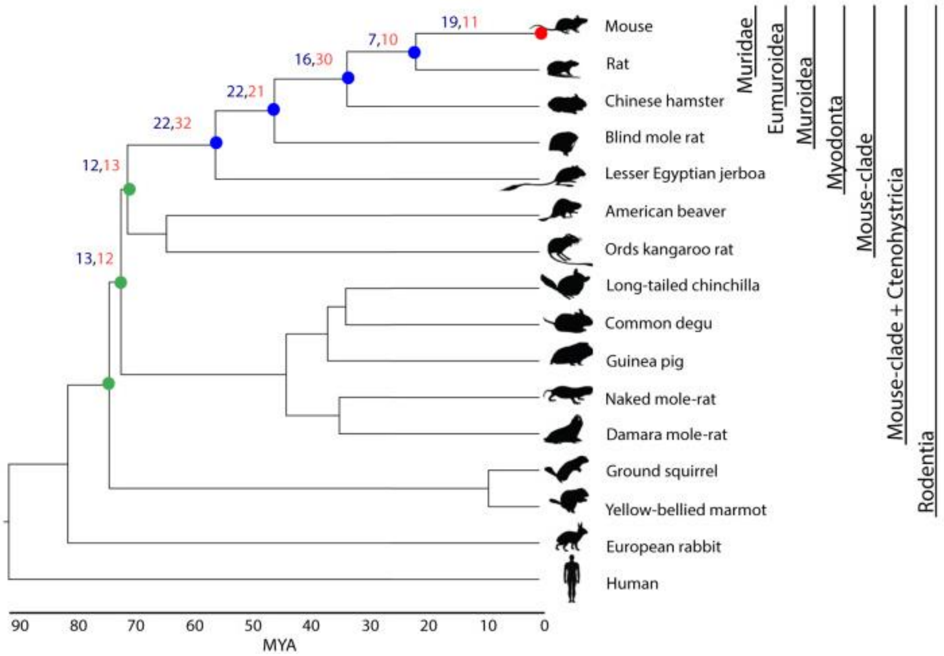


Figure 4-1: Phylogenetic tree of the rodent and outgroup species included in the DESCHRAMBLER analysis. The chromosomal rearrangements between the ancestors are shown for each node: in blue the number of inversions, in red the number of inter-chromosomal rearrangements. Green coloured dots denote ancestral lineages; blue dots represent recent ancestors (Muridae, Eumoroidea, Muroidea and Myodonta, respectively), the red dot depicts the mouse specific. Figure produced by D. Doddamani, taken from the paper “3D chromatin remodelling in the germ line modulates genome evolutionary plasticity” (305).

4.2 Chromatin state dynamics during the transition through spermatogenesis

To investigate the relationship between EBRs and epigenetic context in mouse germ cells, we first analysed the landscape of the higher-order chromatin organisation during spermatogenesis. As spermatogenesis progresses, three main cell types in different developmental stages can be easily isolated using FACS. Here, I focus on adult germinal stem cells (AGSC) or spermatogonia (pre-meiotic cells), spermatocytes (meiotic cells) and spermatids (post-meiotic cells) (Figure 4-3A and Table 2-2) as described in Alverez-Gonzalez *et al*, 2022 (305). Epigenetic data for the three cell types was obtained from one study Hammoud *et al*, 2014 (252) to avoid any methodological bias. These three cell types represent different developmental stages of spermatogenesis and thus allowed me to determine how the chromatin state differed through spermatogenesis. The ChromHMM tool (264) (see methods section 2.9) was used to bioinformatically divide the genome into different states or regions depending on the intensity of the histone marks in the input files. Three histone marks were analysed, two marks of active chromatin (H3K4me3 and H3K27ac), and one repressive (H3K27me3) (Figure 4-2). ChromHMM was run with the concatenated option, which integrated the three histone marks from spermatogonia, spermatocytes and spermatids to produce a single set of states. The percentage of the genome covered by each histone varied slightly between the cell types, H3K4me3 ranged from 1.3 to 4.6% in spermatogonia and round spermatids, H3K27me3 from 1.7% to 3.8% while H3K27ac from 1.1% to 1.3% of the mouse genome, respectively. Using ChromHMM, I defined eight chromatin states in each cell type (Figure 4-2 and Table 4-1). The states were named according to the intensity of the marks present in each state. For example, state E2 with a high concentration of the active mark H3K27ac was termed an active state, whereas state E4 with a high concentration of the repressive chromatin mark H3K27me3 was termed a repressed state. The background state (states E1, E3 and E7), was so called because of low coverage of all three histone marks.

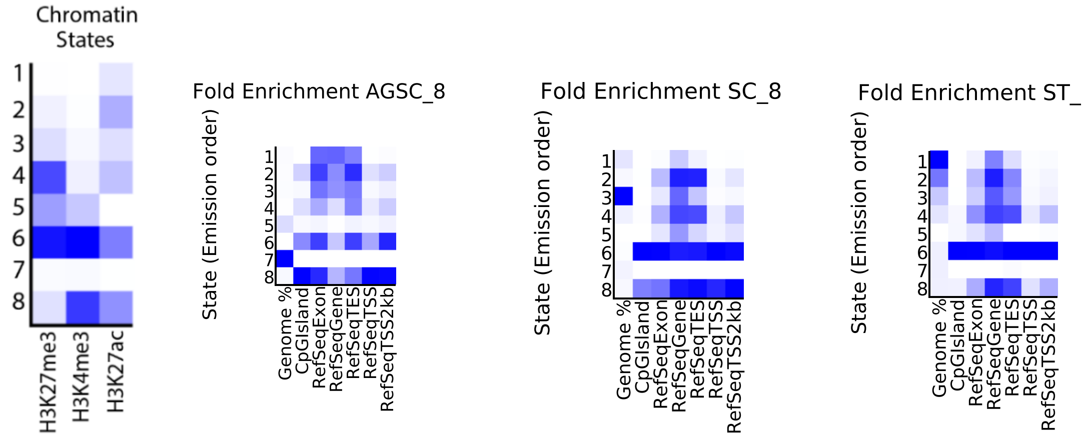


Figure 4-2: ChromHMM overlap plots for the spermatogonia, spermatocyte and spermatid data run with H3K27me3, H3K4me3 and H3k27ac.

The higher the intensity of blue, the higher the intensity of the histone mark.

ChromHMM analysis was run using the concatenated model, learned with 8 states, showing that states 6 and 8 (with high coverage of H3K4me3) are enriched at CpG islands in AGSC and SC. In ST state 6 is also enriched at CpG islands but state 8 is enriched at a lower level than in AGSC and SC.

AGSC= Adult germinal stem cells, SC=spermatocytes, ST=spermatids

Table 4-1: ChromHMM emission state E1-E8 statistics.

State	Spermatogonia			Spermatocytes			Spermatids		
	% Cov	Max (bp)	Average (bp)	% Cov	Max (bp)	Average (bp)	% Cov	Max (bp)	Average (bp)
E1 Background	1.37	72,200	6,537	8.01	234,200	11,895	48.97	281,800	11,910
E2 Active	1.10	35,200	2,129	2.23	61,200	3,282	26.39	121,000	4,876
E3 Background	1.18	62,000	4,560	78.57	769,600	24,808	10.62	120,600	6,208
E4 Repressed	0.31	28,000	2,186	4.44	40,200	1,632	5.11	95,200	1,968
E5 Poised	11.48	23,200	611	0.05	4,400	534	0.01	3,400	550
E6 Trivalent	0.07	4,000	669	0.88	11,400	992	2.94	38,600	1,802
E7 Background	82.30	3,103,000	4,216	3.67	3,106,800	20,900	3.05	3,109,400	50,753
E8 Poised	2.19	16,600	1,827	2.15	23,400	1,882	2.91	20,400	1,379

Cov=coverage. The states are those shown in Figure 4-2. The colours of the states refer to those used in Figure 4-3 and are used to differentiate the different state types.

Calculating the genomic coverage of each state in each cell type showed that the dominant state consists of low histone marks (Table 4-1). This was state E7 in spermatogonia, E3 in spermatocytes and E1 in spermatids, with 82.30, 78.57 and 48.97% coverage, respectively. We then classified the states accordingly to the histone marks, as: E1 background, E2 active (enriched in H3K27ac associated to active enhancers), E3 background (with low coverage of all histone marks), E4 repressed (enriched in the repressive chromatin mark H3K27me3), E5 poised (with both an active and repressive chromatin mark), E6 trivalent (due to a strong signal from all three histone marks), E7 background (low coverage of all histone marks) and E8 poised (with a different combination of histone marks to the posed state of E5). The active state E2 (enriched in H3K27ac), increased from a coverage of 1.1% of the genome in spermatogonia to 26.4% in round spermatids (Figure 4-3 B). State E4, however, was dominated by the repressive chromatin mark H3K27me3, spanning between 0.31% and 5.11% of the genome in spermatogonia and round spermatids, respectively. As for poised chromatin (states E5 and E8 with both H3K27me3 and H3K4me3 marks) it covered from 13.7% in spermatogonia to 2.92% in round spermatids; while state E6 (labelled as trivalent chromatin showing all three histone marks) covered 0.07% to 2.94% in spermatogonia and spermatids (Figure 4-3 B). As all three chromatin states E1, E3 and E7 had low coverage of histone marks they were classified as background state, which we termed E0 (Figure 4-2).

To assess the dynamics of chromatin state transitions throughout spermatogenesis I then compared the transition of chromatin states from spermatogonia to spermatocytes and then to round spermatids for a given genomic region (Figure 4-3C). This allowed the investigation of which regions of the mouse genome change epigenetic status during gametogenesis. A total of 192 combinations of cell type and chromatin state were identified, with 34 combinations covering overall >98% of the genome with at least 0.1% each. All the other combinations were discarded for subsequent analysis. Because of the interesting nature of the trivalent state, although it represented less than 0.1% of the genome, we included it in the following analysis (Table 4-2).

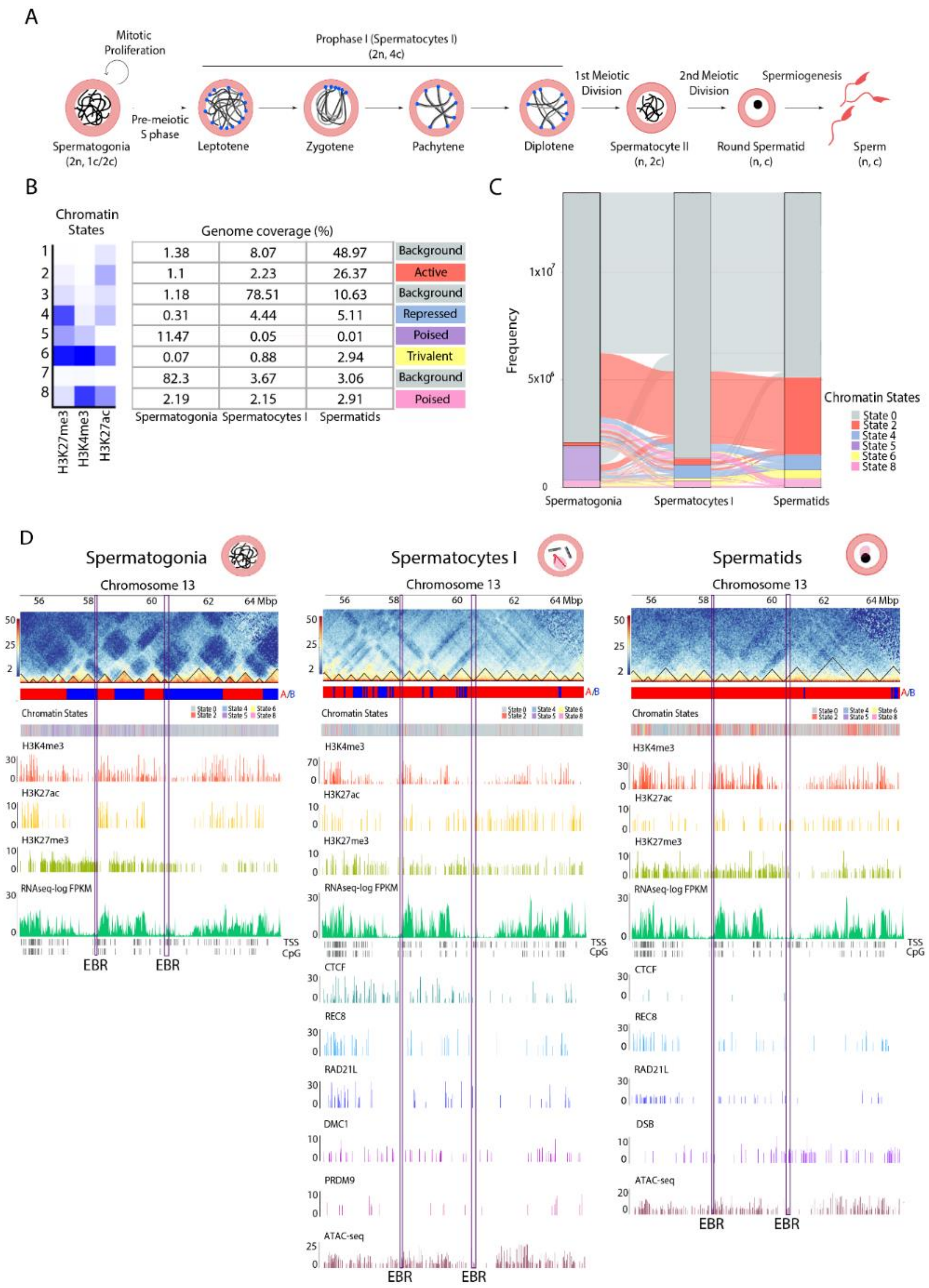


Figure 4-3: Epigenetic landscape dynamics during mouse spermatogenesis.
A) Schematic representation of mouse spermatogenesis. Adapted from (44, 45). Diploid (2n) and haploid (n) numbers are indicated for each cell type as well as the number of chromatids per chromosome (4c, 2c, or c). **B**) ChromHMM chromatin states based on marks H3K27me3, H3K4me3 and H3K27ac. Numbers in the table indicate the percentage of genome coverage for chromatin states in the three cell types analysed (spermatogonia, primary spermatocytes and round spermatids). A total of 6 major chromatin states were found, including background (states 1, 3 and 7; grey), active (state 2; red), repressed (state 4; blue), poised (states 5; purple and 8; pink) and trivalent (state 6; yellow). **C**) Alluvial plots representing chromatin state transitions from spermatogonia to primary spermatocytes and round spermatids. Chromatin states 1, 3 and 7 from panel (**B**) were merged into state 0 (background). **D**) Chromosome 13 region-specific heatmaps at 50 kbp resolution (from 55 Mbp to 65 Mbp), for all three cell types depicting compartment signal (**A**, **B**), chromatin states, H3K4me3, H3K27ac, H3K27me3, RNA-seq (represented as log FPKM), CTCF and cohesin peaks (REC8 and RAD21L) and ATAC-seq. The genomic locations of EBRs are displayed (salmon highlight) in each cell type. Abbreviations – EBRs evolutionary breakpoint regions, FPKM fragments per kilobase of transcript per million fragments mapped. Figure reproduced from (305) .

Table 4-2: Epigenetic transitions during male gametogenesis.
Three-cell type state statistics for the 34 states with >0.1% coverage and the trivalent state E6-E6-E6. Mean region size, median size and max length of a given transition in the mouse genome.

state	Mean (bp)	Median (bp)	Max (bp)	min (bp)	% coverage
E0-E0-E0	3,390	1,800	3,102,800	200	54.45
E0-E0-E2	2,424	1,400	71,400	200	21.86
E5-E0-E0	508	400	9,800	200	5.99
E5-E0-E2	490	400	6,600	200	2.24
other	467	400	23,000	200	1.98
E0-E0-E4	810	600	17,600	200	1.81
E0-E0-E8	886	600	11,400	200	1.49
E0-E4-E4	723	600	11,800	200	1.08
E8-E8-E6	1,654	1,400	14,400	200	1.01
E5-E4-E4	754	600	10,400	200	0.96
E0-E2-E2	1,990	1,200	24,000	200	0.93
E0-E2-E0	1,660	1,200	20,600	200	0.75
E0-E4-E0	723	600	9,600	200	0.70
E5-E0-E4	522	400	7,400	200	0.65
E0-E4-E2	766	600	9,000	200	0.49
E5-E4-E0	606	400	6,400	200	0.36
E2-E0-E2	1,501	1,000	12,800	200	0.31
E5-E0-E8	476	400	3,800	200	0.29
E8-E6-E6	939	800	6,800	200	0.28
E5-E6-E6	589	400	6,600	200	0.21
E0-E8-E6	557	400	4,800	200	0.21
E8-E8-E8	535	400	7,000	200	0.20
E0-E0-E6	467	400	3,400	200	0.19
E0-E8-E8	509	400	3,800	200	0.19
E8-E0-E8	523	400	5,400	200	0.18
E2-E0-E8	997	800	9,200	200	0.15
E5-E4-E6	438	400	4,000	200	0.15
E5-E4-E2	510	400	4,800	200	0.14
E0-E6-E6	473	400	3,200	200	0.14
E2-E8-E6	819	600	6,400	200	0.13
E8-E0-E2	492	400	4,200	200	0.12
E4-E4-E4	1,623	1,000	12,800	200	0.12
E0-E4-E6	395	400	3,000	200	0.11
E5-E8-E6	556	400	4,400	200	0.10
E6-E6-E6	523	400	2,800	200	0.03

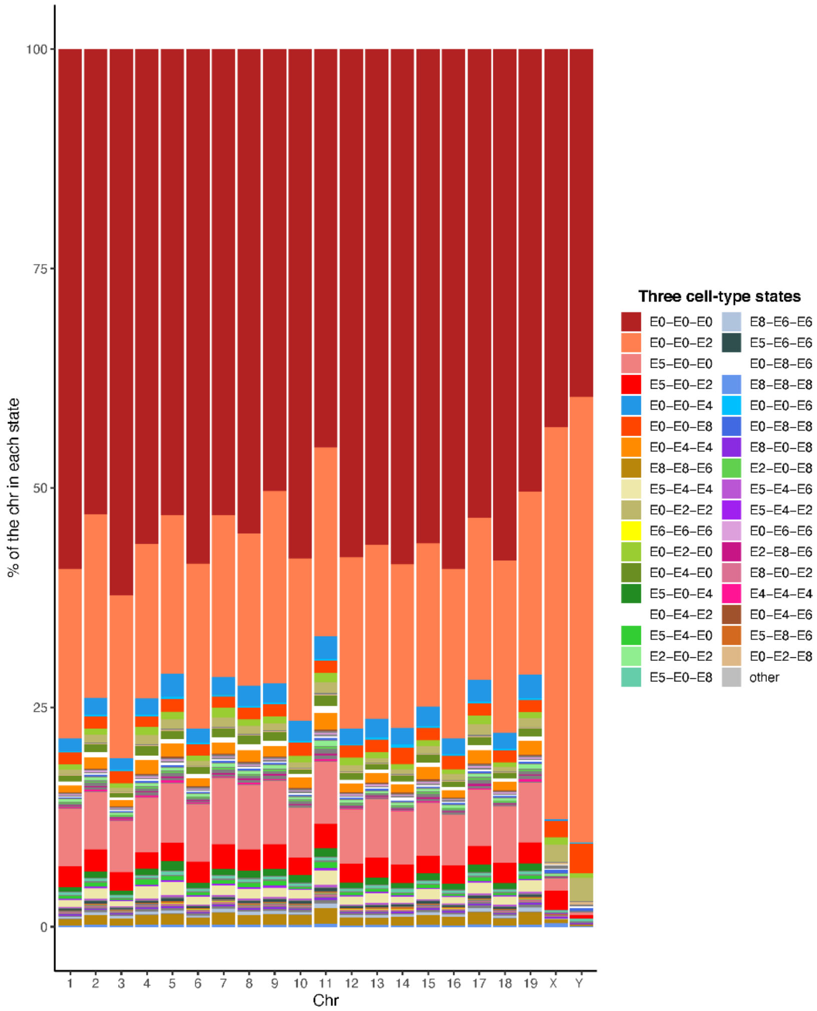


Figure 4-4: Distribution of the 3-cell type states per chromosome.
The autosomes have a different pattern of coverage compared to the sex chromosomes (chr X and chr Y).

When comparing the percentage of each mouse chromosome covered for each of the 34 most frequent chromatin state combinations, we detected that the autosomes have different coverage of the three-cell type states than the sex chromosomes, with state E0-E0-E0 having the largest coverage ranging from 45.43% on chromosome 11 to 62.24% on chromosome 3 (Figure 4-4). For the sex chromosomes, state E0-E0-E2 had the largest coverage (44.71% on chromosome X and 50.76% on chromosome Y). State E5-E0-E0 was lower on the sex chromosomes (X:1.43% and Y:0.31%) compared to an average of 6.51% on the autosomes. State E5-E0-E2 decreased on chromosome Y with 0.4% coverage compared to an average of 2.3% on the autosomes. Instead, state E0-E0-E8 increased on chromosome Y at 3.44% compared to an average of 1.4% on the autosomes.

Most of the genome (54.45%) remained in the same background chromatin state (E0-E0-E0) throughout spermatogenesis. This contrasted with the small proportion of the genome that is maintained active (0.084% in E2-E2-E2), poised (0.20%, E8-E8-E8 and E5-E5-E5 < 0.001%), trivalent (0.029% E6-E6-E6) or repressed (0.12%, E4-E4-E4) in all three cell types. Remarkably, 41.09% of the genome changed chromatin state during spermatogenesis, with E0-E0-E2 being the most common transition (21.9% coverage), followed by E5-E0-E0 (6% coverage). During spermatogenesis, 25.8% of the genome became active, whereas only 4.49% and 1.94% transitioned to repressed or poised states (Figure 4-3C). In contrast, a total of 2.56% of the mouse genome became trivalent in spermatids. As expected, both X and Y chromosomes are enriched in ‘closed’ chromatin states (Figure 4-4) as they are subjected to meiotic sex chromosome inactivation (MSCI) during prophase I and post-meiotic sex chromatin (PMSC) in round spermatids (103, 306).

To investigate labile chromatin landscapes, we analysed the gene content of those three cell type states associated with EBRs (states E6 and E8 in spermatids). A total of 10,925 unique protein-coding genes were present in E6 and E8 regions in spermatids (data obtained using BioMart (279)). Gene ontology (GO) enrichment analysis (≥1.5-fold enrichment and FDR < 0.05) (with the Panther db (268)) identified GO terms related to protein localisation to cell junction and protein dephosphorylation (1.79 and 1.55-fold) as well as dendrite development and regulation of organelle assembly (1.57 and 1.51-fold) (Figure 4-5).

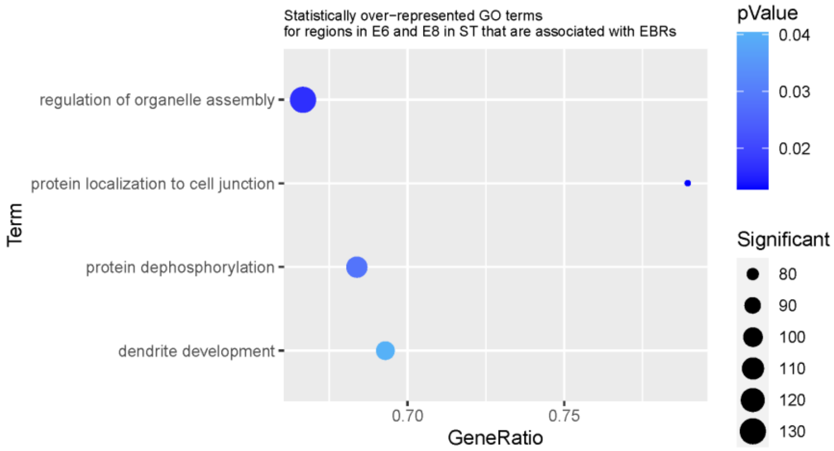


Figure 4-5: Gene ontology bubble plot showing the GO terms with greater than 1.5-fold enriched identified from genomic regions in state E6 (trivalent) and E8 (poised) in spermatids, (states which are positively associated with EBRs).

4.2.1 EBRs are associated to rapid epigenetic turnover in male gametogenesis

After determining the epigenetic landscape of male gametogenesis, we studied its relation to evolutionary chromosome rearrangements. We assessed the co-location of the 35 three-cell type states with the genomic positions of EBRs using a multi-association permutation test (Multi-regioneR (regioneReloaded)) (307). Remarkably, EBRs are negatively associated with the background state (E0-E0-E0), but highly associated with active or poised chromatin (permutation test based on 10,000 permutation, normalised z-score = -0.05 and > 0.01, *p* < 0.05, respectively) (Figure 4-6). This association was stronger with states that transition to E6 and E8 in spermatids (normalised z-score > 0.05, $p < 0.05$), particularly with those EBRs that occurred in the mouse lineage, suggesting that EBRs occur in chromatin environments prone to rapid change during spermatogenesis.

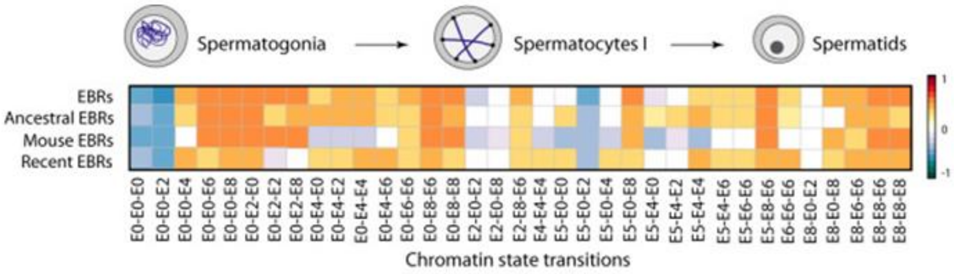


Figure 4-6: Multi-regioneR heatmap displaying correlations between different EBRs (ancestral, recent and mouse specific) and chromatin state transitions between chromatin states (E) in spermatogonia, primary spermatocytes and round spermatids.
Reproduced from*:* (305)*.* Plot produced by L Alvarez-Gonzalez.

The genome positions of EBRs were then integrated with other structural datasets including Hi-C data, CTCF, meiotic cohesins (REC8 and RAD21L), CpG islands, transcription start sites (TSS), ATAC-seq and RNA-seq (Figure 4-3D). Data for this analysis was jointly produced by Lucia Álvarez-González and myself, as part of the paper “3D chromatin remodelling in the germ line modulates genome evolutionary plasticity” (305). The 3D genome folding dynamics (A/B compartments and TADs) was analysed by using published Hi-C maps generated for spermatogonia, primary spermatocytes and round spermatids (103) and compared with the dynamics of the epigenetic landscapes. CTCF and meiotic cohesin binding sites were included for primary spermatocytes and round spermatids (103). Overall, EBRs were associated with regions that changed their state during spermatogenesis (all associations based on multiple permutation test based on 10,000 permutations, normalised z-score > 0.01, $p < 0.05$) (Figure 4-6). Furthermore, EBRs are associated with the ‘closed’ B compartment in pre-meiotic spermatogonia, but with the ‘open’ A compartment in meiotic spermatocytes and post-meiotic spermatids (Figure 4-7). Consistent with this, EBRs are associated with ‘closed’ chromatin environments (ChromHMM states E0, E4, E5) in spermatogonia and ‘open’ chromatin environments (ChromHMM states E2, E6, E8) in both primary spermatocytes and round spermatids. Finally, we see that EBRs are associated with regions that undergo structural remodelling and are associated with TAD boundaries in spermatogonia and spermatocytes but located within TADs in round spermatids (Figure 4-7). This suggests that EBRs are preferentially located in genomic regions that become accessible as spermatogenesis progresses. Evolutionary rearrangements should therefore not disrupt TAD structures in spermatogonia or spermatocytes (as they localise at TAD boundaries) but may do so in round spermatids (as they are located within TADs).

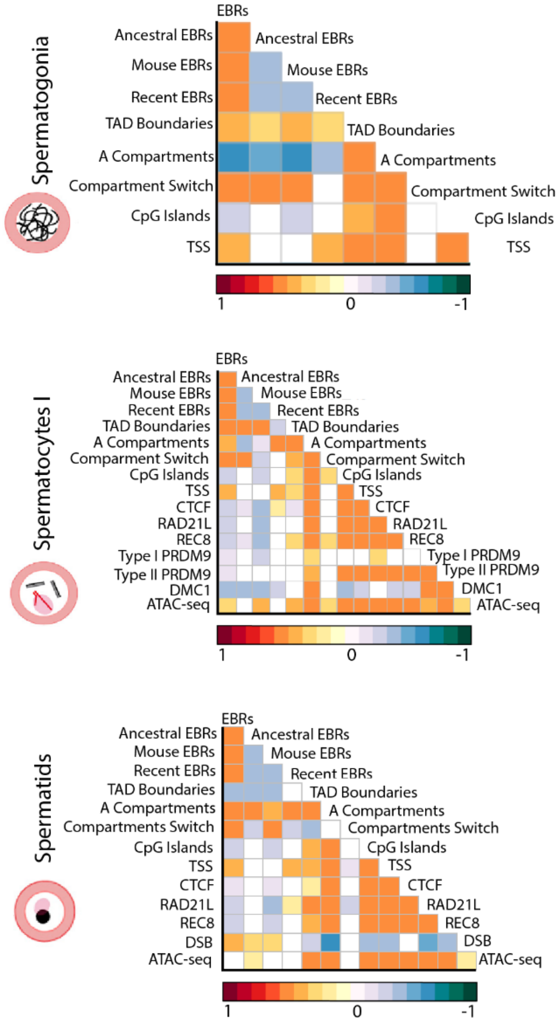


Figure 4-7: Heatmaps obtained by regioneR (multi-comparison) displaying correlations between different EBRs (ancestral, recent and mouse specific) and TAD boundaries, A compartments, compartment switch (from A to B and vice versa), CpG islands, transcription start sites (TSS) in spermatogonia, spermatocytes and spermatids.
CTCF, cohesins (RAD21L and REC8) and ATAC-seq was included for both primary spermatocytes and round spermatids. Primary spermatocytes also included PRDM9 sites (Type I and II) and DMC1 sites. Round spermatids included post-meiotic DSBs. Plots produced by Lucía Álvarez-González (305).

4.3 EBRs are associated with spermatid DSBs but not meiotic DSBs and spermatid DSBs are not associated with meiotic DSBs

As shown in Figure 4-7 spermatid DSBs are positively associated with EBRs (normalised z-score 0.06, $p < 0.05$) and the association was highest when including all the EBRs. It was slightly negative when only including the recent EBRs in the analysis. These results suggest that EBRs are located in the subset of DSBs that occur in open chromatin in round spermatids and indicate that transmissible genomic rearrangements preferentially occur within accessible genomic regions that suffer DNA damage in post-meiotic cells (Figure 4-8).

To search for the evolutionary plasticity of meiotic chromosomal architecture I conducted permutation tests (based on 1,000 permutations) to evaluate the association between spermatid DSBs and the genomic position of DMC1 and PRDM9 (markers of meiotic DSBs), the associations were negative (-23.3 for DMC1 and -8.5 for PRDM9, P-value =0.001). This indicates that the breaks occurring in spermatids are distinct from meiotic DSBs (see Figure 4-8 for a schematic representation of the differing locations of meiotic and post-meiotic DSBs). Therefore, I chose to do a further in-depth analysis of spermatid DSBs to determine their genomic distribution, their chromatin contexts and 3D genome structures associated with these breaks.

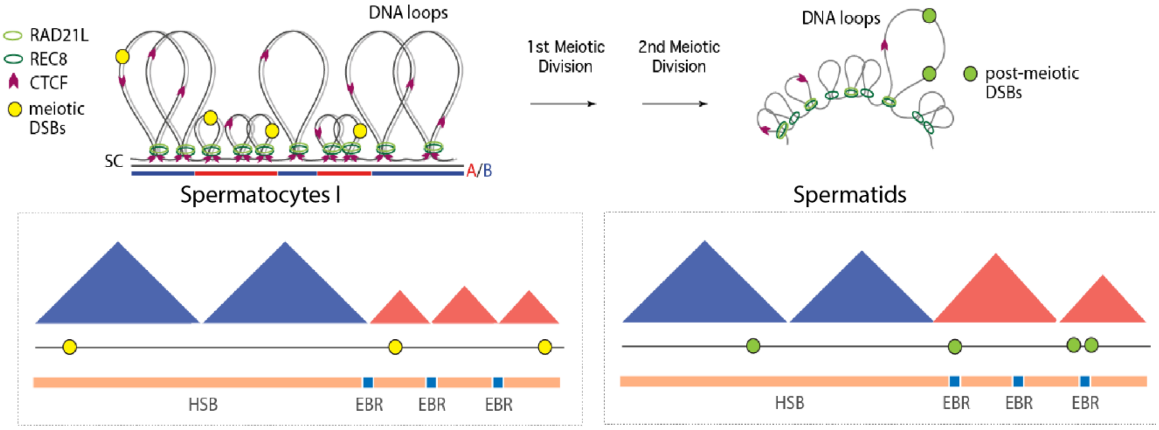


Figure 4-8: Working model depicting the disposition of the genome folding (DNA loops and compartments) in relation to cohesins, CTCF, meiotic DSBs and EBRs.
In the case of primary spermatocytes DNA loops protrude out of the chromosomal axes with meiotic DSBs occurring inside TADs in A compartments; EBRs are associated with TAD boundaries. In the case of round spermatids, EBRs are associated with post-meiotic DSBs inside TADs in A compartments. Abbreviations – EBRs evolutionary breakpoint regions, TADs topological associated domains, DSBs double-strand breaks, FPKM fragments per kilobase of transcript per million fragments mapped. Reproduced from (305).

4.4 Genomic distribution of spermatid DSBs

We used publicly available DSB data, (Table 2-2 and Supplementary Table 8-5) from spermatids: these comprised two replicates of round spermatids at developmental steps 1–9 (DSB18 and 19), and one of condensing spermatids at developmental steps 15–16 (DSB20). We first tested the concordance between the three spermatid files by detecting overlaps in genomic windows of high resolution (1 kb, Figure 4-9) and moderate resolution (5 kb, Supplementary Figure 8-9). This showed strong agreement between all three files. Around half the genomic windows containing a DBrIC signal in any given file also contained a DBrIC signal in at least one of the other files (47.8–56.4% overlap at 1 kb resolution, 55.2– 68.6% overlap at 5 kb resolution). Importantly, the overlap between round and condensing spermatid data was as close as the overlap between the two round spermatid replicates, indicating that DSBs occur in similar genomic locations throughout different stages of spermatid development. We therefore combined the three DBrIC files in all subsequent analyses to yield a single set of spermatid DBrIC peaks. As previously described (305), from the combined files we identified a total of 151,732 post-meiotic DSB locations in spermatids, covering 1.49% of the mouse genome (Figure 4-10C). The DBrIC signal peaks ranged from 146 to 6662 bp, with a mean and a median of 267 and 213 bp respectively. The coverage of DBrIC peaks per chromosome was scaled per Mb and plotted against chromosome length. This showed that chr11 and chrY were outliers with chromosome 11 having the lowest coverage of DBrIC peaks (0.97%) while the Y chromosome had the largest (3.51%) (Figure 4-11). The mean number of DBrIC peaks per Mb genome wide was 56, whereas for chr11 it was 37 and for chrY it was 144. Consistent with published findings (84), the post meiotic DBrIC peaks were also associated with repeat content within each chromosome (r2 = 0.78, P = 0.00003) and were enriched for simple repeats and satellite regions (Figure 4-12). Specifically, using genomic association tester (GAT) we found post-meiotic DBrIC peaks co-localise with transposable elements L1Md T and L1Md A (12.6 and 11.1 fold, P = 0.001). ChrY has the second highest coverage of RepeatMasker elements per Mb and chr11 has the second lowest. Given that spermatid DBrIC peaks are associated with repeat content this may partly explain why chrY has the highest coverage of spermatid DBrIC peaks and chr11 the lowest. For simplicity, we refer to spermatid DBrIC peaks hereafter as ‘DSB locations’ but note that this does not mean the breaks occur at precisely localised sites.

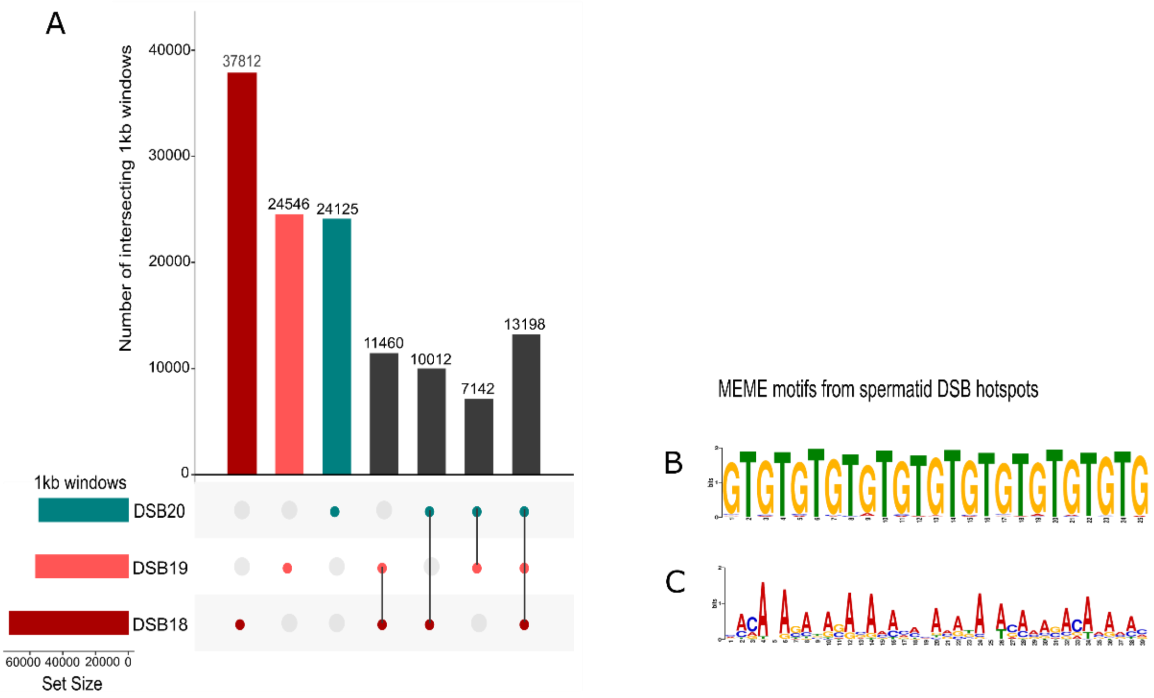


Figure 4-9: UpsetR plots and MEME motifs from the spermatid DSB data.
A) A 1 kb UpSetR plot showing the overlap of the three spermatid DSB files merged to create the spermatid DSB locations track. The number of 1 kb windows with a DSB signal in each file is represented on the left barplot as ‘set size’. The X-axis represents the number of 1 kb windows containing a DSB signal for the different overlap combinations. Different combinations of overlap are represented by the black lines interlinking the coloured circles. The DSB18/19 files are round spermatids stages 1–9 (shown in red and pink bars) and represent the total number of 1 kb windows containing a signal unique to these files. The DSB20 file (shown in the blue bar) is condensing spermatids stage 15–16 and this peak also represents the total number of 1 kb windows containing a DSB unique to this file. Bars in dark grey represent the number of 1 kb windows with signal in more than one file. **B and C**) Spermatid DSB MEME motifs, ordered by decreasing E-value. The E-value for motif B was 1.4e-1464 and the E-value for motif C was 1.1e-491. The input file to MEME contained all spermatid DSB locations.

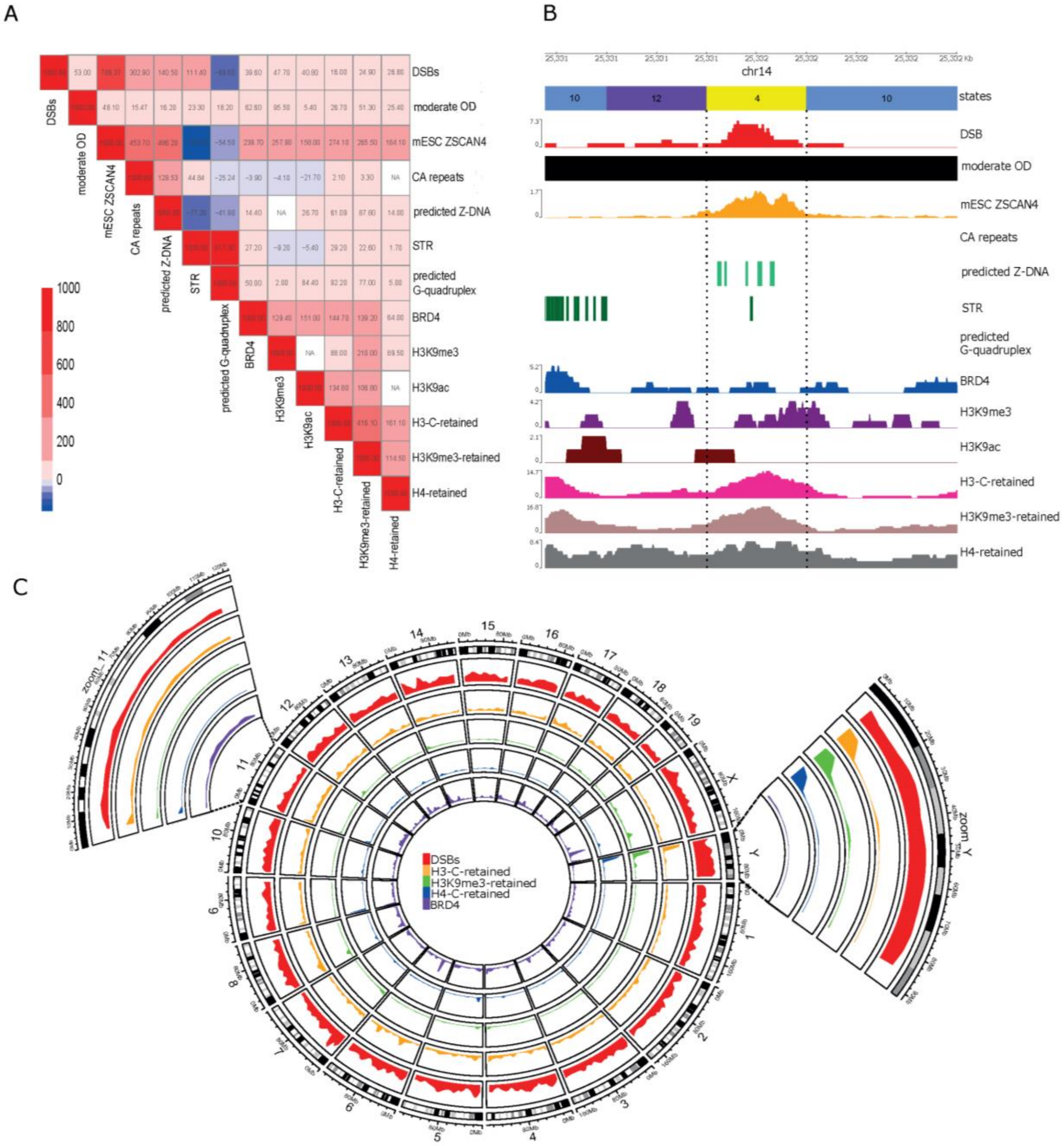


Figure 4-10: Heatmap, pygenomes plot and circos plot showing the associations of spermatid DSBs with different classes of non-B DNA, repeats and histone marks.

A) Heatmap plotting the Z-score of spermatid DSB locations showing the association with different classes of non-B DNA. For plotting the associations in red on the diagonal have been fixed at 1000. Non-significant values are in white, and all coloured cells have a P-value of ≤0.05. Red shading represents a significant positive association and blue shading a significant negative association. The OD sample is the moderate OD damage, top 1% of 50 kb regions. **B)** Example of a genomic region showing association of a spermatid DBrIC peak with ChromHMM states, oxidative damage, mESC ZSCAN4, CA repeats, predicted Z-DNA, STR, predicted G-quadruplex, BRD4, H3K9me3, H3K9ac and retained histones. **C)** Circos plot showing the distribution of spermatid DSB locations across the genome (red track). The orange, green and blue tracks show the distribution of retained histones in sperm and the purple track shows the distribution of spermatid BRD4. The two extra panels are enlarged tracks for chr11, which has the lowest coverage of spermatid DSB locations and chrY with the highest spermatid DSB location coverage.

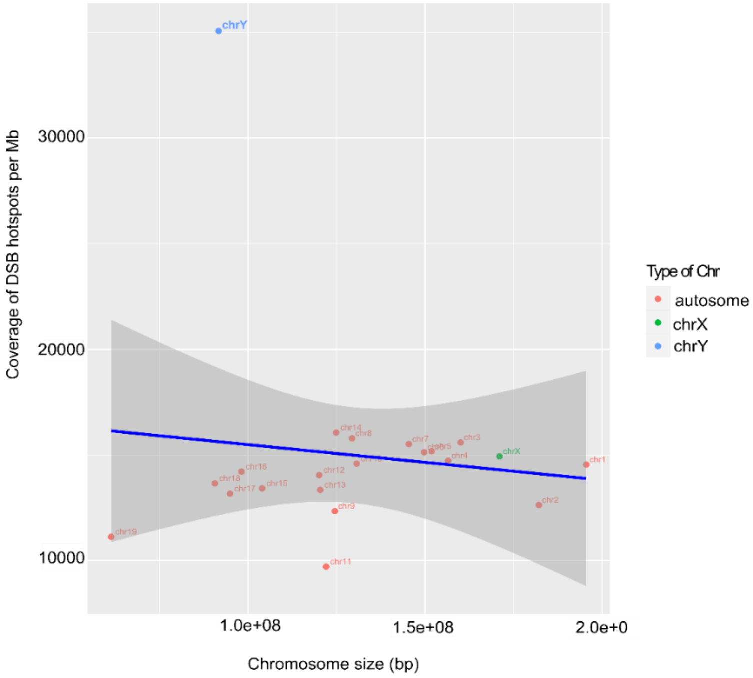


Figure 4-11: Correlation between post-meiotic DSBs and chromosomal size.
Linear regression of the number of post-meiotic DSBs (expressed as total bp coverage per Mb) detected in mouse chromosomes. Autosomes are depicted in red, the X chromosome in green and the Y chromosome in blue. Grey shading represents 95% confidence interval.

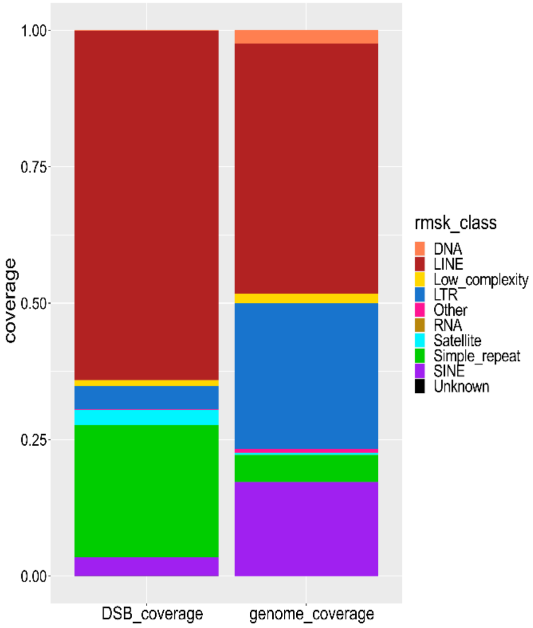


Figure 4-12: Coverage of transposable elements within the post-meiotic mouse spermatid DSBs vs the whole genome.

4.5 Identifying spermatid specific DSBs

We classified spermatid DSB locations as uniquely found in spermatids (spermatid-specific) or shared with other cell types by overlapping spermatid DSB locations with two publicly available files (sBLISS 63 and sBLISS 66) covering DSB localisation in developing enterocytes, as detected via sBLISS (250) (Supplementary Table 8-5). It is important to note that the methodologies used to detect DSBs in each cell type differ: DBrIC is lower resolution and requires an immunoprecipitation step (see Methods) while sBLISS is higher resolution and utilises adapter ligation in situ. Neither allows direct quantitation of the absolute number of DSBs per cell, but both allow analysis of the distribution of DSB locations across the genome. With these caveats noted, we observe that DSBs in spermatids as measured by DBrIC show a more restricted distribution than enterocyte DSBs as measured by sBLISS, with a smaller number of locations present in the genome (Figure 4-13A). Intriguingly, only a small fraction, (3.7% for sBLISS 63 and 3.6% for sBLISS 66) of spermatid DSB locations overlapped with enterocyte sBLISS DSBs, while 96.3% and 96.4% represented spermatid specific DSBs (Figure 4-13A). Conversely, 99.6% of both sBLISS 63/66 1bp DSBs did not overlap spermatid DSB locations. We considered whether this lack of overlap could be caused by the differing resolution of DBrIC versus sBLISS data. To adjust for this factor, we extended the enterocyte sBLISS data ±133 bp either side of the detected DSB sites to match the average length of the spermatid DBrIC peaks and recalculated the overlap (Figure 4-13B). This showed that only 9.1% and 8.8% of the spermatid DBrIC peaks overlapped the extended sBLISS 63 regions and the extended sBLISS 66 regions, respectively. We conclude that DSB locations in both cell types are overwhelmingly cell type specific, and thus that the processes leading to DNA damage in spermatids and enterocytes are likely to be largely distinct, but that a small number of genomic regions are liable to breakage in both cell types.

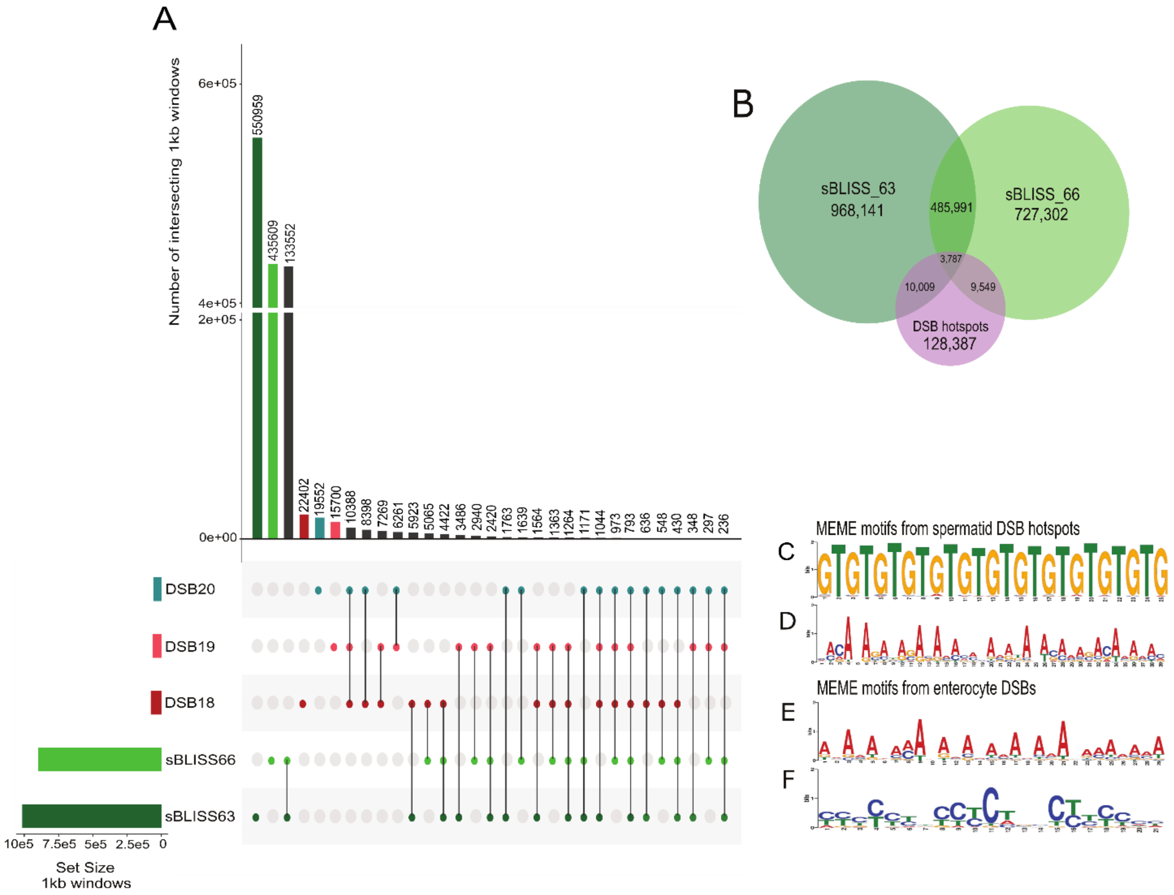


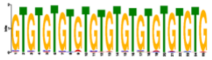
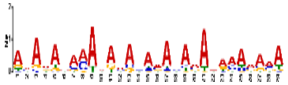
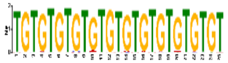
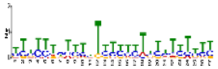
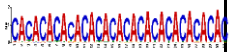
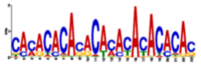
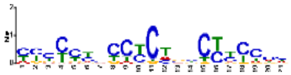
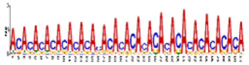
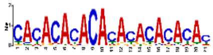
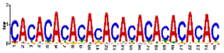
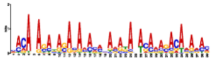
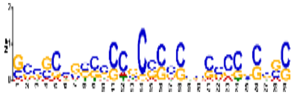
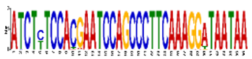
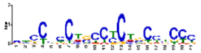
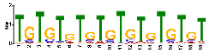
Figure 4-13: A 1kb UpsetR plot and MEME motifs of the spermatid and enterocyte DSBs.
A) A 1 kb UpSetR plot showing the overlap of the three spermatid DSB files 18/19/20 and the enterocyte sBLISS63/sBLISS66 files shown in green. The two green bars represent the total number of 1 kb windows containing a sBLISS63 or sBLISS66 peak unique to these files. The enterocyte files used were the 1 bp DSBs as per the original bed files. **B**) Venn diagram of the spermatid DSB locations file and the extended sBLISS_63/66 files (extended to the mean size of the spermatid DSB locations). Not to scale. **C and D**) Spermatid DSB MEME motifs, ordered by decreasing E-value. The E-value for motif B was 1.4e-1464 and the E-value for motif C was 1.1e-491. The input file to MEME contained all spermatid DSB locations not just the spermatid specific ones. **E and F**) The two most common motifs from MEME for all the extended enterocyte DSBs, ordered by decreasing *E*-value. The E-value for motif E was 3.7e-145 and the E-value for motif F was 1.5e-073.

4.6 Spermatid DSB locations occur in association with $(CA)_n$ and $(NA)_n$ motifs.

Having defined DSB locations for each data set, we used MEME (267) to identify specific DNA sequence motifs associated with the presence of DSBs in each cell type (Table 4-3). For the enterocyte dataset, we used the extended (±133 bp) sBLISS data, to ensure that we were comparing a similar-size genomic region for each DSB location detected in each cell type. This analysis therefore will detect DNA sequence motifs found in the near vicinity of DSBs for each cell type. Considering all spermatid DSB locations together, an alternating purine pyrimidine sequence ($(CA)_n$ or equivalently $(GT)_n$) and a more degenerate alternating $(NA)_n$ motif (Figure 4-9B and C) were identified as being statistically over-represented relative to a random model adjusted for nucleotide frequency. The $(CA)_n$ motif is present in 43.2% of the spermatid DSB location peaks, and the $(NA)_n$ motif is present in 62.4% of the peaks. In contrast, considering all the enterocyte

DSB locations together, neither $(CA)_n$ nor $(NA)_n$ sequences were detectably enriched relative to a random model adjusted for nucleotide frequency. Instead, an A-rich motif was detected in 42.9% of the enterocyte DSB locations and a C-rich motif in 38.4% (Figure 4-13, E-F). This again suggests that there are likely to be distinct damage sources at play in each cell type. We then refined this analysis by searching for motifs specifically within spermatid-specific, enterocyte-specific and shared DSB locations as defined above (Figure 4-13B, Table 4-3). This confirmed the enrichment for $(CA)_n$ and $(NA)_n$ motifs in spermatid-specific DSB locations (40.6% and 32.4% of these locations containing each motif respectively). T-rich and C-rich motifs were also confirmed as enriched in enterocyte-specific DSB locations (30.6% and 25.7% of these locations containing each motif respectively). In this sub analysis, a $(CA)_n$ motif was also detected as enriched in enterocyte-specific DSB locations, but to a much lesser degree, with only 5.3% of enterocyte-specific DSB peaks containing this motif. There was no enrichment for an $(NA)_n$ motif in enterocyte-specific DSB locations. For shared DSB locations found in common across all cell types, the $(CA)_n$ and $(NA)_n$ motifs were very highly enriched (86.9% and 72.9% of locations respectively), but the T-rich and C-rich motifs were not detectably enriched. Overall, 17.2% of all $(CA)_n$ repeats in the genome overlapped with spermatid DSB locations, while 15.1% and 12.8% overlapped with the extended enterocyte sBLISS 63/66 DSBs. Thus, $(CA)_n$ repeats appear to be a common fragility motif in both cell types studied, but slightly more so in spermatids than in enterocytes. Alternating purine-pyrimidine repeats have previously been described as topoisomerase II cleavage sites (308). We therefore tested whether there was a general association between motifs with alternating purine/pyrimidine sequences (positive strand only) and spermatid post-meiotic DSB locations (Supplementary Table 8-6). Using a 9-repeat motif length, $(RY)_9$, there was a significant positive association (Z-score = 29.1, P = 0.001, 1000 permutations). When extending the RY motif to 26 repeat units, $(RY)_{26}$, the mean length of the (CA) repeats in the mouse genome, the association was more significant (Z-score 35.8, P = 0.001, 1000 permutations). The positive association remained when excluding any $(RY)_{26}$ repeats that overlapped any (CA) repeat in the genome (Z-score 33.1 P = 0.001, 1000 permutations). Therefore, we can conclude that in spermatids, DSB locations are associated with (RY) repeats, particularly those with alternating A residues (or equivalently alternating T residues), and most strongly in the context of $(CA)_n$ repeats. However, using permutation testing, we observed a negative association (Z-score = –24.4, P = 0.001, 1000 permutations) between the topoisomerase II consensus motif (RNYNNCNNGYNGKTNYNY) and spermatid post-meiotic DSB locations. Thus, the association with RY repeats is not driven by the canonical topoisomerase II consensus motif.

Table 4-3: Comparison of MEME motifs obtained with different files.

	All spermatid DSB locations	motif as % of all ST DSBs	**All Enterocyte DSBs**	motif as % of EC DSBs	**Spermatid DSB locations NOT overlapping enterocyte DSBs (spermatid specific breaks)**	motif as % of ST specific breaks	**Enterocyte specific breaks**	motif as % of EC specific breaks	**Spermatid DSB locations overlapping sBLISS enterocyte breaks (shared breaks)**	motif as a % of shared ST/ EC breaks
Motif 1		43.2		42.9		40.6		30.6		86.9
Motif 2		49.9		38.4		32.4		5.3		72.9
Motif 3		62.4		3.1		10.2		25.7		76.3

ST=spermatid, EC=enterocyte. The enterocyte file used for all comparisons was the sBLISS63 peaks extended by 133 bp upstream and downstream that overlapped the sBLISS 66 peaks that had also been extended 133bp upstream and downstream, to obtain peaks the same size as the mean spermatid DSB.

4.7 Spermatid DSB locations and enterocyte DSBs are associated with distinct topological configurations of DNA

Our initial analysis (Figure 4-13B–F and Table 4-3) revealed different classes of simple repeat motifs associated with DSBs in each cell type. These repetitive motifs, such as short tandem repeats (STRs) are more likely to fold into non-canonical DNA structures. While DNA in cells typically folds into the widely known B-form with a right-handed helical structure, it is known that $(CA)_n$ sequences can readily undergo a transition to a left-handed Z-DNA conformation when subjected to unwinding torsional stress (309, 310). Therefore, we determined whether each of our DSB categories was associated with STRs or with regions predicted to fold into Z-DNA (Supplementary Table 8-6 and Table 4-4). To disentangle the effect of primary DNA sequence from that of DNA secondary structure, we tested separately for associations between each category of DSB and STRs predicted to form Z-DNA, STRs that are not predicted to form Z-DNA, and non-repetitive regions that are predicted to form Z-DNA. We also tested for any association with G-quadruplexes, as these have also previously been associated with DNA damage (302). Spermatid-specific DSB locations were strongly and independently positively associated with STRs and with predicted regions of Z-DNA, but negatively associated with G-quadruplex forming regions. Enterocyte-specific breaks were positively associated with experimentally determined G-quadruplexes, computationally predicted G-quadruplexes and non-repetitive predicted Z-DNA regions, but negatively associated with STRs and with predicted repetitive Z-DNA forming regions. Shared breaks were positively associated with all features except computationally predicted G-quadruplexes. Given that the spermatid-specific and shared spermatid-enterocyte DSB locations exhibited the same primary sequence motifs, with very similar associations with predicted secondary structures, we therefore pooled these together for subsequent analysis of their epigenetic chromatin context. Enterocyte-specific breaks were not addressed further in this study.

Table 4-4: Correlation between different classes of non-B DNA and spermatid or enterocyte DSBs. Values reported are Z-scores using 1000 permutations *.

Type	All Spermatid DSB locations	Spermatid specific DSB locations	Shared DSBs	Enterocyte specific DSBs	All enterocyte DSBs
predicted Z-DNA No overlap STR	150.7	125.6	61.2	63.8	65.2
STR No overlap predicted Z-DNA	110.9	104.8	42.1	-25.2	-23.8
predicted Z-DNA overlapping STR	1367.7	1303.9	533.2	-22.9	-5.7
predicted G-quadruplex	-72.9	-68.3	-12.5	21.5	19.0
G-quadruplex experimental	-32.5	-35.1	5.6	108.6	109.9

* All P-values were 0.001

4.7.1 R-loops are positively associated with spermatid DSBs

Various literature sources (125, 126) indicate that there may be an association between R-loops and DSBs, either that R-loops lead to stalled polymerase and that this leads to a DSB or that R-loops are formed to stabilise the DNA at a DSB. To determine if there was an association between R-loops and spermatid DSBs we carried out permutation testing between R-loops in mature sperm using data from Rassoulzadegan *et al* 2021 (311) and the spermatid DSBs data. This gave a strong positive association (Figure 4-14), with the association highest at the R-loop site as shown by the local Z-score of 99.263.

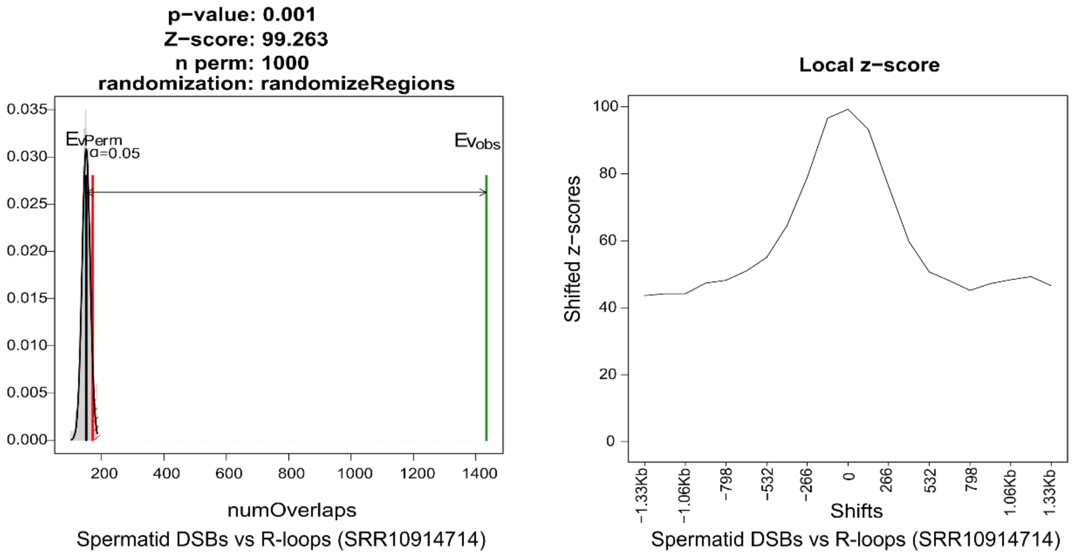


Figure 4-14: RegioneR permutation test results between spermatid DSBs and R-loops in mouse sperm.
The left plot shows a strong positive association between spermatid DSBs and R-loops in mature sperm with a Z-score of 99.3. The right plot shows the local Z-score, which shows that the association is strongest at the R-loop and then drops off as the distance from the R-loop increases. In grey is the number of overlaps of the randomized spermatid DSB regions with the R-loops, clustering around the black bar that represents the mean. Shown in green is the number of overlaps of the original spermatid DSB regions, which is much higher than expected. The red line denotes the significance limit. The local Z-score plots show the strength of the Z-score upstream and downstream from the spermatid DSBs.

4.8 Spermatid DSB locations co-locate with markers of NHEJ but are negatively associated with most other histone modifications

Having investigated the primary and secondary DNA sequence features associated with DSBs, we turned our attention to whether the epigenetic landscape affects the location of DSBs. Using publicly available data for 16 different epigenetic marks in round spermatids (Table 2-2), H3K4me1 showed the highest coverage genome wide (5.9%) while Kac had the lowest (0.001%) (Figure 4-15 and Supplementary Table 8-3).

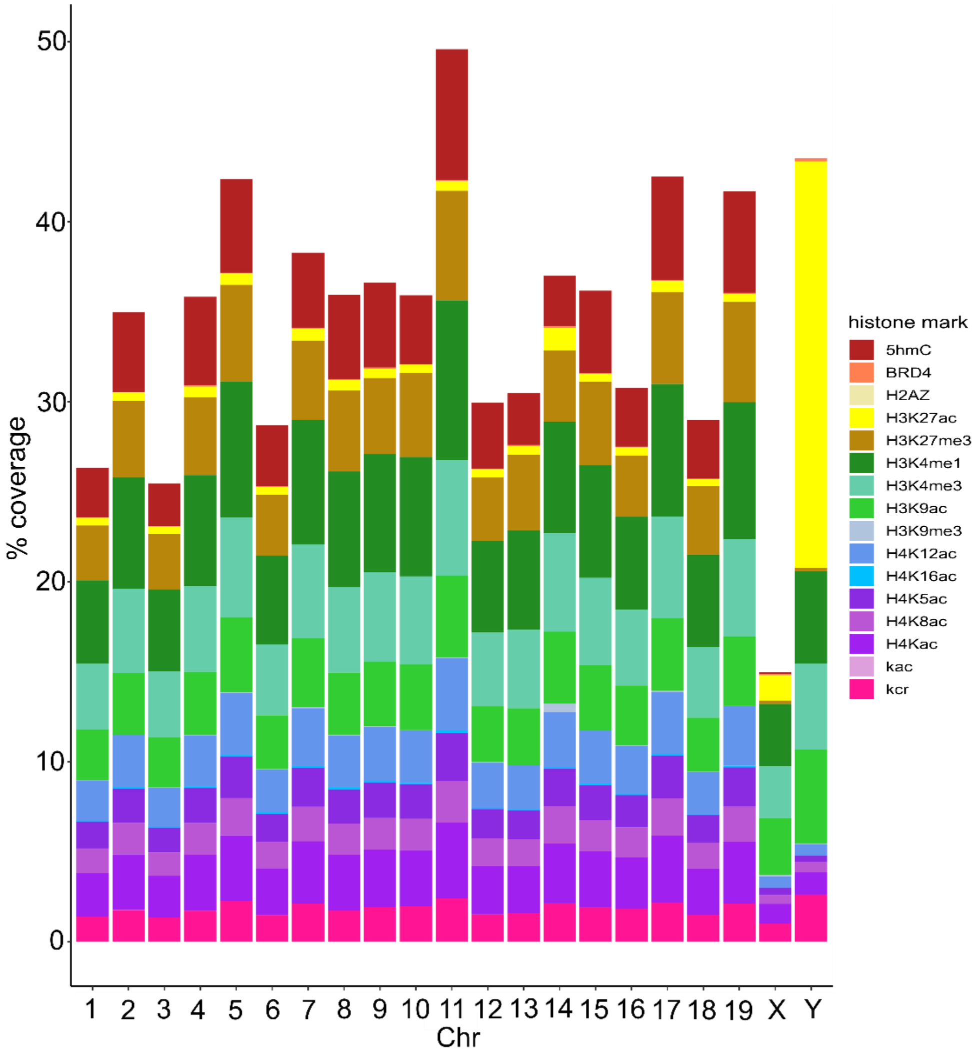
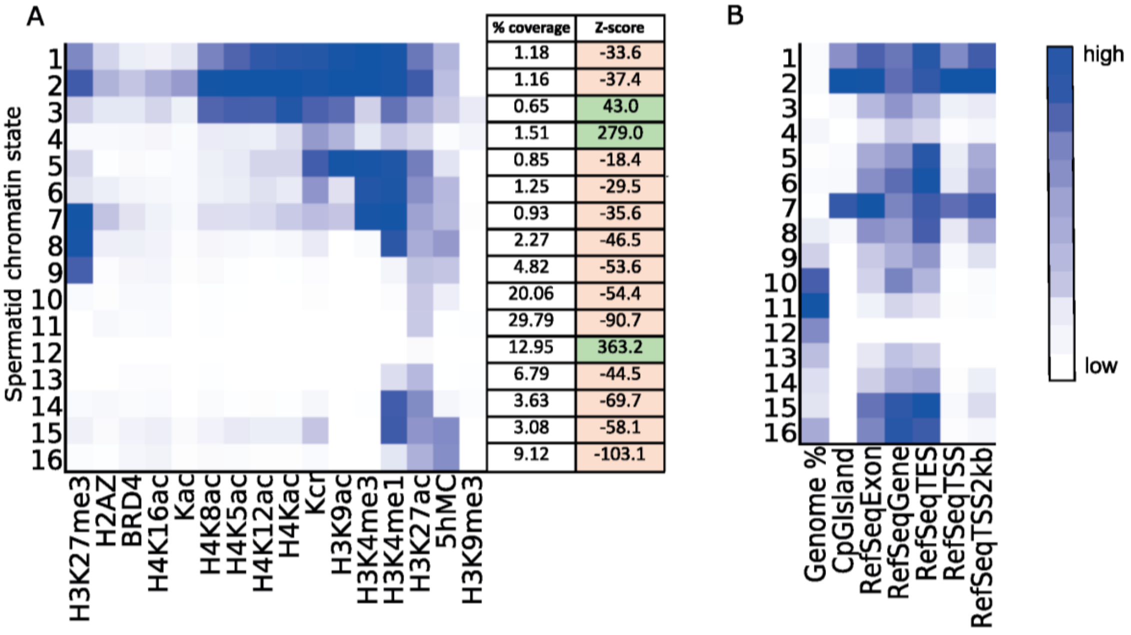


Figure 4-15: Coverage of the 16 histone marks used for ChromHMM analysis in Figure 4-16.

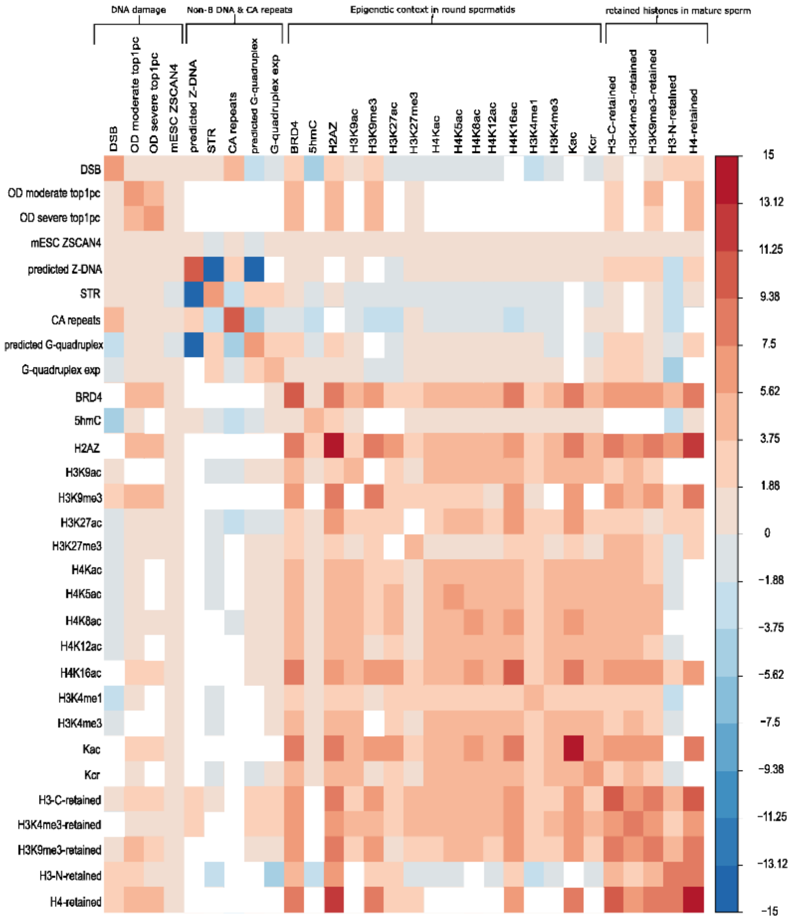
Only BRD4, H3K9me3, H3K9ac, Kac and H2AZ were positively associated with spermatid DSB locations (with Z-scores of 39.6, 47.7, 40.8, 4.9 and 18.0, P = 0.001 with 1000 permutations), while H4K5ac, H4K8ac, H4K12ac, H4Kac, H3K4me3, H3K27me3, Kcr, H3K27ac, H3K4me1 and 5hmC were negatively associated and H4K16ac was not significantly associated (Figure 4-10, Figure 4-18 and Supplementary

Table 8-6). To assess whether spermatid DSB locations occur in a specific chromatin context, we first ran ChromHMM (264) with 16 histone marks. At 200 bp resolution, a total of 16 different chromatin states were identified (Figure 4-16A). The three states with the highest genomic coverage were 10, 11, 12 (20.06, 29.79 and 12.95% coverage) Figure 4-16 B. State E12 was notable for low coverage of all histone marks, while state E2 had the highest coverage of all marks. Coverage of all states varied in each chromosome, with state E12 having the lowest coverage on chr 11 (8.65%) and the highest on chr 3 (15.02%). State E3 had the lowest coverage on chrY (0.20%) and the highest coverage on chr17 (0.83%), while state E4 had the lowest coverage on chr 11 (1.16%) and the highest on chrY (2.19%) (Figure 4-17 and Supplementary Table 8-4). Only states E3, E4 and E12 were positively associated with spermatid DSB locations (Z-scores 43.0, 279.0, 363.2, P = 0.001, 1000 permutations) (Figure 4-16A, Supplementary Table 8-6 and Figure 4-17). Our results thus far suggest that spermatid DSB locations are strongly associated with the (CA)$_n$ motif, tend to occur in genomic regions with low coverage of histone modifications (state E12, Figure 4-16) and are located in regions where BRD4 is found. A positive association between BRD4 and H3K9me3 (both part of the DNA damage response), and spermatid DSB locations provides further supporting evidence that the DSB locations we identify are indeed related to DNA damage and not spurious associations. Notably BRD4 is specifically associated with NHEJ (167) rather than other DSB repair pathways, consistent with the lack of homologous recombination and the requirement for NHEJ in spermatid DNA repair.



% coverage	Z-score
1.18	-33.6
1.16	-37.4
0.65	43.0
1.51	279.0
0.85	-18.4
1.25	-29.5
0.93	-35.6
2.27	-46.5
4.82	-53.6
20.06	-54.4
29.79	-90.7
12.95	363.2
6.79	-44.5
3.63	-69.7
3.08	-58.1
9.12	-103.1

Figure 4-16: ChromHMM emission plot showing spermatid chromatin states E1–E16.
A) The table shows genome coverage (%) and Z-score results of permutation tests between spermatid post-meiotic DSB locations and ChromHMM states. In green are ChromHMM states with a significant positive association with spermatid DSB locations and in pink are states with a significant negative association with spermatid DSB locations. All Z-score P-values were 0.001.
B) ChromHMM overlap plot showing fold enrichment with various RefSeq categories.



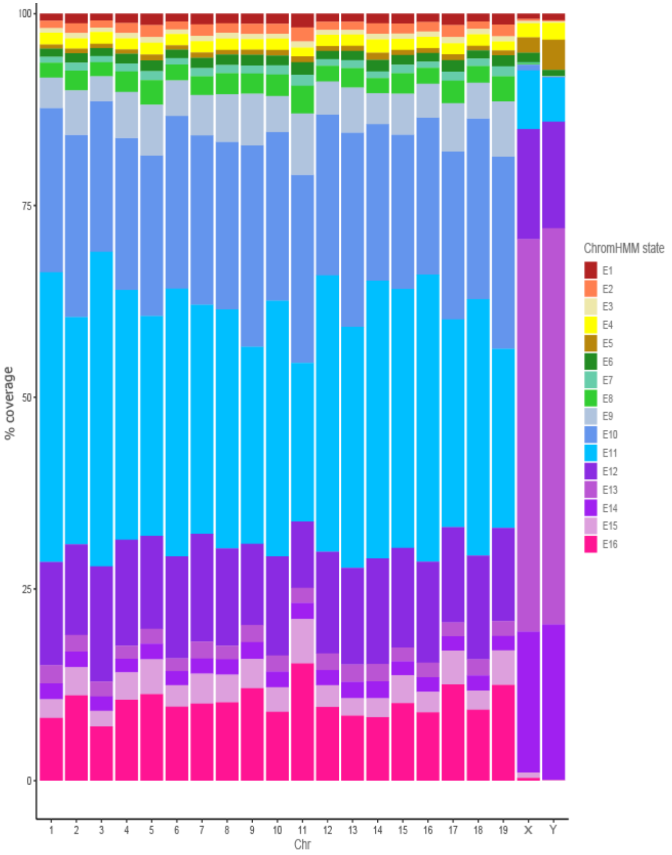


Figure 4-17: Coverage of the 16 spermatid ChromHMM states, related to Figure 4-16. For raw data used to make the plot please see Supplementary Table 8-4.

Figure 4-18: Permutation association of DNA damage, non-B DNA repeats, epigenetic context in round spermatids and retained histones in mature sperm.

Genomic association tester (GAT) heatmap showing the log2 fold change between samples. Shades of red show significant positive log2 fold change & shades of blue show significant negative log2 fold change. Non-significant associations with a P-value of > 0.01 are shown in white. In cases where fold change was 0, the log2 fold change has been fixed on this plot as –14.

4.9 Spermatid DSB locations are associated with histone retention and oxidative damage in mature sperm

The presence of DSBs in spermatids may affect subsequent downstream events in spermatogenesis either directly (if DSB formation and repair interferes with protamination) or indirectly (if both DSB formation and protamination are affected by the same underlying genomic features). Therefore, we investigated the association of spermatid DSB locations with regions retaining histones in mature sperm (Figure 4-18). Our results showed that post-meiotic DSB locations are enriched in retained histones in mature sperm (Z-scores of 16.0, 24.9 and 26.8, P = 0.001, 1000 permutations for H3-C, H3K9me3 and H4 replicate 2 respectively). As such, DSB locations may have a lower rate of histone to protamine replacement or the inability to undergo repackaging. We also investigated the correlation of predicted Z-DNA with regions retaining histones in mature sperm. H3-C, H3K4me3 and H3K9me3 replicate 2 are all positively associated with Z-scores of 61.1, 81.6, 87.8 (P = 0.001, 1000 permutations) (Supplementary Table 8-6). As has been previously described, reactive oxygen species (ROS) such as hydrogen peroxide are required for further chromatin compaction as they are essential for the formation of protamine-to-protamine disulfide bond formation (312). However, altered DNA packaging in round spermatids or sperm may increase its vulnerability to oxidative damage during epididymal maturation. Therefore, using publicly available data from two mouse genotypes with moderately and severely increased susceptibility to oxidative damage (OD) in mature sperm (Table 2-2) (141), we determined whether regions of high OD co-localise with spermatid DSB locations–i.e., whether pre-existing damage in spermatids may precondition mature sperm for further damage. Oxidative damage in sperm is in general not concentrated into tight hotspots but occurs across broader regions reflecting larger scale variations in packaging properties. Therefore, following previous publications (259), we divided the mouse genome into 50 kb windows and identified the top 1% of windows with the highest oxidative damage for both the moderate and severely damaged samples (see Materials and Methods). Chromosome 5 had the highest coverage for both samples while chromosome Y the lowest (Table 4-5). Overall, moderate and severe OD regions correlate with spermatid DSB locations (Z-score = 27.1 and 27.4, P = 0.001, 1000 permutations), as well as predicted Z-DNA and STRs (with Z-scores of 16.2 and 23.3, P = 0.001, 1000 permutations for moderate OD respectively) (Supplementary Table 8-8).

However, OD regions also showed a positive correlation with predicted G-quadruplexes that was not observed for spermatid DSB locations (Z-score = 18.2). To assess whether the associations of OD regions with non-B DNA were related to the co-localization of non-B DNA regions with spermatid DSB locations, we removed any non-B DNA regions that overlapped the spermatid DSB locations and repeated the analysis. Both moderate and severe OD regions still gave positive correlations with predicted Z-DNA, STR and G-quadruplexes (with Z-scores of 10.7, 16.9 and 7.6, P = 0.001, 1000 permutations for moderate OD) (Supplementary Table 8-6).

Table 4-5: Percentage coverage of the chromosomes with the top 1 % of 50kb OD damaged windows for the moderate and severe sample.

chr	moderate OD % coverage	severe OD % coverage
chr1	0.92	0.97
chr2	0.93	0.93
chr3	1.03	1.31
chr4	0.93	1.15
chr5	2.17	1.71
chr6	1.04	1.00
chr7	0.72	0.79
chr8	1.74	1.43
chr9	1.08	0.76
chr10	1.03	1.22
chr11	0.94	0.90
chr12	0.58	0.71
chr13	1.04	1.00
chr14	0.80	0.72
chr15	0.77	1.15
chr16	1.12	1.17
chr17	1.37	1.16
chr18	0.77	0.99
chr19	0.57	0.81
chrX	0.56	0.439
chrY	0.49	0.436

4.10 The transcription factor ZSCAN4 in embryonic stem cells with two-cell like features is positively associated with spermatid DSB locations and regions of high oxidative damage in sperm

In the post-fertilisation embryo, protamine packaging of the paternal DNA must be replaced by histone proteins. This may once again make the genome vulnerable to DNA damage from torsional changes and/or topoisomerase activity. ZSCAN4 is a transcription factor which occupies a subset of $(CA)_n$ microsatellite repeats in their nucleosomal form in mouse two-cell embryos (97), and is thought to help protect these fragile regions from genomic instability. Therefore, we investigated whether ZSCAN4 bound regions in E14Tg2a (E14) mouse embryonic stem cells (mESCs) were associated with spermatid DSB locations and with sperm oxidative damage. Two-cell like ZSCAN4 regions in the mouse genome were not uniformly distributed, with chromosome 11 having the highest coverage of two-cell like ZSCAN4 (36.2%) and chromosome Y the lowest (13.5%) (Table 4-6). Our results indicate a very large positive association between post-meiotic spermatid DSB locations and embryonic stem cell two-cell like ZSCAN4 (Z-score = 706.4, P = 0.001, 1000 permutations, Figure 4-10A and Figure 4-18). Because ZSCAN4 has been shown to be associated with $(CA)_n$ repeats, we tested whether our association is driven solely by these repeats, or whether there is an independent association between spermatid DSB locations and ZSCAN4 binding in ESCs. To do so, we then separated spermatid DSB locations that overlap CA repeats then those that do not overlap and re-ran permutation testing. Both analyses showed a positive association, with the significance of the association being higher for spermatid DSB locations that do not overlap CA repeats (Z-score = 629.2, P = 0.001, 1000 permutations) than for spermatid DSB locations that overlap CA repeats (Z-score = 389.6, P = 0.001, 1000 permutations). Since the association appeared not to depend on the primary sequence motif, we therefore also looked at the association of predicted Z-DNA forming regions to two-cell like ZSCAN4. Predicted Z-DNA regions that overlapped spermatid DSB locations were strongly positively associated with two-cell like ZSCAN4 (Z-score = 356.2, P = 0.001, 1000 permutations). Furthermore, the top 1% of 50 kb moderate OD damaged regions were positively associated with two-cell like ZSCAN4 (Z-score = 48.1, P = 0.001, 1000 permutations) (Supplementary Table 8-6). Mouse embryonic stem cell two-cell like ZSCAN4 regions and regions retaining H3K9me3 in mature sperm are strongly positively associated (Z-score 1131.0, P = 0.001, 1000 permutations). This may suggest that regions retaining histones represent fragile genomic regions in a two-cell embryo.

Table 4-6: Non-B DNA and mESC ZSCAN4 coverage across the mouse chromosomes.
Coverage as a % of each chromosome length.

Chr	mESC ZSCAN4	Predicted Z-DNA	STR	Predicted G-quadruplex	G-quadruplex experimental
1	32.170	0.083	2.033	0.595	4.416
2	33.642	0.093	2.044	0.600	4.812
3	31.359	0.074	1.980	0.594	4.085
4	33.587	0.087	2.043	0.631	4.956
5	34.339	0.110	2.104	**0.658**	5.153
6	32.633	0.084	2.028	0.600	4.481
7	32.885	0.086	1.951	0.644	5.095
8	33.319	0.102	2.110	0.634	4.926
9	34.421	0.105	2.049	0.590	4.989
10	33.447	0.093	2.140	0.623	4.666
11	**36.192**	**0.111**	2.184	0.642	**5.818**
12	31.899	0.085	2.018	0.593	4.398
13	33.372	0.086	1.985	0.569	4.373
14	32.102	0.076	1.980	0.582	4.089
15	31.811	0.087	2.026	0.639	4.909
16	31.611	0.082	2.047	0.584	4.290
17	34.096	0.102	2.060	0.645	5.125
18	30.874	0.086	2.025	0.586	4.419
19	31.820	0.095	2.051	0.606	5.078
X	18.006	0.039	**1.754**	0.622	2.927
Y	**13.523**	**0.029**	**2.373**	**0.318**	**0.322**
Genome coverage	31.389	0.085	2.037	0.602	4.465

mESC ZSCAN4=GSM4175885_GFP-Zscan_GFP_ChIP
predicted Z-DNA =predicted Z-DNA non-B DB (not overlapping STR)
STR =Short tandem repeats non-B DB (not overlapping predicted Z-DNA)
predicted G-quadruplex =predicted G-quadruplex Non-B DB
G-quadruplex experimental =GSM3003548_Mouse_all_w15_th_1_minus.hits.max.PDS.w50.35.

Per column, the value in green represents the highest coverage per chromosome and the value in red represents the lowest coverage. The Y chromosome has the lowest coverage of mESC ZSCAN4, predicted Z-DNA, predicted G-quadruplexes and experimental G-quadruplexes.

4.11 Discussion

This chapter has explored the relationship between 3D chromatin remodelling in mouse germ cells and evolutionary changes in genome structure, as the role of genome folding in the heritability and evolvability of structural variations is not well understood. We identified the location of EBRs in seven ancestral rodent genomes and identified the dynamics of the structural and epigenetic properties of the EBRs we identified through mouse spermatogenesis. From the positive association of EBRs to active or poised chromatin (identified through ChromHMM analysis) we can conclude that EBRs occur in genomic regions that become accessible as meiosis progresses especially in post-meiotic spermatids. EBRs are positively associated with post meiotic DSBs, but negatively associated with marks of meiotic DSBs DMC1 and PRDM9.
As we identified the positive association of EBRs with post-meiotic DSBs, we then carried out an in-depth investigation into post-meiotic DSBs to investigate the genomic and epigenetic contexts associated with these DSBs. Consistent with previously published results we find spermatid DSBs positively associated with short tandem repeats and LINE elements. We further show spermatid DSBs preferentially occur in association with $(CA)_n$, $(NA)_n$ and $(RY)_n$ repeats, in predicted Z-DNA, are not associated with G-quadruplexes, are preferentially found in regions of low histone mark coverage and engage the remodelling/NHEJ factor BRD4. Locations incurring DSBs in spermatids also show distinct epigenetic profiles throughout later developmental stages: regions retaining histones in mature sperm, regions susceptible to oxidative damage in mature sperm, and fragile two-cell like embryonic stem cell regions bound by ZSCAN4 all co-localise with spermatid DSBs and with each other.

The results presented here show the importance of 3D chromatin organisation in the formation of transmissible chromosomal re-organisations within the male germline. Firstly, accurate re-constructing of seven ancestors in the rodent lineage was required to identify chromosome rearrangements and EBRs. We showed that chromatin environments that become accessible as meiosis progresses are positively associated with EBRs, especially in round spermatids (post-meiotic cells) which are susceptible to DNA damage. Although we did not analyse SPO11 sites directly but instead used the genomic distribution of male DMC1 and PRDM9 as markers of meiotic DSB locations, we saw that meiotic DSBs do not co-localise with EBRs. Meiotic DSBs in both males and females are driven by the same mechanisms (98), suggesting that their location will be similar in both sexes. As such, we can predict that DMC1 and PRDM9 sites in both sexes are not co-located with EBRs. Instead, EBRs were positively associated with post-meiotic DSB hotspot locations in spermatids. This is consistent with the fact that spermatids are haploid cells and so lack a template for more accurate repair mechanisms, instead relying on error prone processes such as NHEJ (293, 313). Transmissible chromosome rearrangements are therefore likely more strongly associated with male specific post-meiotic DNA damage locations rather than with meiotic DSB locations (present in both sexes).

It is well known that changes in chromatin accessibility can affect the likelihood of some genomic regions to undergo chromosomal breakage (44, 314, 315). In somatic cells upon DSB induction, TAD boundaries strengthen which helps to allow the chromatin to become more accessible to accommodate the many proteins of the DNA damage repair (DDR) pathway (316).

Moreover, regions of high interaction are normally transcriptionally active (315), with promotor-enhancer interactions enabled by chromatin remodelling (317). Our data indicates a clear association between EBRs and TAD boundaries in spermatogonia and spermatocytes. This ties with the mounting evidence pointing to an association between EBRs and TAD boundaries in both mammalian and bird somatic cells (44, 314, 315, 318). As such, rearrangements that occur at TAD boundaries are less likely to disrupt gene regulation and are therefore not selected against. However, we also found, DSBs that initiate rearrangements occur more frequently within TADs in open chromatin. This contradiction is resolved as we show that EBRs are positively associated with genomic regions that form closed chromatin in spermatogonia (pre-meiotic cells) or primary spermatocytes (meiotic cells), but these regions form open chromatin in spermatids. These regions lie within TADs in spermatids but are closer to TAD boundaries in other types of cells. Rearrangements initiated by DSB formation, and the subsequent DNA rejoining, might be removed by cell cycle checkpoints and/or viability selection on the resulting offspring. During meiosis, three main meiotic checkpoints exist (1) response to unrepaired DSBs, (2) meiotic silencing of unsynapsed chromatin (MSUC) (see section 1.9.5), and (3) the spindle assembly checkpoint (SAC) (306) (see section 1.9.2). Any chromosomal rearrangements that occur before or during meiosis therefore have a high probability of being eliminated via the aforementioned checkpoints. Moreover, intrachromosomal translocations if not eliminated during meiosis often result in aneuploid gametes and non-viable offspring, and therefore not passed to the next generation. This supports our observations that EBRs are negatively associated with programmed meiotic DSBs in primary spermatocytes. Because post meiosis there are no further checkpoints, and the cell will be euploid as the cells are haploid at this point, no elimination of chromosome rearrangements is taking place after meiosis.

Because the post-meiotic stage is longer in males, and chromatin compaction is also exclusively taking place in spermatids, this points to a paternal bias for the appearance of chromosome rearrangements. Linked to this, there is a widely known paternal mutation bias (319–326), widespread across amniotes which was initially ascribed to the higher number of cell divisions in the paternal line (reviewed in ((327)), but which has recently been shown to be partly independent of cell division (318). Paternal specific events during post-replicative stages of gametogenesis and fertilisation are attractive candidate mechanisms for a cell division independent contribution to this paternal mutation bias. These include the extensive chromatin remodelling occurring during paternal genome condensation and decondensation, and also the specific exposure of the paternal genome to oxidative damage during epididymal transit and fertilisation. This paternal mutation bias is mirrored by an increasing appreciation of the role of sperm DNA fragmentation as a cause of male sterility, in particular as a cause of recurrent miscarriage and IVF failure (102, 303). It is therefore important to understand the factors underpinning DNA damage occurrence and localisation in the male germ line, as any damage remaining in mature sperm that cannot be repaired by the oocyte has the potential to affect a developing embryo, with both clinical and evolutionary consequences.

Here, we expand on previous investigations exploring the distribution of spermatid post-meiotic DSB locations. Our investigation first compared spermatid and enterocyte DSBs to establish the tissue-specific nature of spermatid DSB locations and investigated both primary sequence motifs and secondary DNA structure associations in each cell type. This showed that the patterns of breakage are cell type specific, and in particular that spermatid DSB locations are associated with $(CA)_n$ / $(NA)_n$ repeats and predicted Z-DNA, while enterocyte DSBs are associated with poly-TG-rich regions and G-quadruplexes. The small proportion of shared breaks showed a similar motif pattern to spermatid specific breaks. We conclude that the association with predicted Z-DNA and simple repeats is driven by a process that is very prominent in spermatids but less so in enterocytes – most likely the huge torsional strain changes that occur as the genome is remodelled in spermatids. Conversely, DSBs associated with G-quadruplexes are more prominent in enterocytes. This is expected since G-quadruplexes lead to DSBs due to stalled replication when the replisome cannot unwind the quadruplex. Spermatids are post-replicative cells, likely explaining why we do not observe a positive association with G-quadruplexes in this cell type (Table 4-4). This may also explain the much larger number of DSB locations identified in enterocytes compared to spermatids. We considered whether our results might be confounded by technical differences between DBrIC (used in the spermatid study) and sBLISS (used in the enterocyte study). The primary technical difference between the protocols is that DBrIC is carried out on purified chromatin and DNA is immunoprecipitated, while sBLISS is carried out in situ on fixed nuclei with adapter ligation. Conceivably some chromatin regions may therefore be less accessible in the sBLISS enterocyte study. However, the fact that this study detected a larger number of DSB regions than the DBrIC spermatid study indicates that chromatin accessibility is unlikely to greatly compromise DSB detection by sBLISS. While the techniques also differ in resolution, we adjusted for this in our analysis by extending the sBLISS location to match the average DBrIC peak width. Nevertheless, it remains possible that aspects of the differences between the cell types are driven by methodological differences between DSB detection methods

– future work in this area could aim to profile DSB locations in a range of tissues utilising a common methodology to allow systematic exploration of this question.

Following this, we carried out a broad examination of data for 16 histone marks in spermatids in relation to the spermatid DBrIC DSB locations. Finally, we related this to downstream events including histone retention and oxidative damage in mature sperm, and a transcription factor known to bind fragile chromatin in the early embryo. We show that many of these events are strongly associated with each other. These associations in turn explain aspects of the overall chromosomal distribution of both spermatids DSBs and these various chromatin components. Chr11 has the highest coverage of mESC ZSCAN4, predicted Z-DNA (positively associated with spermatid DSB locations) and experimental G-quadruplexes (negatively associated with spermatid DSB locations). Conversely chrY has the lowest coverage of mESC ZSCAN4, predicted Z-DNA and experimental G-quadruplexes. Taken together, our results suggest the presence of a common "vulnerability code" that predisposes specific regions of the paternal genome to damage at several stages of the reproductive cycle. The reasons for these associations remain to be established experimentally. However, the association between spermatid DSB locations and predicted Z-DNA regions is most likely related to the torsional changes discussed above. Torsional strain can also lead to non-B DNA formation and in particular to the formation of Z-DNA due to unwinding stress, such as that generated during protamination. Kim *et al* 2021 (328) have shown that TG repeats preferentially formed Z-DNA over CG repeats, as the free energy barrier of the transition from B to Z-DNA was lower for TG repeats than CG repeats. They also showed that more torsional stress is required for the formation of Z-DNA in TG repeats than that required in CG repeats, as Z-DNA in TG repeats is less stable. The lower free energy barrier to form Z-DNA in TG/CA repeats might account for the enrichment of this specific motif that we observe in the spermatid DSB locations. We therefore hypothesise that B-DNA to Z-DNA transitions in specific regions of the genome may act as a molecular "crumple zone", either buffering the torsional strain until it can be relieved by strand scission and helix unwinding, or simply acting in concert with topoisomerase activity to relieve the accumulated tension. While DSB locations in spermatids do not associate with the canonical topoisomerase II motif, it may be that conversion to Z-DNA facilitates topoisomerase cleavage and/or modulates its binding site specificity. Work with Drosophila topoisomerase II (329, 330) showed that topoisomerase II can bind and cleave Z-DNA and it has a higher affinity for Z-DNA than B-DNA. Choi *et al* 1995 (331) suggested that both Z-DNA and B-DNA appear to be equally attractive as topoisomerase II cleavage sites, however a supercoiled substrate (such as Z-DNA) enhanced the cleavage efficiency of topoisomerase II without altering the specificity. Work by Szlachta *et al*, 2020 (10) has shown that in human cell lines regions of the genome that have a higher potential to form stable DNA secondary structures are more prone to DSBs induced by topoisomerase II compared to random and flanking sequences.

Collectively, this work helps to explain the positive association we have observed with spermatid DSB locations and predicted Z-DNA, in that topoisomerase may preferentially cleave DNA in the Z-form giving the positive association with spermatid DSB locations. Alternatively, non-B DNA structures might lead to DNA damage through a topoisomerase-independent mechanism. The optimal substrates for the mismatch repair (MMR) proteins share similar features with some non-B DNA structures, such as the junction of B to Z-DNA (114). The junctions could be mistaken for regions of damage and an incomplete mis match repair could result in a DSB at these regions. Future experimental work will therefore be required to resolve precisely where and when Z-DNA forms during sperm condensation, and how this relates to topoisomerase activity.

Following our investigation of the factors associated with DSBs during spermatid development, we turned our attention to potential downstream consequences of this damage. We show here for the first time that retention of histones in mature sperm is positively associated with DSB locations arising several days previously, during spermatid elongation (Figure 4-10). This implies that DSB formation during chromatin condensation selectively impairs local replacement of histones with protamines. This could be a direct effect, if the presence of DNA repair factors prevents access by the protamination machinery. Alternatively, it may be an indirect effect mediated by Z-DNA, if Z-DNA is refractory to removal of histones and replacement by protamines (332) and thus protamination in one region of the genome could trigger refolding of nearby “crumple zones” into Z-DNA and prevent them in turn becoming protaminated. Despite the association between spermatid DSB locations and retained histones in sperm, we found that DSBs were negatively associated with many histone modifications in spermatids, with the only positive associations being with lysine acetylation (a core event during histone eviction) and components of the spermatid DNA damage response (BRD4, H3K9me3 and H2AZ). It is possible that high coverage of a chromosome with histone marks may reduce its susceptibility to damage, for example chromosome 11 had the highest coverage of states E1 and E2 (Figure 4-16) and it has the lowest coverage of DSBs. Conversely the Y chromosome has the second lowest coverage of state E1 and the lowest coverage of state E2 and it has the highest coverage of DSBs. Alternatively, histone proteins may be removed at DSB sites to facilitate the repair process (333).

We also showed that regions of high oxidative damage are also positively associated with spermatid post-meiotic DSB locations, predicted Z-DNA, STRs and retained histones in sperm. It remains to be established whether spermatid DSB locations intrinsically prime sperm chromatin for oxidative damage, or whether the effect is mediated indirectly via histone retention, since the DNA of regions retaining histones in mature sperm is less compact and therefore more susceptible to OD. Intriguingly, other work shows that the guanines in Z-DNA are more sensitive to alkylating modifications than in B-DNA (334, 335). Once these modifications have formed on Z-DNA they are resistant to excision by repair enzymes (336). If this applies also to oxidative damage it could provide an alternative mechanism for the co-localisation of sperm oxidative damage with spermatid DSB locations and Z-DNA. There are clearly multiple factors that predispose sperm chromatin to OD since we (and others (337)) also observe a correlation between OD and G-quadruplex regions. It is unclear why G-quadruplexes should be susceptible to OD in mature sperm as these are neither associated with spermatid damage nor histone retention. It may be that this is an artefactual association since OD in this study was defined as the presence of the oxidized base 8-hydroxy-2'-deoxyguanosine (8OHdG), which will thus inevitably be more prevalent in G-rich regions capable of G-quadruplex formation. Susceptibility of the chromosomes to external oxidative damage has been reported to depend upon their position within the sperm head, with a peripheral or basal location being more susceptible to oxidative attack (144).

Finally, we showed that spermatid DSB locations (and other associated features such as predicted Z-DNA and CA repeats) were also strongly associated with genomic regions that bind ZSCAN4 in embryonic stem cells with 2-cell like features. Importantly, the association between spermatid DSB locations and 2-cell like ZSCAN4 was even stronger when we focused on DSB locations occurring outside CA repeats, indicating that this is not driven solely by the proposed CA binding activity of ZSCAN4. Rather, it may be that ZSCAN4 is recognising Z-DNA directly. Regardless of the precise mechanisms, our results demonstrate a continuity of genomic localisation of damage-associated factors from spermatid elongation, through the mature sperm, up to embryonic development. As the earliest event in this chain, it seems possible that chromatin remodelling events in spermatids not only trigger DSBs during chromatin condensation but may also precondition DNA for damage from multiple sources later on in reproduction.

The results presented here are the associations between different marks in spermatids, retained histones in epididymal sperm and predicted non-B DNA. Without single cell data we cannot unequivocally determine whether a DSB occurs in the exact region of a particular histone modification. Even with single cell data, it is impossible to measure the same cell at different stages of its life history without the development of non-destructive techniques for chromatin profiling. Thus, while we show that spermatid DSB locations are associated both with predicted Z-DNA and subsequently with regions retaining histones in epididymal sperm, we cannot say for sure whether histone retention is a consequence of changes in DNA conformation, or simply occurs in the same genomic regions. We note however that DSBs themselves are rare events affecting only a small fraction of cells and thus the downstream events in mature sperm and the early embryo are not direct consequences of DSBs occurring in spermatids. There may be different sequences of events leading to the associations that we have observed. One possibility is that histone retention at specific fragile sites is necessary to retain them in the B-DNA form. In the rare cells where histones are removed from these sites, this triggers refolding into Z-DNA and vulnerability to strand breakage. Under this hypothesis, the signals dictating histone retention at these sites remain to be elucidated. Alternatively, it may be that these sites constitutively fold into Z-DNA during spermiogenesis, and that this directly prevents histone replacement and leads to histone retention. Under this alternative hypothesis, the triggers for refolding into Z-DNA and the kinetics for re-establishment of B-DNA following fertilisation remain to be determined. Direct profiling of Z-DNA at different stages of spermiogenesis may in time resolve some of these questions.

In summary, we have shown that spermatid DSB locations are positively associated with specific primary and secondary DNA structures, and with a limited range of histone modifications. These same genomic regions are then associated with damage and damage-control factors both in the mature sperm and the embryo, signifying the presence of a common “vulnerability code” for the paternal genome throughout reproduction. We conjecture that the switch from B to Z-DNA acts as a molecular crumple zone to help relieve the torsional strain that occurs during chromatin remodelling in spermatids, with the caveat that the Z-DNA is more prone to DSBs than B-DNA possibly by preferential topoisomerase II cleavage (Figure 4-19). Some aspect of the remodelling process –potentially DNA conformational change, or alternatively the repeated severing and re-joining of DNA by topoisomerases - subsequently has impacts on protamination and histone retention in sperm, in turn affecting oxidative damage. Finally, these same regions once again become vulnerable as protamines are replaced by histones in the embryo. Understanding how all classes of non-B DNA and epigenetic marks integrate with the NHEJ pathway and the DNA damage response will further our understanding of DNA damage and paternal genome mutagenesis throughout the reproductive cycle.

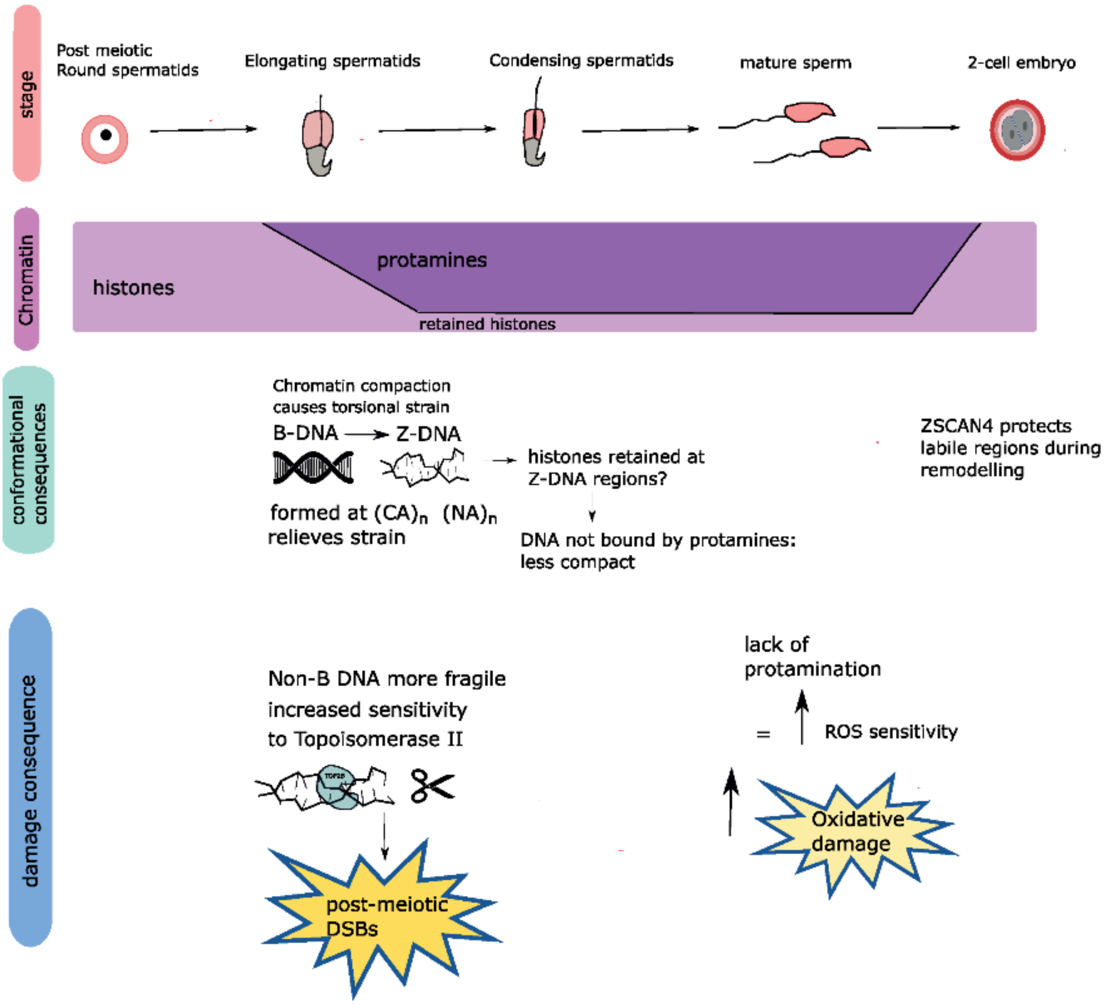

stage
Post meiotic Round spermatids
Elongating spermatids
Condensing spermatids
mature sperm
2-cell embryo
Chromatin
histones
protamines
retained histones
conformational consequences
Chromatin compaction causes torsional strain
B-DNA → Z-DNA
formed at (CA)n (NA)n relieves strain
histones retained at Z-DNA regions?
DNA not bound by protamines: less compact
ZSCAN4 protects labile regions during remodelling
damage consequence
Non-B DNA more fragile increased sensitivity to Topoisomerase II
post-meiotic DSBs
lack of protamination
= ROS sensitivity
Oxidative damage

Figure 4-19: Summary hypothesis schematic of spermiogenesis and the possible role of Z-DNA in DSB formation.
The stage track shows different stages of spermiogenesis. The chromatin track shows the changes that occur in chromatin as spermiogenesis progresses. The conformational consequences track shows the progression of chromatin changes through reproduction, while the damage consequence track shows the consequences for different types of DNA damage at each stage. We propose that DSBs in spermatids effectively act as a 'tracer' for regions undergoing remodelling due to torsional changes, while oxidative damage in mature sperm 'traces' regions that are vulnerable due to histone retention. These regions are substantially shared, because some aspect of the remodelling process triggers histone retention and thus ongoing vulnerability to damage.

The analysis presented in this chapter has only been carried out with *Mus musculus* data, it would be interesting to repeat this analysis with other species such as rat, rhesus money or a marsupial species. This would enable us to determine whether there is a shared "vulnerability code" between different species, or whether the results presented above only apply to mouse. If the EBRs within more species could be identified and data for DSB locations in round spermatids determined, then a similar analysis could be re-capitulated. ChIP-sequencing for oxidative damage in mature sperm from other species would also strengthen the conclusions that could be drawn. Higher resolution Cut&Tag results could be obtained if an alternative protocol such as the recently developed MulTI-Tag (338), (in which you can target 2 marks in the same cell) were used. This may allow an in-depth analysis of say Z-DNA/DSBs or a histone mark in a single round spermatid

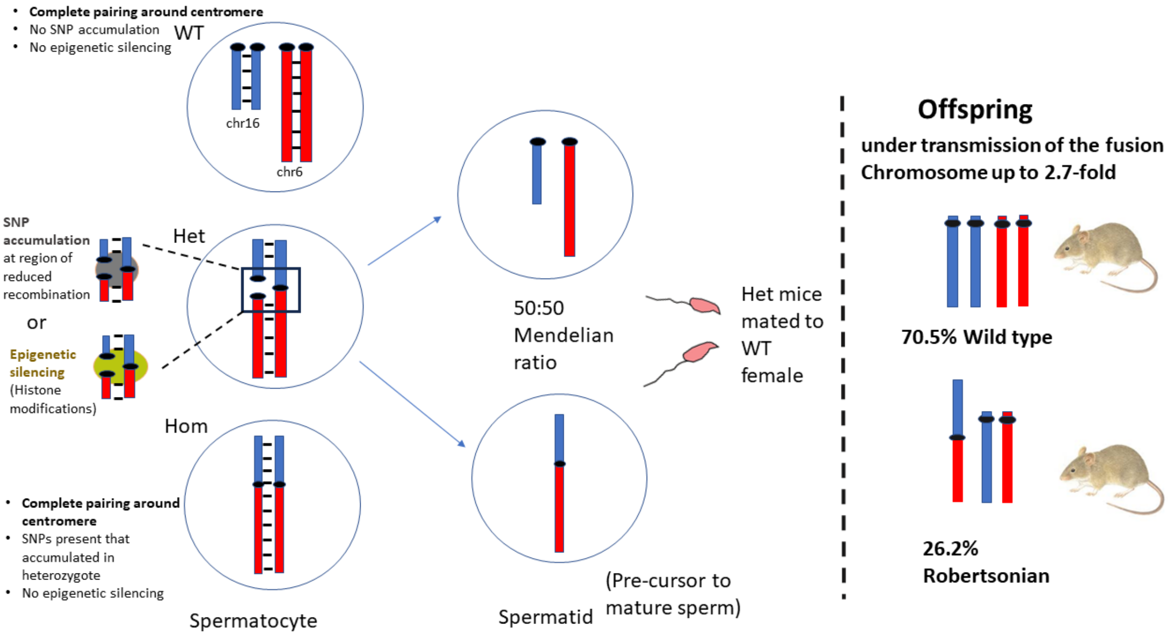
and would resolve some of the uncertainties obtained with our results when we have used predicted Z-DNA data with a population-based measurement of round spermatid DSBs.
The Z-DNA used for permutation testing against the spermatid DBS was predicted data, as at the time of analysis no experimental data in the C57BL/6 mouse strain was available. Ideally ChIP-seq experiments or Cut&Tag would be carried out using antibodies against Z-DNA in the same mouse strain as the round spermatid DSB data (C57BL/6). This would determine if there were any differences in the associations obtained with experimental and predicted data.

.

5. Investigating the role of non-Mendelian inheritance in the spread of chromosome fusions.

In the previous chapter we investigated where and when chromosome rearrangements might occur during male gametogenesis in mice. We found that the locations of DSBs in post-meiotic spermatids are positively associated with Evolutionary breakpoint regions (EBRs), indicating that CRs originate after meiosis has taken place, and during spermatid development. But how are CRs passed to the next generation, and, more importantly, to the rest of the individuals of a population?

To study the mechanisms of inheritance of chromosome rearrangements, I focused on chromosome fusions, particularly Robertsonian fusions. Robertsonian fusions (see introduction section 1.3.2.2) reduce the chromosome number within a species. A Robertsonian chromosome is formed by the creation of DSBs and then a fusion between two acrocentric chromosomes. It has been reported that chromosomal fusions in mice can occur in nature and one such occurrence is the Robertsonian fusion of chr6 and chr16. A mouse line containing this homozygous chromosomal fusion is available from the Jackson laboratory (Rb(6.16)24Bnr), strain number 000885 (see section 2.1.1). This strain was derived from wild mice that were originally captured in Southern Italy and back crossed to produce homozygous Rob mice. Interestingly, this Robertsonian chromosome shows transmission ratio distortion, i.e., it is under transmitted when heterozygous male mice are mated to wild type females (339) (Figure 5-1).



Complete pairing around centromere
No SNP accumulation
No epigenetic silencing
WT
chr16
chr6
SNP accumulation at region of reduced recombination
or
Epigenetic silencing (Histone modifications)
Het
Hom
Complete pairing around centromere
SNPs present that accumulated in heterozygote
No epigenetic silencing
Spermatocyte
50:50 Mendelian ratio
Het mice mated to WT female
Spermatid
(Pre-cursor to mature sperm)
Offspring
under transmission of the fusion Chromosome up to 2.7-fold
70.5% Wild type
26.2% Robertsonian

Figure 5-1: Summary of the transmission ratio distortion observed when mating heterozygous Rob6.16 mice to wild type females.

Using this homozygous Rob6.16 mouse strain we generated heterozygous mice to investigate how chromosomal fusions can influence their own transmission. Two hypotheses might explain the transmission distortion: i) accumulation of SNPs within the region of reduced recombination around the fusion site, if these SNPs occur within key genes involved in fertilisation, then this may reduce the transmission of the fused chromosomes; or ii) epigenetic silencing of genes around the fusion point, again if key genes involved in fertilisation are affected then this could influence the transmission of the fusion chromosome.

To identify the possible mechanism explaining the transmission ratio distortion in these Rob mice, we devised the following aims:

1. To assess the impact of reduced recombination and the accumulation of deleterious SNPs in genes surrounding the fusion points. To do so, we:
    a. determined the size of the region of reduced recombination (called from now on region of interest (ROI)) at the start of the Robertsonian6.16 fusion.
    b. investigated the effect of SNPs in genes within the ROI that have a key role in fertility.

2. To determine the existence of epigenetic silencing of genes nearby the fusion point. We:
    a. Compared gene expression levels in spermatids between wild type, heterozygous Rob6.16 and homozygous Rob6.16 mice.
    b. Identified which allele was expressed in the heterozygous Rob6.16 mice.

5.1 Defining the region of interest in the Robertsonian 6.16 heterozygotes

Any variant – whether genetic or epigenetic – that is capable of triggering transmission ratio distortion and thus biasing the inheritance of the Rob fusion chromosome must necessarily be tightly genetically linked to the fused centromere. Thus, we wished to restrict our search for such causative variants to the immediate vicinity of the fused centromere. In principle, identifying this region of interest can be achieved by looking at the nucleotide diversity in the vicinity of the fused centromeres. When Robertsonian translocations are polymorphic within a population and thus regularly present in heterozygous animals, this will lead to suppressed recombination in the vicinity of the fusion. This in turn will lead to a lack of repair of putative mutations and an increase of SNPs in these regions. Mutations that are tightly linked to (private to) the fused haplotype are candidate causative mutations for non-Mendelian inheritance. However, for this wild-derived lab strain we do not have access to population-level data that would allow us to identify the true window of recombination suppression around the centromere.

Instead, we made use of the fact that the original wild-caught mice with the Rob6.16 fusion were substantially diverged from the *Mus musculus* laboratory strain, with the divergence spread throughout the genome as shown through our SNP and nucleotide diversity analysis (Figure 5-2). This founder individual was then introduced to the lab, followed by back crossing to a laboratory mouse strain. This resulted in mosaic offspring with genomes that were partly derived from the Rob6.16 founder and part lab strain derived. These mosaic offspring were finally inbred by brother-sister mating to establish the final wild-derived line that was homozygous for the fusion chromosome. Thus, by comparing the genome of the final wild-derived Rob6.16 line to the reference genome, we can identify "windows" of high and low divergence (Figure 5-2), representing those parts of the genome that derive from the founder animal or from a laboratory origin respectively.

Since the line was selected to retain the Rob fusion chromosome, the fused centromere present in the final mosaic wild-derived line must ultimately come from the Rob6.16 founder genome. Thus, we predicted that the proximal parts of chr6 and chr16 nearest the fusion would have high divergence from the reference *Mus musculus* genome, and that this window of high divergence must contain the genetic and/or epigenetic variants responsible for non-Mendelian inheritance of the Rob fusion chromosome in laboratory crossing experiments. To identify high confidence SNPs that distinguish the Rob6.16 line from the reference genome, we sequenced at 27X coverage two Rb6.16 homozygous mice and focused on SNPs identified as being homozygously present in both animals and concordant between the two animals sequenced.

Table 5-1: Number of confident SNPs and affected genes in the homozygous Rb6.16 mice.

	Total		**Chr 6**	**Chr 16**
	Mouse 1	**Mouse 2**		
Raw number of SNPs	3,087,583	3,095,062		
High quality SNPs	2,197,679	2,224,878		
Concordant SNPs between animals	1,950,753		217,307	102,661
SNPs in protein-coding genes			1,086	365
No. of protein-coding genes with SNPs			244	141
No. of protein-coding genes with SNPs within ROI			50	125

A total of approximately 3.1 million raw unfiltered SNPs were identified using GATK (Table 5-1). After removing low quality SNPs, we retained ~71 % of the raw data. Concordant SNPs in both the samples were then calculated, representing ~ 1.95 million SNPs. The tool SnpEff (272) was then used to filter the SNPs to only those that were annotated as protein coding SNPs of high, moderate, or low SnpEff category (see methods).

To identify the region of increased SNP density, we used two approaches. First, we created 1Mb non-overlapping windows in the mouse genome and calculated the density of concordant high-quality SNPs in each window (Figure 5-2). Then, we estimated the nucleotide divergency across the genome in 1Mb non-overlapping windows using VCFtools –window-pi. Both analyses showed that at the start of chr6 and chr16 there was approximately a 27Mb region of increased SNP density and high nucleotide divergence compared to the rest of the genome (Figure 5-2). These regions will subsequently be referred to as the region(s) of interest (ROI). To assess whether the ~27Mb regions were significantly enriched in SNPs we performed a simulation test. First, we calculated the mean number of SNPs in the 27Mb regions of interest in chr6 and chr16 from the concordant homozygous GATK hard filtered SNP (56,039 SNPs). Then we randomly simulated two genomic regions of 27Mb across the mouse genome and obtained their mean number of SNPs. This simulation was repeated 10,000 times. Only in 111 permutations did we obtained ≥ 56,039 SNPs in the simulated regions, giving a p-value of 0.01. This shows that there is a statistically significant accumulation of SNPs within the ROI, relative to random regions of the same size in the rest of the mouse genome. At the genome-wide level, using the VCFtools windows-pi data, I estimated the percentage of the Rob6.16 genome that was of low diversity (<10 variants per Mb window). Approximately 50% of the Rob6.16 genome was of low diversity. This indicates that ~50% of the Rob6.16 mouse line genome is lab-derived and 50% is derived from the original founder. We therefore deduced that during establishment, the line was inbred from the F1 generation onwards, immediately after introduction to the lab.

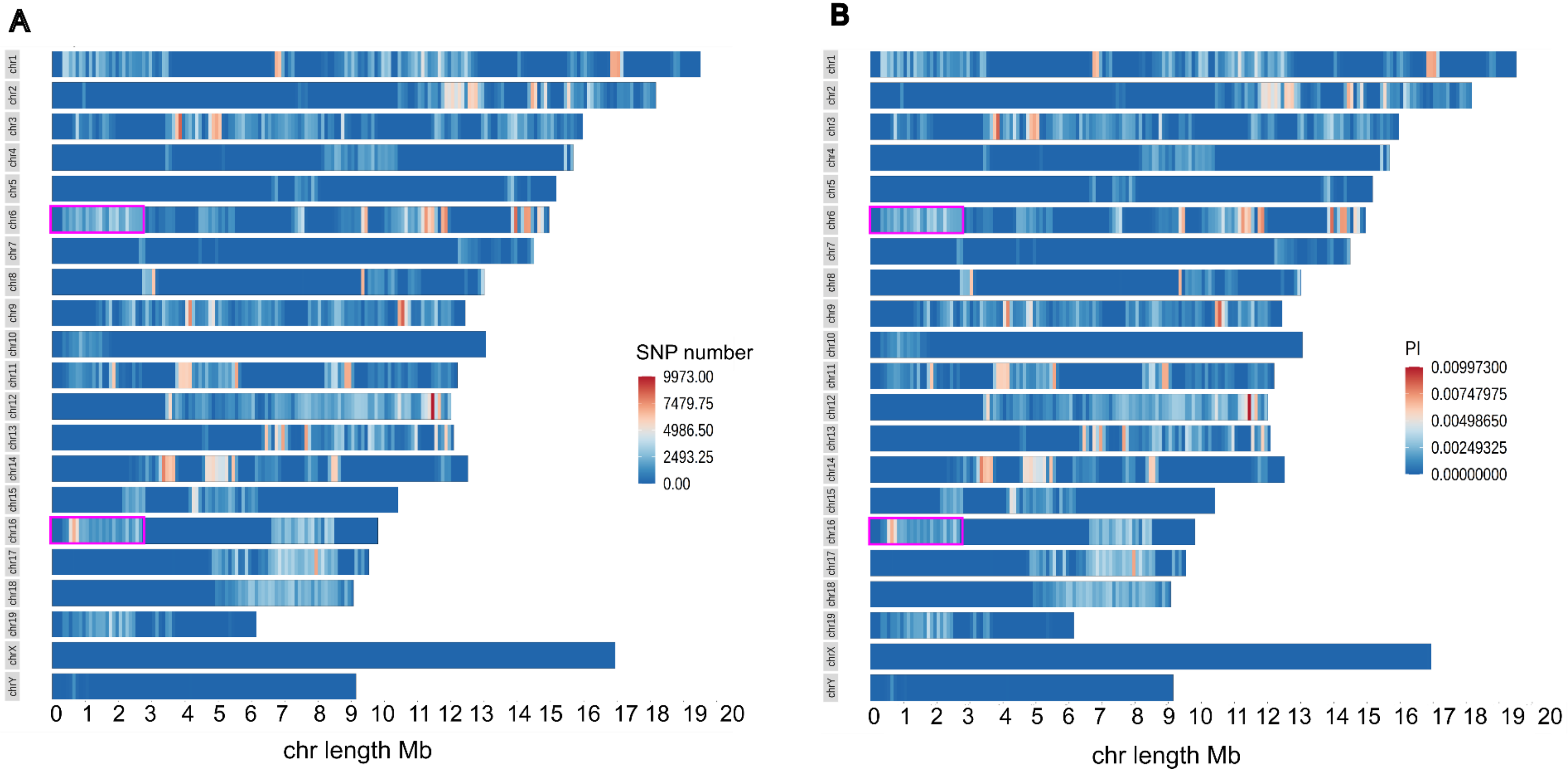


Figure 5-2: Bar plot showing the SNP density and nucleotide diversity (PI) in 1Mb windows for concordant filtered SNPs from two Robertsonian6.16 mice.
SNP density is shown in plot **A** and nucleotide diversity in plot **B**.
Blue shading represents windows of low SNP density/nucleotide diversity and red shading represents 1Mb windows with the highest SNP density/ nucleotide diversity. A ~27Mb region (purple boxes) at the start of chr6 and chr16 (near the centromeres) clearly shows increased SNP density/nucleotide diversity, representing the putative linkage block (region of interest (ROI)). The ROI obtained by the two different methods (bedtools intersect with SNP number and vcftools -windows-pi) is the same size.

5.2 Investigating the impact of deleterious mutations in the Rb6.16 mice

To understand whether the SNPs found in the ROI were affecting genes related to fertility, we identified the SNPs classified as having a high, moderate, or low effect in SnpEff (Supplementary Table 8-11 and Supplementary Table 8-12 show genes within the ROI with missense or stop-gained mutations). On chr6 there were 87 protein coding genes within the ROI. Of these 87 genes none presented high impact SNPs, while 18 genes contained moderate impact SNPs (Figure 5-3) and 53 genes contained low impact SNPs. The ROI on chr16 was more gene dense with 235 protein coding genes. Of these 235 genes there were two genes with high impact SNPs, 63 genes with moderate impact SNPs (Figure 5-3) and 117 genes with low impact SNPs. Interestingly the *Slx4* gene on chr16 has the highest number of missense SNPs (7 in total). Gene ontology enrichment analysis using the 83 genes (within the ROI) with high or moderate impact SNPs in Panther (268) did not provide any statistical overrepresentation of terms related to biological processes.

However, we found that in the homozygous mice some mice had an abnormally small testis. When preparing the three homozygous Robertsonian mice for flow sorting it was observed that out of three mice studied, one testis out of 6 was abnormally small (32mg testis weight, compared to an average of 78.4mg for the other 5 testis), this is potentially due to the missense mutations within *Slx4* as previous research (340) has shown that when this gene is mutated it can lead to disruption of spermatogenesis. Bernstein *et al* 2010 have shown that the sperm produced from *slx4*-/- mice are poorly motile with acrosomal abnormalities (341).

Mammalian spermatogenesis includes a long haploid stage with extensive gene expression, but gene products can be shared through spermatid cytoplasmic bridges. This transcript sharing decreases phenotypic differences between individual haploid sperm. Recently, a study on cow, mouse and human showed that a significant proportion of genes exhibit allelic bias linked to the haploid genotype of the cell, termed genoinformative markers (GIMs) (185). These GIM genes, which are not shared between spermatids, could cause transmission ratio distortion as this may lead to phenotypic differences between sperm carrying different alleles of a gene. From the 322 genes within the ROI of chr6 and chr16, only 22 contained high and moderate effect SNPs and were GIMs (Figure 5-3, Supplementary Table 8-11 and Supplementary Table 8-12). We then investigated the function of the genes within the ROI containing high and moderate effect SNPS using the NCBI gene webpages and GOnet (342) for genes with possible roles in sperm motility or fertilisation. We found seven genes are associated to fertility, including *Pla2g10*, *Prm2*, *Spam1*, *Hyal6*, *Slx4*, *Rimbp3* and *Tekt5*, of which only *Pla2g10* and *Prm2* are GIMs. While individual genes are discussed further in section 5.4, I note at this point that we were able to not only replicate the finding of DeLeon *et al* that Spam1 is mutated in this line, but also implicate other genes with functions during the acrosome reaction. Particularly interesting is the phospholipase A2 group 10 (s*Pla2g10*) gene on chromosome 16, since this is not only known to affect fertility but is also more confidently called as a GIM than *Spam1*. In the Rob6.16 line *Pla2g10* has two missense mutations in exon three (p.Met109Ile and p.Tyr72His). PLA2G10 is a member of the phospholipase A2 family of lipolytic enzymes that hydrolyses glycerophospholipids to produce free fatty acids and lysophospholipids. Escoffier *et al* (2010) (343) have shown that this phospholipase has a role in the acrosome reaction and controls fertility outcomes in mice. Therefore, the presence of mutations in the s*Pla2g10* gene could impact acrosomal reaction efficiency and therefore the fertilising ability of the sperm containing the Rob6.16 fusion.

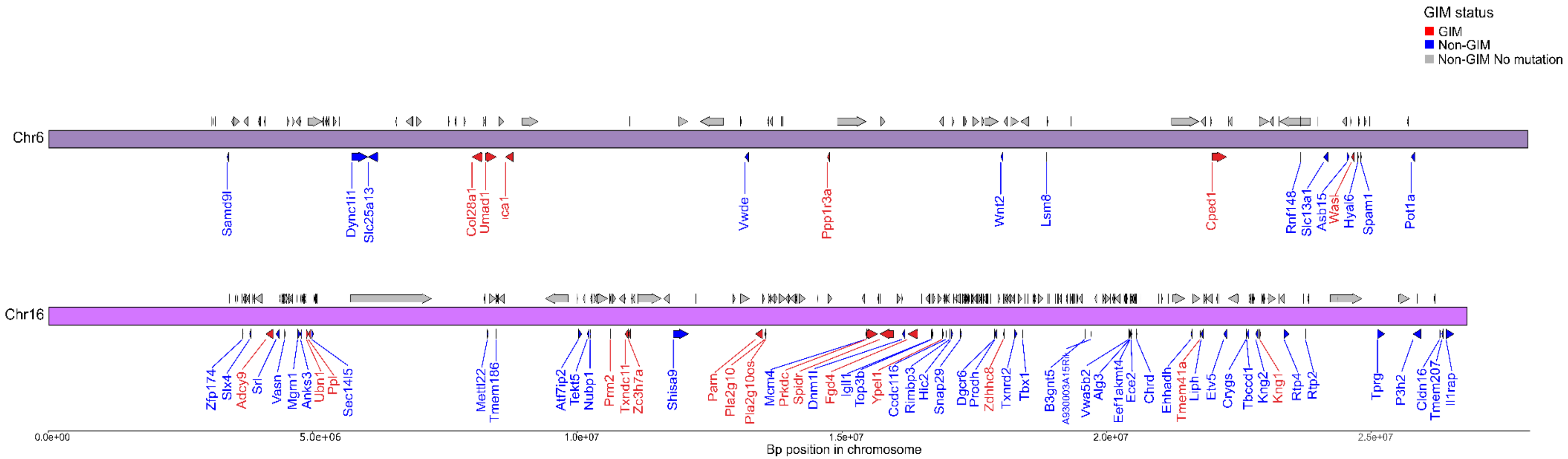


Figure 5-3: The ROI of chr6 and chr16 showing genes with high and moderate effect SNPs and their GIM status (red and blue arrows).
Genes shown in grey are genes which are not genoinformative, but may have low effect SNPs (e.g., synonymous) or no SNPs. Genes shown in red are GIMs and genes shown in blue are Non-GIMs. The total number of protein coding genes within the ROI was 87 for chr6 and 235 for chr16.

Moving on from SNPs to consider other types of mutation, in the ROI, we identified four INDELS affecting three genes in chromosome 6, and seven INDELS within six genes in chromosome 16. Four of these were classified as high, six as moderate and only one as low effect according to SnpEff (Table 5-2). As above, we investigated the role of these genes in sperm motility and fertilisation. From the nine genes with INDELS, two (*Glcci1* and *Prm3*) are related to fertility, and only one is a GIM. *Glcci1* (glucocorticoid induced transcript 1) contains an INDEL in the Rob6.16 mice and is labelled as a GIM (185). This gene is expressed in mouse testis and has a long and a short form (344), however the INDEL is a splice region variant with predicted low effect so was not considered for further investigation. Interestingly, the Protamine-3 (*Prm3*) gene contained an in-frame deletion of three nucleotides. PRM3 has been shown to have a role in sperm motility (345). Therefore, it is possible that this INDEL could affect sperm motility in the sperm carrying the Robertsonian chromosome, but further research or modelling would be required to determine if gene function is disrupted.

To summarise, we found seven candidate genes with possible roles in fertility that contain high or moderate effect SNPs within the ROI of chr6 and chr16. These SNPs could possibly be the cause of the observed TRD when heterozygous males are mated to wild type females. However, it is not clear which gene (or combination of genes) is the strongest candidate and some genes could have a yet undiscovered role in spermatogenesis or fertility. However, as PLA2G10 has already been shown to have a role in the acrosome reaction and control fertility outcomes in mice (343), this gene warrants further consideration.

This preliminary analysis has focused on genes with a role in fertility. However, other genes without a role in fertility could be equally as important and will warrant further investigation, but this is beyond the scope of this thesis.

Table 5-2: Table of protein coding genes containing INDELS within the ROI of chr6/chr16. (*Indicates that the testis RPKM value was obtained from the Kassemann Evo devo app (346) rather than the NCBI website (https://www.ncbi.nlm.nih.gov/).
The Bhutani *et al* 2021 confident GIM data was obtained from: (185).

Chr	**Gene ID**	**Gene name**	**Fertility related**	**Mutation type/ predicted effect**	**Testis RPKM**	**Bhutani data confident GIM**
6	ENSMUSG00000029638	*Glcci1*	yes	Splice region variant/ LOW	19.83	yes
6	ENSMUSG00000041390	*Mdfic*	No	Frameshift variant/ HIGH	0.14	No
6	ENSMUSG00000000416	*Cttnbp2*	No	Two Conservative in frame insertions/ MODERATE	0.35	No
16	ENSMUSG00000039427	*Alg1*	No	Frameshift variant/ HIGH	8.40	No
16	ENSMUSG00000022686	*B3gnt5*	No	Conservative in frame insertion/ MODERATE	0.03*	No
16	ENSMUSG00000022504	*Ciita*	No	Frameshift variant & start lost/ HIGH Disruptive in frame deletion/ MODERATE	0.01*	No
16	ENSMUSG00000068663	*Clec16a*	No	Conservative in frame deletion / MODERATE	2.97	No
16	ENSMUSG00000022534	*Mefv*	No	Frameshift variant/ HIGH	0.01	No
16	ENSMUSG00000050058	*Prm3*	Yes	Disruptive in frame deletion/ MODERATE	470.82*	No

5.3 Investigating epigenetic silencing in heterozygous Rob6.16 mice

Next, we investigated whether there was post-meiotic epigenetic silencing of genes in heterozygous Rob6.16 mice at the pachytene stage that was maintained in round spermatids. Prior to carrying out ChIP-seq or Cut&Tag to look at epigenetic silencing, we first studied bulk gene expression in spermatids to determine whether there was any detectable silencing at the fusion point. To do so, we performed RNA-seq experiments and established gene expression within the ROI between wildtype, heterozygous Rob6.16 and homozygous Rob6.16 mice. Three animals of each genotype were included in further experiments.

First, we confirmed the genotype of the animals chosen for the experiment. Using the genotyping PCR assay designed in section 2.1.4, the expected banding pattern for each sample was obtained (Figure 5-4).

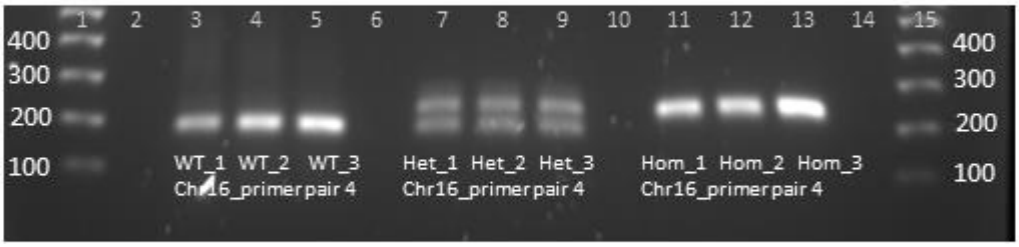


Figure 5-4: PCR genotyping of the spermatid samples sent for RNA-sequencing.
The numbers against the ladder shown on the far left and far right are in increments of 100bp. Lanes 3-5 contain the wild type (WT) C57BL/6 samples 1-3, showing one band with the expected size of 184bp. Lanes 7-9 contain the Heterozygous Rob6.16 samples, (Het) 1-3, with two bands of the expected sizes of 184bp and 224bp. Lanes 11-13 contain the Homozygous Rob6.16 samples, with the expected size of 224bp.

We required highly pure spermatids for the RNA-seq experiments, these were obtained through FACS of a dissociated testis. We sorted a minimum of 1.2 million cells per animal in aliquots of 300,000 cells as well as a smaller aliquot (~200,000 cells) to use for immunofluorescence staining as a purity check.

For the homozygous Rob6.16 samples the purity was between 98-100% round and elongating spermatids. For the heterozygous Rob6.16 samples the purity was between 97.5-98.5% round and elongating spermatids, while a purity of 99-100% round and elongating spermatids was achieved for the wildtype samples. A total of 600,000 spermatids from each sample were used for RNA extraction, giving a yield of 1.3 to 1.6 µg in the homozygous samples, 1.1-1.9 µg in heterozygous Rob6.16 mouse and 1.4-1.7µg in the wildtype samples. The RNA obtained from the sorted spermatids was sent to Novogene (Cambridge, UK) for RNA-seq library prep (see methods section 2.3) and Illumina sequencing. Between 19 and 30 million reads were obtained for all the samples, with more than 89% of the reads mapping uniquely to the mouse genome (Table 5-3).

Between 21,399 and 24,052 genes were expressed in mouse spermatids in homozygous Rob6.16 and heterozygous Rob6.16 mice, representing 37.38% and 42.5% of all genes in the genome, respectively. This high percentage is expected, as testis is a highly transcriptionally active tissue in which a broad range of genes are expressed (347). To cross compare gene expression among samples, we used DESeq2 (278) which internally accounts for differences in library size using the median of ratios method. Then, we further normalised our data by releveling gene expression to the wildtype sample.

Table 5-3: RNA-sequencing statistics.

Spermatid sample	Raw trimmed Reads per sample- (STAR input)	STAR uniquely mapped reads	% uniquely mapped reads	Number of genes expressed*
WT_1	30,114,344	27,081,031	89.93	23,212
WT_2	25,245,826	22,418,407	88.80	22,441
WT_3	21,750,837	20,601,396	94.72	22,153
Het_1	30,091,298	28,504,974	94.73	24,052
Het_2	24,790,642	23,490,142	94.75	23,057
Het_3	20,729,452	19,613,442	94.62	22,347
Hom_1	26,844,264	24,245,226	90.32	22,588
Hom_2	19,406,956	17,647,482	90.93	21,399
Hom_3	28,918,843	26,788,478	92.63	23,109

*A gene was classed as expressed if it had a feature count of >10.

5.3.1 Comparing gene expression in the different genotypes

After normalisation, we performed differential gene expression analysis between the different genotypes (wildtype vs heterozygous and heterozygous vs homozygous Rob6.16 animals) using DESeq2 (Supplementary Table 8-9 and Supplementary Table 8-10). A total of 3,884 genes were significantly differentially expressed when comparing WT to heterozygous Rob6.16 mice, while 4,614 genes were differentially expressed comparing heterozygous to homozygous Rob6.16 mice (Table 5-4). Because we want to study the possible effect of epigenetic silencing in pachytene due to asynapsis in the heterozygous Rob6.16 mice, we particularly focused on those genes whose expression is lower in the heterozygous animals compared to both wild type and homozygous Rob6.16 mice (i.e. where there is potentially transcriptional silencing specific to the heterozygous males). We found 1,994 genes that showed a lower expression in the heterozygous animals compared to wildtype, with seven and 30 of these located in the ROI of chr6 and chr16, respectively. While 2,359 showed a lower expression in heterozygous compared to homozygous

Rob 6.16 mice, only 23 and 31 were located in the ROI of chr6 and chr16, respectively. Looking at both comparisons, we found that overall, 687 genes were significantly lower expressed in the heterozygous Rob6.16 animals than both the other strains, with only one and seven genes found in the ROI of chr6 and chr16, respectively (Table 5-4). We then determined whether these downregulated genes were shared between the cytoplasmic bridges of spermatids i.e., if they were a genoinformative markers (GIM) according to previous publications (185). Only two genes were labelled as confident GIM genes: Protamine2 (*Prm2*) and ubiquitin-conjugating enzyme E2 variant 2 (*Ube2v2*). *Ube2v2* did not contain any missense mutations within the Rob6.16 fusion chromosome but was annotated with upstream, downstream and splice region variants, while *Prm2* contained a missense mutation and was annotated with upstream and downstream variants.

Table 5-4: Summary table of the differentially expressed genes among genotypes.
For a full list of genes, see Supplementary Table 8-9 and Supplementary Table 8-10. The GIM data was taken from Bhutani *et al* 2021 (185).

	Genome-wide		**ROI chr6**		**ROI chr16**	
Comparison	**Total**	**GIM**	**Total**	**GIM**	**Total**	**GIM**
WT-Het DEG (either direction)	3,884	840	28	6	70	15
WT-Het, lower in het	1,994	405	7	1	30	7
Het-Hom DEG (either direction)	4,614	847	31	7	68	10
Het-Hom, lower in Het	2,359	483	23	5	31	8
Het lower in both comparisons	687	127	1	0	7	2

*DEG=differentially expressed gene.

Finally, to determine if the normalized read counts from the DESeq2 data set were lower for the region of interest compared to random genomic regions of the same size, we performed a simulation test. First, we summed the reads for genes within the region of interest on chr6 and chr16 in each Rob6.16 heterozygous replicate. Then we simulated 20 random regions of a mean size of 27 Mb across the genome using the *randomRegions* function from RegioneR (265). BioMart was then used to obtain the genes present within these simulated random regions. From the DESeq2 dataset the number of reads within these simulated regions was summed and the average read count of the simulated regions calculated. No statistical differences were found when comparing observed read count between wildtype and heterozygous Rob6.16 mice (unpaired t-test, p-value= 0.074 and 0.161 for chr6 and chr16, respectively) or between homozygous and heterozygous Rob6.16 mice (p-value= 0.2555 and 0.5363 for chr6 and chr16, respectively). Supporting our findings, no read count differences were found between simulated regions among the three genotypes (Table 5-5).

Overall, our results indicate that there is no widespread decreased gene expression within the ROI in the heterozygous mice. Given that Rob fusions are widely documented to exhibit asynapsis during pachytene and consequent meiotic silencing, there are two possibilities for the lack of silencing in spermatids. Firstly, it may be that asynapsis / silencing is transient, and that synaptic adjustment during late pachytene permits reactivation of the chromatin prior to the round spermatid stage. Alternatively (and non-exclusively), cells which fail to achieve synaptic adjustment and thus show continued silencing may undergo apoptosis and thus be eliminated before they reach the round spermatid stage. Importantly, the initial work showing continued maintenance of silencing into pachytene was performed on animals with a chromosomal insertion – i.e. extra gene copies for which silencing is not expected to be cell-lethal. In either case, with no widespread silencing of the pericentromeric regions in this model, we concluded that the documented TRD is unlikely to be mediated by epigenetic changes and therefore the initially planned ChIP-seq / Cut&Tag work was not necessary.

Table 5-5: DESeq2 counts for the actual and simulated ROI for the different genotypes and unpaired T-test comparisons.

	Actual chr6 ROI	Actual chr16 ROI	Simulated*
WT_1	131,523.6	796,929.6	318,261.6
WT_2	111,947.8	745,994.2	304,895.2
WT_3	116,801.1	787,066.5	305,540.5
Het_1	100,290.6	768,232.1	293,159.2
Het_2	106,929.5	947,012.8	306,541.3
Het_3	107,505.3	904,149.8	302,125.0
Hom_1	106,623.9	881,965.1	318,345.1
Hom_2	95,467.2	829,536.0	309,009.5
Hom_3	94,835.2	784,680.2	309,414.3
P-value WT-Het	0.0743	0.1606	0.2016
P-value Het-Hom	0.2555	0.5363	0.0794

*Values shown are the average count of the 20 simulated regions.

5.3.2 Assessing allelic imbalanced expression within the heterozygous Rob6.16 mice

So far, no gene expressions differences are seen in the region surrounding the fusion point in the heterozygous Rob6.16 mice, indicating that either post-meiotic silencing is not carried over to spermatids or a complete lack of silencing in these surrounding regions. However, this lack of differences in gene expression could be masked by a strong imbalance in allelic expression within the heterozygous Rob6.16 mice, i.e., the wildtype alleles could be favoured and mask the silencing of the alleles present in the fused copy of the chromosomes. To assess this, we identified allelic expression imbalance within the heterozygous mice. We calculated the ASE (allele-specific expression) score for each SNP identified previously (section 5.1) from the heterozygous Rob6.16 RNA-seq allele-specific expression data (see section 2.17). Then we calculated the global ASE score for each gene. An ASE score of 1 indicates complete allelic imbalance while a value of 0.5 indicates equal expression of both alleles. Looking then at the level of expression of each allele, we determined the direction of the imbalance.

First, 1,950,753 high confident SNPs differentiating both copies of the chromosomes were used. To identify which SNPs showed an imbalanced allelic expression we used the GATK ASEReadCounter tool, and a binomial test to assess whether the imbalance is statistically significant from the null expectation of 0.5 (see methods).

There were 364,637 SNPs genome wide in the ASE analysis, of which 8,798 were in the ROI of chr6 and 9,460 in the ROI of chr16. Genome wide 11,748 SNPs had a significant ASE score, within the ROI 180 SNPs presented a significant allelic expression imbalance in chr6 and 352 in chr16 (Figure 5-5). These SNPs were in 22 annotated genes within the ROI of chr6 and 57 annotated genes within the ROI of chr16 (Supplementary Table 8-13 and Supplementary Table 8-14). Genome wide there were 3,985 SNPs with a significant ASE score of 1, 36 were in the ROI of chr6 and 80 within the ROI of chr16.

However, this analysis has limited power when looking at individual SNPs. A more comprehensive picture of skewed transcription can be obtained by analysing together all the SNPs that lie within the same gene transcript. This is visually evident in Figure 5-5 where "stacks" of adjacent SNPs all show ASE at a similar level, indicating preferential expression of the entire transcriptional unit from the WT or Rob fusion haplotype. Genome wide, 188 genes had a significant ASE score > 0.5, with 3 and 8 genes within the ROI of chr6 and chr16 respectively (Table 5-6). Genome wide 26 genes had a statistically significant ASE score of 1 (none of these were in the ROI of chr6 or chr16).

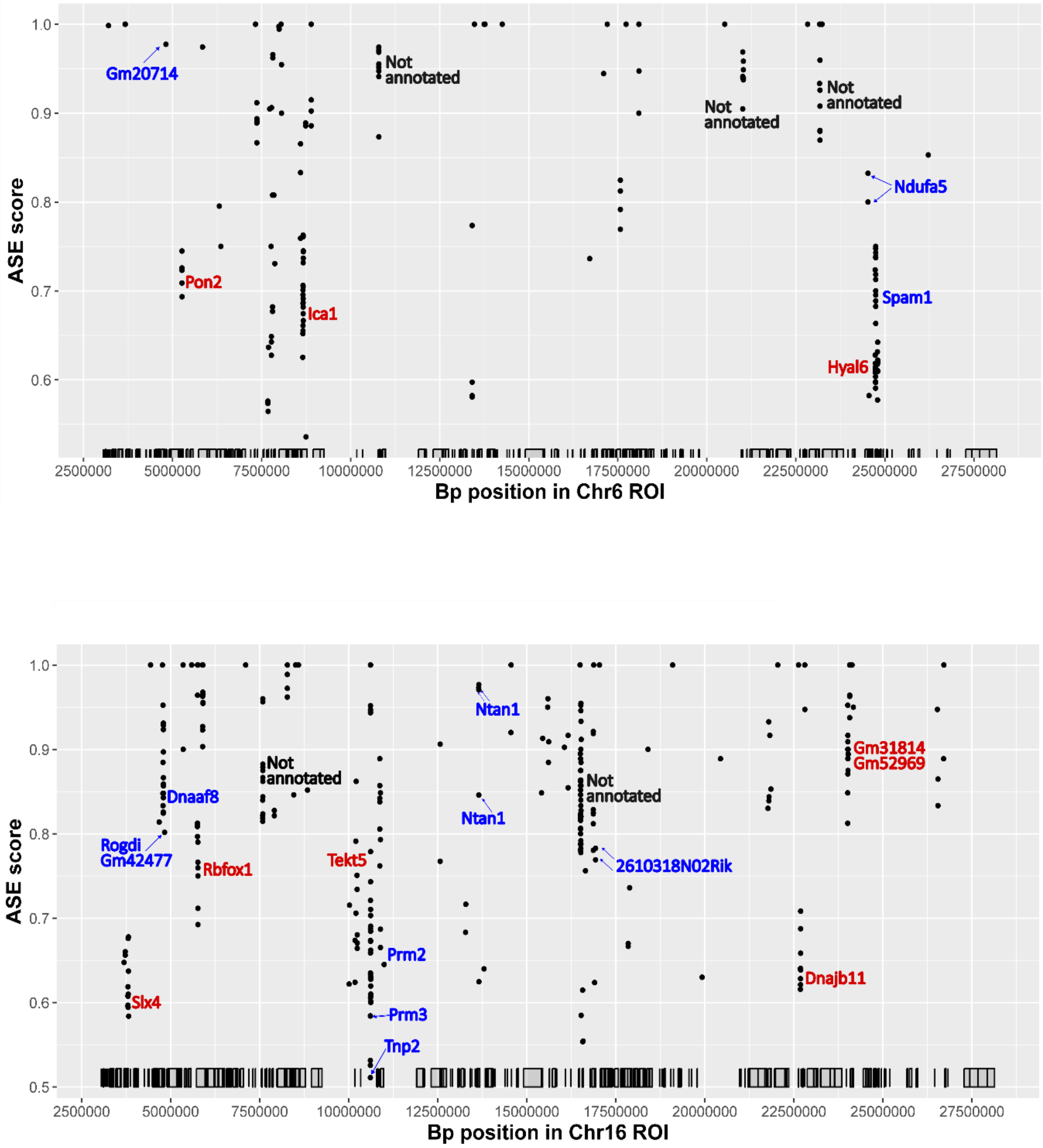


Figure 5-5: Statistically significant ASE scores shown per SNP for the ROI of chr6 and chr16.
Black and grey rectangles depict the genes within the ROI. The genes labelled in blue have significant ASE scores in a per gene ASE analysis. The gene names in blue labelled with arrows are significant on a per gene analysis but have not been deemed an ASE stack. The genes labelled in red contain SNPs with significant ASE scores but are not significant on a per gene ASE analysis. SNP stacks labelled in black text as 'Not annotated' occur in genomic regions without any gene annotation in the mm39 genome. For UCSC genome browser tracks of the different stacks see supplementary Figures 8-12 to 8-19.

We then determined the dominant expressed allele by calculating the ratio of wildtype allele expression versus total gene expression (Table 5-6). Two lncRNA presented a gene ASE score with the wildtype allele predominantly expressed. *Gm20714* (a LncRNA) contained a splice region variant and non-coding transcript exon variant, highly skewed towards the wildtype allele with an ASE score of 0.98; while *Gm42477* showed an ASE score of 0.75, with dominant expression of the wildtype allele. Interestingly, *2610318N02Rik* (RIKEN cDNA 2610318N02 gene) contained two SNPs with significant ASE scores (0.77 and 0.78) in favour of the wildtype and fusion (Rob6.16) alleles respectively, but an overall gene ASE score of 0.76, wildtype skewed.

Ndufa5 (NADH:ubiquinone oxidoreductase subunit A5), involved in the respiratory electron transport chain, presented two SNPs with significant ASE scores of 0.83 and 0.80, with a global gene ASE score of 0.81 in favour of wildtype allele. *Rodgi* (rogdi homolog), located in the nuclear envelope and involved in brain development, contained an intronic SNP with a significant ASE score of 0.8, and a global gene ASE score of 0.69, skewed towards the wildtype allele. *Ntan1* (N-terminal Asn amidase), which is involved in protein degradation, contained three significant SNPs, with ASE scores ranging from 0.85-0.97 in favour of the wildtype allele. When including all the *Ntan1* SNPs in a per gene ASE analysis this gave an ASE score of 0.86.

Spam1 (sperm adhesion molecule 1) contained 9 SNPs, including two missense mutations, with significant ASE scores ranging from 0.58-0.64 (Figure 5-6), all skewed towards the wild type allele, with a global gene ASE score of 0.61. SPAM1 is involved in the acrosome reaction and enables hyaluronoglucosaminidase activity, making it one of our candidate genes.

Finally, *Prm2* (protamine 2) on chr16 contained 20 SNPs with significant ASE scores that ranged from 0.53 to 1, all except one SNP skewed toward wildtype expression. The majority of the *Prm2* SNPs are upstream (Supplementary Figure 8-17), except a missense mutation with an ASE score of 0.53 in favour of the reference allele, a synonymous change with an ASE score of 0.53 and a 5' prime UTR premature start codon gain variant with an ASE of 0.74. Overall, the gene ASE score was 0.54, in favour of the wildtype allele. As PRM2 is a DNA binding protein which is used as a substitute for histone proteins during spermiogenesis, helping to compact the DNA into the sperm head, it was also considered one of our candidate genes.

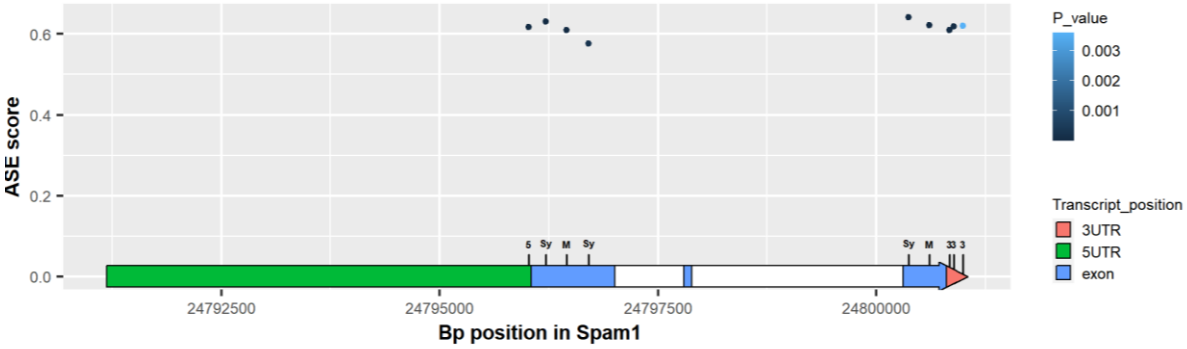


Figure 5-6: Gggenes plot of *Spam1*, showing the position of the SNPs identified through WGS. The labelled SNPs have the following classifications: 5=5'UTR SNP, Sy=Synonymous SNP, M=missense SNP and 3=3'UTR SNP. Shown are all the SNPs within the *Spam1* transcript and their ASE score.

Only three genes showed significant gene expression imbalanced in favour of the fused (Rob6.16) copy, including: *Dnaaf8*, *Tnp2, Prm3* (Table 5-6 and Figure 5-5).

Dnaaf8 (dynein axonemal assembly factor 8), predicted to be located in dynein axonemal particle, contained 12 SNPs with significant ASE scores ranging from 0.83-0.88, all annotated as non-coding transcript exon variants, with expression skewed to the fused (Rob6.16) allele. Globally, the ASE score for the gene was 0.82. *Tnp2* (Transition protein 2), involved in the exchange of histone proteins with protamines, contained one SNP (a synonymous change) with a significant ASE score of 0.51, skewed towards the fused (Rob6.16) allele. And finally, *Prm3* (protamine 3), a protein used to replace histones within the chromatin of haploid spermatids, contained two synonymous SNPs with significant ASE scores of 0.58 (fusion skewed). As both *Tnp2* and *Prm3* only contained synonymous changes, it is likely that the causative SNPs responsible for the higher expression of the Rob fusion allele are upstream or downstream SNPs, likely filtered out due to the SnpEff settings used. Because of that, *Tnp2* and *Prm3* were still considered candidate genes.

There were some SNPs that had significant ASE scores, that resulted in a stacked pattern on the SNP ASE plot (Figure 5-5 (red gene names), but when analysing on a per gene basis did not give a significant result for the gene. There could be other SNPs within these genes that did not reach significance but that could contribute to imbalanced expression. The following genes fell into this category: *Pon2*, *Ica1*, *Hyal6* on chr6 and *Slx4*, *Tekt5*, Dnajb11, *Gm31814* and *Gm52969* on chr16. *Pon2* (paraoxonase 2) may act as antioxidant (348). It contained five SNPs with significant ASE expression ranging from 0.69 to 0.74 in favour of the WT allele (Supplementary Figure 8-12) but did not give a significant ASE score in a per gene ASE analysis. *Hyal6* (hyaluronoglucosaminidase 6) contained 10 SNPs (Supplementary Figure 8-14) with a significant ASE score all in favour of the WT allele, with ASE scores ranging from 0.59 to 0.72. Two of the SNPs were missense, four were synonymous while four were not annotated with the SnpEff settings used. The HYAL6 protein is an extracellular membrane protein which assists sperm penetration through the cumulus-oocyte complex. *Hyal6* (hyaluronoglucosaminidase 6) has been shown through knockout models not to be involved in fertility or sperm characteristics (349).

The *Ical1* gene (islet cell autoantigen 1), human orthologs of which are involved in type 1 diabetes mellitus, contained 24 SNPs (Supplementary Figure 8-12) with significant ASE score. Two of these were skewed towards the fused (Rob6.16) chromosome, a missense variant with an ASE score of 0.63 and a synonymous/ non-coding variant with an ASE score of 0.54. The remaining SNPs were not annotated with the SnpEff settings used.

The *Slx4* gene (structure-specific endonuclease subunit homolog (S. cerevisiae)), located on chr16, encodes a protein which aids in the repair of DNA secondary structure. It contained seven missense SNPs (Supplementary Figure 8-15) with significant ASE scores (all in favour of the reference allele) (Supplementary Table 8-14). Interestingly, no other gene within the ROI contained as many significant missense mutations.

Tekt5 (tektin 5) is located in the sperm flagellum and plays a role in sperm motility (350). It contained five SNPs with significant ASE scores, four SNPs resulted in increased expression of the WT allele of which one was a synonymous change, one was a missense mutation and two were not annotated with the SnpEff settings used. One SNP had an ASE score of 0.86 in favour of the fused (Rob6.16) allele but was not annotated with the SnpEff settings used.

Dnajb11, (DnaJ heat shock protein family (Hsp40) member B11) is a molecular chaperone that stimulates the ATPase activity of Hsp70 heat-shock proteins to promote protein folding and prevent misfolded protein aggregation. It contained nine SNPs with significant ASE scores in favour of the WT allele ranging from 0.62-0.71. Four of the SNPs were synonymous changes and five were not annotated.

Gm31814, a non-coding RNA contained 33 SNPs with significant ASE scores all of which also overlapped with the *Gm52969* SNPs (Supplementary Figure 8-19). A total of 13 SNPs had significant expression of the WT allele (of which five were annotated as non-coding transcript exon variants and eight were not annotated with the SnpEff settings used) while 20 SNPs had significant expression of the fused (Rob6.16) allele (all of which were not annotated with the SnpEff settings used). *Gm52969* contained 38 SNPs with significant ASE scores, five of which did not overlap with the SNPs in *Gm31814*, these five SNPs had dominant expression of the fused allele.

Table 5-6: Genes with statistically significant ASE scores within the ROI of chromosome 6 and chromosome 16, showing the significant SNPs per gene.

Gene name	chr	SNP start	Ref sum	Total count	ASE score per SNP	corrected P-value (per SNP) <=0.05	Dominant allele per SNP	SNP type*	No. of SNPs per gene	Ref sum (Gene)	Total count (Gene)	ASE score per gene	corrected P-value (per Gene) <=0.05	Dominant expressed allele (whole gene)
Gm20714	6	4816230	130	133	0.98	5.58E-32	WT	SR	3	43	45	0.963	5.17E-10	WT
Ndufa5	6	24527627	129	155	0.83	3.93E-15	WT	NTA	5	52	65	0.809	2.90E-05	WT
			128	160	0.80	1.96E-12	WT	NTA						
Spam1	6	24796021	528	855	0.62	1.10E-09	WT	Nc	23	178	290	0.612	7.36E-03	WT
		24796215	545	863	0.63	2.35E-12	WT	Sy,Nc						
		24796457	597	978	0.61	8.87E-10	WT	Mis						
		24796707	523	906	0.58	2.59E-04	WT	Sy,Nc						
		24800369	476	741	0.64	1.88E-12	WT	Sy,Nc						
		24800607	395	635	0.62	1.07E-07	WT	Mis						
		24800840	468	768	0.61	1.83E-07	WT	NTA						
		24800890	276	446	0.62	4.78E-05	WT	NTA						
		24800990	173	279	0.62	3.60E-03	WT	NTA						
Dnaaf8	16	4782352	7	46	0.85	1.4E-04	Fu	Nc	41	40	221	0.817	4.55E-20	Fu
		4783716	3	26	0.88	4.3E-03	Fu	Nc						
		4783812	4	24	0.83	4.8E-02	Fu	Nc						
		4783817	4	24	0.83	4.8E-02	Fu	Nc						
		4783973	1	21	0.95	1.2E-03	Fu	Nc						
		4794042	247	1407	0.82	1.14E-137	Fu	Nc						
		4794085	257	1475	0.83	1.09E-145	Fu	Nc						
		4795886	128	894	0.86	1.14E-107	Fu	Nc						
		4795917	123	809	0.85	6.77E-92	Fu	Nc						
		4795919	128	814	0.84	2.22E-89	Fu	Nc						
		4796112	178	1262	0.86	8.45E-155	Fu	Nc						
		4796626	193	1445	0.87	4.67E-186	Fu	Nc						
Rogdi	16	4830519	93	116	0.80	5.64E-09	WT	NTA	3	51	75	0.688	4.66E-02	WT
Gm42477	16	4830519	93	116	0.80	5.64E-09	WT	NTA	2	52	69	0.752	2.12E-03	WT
Tnp2	16	10606175	36334	74297	0.51	2.83E-07	Fu	Sy	1	36334	74297	0.511	3.44E-07	Fu
Prm3	16	10608574	1523	3661	0.58	1.23E-21	Fu	Sy	2	1360	3270	0.584	2.07E-19	Fu
		10608598	1197	2879	0.58	5.21E-17	Fu	Sy						
Prm2	16	10609683	32436	61699	0.53	2.05E-34	WT	Sy	33	2154	4027	0.535	8.45E-04	WT
		10609736	34156	64277	0.53	7.12E-54	WT	Mis						
		10609988	100	293	0.66	5.73E-06	Fu	NTA						
		10610237	53	53	1.00	6.02E-14	WT	NTA						
		10610951	342	519	0.66	7.62E-11	WT	NTA						
		10610996	287	456	0.63	3.60E-06	WT	NTA						
		10611099	172	284	0.61	1.7E-02	WT	NTA						
		10611195	188	297	0.63	3.6E-04	WT	NTA						
		10611304	196	291	0.67	3.84E-07	WT	NTA						
		10611892	154	224	0.69	2.11E-06	WT	NTA						
		10611926	155	215	0.72	1.07E-08	WT	NTA						
		10611951	160	252	0.63	1.3E-03	WT	NTA						
		10612013	161	170	0.95	3.24E-34	WT	NTA						
		10612118	167	177	0.94	6.34E-35	WT	NTA						
		10612121	170	180	0.94	9.73E-36	WT	NTA						
		10612155	158	166	0.95	2.37E-34	WT	NTA						
		10613571	77	77	1.00	5.86E-21	WT	NTA						
		10613788	236	303	0.78	1.36E-20	WT	NTA						
		10613865	244	347	0.70	5.53E-12	WT	NTA						
		10613874	243	327	0.74	1.50E-16	WT	5P						
Ntan1	16	13651888	36	37	0.97	7.32E-08	WT	NTA	17	56	65	0.858	9.17E-07	WT
		13651912	33	34	0.97	4.72E-07	WT	NTA						
		13652248	851	1006	0.85	4.23E-113	WT	Nc						
2610318N02Rik	16	16931361	3114	3978	0.78	3.28E-291	WT	Sy	19	174	228	0.761	3.32E-13	WT
		16933013	9	39	0.77	3.4E-02	Fu	Sy						

***(SnpEff run with: no downstream, no intergenic, no intron, no upstream, no UTR- to limit output type to Protein coding only).**

Sy= synonymous, Mis=Missense, Nc =non-coding transcript exon variant, SR= splice region variant & non-coding transcript exon variant, 5P=5 prime UTR premature start codon gain variant. NTA = not annotated with the SnpEff settings used.
SNPs with significant ASE scores, where per gene analysis gave a non-significant result are not shown (see supplementary information: Supplementary Table 8-13 and Supplementary Table 8-14).
WT= Wild type, Fu = Rob6.16 fusion within heterozygous Rob6.16 mice.

5.4 Candidate genes that could be the cause of the TRD

In the previous sections we showed that either no widespread epigenetic silencing is occurring or that the silencing in pachytene is not carried over to round spermatids in the heterozygous Rob6.1.6 mice near the fusion point. However, in both the analysis of the effect of point mutations (section 5.2) and gene expression differences between genotypes (section 5.3.1), we have identified individual candidate genes that either have missense mutations, are selectively under expressed in the heterozygous males, have allele-specific expression in the heterozygous males, or a combination of all of these.
Next, for those candidate genes highlighted by each of these different lines of evidence, we determined which genes were genoinformative markers (GIM) using previously published data (185). As discussed previously, GIMs are genes whose transcripts are likely not shared across the cytoplasmic bridges of developing spermatids and are strong candidates for the cause of the TRD if they have a fertility phenotype. This is because genes with transcripts that are not shared between developing spermatids could lead to phenotypic differences in sperm carrying either the wild type or alternative copies, especially for genes harbouring SNPs of high and moderate effect. Finally, we examined whether there was any evidence in the literature for a functional role of the candidate genes in fertility, particularly roles acting between ejaculation and fertilisation – i.e. the stage when the TRD occurs.

5.4.1 Evidence based on identification of consequential point mutations

The ROI of chr6 contained 18 genes with high or moderate effect SNPs of which six were GIM (*Col28a1, Umad1, Ica1, Ppp1r3a, Cped1* and *Wasi)* (Figure 5-3). The ROI of chr16 contained 56 genes with high and moderate effect SNPs of which 16 were GIMs (*Adcy9, Ubn1, Ppi, Prm2*, *Txndc11, Zc3h7a, Pam, Pla2g10, Pla2g10os, Prkdc, Spidr, Fgd4, Ypel1, Zdhhc8*, *Tmem41a* and *Kng1*). Focusing on genes known to play a role in sperm motility or the acrosome reaction further reduced the number to five genes, two of which were GIMs (*Prm2* and *Pla2g10*) and three of which were not (*Spam1, Hyal6* and *Tekt5)*. While *Slx4* is not a GIM and is not known to act during fertilisation, it does however have a role in fertility and it contained seven missense mutations. Functional studies would be required to determine if heterozygous *Slx4* knockout mice bred to wild type females showed transmission ratio distortion (lack of transmission of the mutated copy of SLX4). However, as *Slx4* is not a GIM, as previously determined (185), it is likely shared between the developing spermatids. One would expect a strong candidate for the cause of the TRD to be a GIM i.e., not shared between the spermatids, to account for functional differences between sperm carrying the WT or mutant copy. However, it is possible that if sharing is incomplete then there may still be some functional differences in the spermatids that carry the mutated copy of *Slx4* on the Robertsonian chr6.16 fusion. *Spam1* and *Hyal6* have roles in fertility, but are not GIMs according to Bhutani *et al* 2021 (185). As discussed in the Introduction (section 1.8.3), single knockout mutations of either *Hyal6* (349) or *Spam1* (188) have determined that this does not result in impaired fertility and there were no reports of non-Mendelian inheritance from hemizygous null males, so these genes will not be discussed in detail but are included for completeness and consistency with previous studies in this model. *Tekt5* has recently been identified as being important for flagella formation (350) and sperm from *Tekt5*-/- mice have a lower fraction of motile cells (heterozygous animals were not studied). Sharing of *Tekt5* transcripts in cytoplasmic bridges in sperm was undetermined in published data (185).

5.4.2 Evidence based on selective under expression in heterozygous males

Studying the gene expression in round spermatids of the different genotypes, we identified those genes where the Rob6.16 heterozygous mice showed a significantly reduced expression compared to both the wildtype and the Rb6.16 homozygous mice. A total of eight genes were highlighted, with two being GIMs. *Gm43960* on chr6 had lower expression in heterozygotes, but it was not a GIM, so will not be discussed further. Seven genes within the ROI of chr16 had lower expression in the heterozygote mice (*Ube2v2*, *Gm49521*, *Ephb3*, *Gm15738*, *Tnp2, Prm3* and *Prm2*) but only *Prm2* and *Ube2v2* were GIMs. *Ube2v2, Tnp2, Prm3* and *Prm2* are genes which play a role in fertility. *Ube2v2* was called as a confident GIM by Bhutani *et al* 2021 (185). This gene is involved in the DNA damage response and in the Robertsonian fusion it contains upstream/downstream and splice region variants. While this gene did show under expression in the heterozygote, there was no allele-specific expression (see 5.4.3), thus the downregulation must affect both alleles. Therefore, although it is a GIM, because both alleles are downregulated there is no mechanistic basis for this under expression to lead to TRD. This gene will therefore not be discussed further. The *Prm3* gene contained upstream/ downstream and synonymous mutations as well as an in-frame deletion of one amino acid. It showed significantly decreased expression in a WT-Het and a Het-Hom comparison and had a significant ASE score, with dominant expression of the fused allele. Like *PRM2* and PRM1, PRM3 is a protein with a role in sperm DNA compaction and replaces the histone proteins during spermiogenesis. *Tnp2* also contained upstream/downstream and synonymous mutations and like *Prm3* had dominant expression of the fused allele in heterozygotes (Table 5-6). It has a key role in sperm DNA compaction during spermatogenesis, acting to help the removal of histone proteins and their replacement by protamines in elongating spermatids.

5.4.3 Evidence based on allele-specific expression in heterozygous males

Looking at the allelic-specific expression analysis, we identified 22 annotated genes within the ROI of chr6 and 57 annotated genes within the ROI of chr16 that one of the alleles was significantly expressed more than the other (Supplementary Table 8-13 and Supplementary Table 8-14). After determining their GIM status and whether there was any known fertility phenotype, many of the genes that were highlighted by either the mutation analysis or overall expression analysis were also highlighted by the ASE analysis. Overall, the ASE analysis added only one further candidate gene to our list. *Prm1* was found to be borderline in the ASE analysis, since it contained individual SNPs that showed ASE, but this was not significant when aggregated across the whole gene. The *Spam1* and *Hyal6* genes on chr6 contained SNPs with significant ASE scores and have a role in fertility. However, as outlined above, we will not discuss these further as there is evidence against their involvement in TRD. *Slx4*, *Tekt5, Prm3*, *Prm2* and *Prm1* genes on chr16 also contained SNPs with significant ASE scores and have a role in fertility. *Slx4* and, *Tekt5* are listed as unknown, while *Prm3* and *Prm1* are not listed and *Prm2* is a confident GIM (185).

5.4.4 Final scoring of candidate genes

After winnowing the lists of genes in the regions of interest as described above, we arrived at a final list of candidate genes for involvement in the TRD phenotype, based on the various lines of evidence (Table 5-7).

Table 5-7: Summary table of the RNA-seq and WGS data analysis showing genes with a known fertility phenotype.

Gene name	Known role in:	GIM[1]	High/moderate effect SNPs	Selective downregulation in Het (i.e. Het lower than both WT & Hom)	Significant ASE score (per gene)	Dominant allele (if ASE detected)
Pla2g10	Fertility	Yes	Yes	No	No	N/A
Prm1	Fertility	Unknown	No	No	Borderline*	Fu*
Prm2	Fertility	Yes	Yes	Yes	Yes	WT
Prm3	Fertility	Unknown	No (upstream/ downstream/ synonymous)	Yes	Yes	Fu
Spam1	Fertility	No	Yes	No	Yes	WT
Hyal6	Fertility	No	Yes	No	Borderline*	WT*
Tekt5	Fertility	No	Yes	No	Borderline*	WT*
Ube2v2	DNA damage	Yes	No (upstream/ downstream/ splice region variant)	yes	Borderline*	Fu*
Tnp2	Fertility	Unknown	No (upstream/ downstream/ synonymous)	Yes	Yes	Fu
Slx4	Fertility	No	Yes	No	Borderline*	WT*

[1] As determined by Bhutani *et al* 2021 (185)

* Significant ASE score on a per SNP analysis but not when taking all the SNPs within the gene into consideration.

Fu = Rob6.16 fusion within heterozygous Rob6.16 mice

Of these, the strongest two candidates are *Pla2g10* and *Prm2*, since they are both highlighted as GIMs i.e. escaping transcript sharing and thus more likely to trigger haploid selection effects AND have high or moderate SNP mutations that may affect fertilisation capacity. We therefore attempted a modelling analysis to examine the potential effects of the mutations in these genes.

5.4.5 *sPLA2g10*

The s*Pla2g10* gene contains missense mutations (Tyr72His and Met109Ile), has been classified as a GIM (185) and is involved in the acrosome reaction (343) (see introduction). It has been shown to be expressed in round and elongating spermatids (351) and in accordance with the round spermatid RNA-Seq data generated herein. We identified 2 missense SNPs in exon 3 (Figure 5-7) in the homozygous Rob6.16 mice. The gene expression between genotypes did not show any statistically significant differences between wildtype, homozygous Rob6.16 and heterozygous Rob6.16 mice. Moreover, allelic specific expression analysis showed no differences in expression between the wildtype and the fused copy in the heterozygous Rob6.1.6 mice (Figure 5-7). However, if the transcripts do escape sharing, and the encoded proteins are functionally different, this may be mechanistically sufficient to cause TRD without any effects on gene expression.

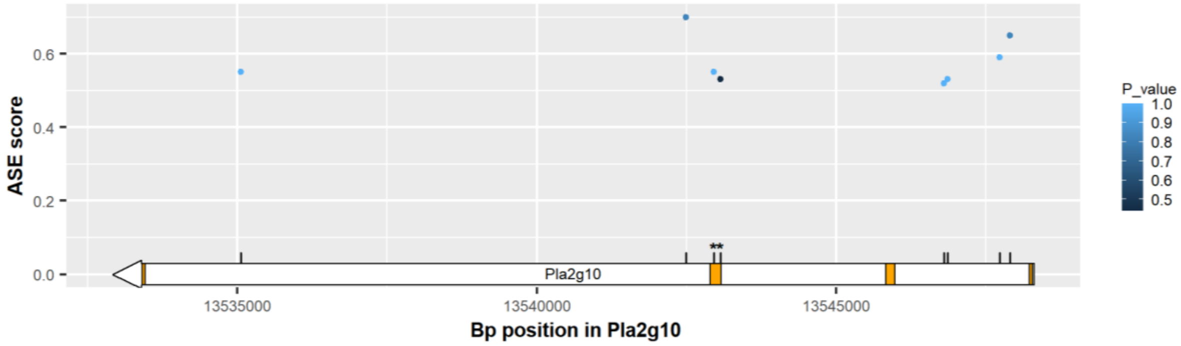


Figure 5-7: Gggenes plot of *Pla2g10* showing the location of SNPs with their ASE score.
The four *Pla2g10* exons are shown in orange. The positions of the SNPs are marked with the black lines. The SNPs labelled with * are the two missense SNPs located within exon 3.

To investigate whether the two missense mutations in exon 3 of *Pla2g10* changed the protein structure, the protein model of the *Mus musculus* reference PLA2G10 was examined within AlphaFold (276) (Figure 5-8). However, the model produced was of low confidence (pLDDT scores <70). Therefore, a protein model, modelling the missense mutations cannot be accurately generated. The structure of PLA2G10 from RoseTTAFold (352), for both the long and the short form of PLA2G10 was examined, but this also showed a low confidence model (not shown), so the effects of the *Pla2g10* SNPs cannot currently be predicted in vitro.

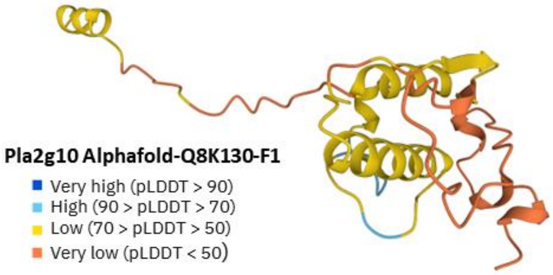


Figure 5-8: The AlphaFold structure of *Mus Musculus* PLA2G10.
The colour key refers to the the pLDDT scores, anything below 70 (yellow or orange in this model) is a low confidence prediction.

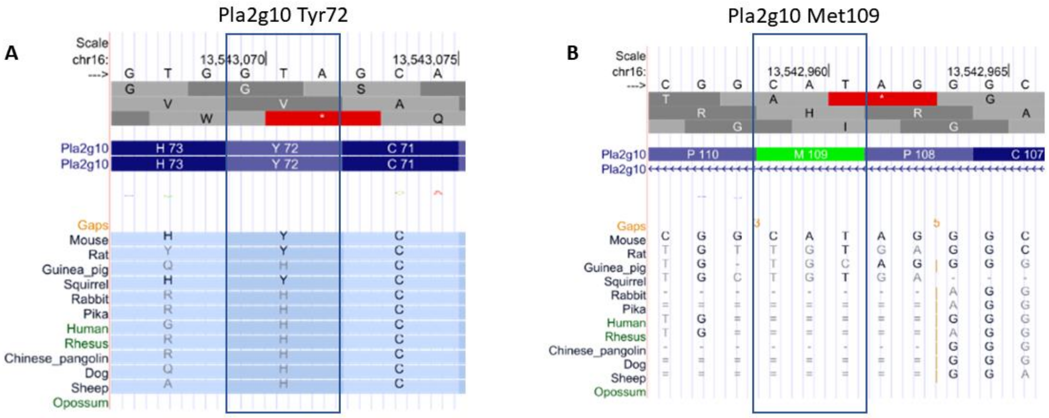


Figure 5-9: UCSC genome browser tracks showing *Pla2g10* missense mutations.
Tyr72 (panel **A**) and *Pla2g10* Met109 (panel **B**) outlined in blue rectangles, for the mm39 genome.
The conservation track is also shown.

To try to understand the possible effect of these two missense mutations (Tyr72His and Met109Ile), we investigated the conservation status of these sequences in other species (Figure 5-9). We see that for the position Tyr72, most of the orthologous proteins present a histidine. Interestingly, this is the amino acid change found in the homozygous Rob6.16 sample, suggesting that the histidine is the ancestral form, and most likely functional. The second missense mutation, Met109, encoded by codon ATG in the mouse reference, presents as the codon ACA in rat and squirrel and GCA in guinea pig. Again, this alternative codon is also the one present in the Rob6.16 mice.

If these mutations result in a non-functional or less functional PLA2G10 protein, then this could result in a scenario where if the *Pla2g10* gene product is not shared between the spermatids as suggested by Bhutani *et al* 2021 (185), then sperm containing the Rob6.16 chromosome may not contain functional s*Pla2g10* whereas sperm containing the non-fused chr6 and chr16 (without the *Pla2g10* mutation) might. This could create differences in the fertilising potential of the sperm, which could lead to under transmission of the fusion chromosome (TRD). This would need to be investigated in a knockout model as there are other secretory group two phospholipases which could compensate for mutated sPLA2G10.

5.4.6 Protamine 2

The protamine 2 gene contained one missense mutation (a T to C change at chr16 position 10609736, resulting in amino acid 43 changing from Threonine to Alanine). Thr43 is a conserved amino acid (Figure 5-10). *Prm2* was differentially expressed between mouse genotypes, with the heterozygous Rob6.16 mice showing a significantly lower expression. The missense mutation had a significant ASE score in favour of the reference allele (Table 5-6). It also contained two 5'UTR variants, as well as one 5'UTR premature start codon variant. Unfortunately, the AlphaFold (276) model of PRM2 was of low confidence and so I could not model the amino acid changes.

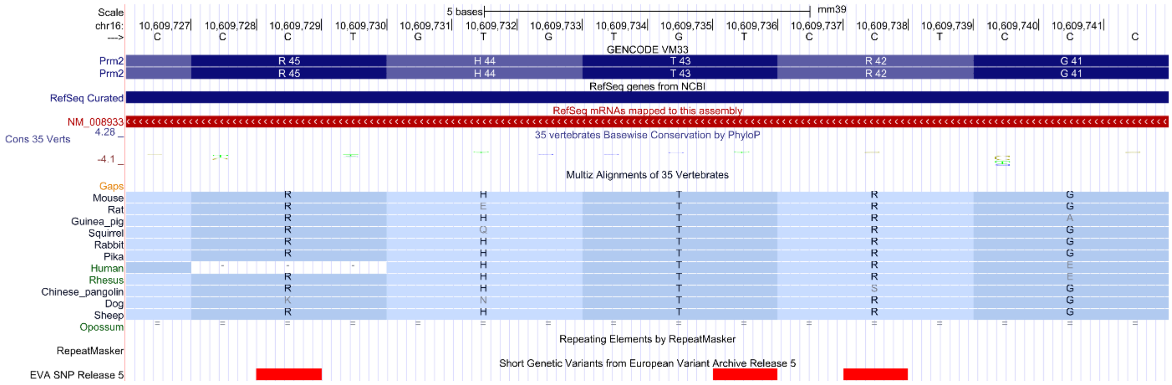


Figure 5-10: UCSC genome browser track showing T43 of Protamine 2.

Interestingly, Schneider *et al* 2016 (353) demonstrated that mice heterozygous for a *Prm2* mutation (a deletion of approx. 100bp) are fertile and have normal sperm motility and morphology, whereas *Prm2*-null mice are infertile due to a complete loss of sperm motility and abnormal sperm head morphology, therefore it seems *a priori* unlikely that the *Prm2* mutations observed are the cause of the TRD. However, depending on the degree of transcript sharing and the rapidity of incorporation of the protein into chromatin it is possible that a mutant protamine could act as a dominant negative. In other words, it is possible that in a *Prm2* hemizygous null animal, the single *Prm2* allele would produce enough transcripts/protein to allow all sperm to condense properly, while in an animal with a mutant *Prm2* allele, the WT protein would be preferentially incorporated into sperm carrying the WT allele, while mutant less functional protein would be preferentially incorporated into sperm carrying the mutant allele. We thus cannot completely rule out *Prm2* at this point.

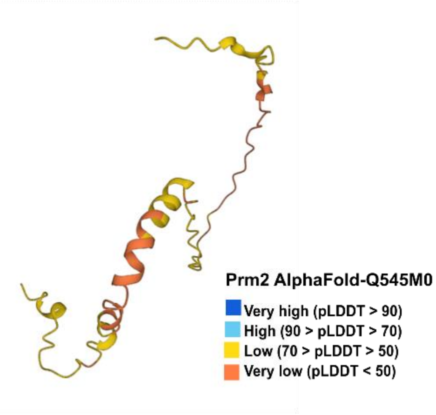


Figure 5-11: The structure of *Mus Musculus* PRM2 from AlphaFold (276).
The colour key refers to the pLDDT scores, anything below 70 (yellow or orange in this model) is a low confidence prediction.

Another interesting point to highlight here is that while *Prm2* carries mutations in the Rob fusion haplotype and the mutant allele was under expressed (with ASE) in the heterozygotes, the opposite was true for *Tnp2* and *Prm3*. These proteins are also involved in sperm packaging, with Tnp2 acting earlier in the process. Under expression of *Prm2* coupled with overexpression of earlier-acting genes could indicate a subtle selective delay in maturation of the sperm carrying the Rob fusion chromosome. In this case, cells carrying the fusion chromosome would appear to show upregulation of “early” genes and downregulation of “late” genes at the final stages of chromatin maturation. More nuanced studies could perhaps be devised to test this hypothesis; however, this would likely require a way to selectively label the different haplotypes *in situ* on testis sections to determine their respective progression at different tubule stages.

5.5 Discussion

5.5.1 A lack of gene silencing within the ROI

To confirm that we did not see silencing of genes within the ROI, we carried out three different analyses: 1) We counted the normalized DESeq2 data set reads within the three wild type samples and the three het samples and compared them using unpaired T-tests, this showed that there was no significant difference in expression between the WT and Het ROIs. 2) We looked at genes within the ROI that were downregulated in a WT-het and a het-hom comparison i.e. genes that may be silenced in heterozygotes. This list contained only eight genes, indicating that there was not widespread silencing of genes within the ROI in heterozygotes.3) We performed allele-specific expression (ASE) analysis for all hard filtered SNP variants within the aligned bam files then filtered them to the region of interest. We further analysed the data by carrying out a binomial test (with p-value adjustment) to determine which SNPs had significant allele-specific expression. We then extended the analysis to calculate ASE scores per gene, again performing a binomial test to determine which ASE scores were significant. This showed that there was not blanket reduction in expression of one allele within the region of interest. This would indicate that there is not complete epigenetic silencing of either the reference or alternative copy of genes within the region of interest i.e., post-meiotic sex chromatin repression (PMSCR). All three methods showed no silencing of genes within the ROI of chr6 and chr16. This was unexpected as Turner *et al* (2005) (199) have previously shown that silencing of unsynapsed chromosome regions can take place in the mouse during meiosis. However, Naumova *et al* 2013 (354) looked at gammaH2AX levels in carriers of Robertsonian translocation and found that "the proportion of spermatocytes with markers of meiotic silencing of unsynapsed chromatin (MSUC) at trivalents depends on both, the stage of meiosis and the number of translocations". This may therefore suggest that the cellular response to asynapsed Robertsonian chromosomes is different from that of the sex chromosomes.

It is also possible that cells with synaptic defects at the end of pachytene do not survive, and therefore we may only be observing spermatids in which synapsis has occurred and repression around the Robertsonian chromosome has been resolved. The model used by Turner *et al* (198) to show post meiotic sex chromatin repression was an insertion (an extra copy of a segment of chromosome 7 was inserted into the X chromosome), so maintaining silencing of this region would not result in lack of expression of essential genes, so the cells may have survived when otherwise they would not have.

When looking at ASE scores per gene, *Spam1* (on chr6) had an ASE score of 0.61, skewed towards the reference allele. This would suggest that the alternative mutated (Robertsonian) copy of

Spam1 is expressed at a lower level than the wild type reference allele. If the mutations render the SPAM1 protein non-functional then this could cause a reduced ability of the sperm to disperse cumulus cells from the cumulus mass, resulting in delayed fertilization solely at the early stages after insemination as described in (355). I.e., the absence of *Spam1* does not render male mice infertile, but just delays the speed at which they can fertilise an egg. Zheng *et al* 2001 (356) claimed that *Spam1* escaped sharing between the spermatids whereas Bhutani *et al* 2021 (185) did not identify *Spam1* as a GIM. Therefore, further studies will be required to determine whether Spam1 is or is not a GIM.

5.5.2 Whole genome sequencing of Homozygous Rob6.16 mice and RNA-seq analysis of wild type, heterozygous and homozygous Rob6.16 mice

In our analysis of the mice, we focused on SNPs, indels and gene expression differences that distinguish the fused vs unfused haplotypes. We tried to evaluate concordant copy number variants within the region of interest for chr6 and chr16 (using the tool CNVpytor), however only one exonic copy number variant (a frameshift deletion) was obtained for chr16 (*Tmem207*) and no copy number variants within the region of interest for chr6. Due the difficulty in obtaining accurate copy number variants from short read sequencing, copy number variants were not investigated further.

Genotyping of the homozygous Rb(6.16)24Lub mice was used to determine if there were any SNPs within the region of interest (ROI) that could be causing the non-Mendelian inheritance of the fusion chromosome from heterozygotes bred to wild type females. From the list of genes within the ROI of chr6 and chr16 we identified those that contained missense mutations, then we searched the literature to determine if any of these genes were currently known to have a fertility phenotype.

The *Pla2g10* gene on chromosome 16 was identified as being a key candidate as it contained two missense mutations within exon three and has been shown to be a component of the acrosome reaction (357) and to improve IVF success rates. Bhutani *et al* 2021 (185) concluded that the *Pla2g10* gene was likely not shared between developing spermatids. If this is the case, then this may lead to phenotype differences between sperm carrying the Robertsonian chromosome and wild type sperm. Sperm carrying the Robertsonian chromosome and a mutated PLA2G10 protein may have a less efficient acrosome reaction than wild type sperm, meaning that the wild type sperm are more likely to fertilise the oocyte. The PLA2G10 protein is secreted, this may help to explain the effect observed by Aranha and DeLeon (1992) (358) of increased transmission of the fusion chromosome when sperm are aged in an epididymis that has been tied off and is then later unblocked. Sperm containing the Robertsonian chromosome and sperm containing the wild type chromosome will be present within the epididymis. If the wild type sperm undergo premature capacitation, then this will release functional PLA2G10 protein into the epididymis, this could then act on the Robertsonian sperm to improve their fertilisation ability. Murase *et al* 2016 (359) have shown that metabolic products of PLA2G10 degradation can improve the fertilising ability of sperm.

Despite identifying the *Pla2g10* gene as a candidate gene for the cause of the transmission ratio distortion with two missense mutations, it was not significantly differentially expressed (p-value ≤ 0.05) between genotypes, and there were no significant allelic expression differences per SNP between the reference and alternative allele. Protein modelling of the impact of the missense mutations using AlphaFold (276) and PyMOL (Schrödinger) could not be carried out as the AlphaFold model was of low confidence. To determine whether the mutated PLA2G10 protein folds correctly NMR studies could be carried out. It is possible that the PLA2G10 mutated protein can fold correctly, but that is has altered kinetics, such as the turnover number. If the mutated PLA2G10 has a lower turnover, then this could possibly explain the reduced transmission ratio distortion observed when sperm from heterozygous Rob6.16 mice are aged in the epididymis. Therefore, to determine if this gene plays a role in causing the transmission ratio distortion observed within the heterozygous Rob6.16 mice, functional studies such as heterozygous knockout models would be required, or mouse models recapitulating the *Pla2g10* mutations found in the Rob6.16 mice. Several groups have mouse models with s*Pla2g10* deletions but might not necessarily have bred heterozygous males to wild type females and recorded the genotypes of all offspring produced. Moreover, functional studies would be required to determine the impact of the missense mutations on the protein function, such as receptor binding.

The protamine 2 gene (*Prm2*) contained missense mutations and was differentially expressed between the genotypes. Schneider *et al* 2016 (353) demonstrated that mice heterozygous for a *Prm2* mutation are fertile and have normal sperm motility and morphology, whereas *Prm2*-null mice are infertile due to a complete loss of sperm motility and abnormal sperm head morphology. Therefore, the mutations within the *Prm2* gene in the Robertsonain6.16 mouse are unlikely to render *Prm2* non-functional, as successful fertilisation in the homozygote can still occur. Sperm carrying the Robertsonian allele (and therefore the mutated copy of *Prm2*) may be less competitive than sperm carrying the WT allele in heterozygotes. Although Bhutani *et al* (185) predicted *Prm2* to be a confident GIM, Schneider *et al* 2016 (353) showed through immunohistochemistry staining that in *Prm2+/− mice,* PRM2 protein is found in all spermatids of seminiferous tubules, indicating that transcript sharing does occur for *Prm2*. However, it is possible that the level of transcript shared between the spermatids remains below that in wild type cells.

GO pathway analysis using the Panther db (268) was carried out for all the genes within the ROI that were differentially expressed in either direction, in WT-Het and Het-Hom comparisons, but this did not yield any significant results. Interestingly, the Spam1 and Hyal5 genes that were previously thought to be the cause of the TRD were not significantly differentially expressed (in either direction) between WT-Het and Het-Hom with a P-value cut off ≤ 0.05.

Proteins released within the epididymis may also act on the sperm to aid their fertilising potential, when sperm are aged by storage in the epididymis. The epididymis releases exosomes called epididysomes. Epididymosomes released from the epididymal epithelium contain proteins, noncoding RNAs and a distinct set of lipids that are transferred to spermatozoa while they pass through the epididymis. Skerget *et al* 2015 (360) have identified the proteins that are potentially added or potentially removed as sperm pass through the epididymis. PLA2G10 was not one of the listed proteins. One or a combination of proteins could potentially be released from the epididysomes within heterozygous Rob6.16 mice that could help to restore the in vivo fertilization potential of sperm stored (aged) with the epididymis, an in-depth analysis of possible candidates is beyond the scope of this thesis.

5.5.3 INDELs with the ROI

There were only 7 genes in total within the region of interest that contained INDELS in protein coding genes (see results). Most are likely not involved in spermatogenesis or fertilisation. However, there was a single amino acid deletion in *Prm3* (Glu59del), the effect of which on the PRM3 protein is unknown. If this disrupts the function of protamine 3 then this could have an impact on chromatin condensation and consequently sperm motility (345). Therefore, sperm carrying the Robertsonian chromosome could have reduced motility due to a mutation in *Prm3*. The AlphaFold model of *Mus musculus* PRM3 is of low confidence, so modelling could not be carried out. However, this *Prm3* mutation alone would not explain the improvement in the fertilising ability of the Robertsonian sperm when aged in the epididymis.

The *Glcci1* gene on chr 6 contained an INDEL which is a splice region variant, Takada *et al*, (2023) (344) have proposed that *Glcci1* short form, which is primarily expressed in spermatids may act as a novel anti-apoptotic mediator in mature murine testis. The *Glcci1* gene is a GIM (185). Therefore, if *Glcci1* is not shared between the spermatids and the INDEL within the Robertsonian sample results in a non-functional GLCCI1 protein, this could potentially lead to apoptosis of the sperm carrying the mutated Glcci1. However, this is unlikely to be the case as when Chayko and Martin-DeLeon 1992 (339) used the sperm from Robertsonian 6.16 mice in *in vitro* fertilization experiments, a 1:1 ratio of chromosomally normal to balanced reciprocal embryos was obtained. This suggests that chromosomally normal sperm and sperm carrying the Robertsonian 6.16 chr were produced in a 1:1 ratio, suggesting apoptosis of sperm carrying the Robertsonian6.16 chromosome does not occur.

5.5.4 Planned for (but abandoned) Cut&Tag work

We initially planned to carry out Cut&Tag work to investigate allele-specific chromatin modifications in the heterozygous males. This however was discontinued when there was no evidence of post-meiotic silencing around the fused centromeres. For the work presented here, RNA was extracted from 600,000 spermatids per mouse over two RNAqueous columns with DNase treatment, this gave ample RNA with high RNA integrity score (RIN) for un-stranded paired end RNA-seq library preparation. Alternative RNA extraction protocols were not tested, as the RNAqueous kit enabled small elution volumes which were suited to the yield of cells (600,000) obtained from the sort.

If later data indicates that chromatin studies in the heterozygotes would be useful, further work could be carried out to sort spermatids from wild type, C57BL/6 and Robertsonian 6.16 mice to carry out Cut&Tag sequencing with repressive chromatin marks, such as H3K27me3. The allelic balance of the immunoprecipitated DNA could then be used to determine whether the fused or the unfused copies of chr6 or chr16 were preferentially marked with repressive chromatin marks or whether there was an even distribution. In heterozygotes a mark of MSCI such as gammaH2AX could be measured in pachytene cells and a mark of PMSC (such as H3k9me3) could be measured in round spermatids. A windowed analysis approach could then be carried out using all the SNPs that distinguish the Robertsonian haplotype from WT C57BL/6 to determine for all reads mapping in each window which SNP are they carrying. Windows of differentiation on other chromosomes could act as negative controls. In these windows we would expect to see an even distribution of reads between the haplotypes. In the region of interest of the fused chromosomes, we may observe a skew with either the fused or the unfused copy of the chromosome being preferentially silenced.

5.5.5 Potential Future improvements to methodology

5.5.5.1 Rob6.16 Genotyping PCR reactions.

I have successfully designed PCR primers to genotype wild type C57BL/6 mice, Homozygous Robertsonian 6.16 mice and Heterozygotes. The PCR primers spanned INDELs in the Robertsonian mice. This represents a quicker approach than karyotyping and may require less starting material. Genotyping PCR was used to confirm that the mice that were sorted were of the correct genotype. The primers used spanned insertions or deletions and so were not specific to a particular combination of SNPs (i.e., they were not allele specific). Allele specific primers could be developed, with the 3’ end of the primers covering adjacent SNPs, with 5’ tails of different length so that the PCR reaction could be multiplexed. Ideally for any new primers developed a karyotype of the starting sample would already have been carried out before genotyping via PCR, so that the karyotype results could be used to corroborate the PCR results.

5.5.5.2 Alternative alignment tools or SnpEff settings

Alternative tools could be used for the RNA-seq read alignment such as MINTIE (361), this does not require a reference gft file but “combines de novo assembly of transcripts with differential expression analysis to identify up-regulated novel variants in a case sample” vs a set of controls. The SnpEff settings chosen (-no-downstream -no-intergenic -no-intron -no-upstream -no-UTR) excluded most variants which were not protein coding from annotation. It is conceivable that there may be intergenic variants present within the Robertsonian line that could contribute to the observed TRD. However, this has not been investigated within this study, but it something that could be done in future analyses.

5.5.6 Moving beyond single candidate genes: deeper biological understanding

It is possible that a single gene is not responsible for the TRD observed. A combination of mutations in several genes could act cumulatively to cause the effect, however, it would need to be tested in single knockout models or knockout combinations. As discussed previously it is likely that secreted proteins may play a role in the observed under-transmission of the fused chromosome in heterozygous mice. As when sperm were aged in the epididymis the under transmission of the fused chromosome when heterozygous male mice were mated to wild type females decreased, possibly due to the action of proteins secreted from the wild-type sperm. Alternatively aging of the sperm may allow mutated proteins with reduced binding affinity or kinetics to get closer to the wild type level of activity.

6. General Discussion

In this thesis I aimed to understand the interplay between gametogenesis, 3D nuclei structure and evolutionary chromosome rearrangements, as well as shed light on the spread of these rearrangements in a population. The key findings derived from my work are that EBRs are associated to post-meiotic spermatid DSBs and not meiotic DSBs. Chromatin states do change as spermatogenesis progresses, and it is these regions changing chromatin state that are associated to EBRs. Moreover, there appears to be a common vulnerability code in spermatids, in that spermatid DSBs are associated with specific chromatin state changes during spermatogenesis, with predicted non-B DNA structures (e.g. Z-DNA, that may regulate DNA tension during sperm head compaction) and with regions prone to oxidative damage in mature sperm.

Unexpectedly, I found no evidence for chromatin silencing in the vicinity of the Rob fusion points in the heterozygous Rb6.16 mice, indicating that Rob fusions may be regulated differently to other structural variants during spermatogenesis. However, non-synonymous gene variants linked to the fusion breakpoint were identified in the Rb6.16 fusion model, several of these were in key genes involved in fertility (such as *Pla2g10* and *Prm2*). These mutations may explain the reported transmission skewing when heterozygous Rb6.16 males are mated to wild type females.

To integrate my findings, I propose a new model of genome evolution Figure 6-1.

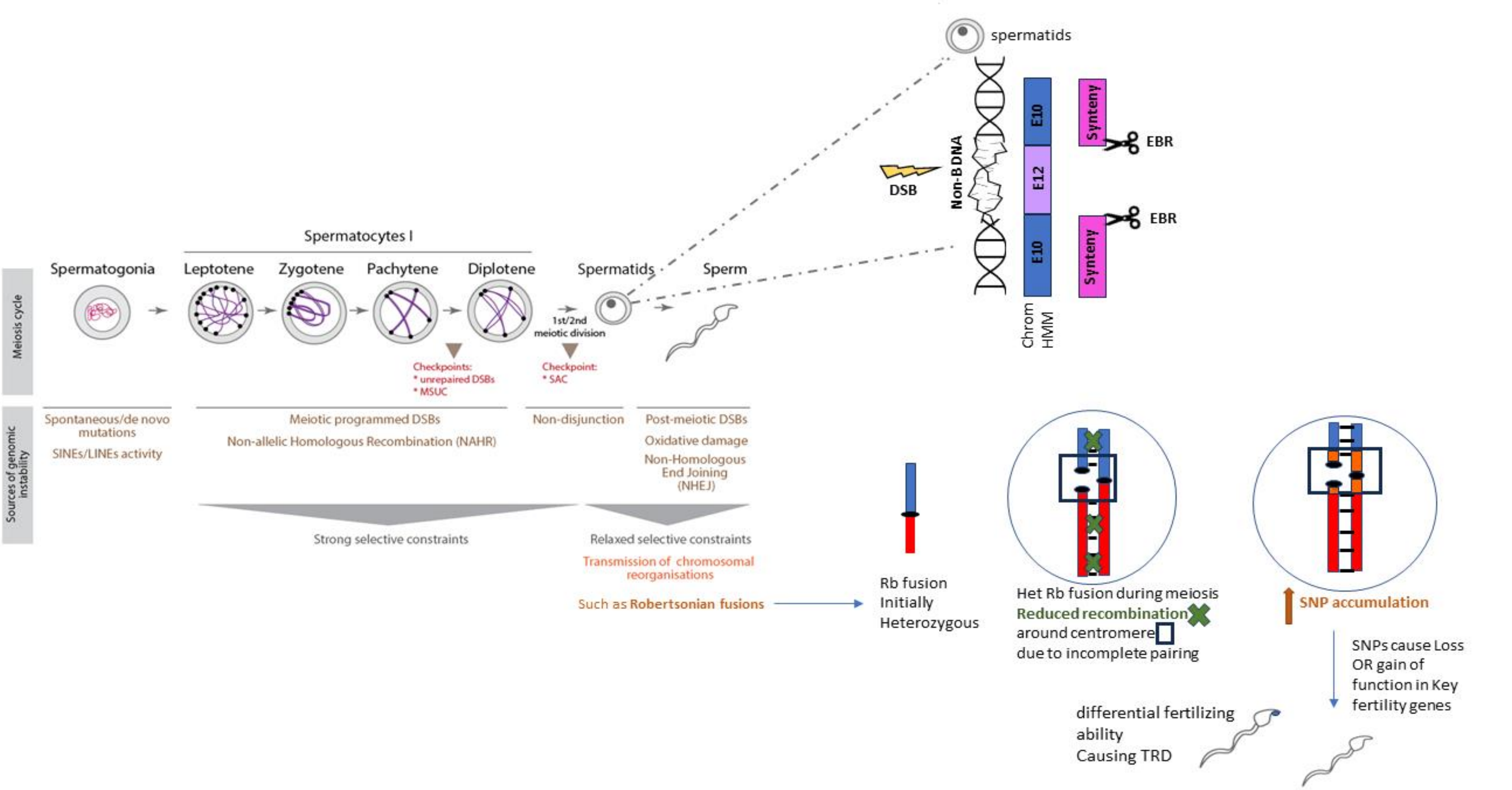
Evolutionary breakpoint regions (represented by the scissor image) are associated with the location of spermatid DSBs (yellow lightning bolt), suggesting that sources of evolutionary novelty arise during spermatogenesis. The 3D structure of the genome within spermatids is also a fundamental consideration as this can impact the propensity of a region to incur a DSB during spermiogenesis when the male genome is remodelled with the replacement of histones with protamines. Chromosomal rearrangements such as Robertsonian fusions must occur due to the formation of DSBs in two chromosomes. Figure 6-1 attempts to illustrate how such a fusion could influence its own transmission. In this instance the Rob6.16 fusion likely shows TRD due to the accumulation of deleterious SNPs within the ROI in key genes involved in fertility. Further research will be required to determine the key aspects of how and where spermatid DSBs occur to prove or disprove this model.

In more detail, I have shown that there is a common vulnerability code in spermatids. There are primary sequence motifs, secondary DNA structures and chromatin contexts associated with spermatid DSBs. The locations of DSBs in spermatids overlap EBRs, and as such I concluded that sources of evolutionary novelty arise during the post-meiotic stage of spermatogenesis. It is widely acknowledged that there is a widespread paternal bias in germline mutation in amniotes (318). Rodríguez-Nuevo *et al* 2022 (362) showed that *Xenopus* and human oocytes can maintain

ROS-free mitochondrial metabolism, whereas this does not occur during spermatogenesis. With this in mind, the effect of ROS in the generation of new DSBs might be greater in male germ cells than in oocytes, therefore contributing to the paternal bias in germline mutations. This paternal bias in germline mutations is similar across species (318), despite exposure to different physical environments and different exogenous mutagens. We postulate that the similar bias might be attributable to vulnerable genomic regions suffering DSBs during genome compaction (i.e. the removal of histone proteins and their replacement by protamines) during spermatogenesis combined with an increased effect of ROS. However, further experimental studies in other species would be required to confirm this.

Focusing on how genomic structural novelty can be selected for or against by transmission ratio distortion, we studied whether genetic polymorphisms or epigenetic regulation was more likely to cause the TRD observed in a specific Rob fusion model system. We did not observe systematic post-meiotic silencing of either allele in the heterozygous Robertsonian mice within the region of interest surrounding the centromere. This conclusively shows that epigenetic regulation is unlikely to be causative for the TRD. A caveat here is that I focused my analysis on post-meiotic cells (round spermatids) and thus did not directly check the extent of asynapsis surrounding the fused centromeres. To do this I could have harvested pachytene cells and carried out immunofluorescent staining with gammaH2AX and SYCP3 (synaptonemal complex protein). This would answer the question of whether we do observe meiotic silencing of the fusion chromosome centromeres (as is typical for Rob fusions). Given that we know there is no remaining silencing after meiosis, such work could also show the point at which any such silencing is lost, and/or whether there is apoptosis of meiotic cells that fail to fully reactivate their chromatin post-meiotically.

Turning to the potential genetic causes of TRD - i.e. mutations in linkage with the fusion that act post-meiotically to skew transmission - The suppressed recombination speciation models (363, 364) predict that rearranged chromosomes will have higher gene divergence than non-rearranged or colinear regions. The model also predicts that sterility genes or new alleles may accumulate in the rearranged region, which may or may not have an adaptive role. I have observed the presence of SNPs within genes that have a key role in fertility such as *Pla2g10* and *Prm2*. Transmission ratio distortion, whether by true meiotic drive or by haploid selection can influence the rate and type of chromosomal changes that are inherited. The Robertsonian 6.16 fusion shows under transmission relative to the non-fused chromosomes, likely due to loss of function SNPs within one or several key fertility genes, this would need to be confirmed through functional studies.



spermatids
DSB
Non-B DNA
E10
E12
E10
Synteny
Synteny
EBR
EBR
Chrom HMM
Spermatocytes I
Spermatogonia
Leptotene
Zygotene
Pachytene
Diplotene
Spermatids
Sperm
Meiosis cycle
1st/2nd meiotic division
Checkpoints: * unrepaired DSBs * MSUC
Checkpoint: * SAC
Sources of genomic instability
Spontaneous/de novo mutations
SINEs/LINEs activity
Meiotic programmed DSBs
Non-allelic Homologous Recombination (NAHR)
Non-disjunction
Post-meiotic DSBs
Oxidative damage
Non-Homologous End Joining (NHEJ)
Strong selective constraints
Relaxed selective constraints
Transmission of chromosomal reorganisations
Such as Robertsonian fusions
Rb fusion Initially Heterozygous
Het Rb fusion during meiosis
Reduced recombination
around centromere
due to incomplete pairing
SNP accumulation
SNPs cause Loss OR gain of function in Key fertility genes
differential fertilizing ability Causing TRD

Figure 6-1: Summary schematic of spermatogenesis showing the meiotic stages, different sources of genomic instability and the association of spermatid DSBs to Non-B DNA, chromatin states and evolutionary breakpoint regions (EBRs). Also shown are the possible mechanisms by which a chromosomal rearrangement may influence its own transmission. Rb=Robertsonian 6.16 fusion.

6.1 Potential pitfalls from my approach

Using a variety of previously published epigenomics data

Chapter 4 mainly used previously published data obtained from different publications. This data was originally analysed using different analytical procedures such as differing bioinformatic pipelines or different versions of the same software, which could impact the results obtained. As we mined ChIP-seq data from many different sources, to remove any bias that could be caused by different analysis pipelines we obtained the raw data and processed the files through the same ChIP-seq analysis pipeline. We obtained the spermatid DSB data from one publication and the spermatogonial, spermatocyte, and spermatid samples (used to produce the alluvial plot) from another publication, but for the histone data it was not possible to obtain all the marks from one publication. Different isolation techniques between different publications may have impacted the results. The spermatids may have been treated differently during the extraction process or the sample purity could have been very different, for example varying purity obtained by different FACS pipelines. By using the DSB data from one publication and the spermatogonial, spermatocyte, and spermatid data from another publication we minimised such bias. However, for the other histone marks it was not possible to reduce such bias as the data was obtained from different papers.

Considering predicted data in the context of experimental data

In silico predictions may over or underestimate the value in question. In Chapter 4 I used both predicted and experimental G-quadruplex data (see section 2.7.3). Although there was a negative correlation between G-quadruplexes and DSBs for both the predicted (z-score -75.9) and experimental data (z-score -32.5), only 24.5% of experimental G-quadruplex regions overlap by at least 1bp with the predicted data, while 45.3% of predicted G-quadruplex overlap experimental data. This suggests that either the prediction is not highly accurate, there is an overestimation of the number of G-quadruplex regions in silico, or conversely the experimental data may only capture a portion of the structures able to form G-quadruplexes within the G-quadruplex form at the time the experiment was carried out. This lack of agreement between predicted and experimental data is worth considering when using other predicted data types, such as Z-DNA used in this thesis. If the predicted Z-DNA and STR data also only predicts a proportion of the regions observed in vitro, then we may be missing key biological associations. The true extent of the mismatch may not be apparent until in vitro spermatid Z-DNA/STR data is published that could be compared to the predicted data.

Accounting for variability between mice

RNA-sequencing and ChIP sequencing data can be highly variable, and the results can be dramatically different depending on the experimental conditions such as the age of an animal or the purity of a cell sample. Therefore, repeats are required to ensure valid biological interpretation (365, 366), either technical or biological. In this thesis, we included both technical and biological replicates as follows: i) for Cut&Tag libraries we sequenced two or three replicates per condition, ii) for several histone marks we analysed replicate ChIP-seq samples to determine if the results were concordant.

Whole genome sequencing is not a variable technique per se, especially at high coverage. We carried out whole genome sequencing on two homozygous Rob6.16 mice as this allowed us to determine concordant high confidence homozygous SNPs shared across animals of the same genotype and exclude SNPs specific to an individual mouse. This allowed us to obtain very high confidence homozygous SNP calls which were used in the allele-specific expression analysis.

For any RNA-seq experiment it is important to produce replicate libraries from age matched animals or cell lines of similar passage. Here, we used aged-matched mice (up to two weeks apart) for the RNA-seq experiments, to minimise any differences in expression between the genotypes that could result from using mice of very different ages. Moreover, where possible the RNA-sequencing data from the WT, Het and Hom Rob6.16 mice was sequenced on the same flow cell to avoid any biases caused by different sequencing runs.

6.2 Future directions

The DSB and histone mark data used in this thesis was from a population of cells and not single cell data, therefore the results we observe are correlations. Ideally, we would move to using single cell data with the analysis of multiple marks within the same cell. New methods such as MulTI-Tag (338) have been developed to allow this, which would greatly improve the resolution of any analysis.

We used the locations of predicted Z-DNA, it would be interesting to carry out ChIP-seq with anti-Z-DNA antibodies in spermatids and compare the results to the locations of spermatid DSBs. The protocols presented here for the flow sorting of spermatids would allow high purity spermatid samples to be isolated which would be required for such experiments.

And, as we did not identify epigenetic silencing within the ROI of the Rob6.16 fusion chromosome, it would be interesting to make knockout models of some of the high effect SNPs, particularly those that are GIM (such as *Pla2g10*) to determine whether they result in any TRD. Finally, the DSB results presented here only used *Mus musculus* spermatid data, it would be interesting to determine if the same correlations (such as those shown in Figure 6-1) occur in other lineages, particularly for those that remodel their genomes with the replacement of histones with protamines during spermiogenesis.

7 References

1. Luger,K., Mäder,A.W., Richmond,R.K., Sargent,D.F. and Richmond,T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature 1997 389:6648*, **389**, 251–260.
2. Bannister,A.J. and Kouzarides,T. (2011) Regulation of chromatin by histone modifications. *Cell Research 2011 21:3*, **21**, 381–395.
3. Wu,J. and Grunstein,M. (2000) 25 years after the nucleosome model: chromatin modifications. *Trends Biochem Sci*, **25**, 619–623.
4. Heitz,E. (1928) Das Heterochromatin der Moose Bornträger.
5. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, **129**, 823–837.
6. Schneider,R., Bannister,A.J., Myers,F.A., Thorne,A.W., Crane-Robinson,C. and Kouzarides,T. (2003) Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nature Cell Biology 2003 6:1*, **6**, 73–77.
7. Ayrapetov,M.K., Gursoy-Yuzugullu,O., Xu,C., Xu,Y. and Price,B.D. DNA double-strand breaks promote methylation of histone H3 on lysine 9 and transient formation of repressive chromatin. 10.1073/pnas.1403565111.
8. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M., Sharp,P.A., *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*, **107**, 21931–21936.
9. Voigt,P., LeRoy,G., Drury,W.J., Zee,B.M., Son,J., Beck,D.B., Young,N.L., Garcia,B.A. and Reinberg,D. (2012) Asymmetrically Modified Nucleosomes. *Cell*, **151**, 181.
10. Kim,Y.Z. (2014) Altered Histone Modifications in Gliomas. *Brain Tumor Res Treat*, **2**, 7.
11. Alpsoy,A., Sood,S. and Dykhuizen,E.C. (2021) At the Crossroad of Gene Regulation and Genome Organization: Potential Roles for ATP-Dependent Chromatin Remodelers in the Regulation of CTCF-Mediated 3D Architecture. *Biology 2021, Vol. 10, Page 272*, **10**, 272.
12. Cremer,T., Cremer,C., Schneider,T., Baumann,H., Hens,L. and Kirsch-Volders,M. (1982) Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments. *Hum Genet*, **62**, 201–209.
13. Boyle,S., Gilchrist,S., Bridger,J.M., Mahy,N.L., Ellis,J.A. and Bickmore,W.A. (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum Mol Genet*, **10**, 211–219.
14. Lieberman-Aiden,E., Van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O., *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
15. Tavares-Cadete,F., Norouzi,D., Dekker,B., Liu,Y. and Dekker,J. (2020) Multi-contact 3C reveals that the human genome during interphase is largely not entangled. *Nature Structural & Molecular Biology 2020 27:12*, **27**, 1105–1114.
16. Tanabe,H., Müller,S., Neusser,M., Von Hase,J., Calcagno,E., Cremer,M., Solovei,I., Cremer,C. and Cremer,T. (2002) Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proc Natl Acad Sci U S A*, **99**, 4424–4429.
17. Gibcus,J.H. and Dekker,J. (2013) Connecting the genome: dynamics and stochasticity in a new hierarchy for chromosome conformation. In.

18. Lajoie,B.R., Dekker,J. and Kaplan,N. (2015) The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods*, **72**, 65–75.
19. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
20. Zhan,Y., Mariani,L., Barozzi,I., Schulz,E.G., Blüthgen,N., Stadler,M., Tiana,G. and Giorgetti,L. (2017) Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Res*, **27**, 479–490.
21. Lupiáñez,D.G., Kraft,K., Heinrich,V., Krawitz,P., Brancati,F., Klopocki,E., Horn,D., Kayserili,H., Opitz,J.M., Laxova,R., *et al.* (2015) Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, **161**, 1012–1025.
22. Splinter,E., Heath,H., Kooren,J., Palstra,R.J., Klous,P., Grosveld,F., Galjart,N. and De Laat,W. (2006) CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev*, **20**, 2349–2354.
23. Phillips-Cremins,J.E., Sauria,M.E.G., Sanyal,A., Gerasimova,T.I., Lajoie,B.R., Bell,J.S.K., Ong,C.T., Hookway,T.A., Guo,C., Sun,Y., *et al.* (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**, 1281–1295.
24. de Wit,E., Vos,E.S.M., Holwerda,S.J.B., Valdes-Quezada,C., Verstegen,M.J.A.M., Teunissen,H., Splinter,E., Wijchers,P.J., Krijger,P.H.L. and de Laat,W. (2015) CTCF Binding Polarity Determines Chromatin Looping. *Mol Cell*, **60**, 676–684.
25. Mérot,C., Oomen,R.A., Tigano,A. and Wellenreuther,M. (2020) A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. *Trends Ecol Evol*, **35**, 561–572.
26. Levan,A., Fredga,K. and Sandberg,A.A. (1964) Nomenclature for centromeric position on chromosomes. *Hereditas*, **52**, 201–220.
27. Canoy,R.J., Shmakova,A., Karpukhina,A., Shepelev,M., Germini,D. and Vassetzky,Y. (2022) Factors That Affect the Formation of Chromosomal Translocations in Cells. *Cancers (Basel)*, **14**, 5110.
28. McKinlay Gardner,R.J. and Amor,D.J. (2018) Robertsonian Translocations. In *Gardner and Sutherland's Chromosome Abnormalities and Genetic Counseling*. Oxford University Press, pp. 142–157.
29. Pylyp,L.Y., Zukin,V.D. and Bilko,N.M. (2013) Chromosomal segregation in sperm of Robertsonian translocation carriers. *J Assist Reprod Genet*, **30**, 1141.
30. Capilla,L., Medarde,N., Alemany-Schmidt,A., Oliver-Bonet,M., Ventura,J. and Ruiz-Herrera,A. (2014) Genetic recombination variation in wild Robertsonian mice: On the role of chromosomal fusions and Prdm9 allelic background. *Proceedings of the Royal Society B: Biological Sciences*, **281**.
31. Medarde,N., López-Fuster,M.J., Mũoz-Mũoz,F. and Ventura,J. (2012) Spatio-temporal variation in the structure of a chromosomal polymorphism zone in the house mouse. *Heredity (Edinb)*, **109**, 78–89.
32. Vara,C., Capilla,L., Ferretti,L., Ledda,A., Sánchez-Guillén,R.A., Gabriel,S.I., Albert-Lizandra,G., Florit-Sabater,B., Bello-Rodríguez,J., Ventura,J., *et al.* (2019) PRDM9 Diversity at Fine Geographical Scale Reveals Contrasting Evolutionary Patterns and Functional Constraints in Natural Populations of House Mice. *Mol Biol Evol*, **36**, 1686–1700.
33. Sans-Fuentes,M.A., Ventura,J., López-Fuster,M.J. and Corti,M. (2009) Morphological variation in house mice from the Robertsonian polymorphism area of Barcelona. *Biological Journal of the Linnean Society*, **97**, 555–570.

34. Longo,M.S., Carone,D.M., Comparative,N., Program,S., Green,E.D., O'neill,M.J. and O'neill,R.J. (2009) Distinct retroelement classes define evolutionary breakpoints demarcating sites of evolutionary novelty. 10.1186/1471-2164-10-334.
35. Murphy,W.J., Larkin,D.M., Everts-van der Wind,A., Bourque,G., Tesler,G., Auvil,L., Beever,J.E., Chowdhary,B.P., Galibert,F., Gatzke,L., *et al.* (2003) Dynamics of Mammalian Chromosome Evolution Inferred from Multispecies Comparative Maps. *16. R. H. Smith, R. Mead, Theor. Popul. Biol*, **13**, 383.
36. Murphy,W.J., Larkin,D.M., Everts-Van Der Wind,A., Bourque,G., Tesler,G., Auvil,L., Beever,J.E., Chowdhary,B.P., Galibert,F., Gatzke,L., *et al.* (2005) Evolution: Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science (1979)*, **309**, 613–617.
37. Elsik,C.G., Tellam,R.L., Worley,K.C., Gibbs,R.A., Muzny,D.M., Weinstock,G.M., Adelson,D.L., Eichler,E.E., Elnitski,L., Guigó,R., *et al.* (2009) The Genome Sequence of Taurine Cattle: A window to ruminant biology and evolution. *Science*, **324**, 522.
38. Damas,J., Corbo,M., Kim,J., Turner-Maier,J., Farre,M., Larkin,D.M., Ryder,O.A., Steiner,C., Houck,M.L., Hall,S., *et al.* (2022) Evolution of the ancestral mammalian karyotype and syntenic regions. *Proc Natl Acad Sci U S A*, **119**, e2209139119.
39. Larkin,D.M., Pape,G., Donthu,R., Auvil,L., Welge,M. and Lewin,H.A. (2009) Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Res*, **19**, 770–777.
40. Ruiz-Herrera,A., Castresana,J. and Robinson,T.J. (2006) Is mammalian chromosomal evolution driven by regions of genome fragility? *Genome Biol*, **7**, R115.
41. Nadeau,J.H. and Taylor,B.A. (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc Natl Acad Sci U S A*, **81**, 814–818.
42. Pevzner,P. and Tesler,G. (2003) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc Natl Acad Sci U S A*, **100**, 7672–7677.
43. Wimmer,R., Kirsch,S., Rappold,G.A. and Schempp,W. (2005) Evolutionary breakpoint analysis on Y chromosomes of higher primates provides insight into human Y evolution. *Cytogenet Genome Res*, **108**, 204–210.
44. Farré,M., Robinson,T.J. and Ruiz-Herrera,A. (2015) An Integrative Breakage Model of genome architecture, reshuffling and evolution: The Integrative Breakage Model of genome evolution, a novel multidisciplinary hypothesis for the study of genome plasticity. *BioEssays*, **37**, 479–488.
45. Zhang,Y., McCord,R.P., Ho,Y.J., Lajoie,B.R., Hildebrand,D.G., Simon,A.C., Becker,M.S., Alt,F.W. and Dekker,J. (2012) Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, **148**, 908–921.
46. Henikoff,S., Ahmad,K. and Malik,H.S. (2001) The centromere paradox: Stable inheritance with rapidly evolving DNA. *Science (1979)*, **293**, 1098–1102.
47. Chmátal,L., Gabriel,S.I., Mitsainas,G.P., Martínez-Vargas,J., Ventura,J., Searle,J.B., Schultz,R.M. and Lampson,M.A. (2014) Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. *Curr Biol*, **24**, 2295–2300.
48. Painter,T.S. (1928) A Comparison of the Chromosomes of the Rat and Mouse with Reference to the Question of Chromosome Homology in Mammals. *Genetics*, **13**, 180.
49. Gallardo,M.H., Bickham,J.W., Kausel,G., Köhler,N. and Honeycutt,R.L. (2003) Gradual and quantum genome size shifts in the hystricognath rodents. *J Evol Biol*, **16**, 163–169.
50. Garagna,S., Page,J., Fernandez-Donoso,R., Zuccotti,M. and Searle,J.B. (2014) The robertsonian phenomenon in the house mouse: mutation, meiosis and speciation. *Chromosoma*, **123**, 529–544.

51. Harper,L., Golubovskaya,I. and Cande,W.Z. (2004) A bouquet of chromosomes. *J Cell Sci*, **117**, 4025–4032.
52. Lane,S. and Kauppi,L. (2019) Meiotic spindle assembly checkpoint and aneuploidy in males versus females. *Cell Mol Life Sci*, **76**, 1135—1150.
53. Aranha,I.P. and Martin-DeLeon,P.A. (1991) The murine Rb(6.16) translocation: evidence for sperm selection and a modulating effect of aging. *Hum Genet*, **87**, 278–284.
54. Bolcun-Filas,E. and Handel,M.A. (2018) Meiosis: the chromosomal foundation of reproduction. *Biol Reprod*, **99**, 112–126.
55. Baudat,F., Imai,Y. and De Massy,B. (2013) Meiotic recombination in mammals: localization and regulation. *Nature Reviews Genetics 2013 14:11*, **14**, 794–806.
56. Conklin,E.G. and Conklin,E.G. (1915) Why Polar Bodies Do Not Develop. *Proceedings of the National Academy of Sciences*, **1**, 491–496.
57. Russell,L., Ettlin,R., SinhaHikin,A. and Clegg,E. (1990) Histological and Histopathological Evaluation of the Testis 1st edition. Cache River Press.
58. Meistrich,M.L., Mohapatra,B., Shirley,C.R., Zhao,M., Meistrich,M.L., Mohapatra,~) - B, Shirley,• C R, Zhao,• M and Shirley,C.R. (2003) Roles of transition nuclear proteins in spermiogenesis. *Chromosoma*, **111**, 483–488.
59. Marcon,L. and Boissonneault,G. (2004) Transient DNA Strand Breaks During Mouse and Human Spermiogenesis:New Insights in Stage Specificity and Link to Chromatin Remodeling. *Biol Reprod*, **70**, 910–918.
60. Caldwell,K.A. and Handel,M.A. (1991) Protamine transcript sharing among postmeiotic spermatids (haploid gene expression/spermatogenesis/aneusonmic sperm/in situ hybridization). *Proc. Natl. Acad. Sci. USA*, **88**, 2407–2411.
61. Huckins,C. (1971) The spermatogonial stem cell population in adult rats. I. Their morphology, proliferation and maturation. *Anat Rec*, **169**, 533–557.
62. de Rooij DG and Russell LD (2000) All You Wanted to Know About Spermatogonia but Were Afraid to Ask. *J Androl*, **21**, 776–798.
63. Licatalosi,D.D. (2016) Roles of RNA-binding Proteins and Post-transcriptional Regulation in Driving Male Germ Cell Development in the Mouse. *Adv Exp Med Biol*, **907**, 123.
64. Griswold,M.D. (2016) Spermatogenesis: The Commitment to Meiosis. *Physiol Rev*, **96**, 1.
65. Nishimura,H. and L'Hernault,S.W. (2017) Spermatogenesis. *Current Biology*, **27**, R988–R994.
66. De Rooij,D.G. and Grootegoed,J.A. (1998) Spermatogonial stem cells. *Curr Opin Cell Biol*, **10**, 694–701.
67. Oakberg,E.F. (1971) Spermatogonial stem-cell renewal in the mouse. *Anat Rec*, **169**, 515–531.
68. Keeney,S., Giroux,C.N. and Kleckner,N. (1997) Meiosis-Specific DNA Double-Strand Breaks Are Catalyzed by Spo11, a Member of a Widely Conserved Protein Family. *Cell*, **88**, 375–384.
69. Baudat,F., Buard,J., Grey,C., Fledel-Alon,A., Ober,C., Przeworski,M., Coop,G. and De Massy,B. (2010) PRDM9 is a Major Determinant of Meiotic Recombination Hotspots in humans and mice. *Science*, **327**, 836.
70. Baker,C.L., Walker,M., Kajita,S., Petkov,P.M. and Paigen,K. (2014) PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration. *Genome Res*, **24**, 724–732.
71. Grey,C., Barthès,P., Friec,C.-L., Langa,G. and Baudat,F. (2011) Mouse PRDM9 DNA-Binding Specificity Determines Sites of Histone H3 Lysine 4 Trimethylation for Initiation of Meiotic Recombination. *PLoS Biol*, **9**, 1001176.

72. Jeffreys,A.J., Murray,J. and Neumann,R. (1998) High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol Cell*, **2**, 267–273.

73. Hayashi,K., Yoshida,K. and Matsui,Y. (2005) A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. *Nature 2005 438:7066*, **438**, 374–378.

74. Wu,H., Mathioudakis,N., Diagouraga,B., Dong,A., Dombrovski,L., Baudat,F., Cusack,S., DeMassy,B. and Kadlec,J. (2013) Molecular basis for the regulation of the H3K4 methyltransferase activity of PRDM9. *Cell Rep*, **5**, 13–20.

75. Neale,M.J. and Keeney,S. (2006) Clarifying the mechanics of DNA strand exchange in meiotic recombination. *Nature*, **442**, 153.

76. Kobayashi,W., Takaku,M., Machida,S., Tachiwana,H., Maehara,K., Ohkawa,Y. and Kurumizaka,H. (2016) Chromatin architecture may dictate the target site for DMC1, but not for RAD51, during homologous pairing. *Sci Rep*, **6**.

77. Cole,F., Kauppi,L., Lange,J., Roig,I., Wang,R., Keeney,S. and Jasin,M. (2012) Homeostatic control of recombination is implemented progressively in mouse meiosis. *Nat Cell Biol*, **14**, 424.

78. Kotaja,N. (2013) Spermatogenesis, Mouse. *Brenner's Encyclopedia of Genetics: Second Edition*, 10.1016/B978-0-12-374984-0.01461-3.

79. Sprando,R.L. and Russell,L.D. (1987) Comparative study of cytoplasmic elimination in spermatids of selected mammalian species. *American Journal of Anatomy*, **178**, 72–80.

80. Erkek,S., Hisano,M., Liang,C.Y., Gill,M., Murr,R., Dieker,J., Schübeler,D., Vlag,J. Van Der, Stadler,M.B. and Peters,A.H.F.M. (2013) Molecular determinants of nucleosome retention at CpG-rich sequences in mouse spermatozoa. *Nature Structural & Molecular Biology 2013 20:7*, **20**, 868–875.

81. Yamaguchi,K., Hada,M., Fukuda,Y., Inoue,E., Makino,Y., Katou,Y., Shirahige,K. and Okada,Y. (2018) Re-evaluating the Localization of Sperm-Retained Histones Revealed the Modification-Dependent Accumulation in Specific Genome Regions. *Cell Rep*, **23**, 3920–3932.

82. Risley,M.S., Einheber,S. and Bumcrot,D.A. (1986) Changes in DNA topology during spermatogenesis. *Chromosoma*, **94**, 217–227.

83. Hud,N. V., Allen,M.J., Downing,K.H., Lee,J. and Balhorn,R. (1993) Identification of the Elemental Packing Unit of DNA in Mammalian Sperm Cells by Atomic Force Microscopy. *Biochem Biophys Res Commun*, **193**, 1347–1354.

84. Grégoire,M.C., Leduc,F., Morin,M.H., Cavé,T., Arguin,M., Richter,M., Jacques,P.É. and Boissonneault,G. (2018) The DNA double-strand "breakome" of mouse spermatids. *Cellular and Molecular Life Sciences*, **75**, 2859–2872.

85. Ward,W.S. (1993) Deoxyribonucleic acid loop domain tertiary structure in mammalian spermatozoa. *Biol Reprod*, **48**, 1193–1201.

86. Kimura,Y. and Yanagimachi,R. (1995) Intracytoplasmic Sperm Injection in the Mouse. *Biol Reprod*, **52**, 709–720.

87. Miller,D., Brinkworth,M. and Iles,D. (2010) Paternal DNA packaging in spermatozoa: more than the sum of its parts? DNA, histones, protamines and epigenetics. *Reproduction*, **139**, 287–301.

88. Ward,W.S. (2011) Regulating DNA Supercoiling: Sperm Points the Way. *Biol Reprod*, **84**, 841.

89. Leduc,F., Maquennehan,V., Nkoma,G.B. and Boissonneault,G. (2008) DNA Damage Response During Chromatin Remodeling in Elongating Spermatids of Mice. *Biol Reprod*, **78**, 324–332.

90. Siklenka,K., Erkek,S., Godmann,M., Lambrot,R., McGraw,S., Lafleur,C., Cohen,T., Xia,J., Suderman,M., Hallett,M., *et al.* (2015) Disruption of histone methylation in developing sperm impairs offspring health transgenerationally. *Science (1979)*, **350**.

91. Teperek,M., Simeone,A., Gaggioli,V., Miyamoto,K., Allen,G.E., Erkek,S., Kwon,T., Marcotte,E.M., Zegerman,P., Bradshaw,C.R., *et al.* (2016) Sperm is epigenetically programmed to regulate gene transcription in embryos. *Genome Res*, **26**, 1034–1046.
92. Hammoud,S.S., Nix,D.A., Zhang,H., Purwar,J., Carrell,D.T. and Cairns,B.R. (2009) Distinctive Chromatin in Human Sperm Packages Genes for Embryo Development. *Nature*, **460**, 473.
93. Miller,D. and Ostermeier,G.C. (2006) Towards a better understanding of RNA carriage by ejaculate spermatozoa. *Hum Reprod Update*, **12**, 757–767.
94. MacLaughlin,J. and Terner,C. (1973) Ribonucleic acid synthesis by spermatozoa from the rat and hamster. *Biochem J*, **133**, 635–639.
95. Cho T., Sakai S., Nagata M. and Aoki F. (2002) Involvement of chromatin structure in the regulation of mouse zygotic gene activation. *Animal Science Journal*, **73**, 113–122.
96. Falco,G., Lee,S.L., Stanghellini,I., Bassey,U.C., Hamatani,T. and Ko,M.S.H. (2007) Zscan4: a novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. *Dev Biol*, **307**, 539.
97. Srinivasan,R., Nady,N., Arora,N., Hsieh,L.J., Swigut,T., Narlikar,G.J., Wossidlo,M. and Wysocka,J. (2020) Zscan4 binds nucleosomal microsatellite DNA and protects mouse two-cell embryos from DNA damage. *Sci Adv*, **6**.
98. Brick,K., Thibault-Sennett,S., Smagulova,F., Lam,K.W.G., Pu,Y., Pratto,F., Camerini-Otero,R.D. and Petukhova,G. V. (2018) Extensive sex differences at the initiation of genetic recombination. *Nature*, **561**, 338–342.
99. Lam,I. and Keeney,S. (2014) Mechanism and regulation of meiotic recombination initiation. *Cold Spring Harb Perspect Biol*, **7**.
100. Ahmed,E.A., de Boer,P., Philippens,M.E.P., Kal,H.B. and de Rooij,D.G. (2010) Parp1-XRCC1 and the repair of DNA double strand breaks in mouse round spermatids. *Mutat Res*, **683**, 84–90.
101. Khokhlova,E. V, Fesenko,Z.S., Sopova,J. V and Leonova,E.I. (2020) Features of DNA Repair in the Early Stages of Mammalian Embryonic Development. *Genes (Basel)*, **11**.
102. Haddock,L., Gordon,S., Lewis,S.E.M., Larsen,P., Shehata,A. and Shehata,H. (2021) Sperm DNA fragmentation is a novel biomarker for early pregnancy loss. *Reprod Biomed Online*, **42**, 175–184.
103. Vara,C., Paytuví-Gallart,A., Cuartero,Y., Le Dily,F., Garcia,F., Salvà-Castro,J., Gómez-H,L., Julià,E., Moutinho,C., Aiese Cigliano,R., *et al.* (2019) Three-Dimensional Genomic Structure and Cohesin Occupancy Correlate with Transcriptional Activity during Spermatogenesis. *Cell Rep*, **28**, 352-367.e9.
104. Flyamer,I.M., Gassler,J., Imakaev,M., Brandão,H.B., Ulianov,S. V, Abdennur,N., Razin,S. V, Mirny,L.A. and Tachibana-Konwalski,K. (2017) Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, **544**, 110–114.
105. Mark,M., Teletin,M., Vernet,N. and Ghyselinck,N.B. (2015) Role of retinoic acid receptor (RAR) signaling in post-natal male germ cell differentiation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, **1849**, 84–93.
106. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P., *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature 2003 420:6915*, **420**, 520–562.
107. McClintock,B. (1950) The Origin and Behavior of Mutable Loci in Maize. *Proc Natl Acad Sci U S A*, **36**, 344.
108. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., Fitzhugh,W., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature 2001 409:6822*, **409**, 860–921.

109. Kazazian,H.H., Wong,C., Youssoufian,H., Scott,A.F., Phillips,D.G. and Antonarakis,S.E. (1988) Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, **332**, 164–166.
110. Han,J.S., Szak,S.T. and Boeke,J.D. (2004) Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature*, **429**, 268–274.
111. Startek,M., Szafranski,P., Gambin,T., Campbell,I.M., Hixson,P., Shaw,C.A., Stankiewicz,P. and Gambin,A. (2015) NAR Breakthrough Article: Genome-wide analyses of LINE–LINE-mediated nonallelic homologous recombination. *Nucleic Acids Res*, **43**, 2188.
112. Mätlik,K., Redik,K. and Speek,M. (2006) L1 antisense promoter drives tissue-specific transcription of human genes. *J Biomed Biotechnol*, **2006**.
113. Yudkin,D., Hayward,B.E., Aladjem,M.I., Kumari,D. and Usdin,K. (2014) Chromosome fragility and the abnormal replication of the FMR1 locus in fragile X syndrome. *Hum Mol Genet*, **23**, 2940.
114. Wang,G. and Vasquez,K.M. (2014) Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA Repair (Amst)*, **19**, 143–151.
115. Rich,A. and Zhang,S. (2003) Z-DNA: the long road to biological function. *Nature Reviews Genetics 2003 4:7*, **4**, 566–572.
116. Harteis,S. and Schneider,S. (2014) Making the Bend: DNA Tertiary Structure and Protein-DNA Interactions. *Int J Mol Sci*, **15**, 12335.
117. Rich,A., Nordheim,A. and Wang,A.H.J. (1984) The chemistry and biology of left-handed Z-DNA. *Annu Rev Biochem*, **53**, 791–846.
118. Nordheim,A. and Rich,A. (1983) The sequence (dC-dA)R(dG-dT). forms left-handed Z-DNA in negatively supercoiled plasmids (Z-DNA antibodies/filter binding assay/DNA-protein crosslinking/chromatin structure/recombination). *Proc. Nati. Acad. Sci. USA*, **80**, 1821–1825.
119. Haniford,D.B. and Pulleyblank,D.E. (1983) Facile transition of poly[d(TG)·d(CA)] into a left-handed helix in physiological conditions. *Nature 1983 302:5909*, **302**, 632–634.
120. Peck,L.J., Nordheim,A., Rich,A. and Wang,J.C. (1982) Flipping of cloned d(pCpG)n.d(pCpG)n DNA sequences from right- to left-handed helical structure by salt, Co(III), or negative supercoiling. *Proc Natl Acad Sci U S A*, **79**, 4560.
121. Sen,D. and Gilbert,W. (1988) Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature 1988 334:6180*, **334**, 364–366.
122. Zheng,K.W., Zhang,J.Y., He,Y. De, Gong,J.Y., Wen,C.J., Chen,J.N., Hao,Y.H., Zhao,Y. and Tan,Z. (2020) Detection of genomic G-quadruplexes in living cells using a small artificial protein. *Nucleic Acids Res*, **48**, 11706–11720.
123. Gray,L.T., Vallur,A.C., Eddy,J. and Maizels,N. (2014) G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD. *Nat Chem Biol*, **10**, 313–318.
124. Wu,Y., Shin-ya,K. and Brosh,R.M. (2008) FANCJ helicase defective in Fanconia anemia and breast cancer unwinds G-quadruplex DNA to defend genomic stability. *Mol Cell Biol*, **28**, 4116–4128.
125. Gan,W., Guan,Z., Liu,J., Gui,T., Shen,K., Manley,J.L. and Li,X. (2011) R-loop-mediated genomic instability is caused by impairment of replication fork progression. *Genes Dev*, **25**, 2041.
126. Bader,A.S. and Bushell,M. (2020) DNA:RNA hybrids form at DNA double-strand breaks in transcriptionally active loci. *Cell Death & Disease 2020 11:4*, **11**, 1–7.
127. Sanz,L.A., Hartono,S.R., Lim,Y.W., Steyaert,S., Rajpurkar,A., Ginno,P.A., Xu,X. and Chédin,F. (2016) Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol Cell*, **63**, 167–178.

128. Kouzine,F., Wojtowicz,D., Baranello,L., Yamane,A., Nelson,S., Resch,W., Kieffer-Kwon,K.R., Benham,C.J., Casellas,R., Przytycka,T.M., *et al.* (2017) Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian Genome. *Cell Syst*, **4**, 344-356.e7.
129. Yu,K., Chedin,F., Hsieh,C.L., Wilson,T.E. and Lieber,M.R. (2003) R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nat Immunol*, **4**, 442–451.
130. Duquette,M.L., Handa,P., Vincent,J.A., Taylor,A.F. and Maizels,N. (2004) Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes Dev*, **18**, 1618–1629.
131. Stein,H. and Hausen,P. (1969) Enzyme from Calf Thymus Degrading the RNA Moiety of DNA-RNA Hybrids: Effect on DNA-Dependent RNA Polymerase. *Science (1979)*, **166**, 393–395.
132. Cerritelli,S.M. and Crouch,R.J. (2009) Ribonuclease H: the enzymes in Eukaryotes. *FEBS J*, **276**, 1494.
133. Santos-Pereira,J.M., Herrero,A.B., García-Rubio,M.L., Marín,A., Moreno,S. and Aguilera,A. (2013) The Npl3 hnRNP prevents R-loop-mediated transcription-replication conflicts and genome instability. *Genes Dev*, **27**, 2445–2458.
134. Ohle,C., Tesorero,R., Schermann,G., Dobrev,N., Sinning,I. and Fischer,T. (2016) Transient RNA-DNA Hybrids Are Required for Efficient Double-Strand Break Repair. *Cell*, **167**, 1001-1013.e7.
135. Li,R., Jia,Z. and Trush,M.A. (2016) Defining ROS in Biology and Medicine. *Reactive oxygen species*, **1 1**, 9–21.
136. Chabory,E., Damon,C., Lenoir,A., Kauselmann,G., Kern,H., Zevnik,B., Garrel,C., Saez,F., Cadet,R., Henry-Berger,J., *et al.* (2009) Epididymis seleno-independent glutathione peroxidase 5 maintains sperm DNA integrity in mice. *Journal of Clinical Investigation*, **119**, 2074–2085.
137. Noblanc,A., Kocer,A. and Drevet,J.R. (2014) Recent knowledge concerning mammalian sperm chromatin organization and its potential weaknesses when facing oxidative challenge. *Basic Clin Androl*, **24**.
138. Toppo,S., Flohé,L., Ursini,F., Vanin,S. and Maiorino,M. (2009) Catalytic mechanisms and specificities of glutathione peroxidases: Variations of a basic scheme. *Biochimica et Biophysica Acta (BBA) - General Subjects*, **1790**, 1486–1500.
139. Imai,H., Hakkaku,N., Iwamoto,R., Suzuki,J., Suzuki,T., Tajima,Y., Konishi,K., Minami,S., Ichinose,S., Ishizaka,K., *et al.* (2009) Depletion of Selenoprotein GPx4 in Spermatocytes Causes Male Infertility in Mice. *J Biol Chem*, **284**, 32522.
140. Zhang,T., Chabory,E., Britan,A., Grignard,E., Pitiot,O., Saez,F., Cadet,R., Henry-Berger,J., Vernet,P., Drevet,J.R., *et al.* (2008) GPX5, the selenium-independent glutathione peroxidase-encoding single copy gene is differentially expressed in mouse epididymis. *Reprod Fertil Dev*, **20**, 615–625.
141. Kocer,A., Henry-Berger,J., Noblanc,A., Champroux,A., Pogorelcnik,R., Guiton,R., Janny,L., Pons-Rejraji,H., Saez,F., Johnson,G.D., *et al.* (2015) Oxidative DNA damage in mouse sperm chromosomes: Size matters. *Free Radic Biol Med*, **89**, 993–1002.
142. Conrad,M., Moreno,S.G., Sinowatz,F., Ursini,F., Kölle,S., Roveri,A., Brielmeier,M., Wurst,W., Maiorino,M. and Bornkamm,G.W. (2005) The Nuclear Form of Phospholipid Hydroperoxide Glutathione Peroxidase Is a Protein Thiol Peroxidase Contributing to Sperm Chromatin Stability. *Mol Cell Biol*, **25**, 7637–7644.
143. Champroux,A., Torres-Carreira,J., Gharagozloo,P., Drevet,J.R. and Kocer,A. (2016) Mammalian sperm nuclear organization: resiliencies and vulnerabilities. *Basic and Clinical Andrology 2016 26:1*, **26**, 1–22.
144. Noblanc,A., Damon-Soubeyrand,C., Karrich,B., Henry-Berger,J., Cadet,R., Saez,F., Guiton,R., Janny,L., Pons-Rejraji,H., Alvarez,J.G., *et al.* (2013) DNA oxidative

damage in mammalian spermatozoa: Where and why is the male nucleus affected? *Free Radic Biol Med*, **65**, 719–723.

145. Szlachta,K., Manukyan,A., Raimer,H.M., Singh,S., Salamon,A., Guo,W., Lobachev,K.S. and Wang,Y.H. (2020) Topoisomerase II contributes to DNA secondary structure-mediated double-stranded breaks. *Nucleic Acids Res*, **48**, 6654.
146. Walker,J.R., Corpina,R.A. and Goldberg,J. (2001) Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. *Nature*, **412**, 607–614.
147. Smith,G.C.M. and Jackson,S.P. (1999) The DNA-dependent protein kinase. *Genes Dev*, **13**, 916–934.
148. Meek,K., Douglas,P., Cui,X., Ding,Q. and Lees-Miller,S.P. (2007) trans Autophosphorylation at DNA-dependent protein kinase's two major autophosphorylation site clusters facilitates end processing but not end joining. *Mol Cell Biol*, **27**, 3881–3890.
149. Soubeyrand,S., Pope,L., De Chasseval,R., Gosselin,D., Dong,F., de Villartay,J.P. and Haché,R.J.G. (2006) Artemis Phosphorylated by DNA-dependent Protein Kinase Associates Preferentially with Discrete Regions of Chromatin. *J Mol Biol*, **358**, 1200–1211.
150. De Ioannes,P., Malu,S., Cortes,P. and Aggarwal,A.K. (2012) Structural Basis of DNA Ligase IV-Artemis Interaction in Nonhomologous End-Joining. *Cell Rep*.
151. Ahnesorg,P., Smith,P. and Jackson,S.P. (2006) XLF interacts with the XRCC4-DNA Ligase IV complex to promote DNA nonhomologous end-joining. *Cell*, **124**, 301–313.
152. Buck,D., Malivert,L., De Chasseval,R., Barraud,A., Fondanèche,M.C., Sanal,O., Plebani,A., Stéphan,J.L., Hufnagel,M., Le Deist,F., *et al.* (2006) Cernunnos, a novel nonhomologous end-joining factor, is mutated in human immunodeficiency with microcephaly. *Cell*, **124**, 287–299.
153. Van Gent,D.C. and Van Der Burg,M. (2007) Non-homologous end-joining, a sticky affair. *Oncogene 2007 26:56*, **26**, 7731–7740.
154. Van Heemst,D., Brugmans,L., Verkaik,N.S. and Van Gent,D.C. (2004) End-joining of blunt DNA double-strand breaks in mammalian fibroblasts is precise and requires DNA-PK and XRCC4. *DNA Repair (Amst)*, **3**, 43–50.
155. Yu,X. and Gabriel,A. (2003) Ku-dependent and Ku-independent end-joining pathways lead to chromosomal rearrangements during double-strand break repair in Saccharomyces cerevisiae. *Genetics*, **163**, 843–856.
156. Liang,F., Romanienko,P.J., Weaver,D.T., Jeggo,P.A. and Jasin,M. (1996) Chromosomal double-strand break repair in Ku80-deficient cells. *Proc Natl Acad Sci U S A*, **93**, 8929.
157. Wang,Y., Chen,Y., Wang,C., Yang,M., Wang,Y., Bao,L., Wang,J.E., Kim,B.W., Chan,K.Y., Xu,W., *et al.* (2021) MIF is a 3' flap nuclease that facilitates DNA replication and promotes tumor growth. *Nature Communications 2021 12:1*, **12**, 1–17.
158. Ferguson,D.O., Sekiguchi,J.M., Chang,S., Frank,K.M., Gao,Y., DePinho,R.A. and Alt,F.W. (2000) The nonhomologous end-joining pathway of DNA repair is required for genomic stability and the suppression of translocations. *Proc Natl Acad Sci U S A*, **97**, 6630.
159. Villarreal,D.D., Lee,K., Deem,A., Shim,E.Y., Malkova,A. and Lee,S.E. (2012) Microhomology Directs Diverse DNA Break Repair Pathways and Chromosomal Translocations. *PLoS Genet*, **8**, 1003026.
160. Chen,R. and Wold,M.S. (2014) Replication Protein A: Single-stranded DNA's first responder : Dynamic DNA-interactions allow Replication Protein A to direct single-strand DNA intermediates into different pathways for synthesis or repair. *Bioessays*, **36**, 1156.

161. Davies,A.A., Masson,J.Y., McIlwraith,M.J., Stasiak,A.Z., Stasiak,A., Venkitaraman,A.R. and West,S.C. (2001) Role of BRCA2 in control of the RAD51 recombination and DNA repair protein. *Mol Cell*, **7**, 273–282.
162. Holliday,R. (1974) Molecular aspects of genetic exchange and gene conversion. *Genetics*, **78**, 273–287.
163. Sasaki,M., Lange,J. and Keeney,S. (2010) Genome destabilization by homologous recombination in the germline. *Nat Rev Mol Cell Biol*, **11**, 182.
164. Dhalluin,C., Carlson,J.E., Zeng,L., He,C., Aggarwal,A.K. and Zhou,M.M. (1999) Structure and ligand of a histone acetyltransferase bromodomain. *Nature*, **399**, 491–496.
165. Wu,S.Y. and Chiang,C.M. (2007) The double bromodomain-containing chromatin adaptor Brd4 and transcriptional regulation. *Journal of Biological Chemistry*, **282**, 13141–13145.
166. Zeng,L. and Zhou,M.M. (2002) Bromodomain: an acetyl-lysine binding domain. *FEBS Lett*, **513**, 124–128.
167. Li,X., Baek,G.H., Ramanand,S.G., Sharp,A., Gao,Y., Yuan,W., Welti,J., Rodrigues,D.N., Dolling,D., Figueiredo,I., *et al.* (2018) BRD4 Promotes DNA Repair and Mediates the Formation of TMPRSS2-ERG Gene Rearrangements in Prostate Cancer. *Cell Rep*, **22**, 796–808.
168. Donati,B., Lorenzini,E. and Ciarrocchi,A. (2018) BRD4 and Cancer: going beyond transcriptional regulation. *Mol Cancer*, **17**.
169. Rogakou,E.P., Pilch,D.R., Orr,A.H., Ivanova,V.S. and Bonner,W.M. (1998) DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J Biol Chem*, **273**, 5858–5868.
170. Celeste,A., Petersen,S., Romanienko,P.J., Fernandez-Capetillo,O., Chen,H.T., Sedelnikova,O.A., Reina-San-Martin,B., Coppola,V., Meffre,E., Difilippantonio,M.J., *et al.* (2002) Genomic Instability in Mice Lacking Histone H2AX. *Science*, **296**, 922.
171. Ström,L., Lindroos,H.B., Shirahige,K. and Sjögren,C. (2004) Postreplicative Recruitment of Cohesin to Double-Strand Breaks Is Required for DNA Repair. *Mol Cell*, **16**, 1003–1015.
172. Ünal,E., Arbel-Eden,A., Sattler,U., Shroff,R., Lichten,M., Haber,J.E. and Koshland,D. (2004) DNA damage response pathway uses histone modification to assemble a double-strand break-specific cohesin domain. *Mol Cell*, **16**, 991–1002.
173. Banáth,J.P., MacPhail,S.H. and Olive,P.L. (2004) Radiation Sensitivity, H2AX Phosphorylation, and Kinetics of Repair of DNA Strand Breaks in Irradiated Cervical Cancer Cell Lines. *Cancer Res*, **64**, 7144–7149.
174. Bouquet,F., Muller,C. and Salles,B. (2006) Cell Cycle The Loss of &gamma;H2AX Signal is a Marker of DNA Double Strand Breaks Repair Only at Low Levels of DNA Damage. 10.4161/cc.5.10.2799.
175. Xie,A., Puget,N., Shim,I., Odate,S., Jarzyna,I., Bassing,C.H., Alt,F.W. and Scully,R. (2005) Control of Sister Chromatid Recombination by Histone H2AX. *Mol Cell*, **16**, 1017.
176. Noubissi,F.K., McBride,A.A., Leppert,H.G., Millet,L.J., Wang,X. and Davern,S.M. (2021) Detection and quantification of γ-H2AX using a dissociation enhanced lanthanide fluorescence immunoassay. *Sci Rep*, **11**, 8945.
177. Dobzhansky,T. (1936) Studies on Hybrid Sterility. II. Localization of Sterility Factors in Drosophila Pseudoobscura Hybrids. *Genetics*, **21**, 113–135.
178. Brown Judith D and O'Neill,R.J. (2014) The Evolution of Centromeric DNA Sequences. *Wiley Online Library*.
179. Rhoades MM (1942) Preferential Segregation in Maize. *Genetics*, **27**, 641.
180. Haldane,J.B.S. (1926) A mathematical theory of natural and artificial selection. *Mathematical Proceedings of the Cambridge Philosophical Society*, **23**, 363–372.

181. Burgos,M.H. and Fawcett,D.W. (1955) Studies on the fine structure of the mammalian testis. I. Differentiation of the spermatids in the cat (Felis domestica). *J Biophys Biochem Cytol*, **1**, 287–300.
182. Willison,K., Marsh,M. and Lyon,M. (1988) Biochemical evidence for sharing of gene prodiucts in spermatogenesis. *Genet Res*, **52**, 63–63.
183. Braun,R.E., Behringer,R.R., Peschon,J.J., Brinster,R.L. and Palmiter,R.D. (1989) Genetically haploid spermatids are phenotypically diploid. *Nature*, **337**, 373–376.
184. Véron,N., Bauer,H., Weiße,A.Y., Lüder,G., Werber,M. and Herrmann,B.G. (2009) Retention of gene products in syncytial spermatids promotes non-Mendelian inheritance as revealed by the t complex responder. *Genes Dev*, **23**, 2705–2710.
185. Bhutani,K., Stansifer,K., Ticau,S., Bojic,L., Villani,A.C., Slisz,J., Cremers,C.M., Roy,C., Donovan,J., Fiske,B., *et al.* (2021) Widespread haploid-biased gene expression enables sperm-level natural selection. *Science (1979)*, **371**.
186. Immler,S., Hotzy,C., Alavioon,G., Petersson,E. and Arnqvist,G. (2014) Sperm variation within a single ejaculate affects offspring development in Atlantic salmon. *Biol Lett*, **10**.
187. Dobrovolskaia-Zavadskaia,N. and Kobozieff,N. (1932) Les souris anoures et a queue filiforme qui se reproduisent entre elles sans disjonction. *Société de biologie (Paris, France)*, **110**, 782–784.
188. Park,S., Kim,Y.H., Jeong,P.S., Park,C., Lee,J.W., Kim,J.S., Wee,G., Song,B.S., Park,B.J., Kim,S.H., *et al.* (2019) SPAM1/HYAL5 double deficiency in male mice leads to severe male subfertility caused by a cumulus-oocyte complex penetration defect. *FASEB J*, **33**, 14440–14449.
189. Cocquet,J., Ellis,P.J.I., Mahadevaiah,S.K., Affara,N.A., Vaiman,D. and Burgoyne,P.S. (2012) A Genetic Basis for a Postmeiotic X Versus Y Chromosome Intragenomic Conflict in the Mouse. *PLoS Genet*, **8**, e1002900.
190. Alavioon,G., Hotzy,C., Nakhro,K., Rudolf,S., Scofield,D.G., Zajitschek,S., Maklakov,A.A. and Immler,S. (2017) Haploid selection within a single ejaculate increases offspring fitness. *Proceedings of the National Academy of Sciences*, **114**, 8053–8058.
191. Schwander,T., Libbrecht,R. and Keller,L. (2014) Supergenes and Complex Phenotypes. *Current Biology*, **24**, R288–R294.
192. Lyon,M.F. (2003) Transmission Ratio Distortion in Mice. *Annu Rev Genet*, **37**, 393–408.
193. Lyon,M.F., Evans,E.P., Jarvis,S.E. and Sayers,I. (1979) t-Haplotypes of the mouse may involve a change in intercalary DNA. *Nature*, **279**, 38–42.
194. Lyon,M.F. (1984) Transmission ratio distortion in mouse t-haplotypes is due to multiple distorter genes acting on a responder locus. *Cell*, **37**, 621–628.
195. Berdan,E.L., Blanckaert,A., Butlin,R.K. and Bank,C. (2021) Deleterious mutation accumulation and the long-term fate of chromosomal inversions. *PLoS Genet*, **17**, e1009411.
196. Panithanarak,T., Hauffe,H.C., Dallas,J.F., Glover,A., Ward,R.G. and Searle,J.B. (2004) Linkage-dependent gene flow in a house mouse chromosomal hybrid zone. *Evolution*, **58**, 184–192.
197. Franchini,P., Colangelo,P., Solano,E., Capanna,E., Verheyen,E. and Castiglia,R. (2010) Reduced gene flow at pericentromeric loci in a hybrid zone involving chromosomal races of the house mouse Mus musculus domesticus. *Evolution*, **64**, 2020–2032.
198. Turner,J.M.A., Mahadevaiah,S.K., Ellis,P.J.I., Mitchell,M.J. and Burgoyne,P.S. (2006) Pachytene asynapsis drives meiotic sex chromosome inactivation and leads to substantial postmeiotic repression in spermatids. *Dev Cell*, **10**, 521–529.

199. Turner,J.M.A., Mahadevaiah,S.K., Fernandez-Capetillo,O., Nussenzweig,A., Xu,X., Deng,C.X. and Burgoyne,P.S. (2005) Silencing of unsynapsed meiotic chromosomes in the mouse. *Nat Genet*, **37**, 41–47.

200. Namekawa,S.H., Park,P.J., Zhang,L.F., Shima,J.E., McCarrey,J.R., Griswold,M.D. and Lee,J.T. (2006) Postmeiotic sex chromatin in the male germline of mice. *Curr Biol*, **16**, 660–667.

201. Fayer,S., Yu,Q., Kim,J., Moussette,S., Camerini-Otero,R.D. and Naumova,A.K. (2016) Robertsonian translocations modify genomic distribution of γH2AFX and H3.3 in mouse germ cells. *Mammalian Genome*, **27**, 225–236.

202. MacQueen,A.J. and Hochwagen,A. (2011) Checkpoint mechanisms: The puppet masters of meiotic prophase. *Trends Cell Biol*, **21**, 393–400.

203. Michaelis,C., Ciosk,R. and Nasmyth,K. (1997) Cohesins: chromosomal proteins that prevent premature separation of sister chromatids. *Cell*, **91**, 35–45.

204. Van Heemst,D. and Heyting,C. (2000) Sister chromatid cohesion and recombination in meiosis. *Chromosoma*, **109**, 10–26.

205. Brown,M.S. and Bishop,D.K. (2015) DNA Strand Exchange and RecA Homologs in Meiosis. *Cold Spring Harb Perspect Biol*, **7**, a016659.

206. Lao,J.P. and Hunter,N. (2010) Trying to Avoid Your Sister. *PLoS Biol*, **8**, e1000519.

207. Latos-Bielenska,A. and Vogel,W. (1990) Frequency and distribution of chiasmata in Syrian hamster spermatocytes studied by the BrdU antibody technique. *Chromosoma*, **99**, 267–272.

208. Hunter,N. and Kleckner,N. (2001) The single-end invasion: an asymmetric intermediate at the double-strand break to double-holliday junction transition of meiotic recombination. *Cell*, **106**, 59–70.

209. Shiotani,B. and Zou,L. (2009) Single-Stranded DNA Orchestrates an ATM-to-ATR Switch at DNA Breaks. *Mol Cell*, **33**, 547.

210. Zou,L. and Elledge,S.J. (2003) Sensing DNA damage through ATRIP recognition of RPA-ssDNA complexes. *Science*, **300**, 1542–1548.

211. Ball,H.L., Myers,J.S. and Cortez,D. (2005) ATRIP Binding to Replication Protein A-Single-stranded DNA Promotes ATR–ATRIP Localization but Is Dispensable for Chk1 Phosphorylation. *Mol Biol Cell*, **16**, 2372.

212. Lee,J.H. and Paull,T.T. (2004) Direct Activation of the ATM Protein Kinase by the Mre11/Rad50/Nbs1 Complex. *Science (1979)*, **304**, 93–96.

213. Namiki,Y. and Zou,L. (2006) ATRIP associates with replication protein A-coated ssDNA through multiple interactions. *Proc Natl Acad Sci U S A*, **103**, 580–585.

214. ElInati,E., Russell,H.R., Ojarikre,O.A., Sangrithi,M., Hirota,T., De Rooij,D.G., McKinnon,P.J. and Turner,J.M.A. (2017) DNA damage response protein TOPBP1 regulates X chromosome silencing in the mammalian germ line. *Proc Natl Acad Sci U S A*, **114**, 12536–12541.

215. Smith,J., Mun Tho,L., Xu,N. and A. Gillespie,D. (2010) The ATM-Chk2 and ATR-Chk1 pathways in DNA damage signaling and cancer. *Adv Cancer Res*, **108**, 73–112.

216. Tonami,Y., Murakami,H., Shirahige,K. and Nakanishi,M. (2005) A checkpoint control linking meiotic S phase and recombination initiation in fission yeast. *Proc Natl Acad Sci U S A*, **102**, 5797.

217. Sartori,A.A., Lukas,C., Coates,J., Mistrik,M., Fu,S., Bartek,J., Baer,R., Lukas,J. and Jackson,S.P. (2007) Human CtIP promotes DNA end resection. *Nature*, **450**, 509.

218. Krasner,D.S., Daley,J.M., Sung,P. and Niu,H. (2015) Interplay between Ku and Replication Protein A in the Restriction of Exo1-mediated DNA Break End Resection. *J Biol Chem*, **290**, 18806.

219. Garcia,V., Phelps,S.E.L., Gray,S. and Neale,M.J. (2011) Bidirectional resection of DNA double-strand breaks by Mre11 and Exo1. *Nature*, **479**, 241.

220. Burgoyne,P.S., Mahadevaiah,S.K. and Turner,J.M.A. (2009) The consequences of asynapsis for mammalian meiosis. *Nature Reviews Genetics 2009 10:3*, **10**, 207–216.

221. Manterola,M., Page,J., Vasco,C., Berríos,S., Parra,M.T., Viera,A., Rufas,J.S., Zuccotti,M., Garagna,S. and Fernández-Donoso,R. (2009) A High Incidence of Meiotic Silencing of Unsynapsed Chromatin Is Not Associated with Substantial Pachytene Loss in Heterozygous Male Mice Carrying Multiple Simple Robertsonian Translocations. *PLoS Genet*, **5**, 1000625.
222. Akiyoshi,B., Sarangapani,K.K., Powers,A.F., Nelson,C.R., Reichow,S.L., Arellano-Santoyo,H., Gonen,T., Ranish,J.A., Asbury,C.L. and Biggins,S. (2010) Tension directly stabilizes reconstituted kinetochore-microtubule attachments. *Nature*, **468**, 576–579.
223. Cheeseman,I.M. (2014) The kinetochore. *Cold Spring Harb Perspect Biol*, **6**.
224. Gorbsky,G.J. (2015) The spindle checkpoint and chromosome segregation in meiosis. *FEBS J*, **282**, 2471–2487.
225. Vara,C., Paytuví-Gallart,A., Cuartero,Y., Álvarez-González,L., Marín-Gual,L., Garcia,F., Florit-Sabater,B., Capilla,L., Sanchéz-Guillén,R.A., Sarrate,Z., *et al.* (2021) The impact of chromosomal fusions on 3D genome folding and recombination in the germ line. *Nat Commun*, **12**.
226. Vasco,C., Manterola,M., Page,J., Zuccotti,M., De La Fuente,R., Redi,C.A., Fernandez-Donoso,R. and Garagna,S. (2012) The frequency of heterologous synapsis increases with aging in Robertsonian heterozygous male mice. *Chromosome Research*, **20**, 269–278.
227. Wallace,B.M.N., Searle,J.B. and Everett,C.A. (1992) Male meiosis and gametogenesis in wild house mice (Mus musculus domesticus) from a chromosomal hybrid zone; a comparison between ‘simple’ Robertsonian heterozygotes and homozygotes. *Cytogenet Cell Genet*, **61**, 211–220.
228. Vara,C., Paytuví-Gallart,A., Cuartero,Y., Álvarez-González,L., Marín-Gual,L., Garcia,F., Florit-Sabater,B., Capilla,L., Sanchéz-Guillén,R.A., Sarrate,Z., *et al.* (2021) The impact of chromosomal fusions on 3D genome folding and recombination in the germ line. *Nat Commun*, **12**.
229. Abe,H., Alavattam,K.G., Hu,Y.C., Pang,Q., Andreassen,P.R., Hegde,R.S. and Namekawa,S.H. (2020) The Initiation of Meiotic Sex Chromosome Inactivation Sequesters DNA Damage Signaling from Autosomes in Mouse Spermatogenesis. *Current Biology*, **30**, 408-420.e5.
230. Shiu,P.K.T., Raju,N.B., Zickler,D. and Metzenberg,R.L. (2001) Meiotic silencing by unpaired DNA. *Cell*, **107**, 905–916.
231. Bean,C.J., Schaner,C.E. and Kelly,W.G. (2004) Meiotic pairing and imprinted X chromatin assembly in Caenorhabditis elegans. *Nat Genet*, **36**, 100–105.
232. Mahadevaiah,S.K., Turner,J.M.A., Baudat,F., Rogakou,E.P., De Boer,P., Blanco-Rodríguez,J., Jasin,M., Keeney,S., Bonner,W.M. and Burgoyne,P.S. (2001) Recombinational DNA double-strand breaks in mice precede synapsis. *Nat Genet*, **27**, 271–276.
233. Fernandez-Capetillo,O., Mahadevaiah,S.K., Celeste,A., Romanienko,P.J., Camerini-Otero,R.D., Bonner,W.M., Manova,K., Burgoyne,P. and Nussenzweig,A. (2003) H2AX Is Required for Chromatin Remodeling and Inactivation of Sex Chromosomes in Male Mouse Meiosis. *Dev Cell*, **4**, 497–508.
234. Turner,J.M.A., Aprelikova,O., Xu,X., Wang,R., Kim,S., Chandramouli,G.V.R., Barrett,J.C., Burgoyne,P.S. and Deng,C.X. (2004) BRCA1, Histone H2AX Phosphorylation, and Male Meiotic Sex Chromosome Inactivation. *Current Biology*, **14**, 2135–2142.
235. Rinaldi,V.D., Bolcun-Filas,E., Kogo,H., Kurahashi,H. and Schimenti,J.C. (2017) The DNA damage checkpoint eliminates mouse oocytes with chromosome synapsis failure. *Mol Cell*, **67**, 1026.

236. Cloutier,J.M., Mahadevaiah,S.K., ElInati,E., Nussenzweig,A., Tóth,A. and Turner,J.M.A. (2015) Histone H2AFX Links Meiotic Chromosome Asynapsis to Prophase I Oocyte Loss in Mammals. *PLoS Genet*, **11**.
237. Edelmann,W., Cohen,P.E., Kane,M., Lau,K., Morrow,B., Bennett,S., Umar,A., Kunkel,T., Cattoretti,G., Chaganti,R., *et al.* (1996) Meiotic pachytene arrest in MLH1-deficient mice. *Cell*, **85**, 1125–1134.
238. Eaker,S., Cobb,J., Pyle,A. and Handel,M.A. (2002) Meiotic prophase abnormalities and metaphase cell death in MLH1-deficient mouse spermatocytes: Insights into regulation of spermatogenic progress. *Dev Biol*, **249**, 85–95.
239. Eaker,S., Pyle,A., Cobb,J. and Handel,M.A. (2001) Evidence for meiotic spindle checkpoint from analysis of spermatocytes from Robertsonian-chromosome heterozygous mice. *J Cell Sci*, **114**, 2953–2965.
240. Charalambous,C., Webster,A. and Schuh,M. (2022) Aneuploidy in mammalian oocytes and the impact of maternal ageing. *Nature Reviews Molecular Cell Biology 2022 24:1*, **24**, 27–44.
241. Nyrén,P. and Lundin,A. (1985) Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal Biochem*, **151**, 504–509.
242. Chaitankar,V., Karakülah,G., Ratnapriya,R., Giuste,F.O., Brooks,M.J. and Swaroop,A. (2016) Next Generation Sequencing Technology and Genomewide Data Analysis: Perspectives for Retinal Research. *Prog Retin Eye Res*, **55**, 1.
243. Van der Auwera,G.A., Carneiro,M.O., Hartl,C., Poplin,R., del Angel,G., Levy-Moonshine,A., Jordan,T., Shakir,K., Roazen,D., Thibault,J., *et al.* (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*, **43**.
244. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792–1797.
245. Iacovoni,J.S., Caron,P., Lassadi,I., Nicolas,E., Massip,L., Trouche,D. and Legube,G. (2010) High-resolution profiling of γH2AX around DNA double strand breaks in the mammalian genome. *EMBO J*, **29**, 1446.
246. Massip,L., Caron,P., Iacovoni,J.S., Trouche,D. and Legube,G. (2010) Deciphering the chromatin landscape induced around DNA double strand breaks. *http://dx.doi.org/10.4161/cc.9.15.12412*, **9**, 3035–3044.
247. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
248. Ramírez,F., Ryan,D.P., Grüning,B., Bhardwaj,V., Kilpert,F., Richter,A.S., Heyne,S., Dündar,F. and Manke,T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*, **44**, W160–W165.
249. Clouaire,T., Rocher,V., Lashgari,A., Arnould,C., Aguirrebengoa,M., Biernacka,A., Skrzypczak,M., Aymard,F., Fongang,B., Dojer,N., *et al.* (2018) Comprehensive Mapping of Histone Modifications at DNA Double-Strand Breaks Deciphers Repair Pathway Chromatin Signatures. *Mol Cell*, **72**, 250.
250. Bouwman,B.A.M., Agostini,F., Garnerone,S., Petrosino,G., Gothe,H.J., Sayols,S., Moor,A.E., Itzkovitz,S., Bienko,M., Roukos,V., *et al.* (2020) Genome-wide detection of DNA double-strand breaks by in-suspension BLISS. *Nat Protoc*, **15**, 3894–3941.
251. Colgan,S.P., Eltzschig,H.K., Eckle,T. and Thompson,L.F. (2006) Physiological roles for ecto-5'-nucleotidase (CD73). *Purinergic Signal*, **2**, 351–360.
252. Hammoud,S.S., Low,D.H.P., Yi,C., Carrell,D.T., Guccione,E. and Cairns,B.R. (2014) Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. *Cell Stem Cell*, **15**, 239–253.
253. Tan,M., Luo,H., Lee,S., Jin,F., Yang,J.S., Montellier,E., Buchou,T., Cheng,Z., Rousseaux,S., Rajagopal,N., *et al.* (2011) Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell*, **146**, 1016–1028.

254. Bryant,J.M., Donahue,G., Wang,X., Meyer-Ficca,M., Luense,L.J., Weller,A.H., Bartolomei,M.S., Blobel,G.A., Meyer,R.G., Garcia,B.A., *et al.* (2015) Characterization of BRD4 during mammalian postmeiotic sperm development. *Mol Cell Biol*, **35**, 1433–1448.
255. Maezawa,S., Yukawa,M., Alavattam,K.G., Barski,A. and Namekawa,S.H. (2018) Dynamic reorganization of open chromatin underlies diverse transcriptomes during spermatogenesis. *Nucleic Acids Res*, **46**, 593–608.
256. Akatsuka,S. and Toyokuni,S. (2012) Genome-wide assessment of oxidatively generated DNA damage. *https://doi.org/10.3109/10715762.2011.633212*, **46**, 523–530.
257. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, **9**.
258. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R., *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature 2012 489:7414*, **489**, 57–74.
259. Poetsch,A.R., Boulton,S.J. and Luscombe,N.M. (2018) Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis. *Genome Biol*, **19**, 215–215.
260. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
261. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
262. Marsico,G., Chambers,V.S., Sahakyan,A.B., McCauley,P., Boutell,J.M., Antonio,M. Di and Balasubramanian,S. (2019) Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res*, **47**, 3862–3874.
263. Conway,J.R., Lex,A. and Gehlenborg,N. (2017) UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, **33**, 2938–2940.
264. Ernst,J. and Kellis,M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc*, **12**, 2478–2492.
265. Gel,B., Díez-Villanueva,A., Serra,E., Buschbeck,M., Peinado,M.A. and Malinverni,R. (2016) regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*, **32**, 289–291.
266. Heger,A., Webber,C., Goodson,M., Ponting,C.P. and Lunter,G. (2013) GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics*, **29**, 2046–2048.
267. Bailey,T.L. (2003) Discovering Novel Sequence Motifs with MEME . *Curr Protoc Bioinformatics*, **00**.
268. Thomas,P.D., Ebert,D., Muruganujan,A., Mushayahama,T., Albou,L.P. and Mi,H. (2022) PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Science*, **31**, 8–22.
269. Gu,Z., Gu,L., Eils,R., Schlesner,M. and Brors,B. (2014) circlize Implements and enhances circular visualization in R. *Bioinformatics*, **30**, 2811–2812.
270. Ramírez,F., Bhardwaj,V., Arrigoni,L., Lam,K.C., Grüning,B.A., Villaveces,J., Habermann,B., Akhtar,A. and Manke,T. (2018) High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun*, **9**, 189.
271. Lopez-Delisle,L., Rabbani,L., Wolff,J., Bhardwaj,V., Backofen,R., Grüning,B., Ramírez,F. and Manke,T. (2021) pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics*, **37**, 422–423.
272. Cingolani,P., Platts,A., Wang,L.L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single

nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80.

273. Andrews,S. FastQC: a quality control tool for high throughput sequence data.
274. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
275. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*, **50**, D439–D444.
276. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873*, **596**, 583–589.
277. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
278. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**, 1–21.
279. Martin,F.J., Amode,M.R., Aneja,A., Austine-Orimoloye,O., Azov,A.G., Barnes,I., Becker,A., Bennett,R., Berry,A., Bhai,J., *et al.* (2023) Ensembl 2023. *Nucleic Acids Res*, **51**, D933–D941.
280. Szklarczyk,D., Kirsch,R., Koutrouli,M., Nastou,K., Mehryary,F., Hachilif,R., Gable,A.L., Fang,T., Doncheva,N.T., Pyysalo,S., *et al.* (2023) The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res*, **51**, D638–D646.
281. Kaya-Okur,H.S., Wu,S.J., Codomo,C.A., Pledger,E.S., Bryson,T.D., Henikoff,J.G., Ahmad,K. and Henikoff,S. (2019) CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications 2019 10:1*, **10**, 1–10.
282. Zhang,H., Lu,T., Liu,S., Yang,J., Sun,G., Cheng,T., Xu,J., Chen,F. and Yen,K. (2021) Comprehensive understanding of Tn5 insertion preference improves transcription regulatory element identification. *NAR Genom Bioinform*, **3**.
283. Yang,G., Chen,Y., Wu,J., Chen,S.-H., Liu,X., Singh,A.K. and Yu,X. (2020) Poly(ADP-ribosyl)ation mediates early phase histone eviction at DNA lesions. *Nucleic Acids Res*, **48**, 3001–3013.
284. Iacovoni,J.S., Caron,P., Lassadi,I., Nicolas,E., Massip,L., Trouche,D. and Legube,G. (2010) High-resolution profiling of γH2AX around DNA double strand breaks in the mammalian genome. *EMBO J*, **29**, 1446–1457.
285. Lu,B., Dong,L., Yi,D., Zhang,M., Zhu,C., Li,X. and Yi,C. (2020) Transposase-assisted tagmentation of RNA/DNA hybrid duplexes. *Elife*, **9**, 1–16.
286. Moretti,C., Vaiman,D., Tores,F. and Cocquet,J. (2016) Expression and epigenomic landscape of the sex chromosomes in mouse post-meiotic male germ cells. *Epigenetics Chromatin*, **9**, 47.
287. Iannelli,F., Galbiati,A., Capozzo,I., Nguyen,Q., Magnuson,B., Michelini,F., D'Alessandro,G., Cabrini,M., Roncador,M., Francia,S., *et al.* (2017) A damaged genome's transcriptional landscape through multilayered expression profiling around in situ-mapped DNA double-strand breaks. *Nat Commun*, **8**, 15656.
288. Ernst,C., Eling,N., Martinez-Jimenez,C.P., Marioni,J.C. and Odom,D.T. (2019) Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. *Nature Communications 2019 10:1*, **10**, 1–20.
289. Lima,A.C., Jung,M., Rusch,J., Usmani,A., Lopes,A.M. and Conrad,D.F. (2017) A Standardized Approach for Multispecies Purification of Mammalian Male Germ Cells by Mechanical Tissue Dissociation and Flow Cytometry. *J Vis Exp*, **125**.

290. Farré,M., Kim,J., Proskuryakova,A.A., Zhang,Y., Kulemzina,A.I., Li,Q., Zhou,Y., Xiong,Y., Johnson,J.L., Perelman,P.L., *et al.* (2019) Evolution of gene regulation in ruminants differs between evolutionary breakpoint regions and homologous synteny blocks. *Genome Res*, **29**, 576–589.
291. Capilla,L., Sánchez-Guillén,R.A., Farré,M., Paytuví-Gallart,A., Malinverni,R., Ventura,J., Larkin,D.M. and Ruiz-Herrera,A. (2016) Mammalian Comparative Genomics Reveals Genetic and Epigenetic Features Associated with Genome Reshuffling in Rodentia. *Genome Biol Evol*, **8**, 3703–3717.
292. Keeney, Giroux,C.N. and Kleckner,N. (1997) Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell*, **88**.
293. Ahmed,E.A., Scherthan,H. and de Rooij,D.G. (2015) DNA Double Strand Break Response and Limited Repair Capacity in Mouse Elongated Spermatids. *Int J Mol Sci*, **16**, 29923–29935.
294. Wright,C., Milne,S. and Leeson,H. (2014) Sperm DNA damage caused by oxidative stress: modifiable clinical, lifestyle and nutritional factors in male infertility. *Reprod Biomed Online*, **28**, 684–703.
295. Alavattam,K.G., Maezawa,S., Sakashita,A., Khoury,H., Barski,A., Kaplan,N. and Namekawa,S.H. (2019) Attenuated chromatin compartmentalization in meiosis and its maturation in sperm development. *Nat Struct Mol Biol*, **26**, 175–184.
296. Patel,L., Kang,R., Rosenberg,S.C., Qiu,Y., Raviram,R., Chee,S., Hu,R., Ren,B., Cole,F. and Corbett,K.D. (2019) Dynamic reorganization of the genome shapes the recombination landscape in meiotic prophase. *Nat Struct Mol Biol*, **26**, 164–174.
297. Russell,L.D., Ettlin,R.A., Hikim,A.P.S. and Clegg,E.D. (1990) A Histological and Histopathological evaluation of the testis 1st edition. Cache River Press.
298. Pogany,G.C., Corzett,M., Weston,S. and Balhorn,R. (1981) DNA and protein content of mouse sperm: Implications regarding sperm chromatin structure. *Exp Cell Res*, **136**, 127–136.
299. Hud,N. V., Allen,M.J., Downing,K.H., Lee,J. and Balhorn,R. (1993) Identification of the elemental packing unit of DNA in mammalian sperm cells by atomic force microscopy. *Biochem Biophys Res Commun*, **193**, 1347–1354.
300. Roca,J. and Mezquita,C. (1989) DNA topoisomerase II activity in nonreplicating, transcriptionally inactive, chicken late spermatids. *EMBO J*, **8**, 1855.
301. Chen,J., and,F.L.-M.R. and 1996, undefined Expression and localization of DNA topoisomerase II during rat spermatogenesis. *Wiley Online Library*.
302. Szlachta,K., Manukyan,A., Raimer,H.M., Singh,S., Salamon,A., Guo,W., Lobachev,K.S. and Wang,Y.H. (2020) Topoisomerase II contributes to DNA secondary structure-mediated double-stranded breaks. *Nucleic Acids Res*, **48**, 6654–6671.
303. McQueen,D.B., Zhang,J. and Robins,J.C. (2019) Sperm DNA fragmentation and recurrent pregnancy loss: a systematic review and meta-analysis. *Fertil Steril*, **112**, 54-60.e3.
304. Damas,J., Kim,J., Farré,M., Griffin,D.K. and Larkin,D.M. (2018) Reconstruction of avian ancestral karyotypes reveals differences in the evolutionary history of macro- and microchromosomes. *Genome Biol*, **19**, 1–16.
305. Álvarez-González,L., Burden,F., Doddamani,D., Malinverni,R., Leach,E., Marín-García,C., Marín-Gual,L., Gubern,A., Vara,C., Paytuví-Gallart,A., *et al.* (2022) 3D chromatin remodelling in the germ line modulates genome evolutionary plasticity. *Nat Commun*, **13**.
306. Subramanian,V. V. and Hochwagen,A. (2014) The meiotic checkpoint network: step-by-step through meiotic prophase. *Cold Spring Harb Perspect Biol*, **6**.
307. Malinverni,R., Corujo,D., Gel,B. and Buschbeck,M. (2023) regioneReloaded: evaluating the association of multiple genomic region sets. *Bioinformatics*, **39**.

308. Spitzner,J.R., Chung,I.K. and Muller,M.T. (1990) Eukaryotic topoisomerase II preferentially cleaves alternating purine-pyrimidine repeats. *Nucleic Acids Res*, **18**, 1–11.
309. Wang,A.H.J., Quigley,G.J., Kolpak,F.J., Crawford,J.L., Van Boom,J.H., Van Der Marel,G. and Rich,A. (1979) Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature 1979 282:5740*, **282**, 680–686.
310. Haniford,D.B. and Pulleyblank,D.E. (1983) Facile transition of poly[d(TG) x d(CA)] into a left-handed helix in physiological conditions. *Nature*, **302**, 632–634.
311. Rassoulzadegan,M., Sharifi-Zarchi,A. and Kianmehr,L. (2021) Dna-rna hybrid (R-loop): From a unified picture of the mammalian telomere to the genome-wide profile. *Cells*, **10**, 1556.
312. Loir,M. and Lanneau,M. (1984) Structural function of the basic nuclear proteins in ram spermatids. *J Ultrastruct Res*, **86**, 262–272.
313. Cavé,T., Desmarais,R., Lacombe-Burgoyne,C. and Boissonneault,G. (2019) Genetic Instability and Chromatin Remodeling in Spermatids. *Genes 2019, Vol. 10, Page 40*, **10**, 40.
314. Wang,S., Hassold,T., Hunt,P., White,M.A., Zickler,D., Kleckner,N. and Zhang,L. (2017) Inefficient Crossover Maturation Underlies Elevated Aneuploidy in Human Female Meiosis. *Cell*, **168**, 977-989.e17.
315. Sobhy,H., Kumar,R., Lewerentz,J., Lizana,L. and Stenberg,P. (2019) Highly interacting regions of the human genome are enriched with enhancers and bound by DNA repair proteins. *Sci Rep*, **9**.
316. Sanders,J.T., Freeman,T.F., Xu,Y., Golloshi,R., Stallard,M.A., Hill,A.M., San Martin,R., Balajee,A.S. and McCord,R.P. (2020) Radiation-induced DNA damage and repair effects on 3D genome organization. *Nat Commun*, **11**.
317. Cho,W.K., Spille,J.H., Hecht,M., Lee,C., Li,C., Grube,V. and Cisse,I.I. (2018) Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, **361**, 412–415.
318. Manuel,M. de, Wu,F.L. and Przeworski,M. (2022) A paternal bias in germline mutation is widespread across amniotes and can arise independently of cell divisions. *bioRxiv*, 10.1101/2022.02.07.479417.
319. Makova,K.D. and Li,W.H. (2002) Strong male-driven evolution of DNA sequences in humans and apes. *Nature*, **416**, 624–626.
320. Wolfe,K.H. and Li,W.H. (2003) Molecular evolution meets the genomics revolution. *Nature Genetics 2003 33:3*, **33**, 255–265.
321. Li,W.H., Ellsworth,D.L., Krushkal,J., Chang,B.H.J. and Hewett-Emmett,D. (1996) Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol Phylogenet Evol*, **5**, 182–187.
322. Presgraves,D.C. and Yi,S. V. (2009) Doubts about complex speciation between humans and chimpanzees. *Trends Ecol Evol*, **24**, 533–540.
323. Nachman,M.W. and Crowell,S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297–304.
324. Huang,W., Chang,B.H.J., Gu,X., Hewett-Emmett,D. and Li,W.H. (1997) Sex differences in mutation rate in higher primates estimated from AMG intron sequences. *J Mol Evol*, **44**, 463–465.
325. Chang,B.H.J., Shimmin,L.C., Shyue,S.K., Hewett-Emmett,D. and Li,W.H. (1994) Weak male-driven molecular evolution in rodents. *Proc Natl Acad Sci U S A*, **91**, 827–831.
326. Shimmin,L.C., Chang,B.H.-J. and Li,W.-H. (1993) Male-driven evolution of DNA sequences. *Nature*, **362**, 745–747.
327. Crow,J.F. (2000) The origins, patterns and implications of human spontaneous mutation. *Nature Reviews Genetics 2000 1:1*, **1**, 40–47.

328. Kim,S.H., Jung,H.J., Lee,I.B., Lee,N.K. and Hong,S.C. (2021) Sequence-dependent cost for Z-form shapes the torsion-driven B–Z transition via close interplay of Z-DNA and DNA bubble. *Nucleic Acids Res*, **49**, 3651.
329. Glikin,G.C., Jovin,T.M. and Arndt-jovin,D.J. (1991) Interactions of Drosophila DNA topoisomerase II with left-handed Z-DNA in supercoiled minicircles. *Nucleic Acids Res*, **19**, 7139.
330. Arndt-Jovin,D.J., Udvardy,A., Garner,M.M., Ritter,S. and Jovin,T.M. (1993) Z-DNA binding and inhibition by GTP of Drosophila topoisomerase II. *Biochemistry*, **32**, 4862–4872.
331. In Young Choi, In Kwon Chung and Muller,M.T. (1995) Eukaryotic topoisomerase II cleavage is independent of duplex DNA conformation. *Biochim Biophys Acta*, **1264**, 209–214.
332. Mulholland,N., Xu,Y., Sugiyama,H. and Zhao,K. (2012) SWI/SNF-mediated chromatin remodeling induces Z-DNA formation on a nucleosome. *Cell Biosci*, **2**, 3.
333. Li,X. and Tyler,J.K. (2016) Nucleosome disassembly during human non-homologous end joining followed by concerted HIRA- and CAF-1-dependent reassembly. *Elife*, **5**.
334. Zimmerman,S.B. (1982) The three-dimensional structure of DNA. *Annu Rev Biochem*, **51**, 395–427.
335. Johnston,B.H. and Rich,A. (1985) Chemical probes of DNA conformation: detection of Z-DNA at nucleotide resolution. *Cell*, **42**, 713–724.
336. Lagravère,C., Malfoy,B., Leng,M. and Laval,J. (1984) Ring-opened alkylated guanine is not repaired in Z-DNA. *Nature*, **310**, 798–800.
337. Ding,Y., Fleming,A.M. and Burrows,C.J. (2017) Sequencing the Mouse Genome for the Oxidatively Modified Base 8-Oxo-7,8-dihydroguanine by OG-Seq. *J Am Chem Soc*, **139**, 2569–2572.
338. Meers,M.P., Llagas,G., Janssens,D.H., Codomo,C.A. and Henikoff,S. (2022) Multifactorial profiling of epigenetic landscapes at single-cell resolution using MulTI-Tag. *Nature Biotechnology 2022 41:5*, **41**, 708–716.
339. Chayko,C.A. and Martin-DeLeon,P.A. (1992) The murine Rb(6.16) translocation: alterations in the proportion of alternate sperm segregants effecting fertilization in vitro and in vivo. *Hum Genet*, **90**, 79–85.
340. Crossan,G.P., Van Der Weyden,L., Rosado,I. V., Langevin,F., Gaillard,P.H.L., McIntyre,R.E., Gallagher,F., Kettunen,M.I., Lewis,D.Y., Brindle,K., *et al.* (2011) Disruption of mouse Slx4, a regulator of structure-specific nucleases, phenocopies Fanconi Anemia. *Nat Genet*, **43**, 147.
341. Bernstein,I.R., McLaughlin,E.A., O’Bryan,M.K., Bernstein,I.R., McLaughlin,E.A. and O’Bryan,M.K. (2010) 324. SLX4, A KEY REGULATOR OF MEIOSIS AND DNA REPAIR IN THE MALE GERMLINE. *Reprod Fertil Dev*, **22**, 124–124.
342. Pomaznoy,M., Ha,B. and Peters,B. (2018) GOnet: a tool for interactive Gene Ontology analysis. *BMC Bioinformatics*, **19**, 470.
343. Escoffier,J., Jemel,I., Tanemoto,A., Taketomi,Y., Payre,C., Coatrieux,C., Sato,H., Yamamoto,K., Masuda,S., Pernet-Gallay,K., *et al.* (2010) Group X phospholipase A2 is released during sperm acrosome reaction and controls fertility outcome in mice. *J Clin Invest*, **120**, 1415.
344. Takada,M., Fukuhara,D., Takiura,T., Nishibori,Y., Kotani,M., Kiuchi,Z., Kudo,A., Beltcheva,O., Ito-Nitta,N., Nitta,K.R., *et al.* (2023) Involvement of GLCCI1 in mouse spermatogenesis. *FASEB J*, **37**.
345. Grzmil,P., Boinska,D., Kleene,K.C., Adham,I., Schlüter,G., Kämper,M., Buyandelger,B., Meinhardt,A., Wolf,S. and Engel,W. (2008) Prm3, the fourth gene in the mouse protamine gene cluster, encodes a conserved acidic protein that affects sperm motility. *Biol Reprod*, **78**, 958–967.

346. Cardoso-Moreira,M., Halbert,J., Valloton,D., Velten,B., Chen,C., Shao,Y., Liechti,A., Ascenção,K., Rummel,C., Ovchinnikova,S., *et al.* (2019) Gene expression across mammalian organ development. *Nature 2019 571:7766*, **571**, 505–509.
347. Soumillon,M., Necsulea,A., Weier,M., Brawand,D., Zhang,X., Gu,H., Barthès,P., Kokkinaki,M., Nef,S., Gnirke,A., *et al.* (2013) Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Rep*, **3**, 2179–2190.
348. Ng,C.J., Wadleigh,D.J., Gangopadhyay,A., Hama,S., Grijalva,V.R., Navab,M., Fogelman,A.M. and Reddy,S.T. (2001) Paraoxonase-2 is a ubiquitously expressed protein with antioxidant properties and is capable of preventing cell-mediated oxidative modification of low density lipoprotein. *J Biol Chem*, **276**, 44444–44449.
349. Bang,H., Lee,S., Jeong,P.S., Seol,D.W., Son,D., Kim,Y.H., Song,B.S., Sim,B.W., Park,S., Lee,D.M., *et al.* (2022) Hyaluronidase 6 Does Not Affect Cumulus-Oocyte Complex Dispersal and Male Mice Fertility. *Genes (Basel)*, **13**.
350. Cao,W., Ijiri,T.W., Huang,A.P. and Gerton,G.L. (2011) Characterization of a novel tektin member, TEKT5, in mouse sperm. *J Androl*, **32**, 55–69.
351. Deon,G.A., Glugoski,L., Hatanaka,T., Cavalcante Sassi,F. de M., Nogaroto,V., Bertollo,L.A.C., Liehr,T., Al-Rikabi,A., Moreira-Filho,O., Cioffi,M. de B., *et al.* (2022) Evolutionary breakpoint regions and chromosomal remodeling in Harttia (Siluriformes: Loricariidae) species diversification. *Genet Mol Biol*, **45**, 20210170.
352. Baek,M., DiMaio,F., Anishchenko,I., Dauparas,J., Ovchinnikov,S., Lee,G.R., Wang,J., Cong,Q., Kinch,L.N., Dustin Schaeffer,R., *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.
353. Schneider,S., Balbach,M., Jikeli,J.F., Fietz,D., Nettersheim,D., Jostes,S., Schmidt,R., Kressin,M., Bergmann,M., Wachten,D., *et al.* (2016) Re-visiting the Protamine-2 locus: deletion, but not haploinsufficiency, renders male mice infertile. *Sci Rep*, **6**, 36764.
354. Naumova,A.K., Naumova,A.K., Fayer,S., Leung,J.K., Boateng,K.A., Camerini-Otero,R.D., Taketo,T. and Taketo,T. (2013) Dynamics of Response to Asynapsis and Meiotic Silencing in Spermatocytes from Robertsonian Translocation Carriers. *PLoS One*, **8**.
355. Baba,D., Kashiwabara,S. ichi, Honda,A., Yamagata,K., Wu,Q., Ikawa,M., Okabe,M. and Baba,T. (2002) Mouse sperm lacking cell surface hyaluronidase PH-20 can pass through the layer of cumulus cells and fertilize the egg. *J Biol Chem*, **277**, 30310–30314.
356. Zheng,Y., Deng,X. and Martin-DeLeon,P.A. (2001) Lack of Sharing of Spam1 (Ph-20) among Mouse Spermatids and Transmission Ratio Distortion. *Biol Reprod*, **64**, 1730–1738.
357. Abi Nahed,R., Dhellemmes,M., Payré,C., Le Blévec,E., Perrier,J.P., Hennebicq,S., Escoffier,J., Ray,P.F., Loeuillet,C., Lambeau,G., *et al.* (2022) Treatment of Mouse Sperm with a Non-Catalytic Mutant of PLA2G10 Reveals That PLA2G10 Improves In Vitro Fertilization through Both Its Enzymatic Activity and as Ligand of PLA2R1. *Int J Mol Sci*, **23**, 8033.
358. Aranha,I.P. and Martin-DeLeon,P.A. (1992) Evidence for differential maturation of reciprocal sperm segregants in the murine Rb(6.16) translocation heterozygote. *Mol Reprod Dev*, **32**, 394–398.
359. Murase,R., Sato,H., Yamamoto,K., Ushida,A., Nishito,Y., Ikeda,K., Kobayashi,T., Yamamoto,T., Taketomi,Y. and Murakami,M. (2016) Group X Secreted Phospholipase A2 Releases ω3 Polyunsaturated Fatty Acids, Suppresses Colitis, and Promotes Sperm Fertility. *Journal of Biological Chemistry*, **291**, 6895–6911.
360. Skerget,S., Rosenow,M., Petritis,K. and Karr,T. (2015) Proteome Maturation in the Mouse Epididymis. *PLoS One*.

361. Cmero,M., Schmidt,B., Majewski,I.J., Ekert,P.G., Oshlack,A. and Davidson,N.M. (2021) MINTIE: identifying novel structural and splice variants in transcriptomes using RNA-seq data. *Genome Biol*, **22**, 296.
362. Rodríguez-Nuevo,A., Torres-Sanchez,A., Duran,J.M., De Guirior,C., Martínez-Zamora,M.A. and Böke,E. (2022) Oocytes maintain ROS-free mitochondrial metabolism by suppressing complex I. *Nature*, **607**, 756–761.
363. Ayala,F.J., and Coluzzi,M., (2005) Chromosome speciation: humans, Drosophila, and mosquitoes. *Proc Natl Acad Sci U S A*, **Suppl 1**, 6535–6542.
364. Coluzzi M. (1982) Spatial distribution of chromosomal inversions and speciation in Anopheline mosquitoes. *Prog Clin Biol Res*, **96**, 143-153.
365. Yang,Y., Fear,J., Hu,J., Haecker,I., Zhou,L., Renne,R., Bloom,D. and McIntyre,L.M. (2014) Leveraging biological replicates to improve analysis in ChIP-seq experiments. *Comput Struct Biotechnol J*, **9**, e201401002.
366. Schurch,N.J., Schofield,P., Gierliński,M., Cole,C., Sherstnev,A., Singh,V., Wrobel,N., Gharbi,K., Simpson,G.G., Owen-Hughes,T., *et al.* (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, **22**, 839–851.
367. He,J., Xia,M., Tsang,W.H., Chow,K.L. and Xia,J. (2015) ICA1L forms BAR-domain complexes with PICK1 and is crucial for acrosome formation in spermiogenesis. *J Cell Sci*, **128**, 3822–3836.
368. Nozawa,K., Fujihara,Y., Devlin,D.J., Deras,R.E., Kent,K., Larina,I. V., Umezu,K., Yu,Z., Sutton,C.M., Ye,Q., *et al.* (2022) The testis-specific E3 ubiquitin ligase RNF133 is required for fecundity in mice. *BMC Biol*, **20**.
369. Zhang,R., Xu,J., Shen,C., Zhang,X., Li,S., Lv,J., Xu,D., Huang,X., Zheng,B., Liu,M., *et al.* (2022) Testis-enriched Asb12 is not required for spermatogenesis and fertility in mice. *Transl Androl Urol*, **11**, 168–178.
370. Zhou,J., Du,Y.R., Qin,W.H., Hu,Y.G., Huang,Y.N., Bao,L., Han,D., Mansouri,A. and Xu,G.L. (2009) RIM-BP3 is a manchette-associated protein essential for spermiogenesis. *Development*, **136**, 373–382.

8. Supplementary Information

Supplementary Table 8-1: Details of the male mice used in these studies.

Mouse number	mouse age	sample name	Used for:	Comments	generation
57830	9w6d	Rb1	Whole genome sequencing	sibling of Rb2	first generation from the mice shipped to Charles River
57831	9w6d	Rb2	Whole genome sequencing	sibling of Rb1	
57883	23w3d	Hom_1	Spermatid sorting &RNA-seq	siblings	
57884	23w3d	Hom_2	Spermatid sorting &RNA-seq		
57885	23w3d	Hom_3	Spermatid sorting &RNA-seq		
BRYB134.6e	22w1d	WT_1	Spermatid sorting &RNA-seq	siblings	
BRYB134.6f	22w1d	WT_2	Spermatid sorting &RNA-seq		
BRYB134.6g	22w1d	WT_3	Spermatid sorting &RNA-seq		
M2004833	24w	Het_1	Spermatid sorting &RNA-seq	siblings	
M2004834	24w	Het_2	Spermatid sorting &RNA-seq		
M2004835	24w	Het_3	Spermatid sorting &RNA-seq		

Supplementary Table 8-2: Samtools coverage statistics of the WGS results of samples Rb1 and Rb2.

Rb1	start position	End position	Number of reads	Covered bases	% coverage	mean depth	mean baseq	mean mapq
chr1	1	1.95E+08	35,257,799	1.91E+08	98.0	26.8	35.3	56.4
chr2	1	1.82E+08	56,911,003	1.77E+08	97.5	32.4	35.3	42.3
chr3	1	1.6E+08	28,159,020	1.56E+08	97.6	26.1	35.3	57.1
chr4	1	1.57E+08	28,321,544	1.53E+08	97.2	26.8	35.3	55.5
chr5	1	1.52E+08	27,404,878	1.48E+08	97.5	26.8	35.3	56.0
chr6	1	1.5E+08	28,076,914	1.46E+08	97.3	27.1	35.3	55.8
chr7	1	1.45E+08	26,133,231	1.42E+08	97.7	26.8	35.3	54.2
chr8	1	1.3E+08	23,221,914	1.25E+08	96.4	26.5	35.3	56.4
chr9	1	1.24E+08	40,419,868	1.21E+08	97.1	39.0	35.3	35.4

Rb1	start position	End position	Number of reads	Covered bases	% coverage	mean depth	mean baseq	mean mapq
chr10	1	1.31E+08	23,363,873	1.27E+08	97.4	26.6	35.3	56.6
chr11	1	1.22E+08	21,843,264	1.18E+08	97.1	26.5	35.3	57.8
chr12	1	1.2E+08	22,634,279	1.15E+08	96.2	27.3	35.3	53.1
chr13	1	1.21E+08	21,696,503	1.17E+08	96.5	26.6	35.3	55.7
chr14	1	1.25E+08	22,768,199	1.21E+08	96.4	26.7	35.3	53.1
chr15	1	1.04E+08	18,257,908	1.01E+08	96.8	26.1	35.3	57.6
chr16	1	9.80E+07	17,172,180	9.45E+07	96.4	26.0	35.3	57.5
chr17	1	9.53E+07	16,984,908	9.15E+07	96.1	26.4	35.3	56.7
chr18	1	9.07E+07	15,830,212	8.74E+07	96.4	25.9	35.3	57.7
chr19	1	6.14E+07	10,597,310	5.80E+07	94.5	25.6	35.3	57.9
chrX	1	1.69E+08	16,234,693	1.64E+08	96.8	14.2	35.3	48.1
chrY	1	9.15E+07	5,688,973	5.81E+07	63.5	9.1	35.3	13.6

Rb2	start position	End position	Number of reads	Covered bases	% coverage	mean depth	mean baseq	mean mapq
chr1	1	1.95E+08	35,137,409	1.91E+08	98.0	26.7	35.5	56.3
chr2	1	1.82E+08	55,644,164	1.77E+08	97.5	32.3	35.5	42.8
chr3	1	1.6E+08	27,955,235	1.56E+08	97.6	26.0	35.5	57.1
chr4	1	1.57E+08	28,330,487	1.53E+08	97.2	26.8	35.5	55.5
chr5	1	1.52E+08	27,409,597	1.48E+08	97.5	26.8	35.5	56.0
chr6	1	1.5E+08	27,921,712	1.46E+08	97.3	27.0	35.5	55.8
chr7	1	1.45E+08	26,232,455	1.42E+08	97.7	26.9	35.5	54.2
chr8	1	1.3E+08	23,237,529	1.25E+08	96.4	26.5	35.5	56.4
chr9	1	1.24E+08	39,478,868	1.21E+08	97.1	38.6	35.5	36.1
chr10	1	1.31E+08	23,307,301	1.27E+08	97.4	26.5	35.5	56.5
chr11	1	1.22E+08	21,940,226	1.18E+08	97.1	26.7	35.5	57.8
chr12	1	1.2E+08	22,451,013	1.15E+08	96.2	27.3	35.5	53.2
chr13	1	1.21E+08	21,644,825	1.17E+08	96.5	26.6	35.5	55.7
chr14	1	1.25E+08	22,622,989	1.21E+08	96.4	26.6	35.5	53.1
chr15	1	1.04E+08	18,220,930	1.01E+08	96.8	26.0	35.5	57.6
chr16	1	9.80E+07	17,077,237	9.45E+07	96.4	25.9	35.5	57.5
chr17	1	9.53E+07	17,012,630	9.15E+07	96.1	26.4	35.5	56.7
chr18	1	9.07E+07	15,782,921	8.74E+07	96.4	25.8	35.5	57.7
chr19	1	6.14E+07	10,609,472	5.80E+07	94.5	25.6	35.5	57.9
chrX	1	1.69E+08	16,268,622	1.64E+08	96.8	14.2	35.5	47.7
chrY	1	9.15E+07	6,011,789	5.77E+07	63.1	9.6	35.4	14.4

.

```
Chr6_pair1: 3193727-3194012

WT              -----------------------------------------------------ACACACC
Rob             CAGAGAGTGAAATCACACAAGATGATTGGATAGTAACAGAGCCTGCTGGAGAAACACACC
                                                                     *******
                                                                     >>>>>>>

WT              AAGCCAGAGAAAATGGTTCTCTCCGAGTACCTTAGGGGCAAACTGAGTGGGCAGCCCTAA
Rob             AAGCCAGAGAAAATGGTTCTCTCCGAGTACCTTAGGGGCAAACTGAGTGGGCAGCCCTAA
                ************************************************************
                >>>>>>>>>>>>>

WT              GGCTTCGGGGCATGTGCTGAATGTGAGTGCTGGG--------------------------
Rob             GGCTTCGGGGCATGTGCTGAATGTGAGTGCTGGGAACTGGAGCCAGGTTCAACTGGAAGT
                **********************************

WT              ------------------------------------------------------------
Rob             GGGTAAGCTTTGCTGGATTCCCTGAGGAGAAGGCTTTTGGGTATGCACTGAATGTGAGTG


WT              -----AACTGGAGCCAGGTTCAACTGGAAGTGGGTAAGCTTTGCTGGATTCCCTGAGGAG
Rob             CTGGAAACTGGAGCCAGGTTCAACTGGAAGTGGGTAAGCTTTGCTGGATTCCCTGAGGAG
                     *******************************************************

WT              AAGGCTTCAGGGTATGCGCTGAACGTGAGTGCTGGGAACTGGAGCCGGGTTCAACTGGAA
Rob             AAGGCTTCAGGGTATGCGCTGAACGTGAGTGCTGGGAACTGGAGCCGGGTTCAACTGGAA
                ************************************************************

WT              GTGGGTAAGCATTGCTGGATTCCCTGAGGAGAAGGCTGAGGGGATGCATGCCTGTTGGCG
Rob             GTGGGTAAGCATTGCTGGATTCCCTGAGGAGAAGGCTGAGGGGATGCATGCCTGTTGGCG
                ************************************************************
                                                                    <<<<<<<<

WT              CAGTAACCAA--------------------------------------------------
Rob             CAGTAACCAAGCTGACGGTTTTCCATTTCTTGCTGAAGGCTGGATGTCTGTTTCAGCTCC
                **********
                <<<<<<<<<<

The insertion in the Robertsonian sample is shown in bold.
```

Supplementary Figure 8-1: Chr6 primer pair 1 locations.

```
Chr6_pair2: 3319588-3319792

WT              ------GCAGATATCACAGCAGCACCAGCAGCAGAGGCAGCAGAGGCAGCAGAGGCAGCA
Rob             AGCTTCGCAGATATCACAGCAGCACCAGCAGCAGAGGCAGCAGAGGCAGCAGAGGCAGCA
                      ******************************************************
                      >>>>>>>>>>>>>>>>>>>>

WT              CCAGCAGCAGCAGAGGCAGCAGAGGCAGCAGCAGCAGAGGCAGCAGAGGCAGCA------
Rob             CCAGCAGCAGCAGAGGCAGCAGAGGCAGCAGCAGCAGAGGCAGCAGAGGCAGCAGAGGCA
                ******************************************************

WT              ------------------------------------------------------GCAGCA
Rob             GCAGCAGCAGAGGCAGCAGAGGCAGCAGAGGCAGCAGCAGCAGAGGCAGCAGAGGCAGCA
                                                                      ******

WT              GAGGCAGCACCAGCACCAGCAGCCGCCACCTCCACAACCACCACACTTCCAGTCTCCTGG
Rob             GAGGCAGCACCAGCACCAGCAGCTGCCACCTCCACAACCACCACACTTCCAGTCTCCTGG
                *********************** ************************************

WT              GGCAGCTCCCCAAGGAGGGAGTGGTGGGGACAGAAACCTCACCCCTCCAT----------
Rob             GGCAGCTCCCCAAGGAGGGAGTGGTGGGGACAGAAACCTCACCCCTCCATCCCAGTGTCC
                **************************************************
                                              <<<<<<<<<<<<<<<<<<<<
The insertion in the Robertsonian sample is shown in bold.
```

Supplementary Figure 8-2: Chr 6 primer pair 2 locations.

```
Chr6_pair3: 3488080-3488239

WT              AGGAGACCAGACTGAACACTAAAGTGTCTTAGGGAGTGACTGCAACAAAGCTGGTAAAAG
Rob             AGGAGACCAGACTGAACACTAAAGTGTCTTAGGGAGTGACTGCAACAAAGCTGGTAAAAG
                ************************************************************
                >>>>>>>>>>>>>>>>>>>>


WT              ATCGAGGTGTGCTGAGCCATGGTGCAAGAGATGT--------------------------
Rob             ATCGAGGTGTGCTGAGCCATGGTGCAAGAGATGTGGGGTGTGCTGAGCCATGGGGCAAGA
                **********************************

WT              ---------GGGGTTTGCTGAGCCATGGGGCAAAAGATGGAGGGGTGCTAAGGCATGGTG
Rob             GATGGGGGAGGGGTTTGCTGAGCCATGGGGCAAAAGATGGAGGGGTGCTAAGGCATGGTG
                         ***************************************************
                                                                        <<<<

WT              CAAGAGAGGAGTGGG---------------------------------------------
Rob             CAAGAGAGGAGTGGGGGTGCTGAGCAATGTGTGCAAAGGCTGGTGCTACCTAACACATCA
                ***************
                <<<<<<<<<<<<<<<
The insertion in the Robertsonian sample is shown in bold.
```

Supplementary Figure 8-3: Chr 6 primer pair 3 locations.

```
Chr6_pair4: 6063449-6063601

WT              -----------------------------------AGTTCAAATCCCAGCAACCACATGG
Rob             GGTTAAAAGCACTGGCTACTCTTCCAGAGGTCCAGAGTTCAAATCCCAGCAACCACATGG
                                                   *************************
                                                   >>>>>>>>>>>>>>>>>>>>>

WT              TGGCTCACAACTATCTGTAATGGGATCCAAGGCCTT------------------------
Rob             TGGCTCACAACTATCTGTAATGGGATCCAAGGCCTTCTTCTGGTGTGTCAGCAACAGCAA
                ************************************

WT              ---------------------------TATGTCTCTCTCTCTGTCTCTCTCTCTCTCTGT
Rob             ACTCATATACATTAAATGTGTGTCTTGTATGTCTCTCTCTCTGTCTCTCTCTCTCTCTGT
                                           *********************************

WT              CTGTCTCTGTCTCTCTCTCTTCCAGTGCCATGTCTGTCTACGTCCTGCCATGATGATCA-
Rob             CTGTCTCTGTCTCTCTCTCTTCCAGTGCCATGTCTGTCTACGTCCTGCCATGATGATCAT
                ***********************************************************
                                                      <<<<<<<<<<<<<<<<<<<<<
The insertion in the Robertsonian sample is shown in bold.
```

Supplementary Figure 8-4: Chr6 primer pair 4 locations.

```
Chr16_pair1:3822677-3822839
WT              ------------------------------TGGAGAAAGAGGGGAGAGGGGGGATAAAGG
Rob             AATCGAGGAGAGAGGCCAGCCAGGAATACATGGAGAAAGAGGGGAGAGGGGGGATAAAGG
                                  ******************************************
                                              >>>>>>>>>>>>>>>>>>>>

WT              GATCAGGGAGAGAAGAGAGTCAGAGAGAGAGAAGGGGCAGAGAGATCAAGGAGGGGTTT-
Rob             GATCAGGGAGAGAAGAGAGTCAGAGAGAGAGAAGGGGCAGAGAGATCAAGGAGGGGTTTG
                ***********************************************************

WT              ----------------------------------GGTTTGGTTTGGTTTGGTTTGGTTTT
Rob             GTTTGGTTTGGTTTGGTTTGGTTTGGTTTGGTTGGGTTTGGTTTGGTTTGGTTTGGTTTT
                                                  **************************

WT              TGGAGACAGGGTTTCTCTGTATAGCCCTGGCTGTCCTGGAACTCATTCTGTAGACCAGGC
Rob             TGGAGACAGGGTTTCTCTGTATAGCCCTGGCTGTCCTGGAACTCATTCTGTAGACCAGGC
                ************************************************************
                                                                         <<<
WT              TGGCCTCAAACTCA----------------------------------------------
Rob             TGGCCTCAAACTCAGAAATCCACCTGCCTTTCCCTCCCAAGTACTGGGATTAAAGGAGTG
                **************
                <<<<<<<<<<<<<<
The insertion in the Robertsonian sample is shown in bold.
```

Supplementary Figure 8-5: Chr16 primer pair 1 locations.

```
Chr16_pair2:4881766-4881944
WT              ----ACCCTCCTCTCCAGTCTCTTCTGACAGGGCTGTTCTGGGCAAGTCAGCCATTCTCA
Rob             CCTGACCCTCCTCTCCAGTCTCTTCTGACAGGGCTGTTCTGGGCAAGTCAGCCATTCTCA
                    ********************************************************
                    >>>>>>>>>>>>>>>>>>>>

WT              TCTCTCTGTGTTGGAGAGCCTGTGGGCCTGCATTCCTCT---------------------
Rob             TCTCTCTGTGTTGGAGAGCCTGTGGGCCTGCATTCCTCTGTCAGTGTCAGCAGCTATCTC
                ***************************************

WT              ------------------------------------------------------------
Rob             TGTCACCATATGATTGCTGATGTTCCTTGTTCTCTTGGCACTCCTCTCTGCCTCTCCTGT


WT              ------------------------------------------------GTCAGCCCAGTC
Rob             TTCTGTTATTTTCTTTTGTTTTCTATGGTCATGATTCCTCCTTACCCAGTCAGCCCAGTC
                                                                ************

WT              GCCTTCAAGTTGCTGCTGTTCTCTGCTTTCCATGTGTTCTTCCTTGAAATGGTTTTCCTG
Rob             GCCTTCAAGTTGCTGCTGTTCTCTGCTTTCCATGTGTTCTTCCTTGAAATGGTTTTCCTG
                ************************************************************
                                                                    <<<<<<<<

WT              CCCCTCCCATTT------------------------------------------------
Rob             CCCCTCCCATTTTTTGCAACTTTTATTTTAGCCTGGTAGTGGTTCTTAAAAAACCCATTG
                ************
                <<<<<<<<<<<<
```

The insertion in the Robertsonian sample is shown in bold.

Supplementary Figure 8-6: Chr 16 primer pair 2 locations.

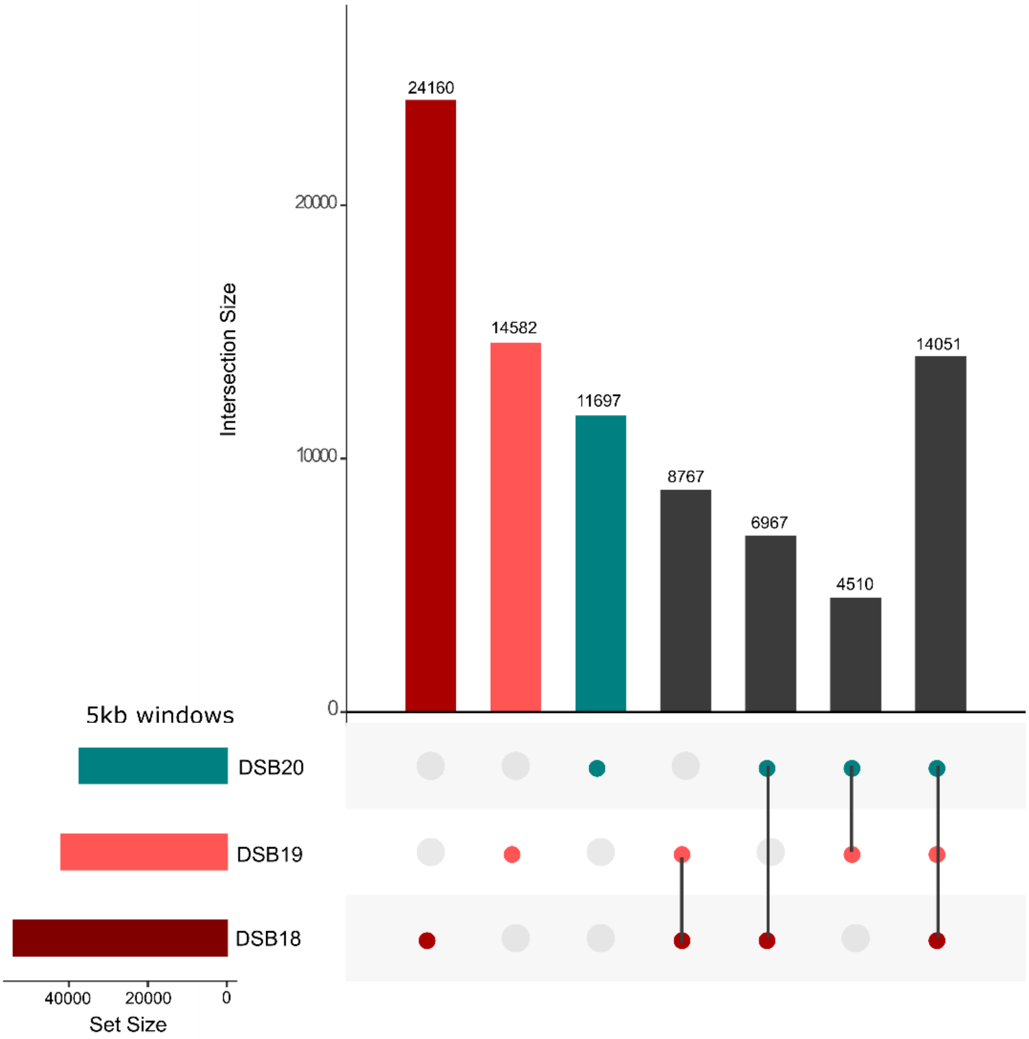
```
chr16_pair3:8287394-8287613

WT              GGCCTCCCATTACCTGTGCTCTAGCTCCCACAGTTCCAAAGCAATGGAGCTAGCAATGCT
Rob             GGCCTCCCATTACCTGTGCTCTAGCTCCCACAGTTCCAAAGCAATGGAGCTAGCAATGCT
                ************************************************************
                >>>>>>>>>>>>>>>>>>

WT              GGACCTCTGAAAGCATGAGACAAAAGAAGTCATTATTTCCTTAGGTGGTTTCTGAAAGCA
Rob             GGACCTCTGAAAGCATGAGACAAAAGAAGTCATTATTTCCTTAG----------------
                ********************************************

WT              TGAGACAAAAGAAGTCATTATTTCCTTAAGTGGTTTTTCCTGAGTATTTATCACAGCCAC
Rob             -----------------------------GTGGTTTTTCCTGAGTATTTATCACAGCCAC
                                             *******************************

WT              AAAAATAAGTGACTAAATACAGCCTCCAACCTTAAGCACT
Rob             AAAAATAAGTGACTAAATACAGCCTCCAACCTTAAGCACT
                ****************************************
                                    <<<<<<<<<<<<<<<<<<<<
```

The deletion in the Robertsonian sample is shown in bold.

Supplementary Figure 8-7: Chr 16 primer pair 3 locations.

```
Chr16_pair4:10239690-10239873

WT              --------------------------------------------CAGGCACCTCATATCC
Rob             CCTTAGAGGCTGCAGCCACAGGAGAGAAGCAGTGAGGCCTGGGGCAGGCACCTCATATCC
                                                            ****************
                                                            >>>>>>>>>>>>>>>>

WT              TCCATGAAATTCCAGTTATGTGTCCCTAGGACTGCAGGGCCCTTGGTCACTCTGCTCCTG
Rob             TCCATGAAATTCCAGTTATGTGTCCCTAGGACTGCAGGGCCCTTGGTCACTCTGCTCCTG
                ************************************************************
                >>>>

WT              AGCTGGGCTC----------------------------------------TGACAATTGT
Rob             TGCTGGGCTCTGACATACTAGTCGGGACCATGTTGAATGTGCTAAGATCATGACAATTGT
                 *********                                        **********

WT              GCGCTGAATGGTCTTCAGAACTGAGGGAGTGGTGTGCAACTAACTCTGTCCCAGGCTCAT
Rob             GCGCTGAATGGTCTTCAGAACTGAGGGAGTGGTGTGCAACTAACTCTGTCCCAGGCTCAT
                ************************************************************

WT              CCAGGTATTGGTTTGCTGTGAGGAGACA--------------------------------
Rob             CCAGGTATTGGTTTGCTGTGAGGAGACAAGGGCTCTTCTGCCCTGTAGATTCCACTTTCC
                ****************************
                        <<<<<<<<<<<<<<<<<<<<
```

Supplementary Figure 8-8: Chr16 prmer pair 4 locations. The Robertsonian insertion is shown in bold.



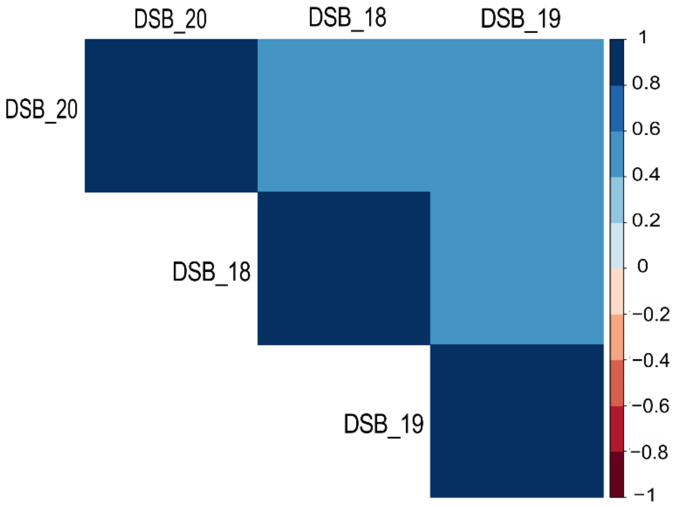
Supplementary Figure 8-9: UpSetR plot showing the overlapping and unique **5kb windows** between the round spermatid DSB files 18/19 and the condensing spermatid file DSB20, related to Figure 4-9.
The number of 5kb windows with a DSB signal in each file is represented on the left barplot as 'set size'. The X-axis represents the number of 5kb windows containing a DSB signal for the different overlap combinations. Different combinations of overlap are represented by the black lines interlinking the coloured circles. The DSB18/19 files are round spermatids stages 1-9 (shown in red and pink bars) and represent the total number of 5kb windows containing a signal unique to these files. The DSB20 file (shown in the blue bar) is condensing spermatids stage 15-16, and this peak also represents the total number of 5kb windows containing a DSB unique to this file. Bars in dark grey represent the number of 5kb windows with signal in more than one file.

Supplementary Table 8-3: % Coverage of the 16 histone marks per chromosome as a percentage of the chromosome length.

Mark	5hmC	BRD4	H2AZ	H3K27ac	H3K27me3	H3K4me1	H3K4me3	H3K9ac	H3K9me3	H4K12ac	H4K16ac	H4K5ac	H4K8ac	H4Kac	kac	kcr
coverage genome wide	3.708	0.052	0.003	1.335	3.828	5.912	4.594	3.521	0.057	2.63	0.077	1.742	1.607	2.87	0.001	1.777
chr1	2.756	0.034	0.003	0.428	3.045	4.634	3.661	2.804	0.019	2.247	0.055	1.473	1.360	2.417	0.0007	1.398
chr2	4.419	0.037	0.003	0.449	4.266	6.187	4.667	3.433	0.012	2.907	0.079	1.919	1.742	3.090	0.0013	1.748
chr3	2.352	0.033	0.003	0.415	3.088	4.547	3.658	2.784	0.029	2.155	0.062	1.369	1.302	2.307	0.0014	1.337
chr4	4.929	0.056	0.002	0.584	4.336	6.177	4.754	3.486	0.037	2.842	0.096	1.945	1.746	3.127	0.0010	1.706
chr5	5.187	0.043	0.002	0.629	5.365	7.572	5.543	4.142	0.064	3.428	0.096	2.309	2.069	3.651	0.0022	2.252
chr6	3.352	0.057	0.002	0.455	3.379	4.928	3.929	2.998	0.011	2.380	0.086	1.568	1.466	2.601	0.0015	1.463
chr7	4.192	0.041	0.002	0.649	4.417	6.896	5.225	3.821	0.071	3.206	0.112	2.154	1.926	3.460	0.0014	2.105
chr8	4.675	0.058	0.001	0.592	4.496	6.421	4.751	3.461	0.048	2.867	0.117	1.892	1.730	3.094	0.0010	1.732
chr9	4.724	0.074	0.013	0.495	4.242	6.546	4.969	3.606	0.028	3.022	0.080	1.961	1.767	3.189	0.0008	1.910
chr10	3.828	0.032	0.002	0.454	4.675	6.624	4.893	3.659	0.009	2.927	0.065	1.920	1.749	3.106	0.0006	1.971
chr11	7.267	0.073	0.009	0.524	6.087	8.879	6.400	4.575	0.031	4.025	0.136	2.665	2.319	4.192	0.0012	2.406
chr12	3.654	0.041	0.002	0.471	3.517	5.092	4.098	3.099	0.037	2.492	0.076	1.627	1.526	2.700	0.0013	1.507
chr13	2.867	0.063	0.003	0.494	4.198	5.508	4.368	3.156	0.027	2.426	0.070	1.594	1.490	2.631	0.0016	1.566
chr14	2.807	0.065	0.003	1.260	3.969	6.164	5.500	3.979	0.483	3.053	0.078	2.106	2.055	3.336	0.0003	2.126
chr15	4.590	0.028	0.002	0.435	4.633	6.261	4.846	3.641	0.009	2.934	0.090	1.932	1.735	3.128	0.0010	1.896
chr16	3.255	0.044	0.003	0.445	3.412	5.151	4.224	3.331	0.043	2.627	0.091	1.791	1.651	2.883	0.0014	1.804
chr17	5.767	0.052	0.007	0.618	5.090	7.364	5.642	4.050	0.052	3.465	0.086	2.362	2.080	3.716	0.0013	2.163
chr18	3.247	0.026	0.003	0.379	3.802	5.141	3.936	2.995	0.013	2.347	0.048	1.539	1.425	2.579	not in file	1.484
chr19	5.638	0.059	0.003	0.45	5.568	7.606	5.421	3.847	0.008	3.317	0.098	2.175	1.941	3.451	0.0012	2.106
chrX	0.070	0.084	5E-04	1.401	0.191	3.474	2.866	3.146	0.107	0.621	0.012	0.399	0.483	1.094	not in file	1.000
chrY	0.03	0.117	not in file	22.595	0.180	5.118	4.812	5.212	0.045	0.624	0.003	0.334	0.594	1.231	not in file	2.608

Supplementary Table 8-4: Percentage coverage of each chromosome by the 16 spermatid ChromHMM states related to Figure 4-17.
cov= Percentage coverage.

state	cov chr1	cov chr2	cov chr3	cov chr4	cov chr5	cov chr6	cov chr7	cov chr8	cov chr9	cov chr10	cov chr11	cov chr12	cov chr13	cov chr14	cov chr15	cov chr16	cov chr17	cov chr18	cov chr19	cov chrX	cov chrY
E1	0.93	1.25	0.91	1.23	1.48	1.01	1.41	1.24	1.31	1.31	1.78	1.05	1.06	1.26	1.28	1.10	1.49	1.03	1.41	0.68	0.86
E2	0.94	1.27	0.92	1.30	1.57	1.04	1.48	1.25	1.34	1.31	1.84	1.06	1.07	1.37	1.31	1.21	1.55	0.97	1.55	0.27	0.13
E3	0.63	0.69	0.55	0.69	0.75	0.62	0.71	0.77	0.68	0.63	0.80	0.67	0.59	0.73	0.68	0.70	0.83	0.68	0.66	0.31	0.20
E4	1.51	1.34	1.54	1.42	1.53	1.47	1.48	1.48	1.37	1.48	1.16	1.47	1.49	1.74	1.46	1.49	1.44	1.52	1.16	1.84	2.19
E5	0.55	0.59	0.58	0.59	0.77	0.57	0.69	0.60	0.66	0.73	0.75	0.62	0.67	0.95	0.66	0.64	0.74	0.58	0.65	1.98	3.97
E6	1.09	1.25	1.02	1.25	1.39	1.10	1.35	1.34	1.40	1.27	1.64	1.08	1.18	1.50	1.26	1.12	1.52	1.07	1.43	1.25	0.77
E7	0.76	1.04	0.82	1.03	1.18	0.82	1.05	1.08	1.04	1.21	1.43	0.86	1.05	0.88	1.12	0.81	1.15	0.97	1.29	0.14	0.01
E8	1.92	2.57	1.84	2.69	3.19	2.08	2.47	2.74	2.63	2.80	3.59	2.06	2.51	1.95	2.65	2.08	2.94	2.17	3.30	0.12	0.04
E9	3.94	5.78	3.19	5.97	6.62	4.61	5.22	6.19	6.70	4.68	7.98	4.29	5.85	4.01	5.32	4.40	6.30	4.65	7.15	0.06	0.03
E10	21.50	23.80	19.70	19.84	20.90	22.50	22.10	21.90	26.30	22.00	24.60	21.00	25.40	20.46	20.14	20.46	21.90	23.57	25.10	0.74	0.15
E11	37.80	29.60	41.00	32.56	28.70	34.90	29.90	31.10	25.70	33.3	20.70	36.00	31.40	36.20	33.77	37.44	27.10	33.43	23.30	7.67	5.73
E12	13.40	11.80	15.00	13.82	12.10	13.30	14.10	12.70	10.6	12.9	8.65	13.30	12.60	13.76	12.99	13.16	12.40	13.54	12.20	14.40	13.90
E13	2.37	2.15	1.94	1.72	1.95	1.66	2.18	1.78	2.26	2.10	1.96	2.16	2.26	2.27	1.78	1.92	1.81	2.11	1.96	51.20	51.70
E14	2.06	2.06	1.89	1.77	2.03	1.83	1.98	1.99	2.20	2.02	2.13	1.97	2.15	2.19	1.82	1.89	1.85	1.98	1.84	18.40	20.30
E15	2.46	3.65	2.04	3.58	4.52	2.82	3.91	3.59	3.82	3.20	5.76	2.86	2.28	2.47	3.65	2.67	4.41	2.50	4.53	0.72	0.06
E16	8.16	11.10	7.06	10.52	11.30	9.63	10.00	10.2	12.00	8.98	15.30	9.59	8.46	8.27	10.09	8.91	12.60	9.22	12.50	0.35	0.02



Supplementary Figure 8-10: Corelation of the post-meiotic spermatid DSB files using 1kb windows. DSB18/19=round spermatid samples and DSB20=condensing spermatid sample.

Supplementary Table 8-5: Peak statistics of the publicly available files used.

File	source	No. of peaks	whole genome coverage (%)	Mean (bp)	Median (bp)	Max (bp)	min (bp)
amerged spermatid DSB ERR1886418/19/20	PRJEB20038	151732	1.490	267	213	6662	146
mESC ZSCAN4	GSM4175885_GFP-Zscan_GFP_ChIP	3.33E+07	31.389	26	22	163	1
predicted Z-DNA no overlap with STR	Non-B DNA DB	225495	0.085	10	9	58	9
STR no overlap predicted Z-DNA	Non-B DNA DB	3.07E+06	2.037	18	13	2472	9
predicted G-quadruplex (Non-B DB)	Non-B DNA DB	490971	0.602	33	27	1029	14
G-quadruplex experimental: GSM3003548_Mouse_all_w15_th-1_minus.hits.max.PDS.w50.35	PMID: 30892612	873010	4.465	139	135	2340	15
SRR625513_H3K4me3_rep1	GSM1046840	85536	2.840	906	418	14961	155
SRR625514_H3K4me3_rep2	GSM1046841	133994	5.000	1017	505	33106	167
SRR625515_H3K27me3_rep1	GSM1046842	57308	2.560	1215	557	149116	155
SRR625516_H3K27me3_rep2	GSM1046843	144925	5.610	1056	481	150067	156
SRR948800.2_H3K4me3	GSM1202707	76720	4.594	1632	785	27888	276
SRR948805.2_H3K27me3	GSM1202710	75550	3.828	1381	782	149722	272
SRR948807_H3K4me1	GSM1202712	136147	5.912	1183	734	35558	260
SRR948811.2_H3K27ac	GSM1202715	112712	1.335	323	247	7625	181
SRR948814.1_5hmC	GSM1202718	241283	3.708	419	321	9895	179
SRR948819/SRR948820_H2AZ	GSM1202722	820	0.003	106	74	1876	47
SRR350960.2_Kac	GSM810677	213	1.03E-03	132	112	561	36
SRR350907.2_Kcr	GSM810678	64131	1.777	755	455	11473	179
SRR1596612.1_BRD4	GSM1519002	4072	0.052	348	220	47342	87
SRR1596613.1_H3K9me3	GSM1519003	4883	0.057	317	221	11529	70
SRR1596614.1H3K9ac	GSM1519004	79466	3.521	1208	485	30213	159
SRR1596615.1_H4K5ac	GSM1519005	43975	1.742	1080	631	21109	148
SRR1596616.1_H4K8ac	GSM1519006	44629	1.607	981	608	29196	135
SRR1596617.1_H4K12ac	GSM1519007	58141	2.630	1233	641	61024	145
SRR1596618.1_H4K16ac	GSM1519008	5291	0.077	397	266	4568	86
SRR1596619.1_H4Kac	GSM1519009	62497	2.870	1252	741	60970	159
DRR124323.1_H3-C#1	PRJDB6711	266	0.005	526	358	5711	252
DRR124324.1_H3-C#2		2418	0.048	545	407	6645	248
DRR124330.1_H3-N#1		22	0.001	1308	868	4174	256
DRR124335.1_H3-N#2		4547	0.184	1100	769	7997	255
DRR124328.1_H3K4me3#1		3273	0.057	474	404	4090	247
DRR124333.1_H3K4me3#2		18901	0.458	660	533	4301	254
DRR124329.1_H3K9me3#1		767	0.019	686	409	10897	253

File	source	No. of peaks	whole genome coverage (%)	Mean (bp)	Median (bp)	Max (bp)	min (bp)
DRR124334.1_H3K9me3#2		6117	0.164	729	465	15858	244
DRR124331.1_H4-#1		14	0.000	841	547.5	2419	281
DRR124336.1_H4-#2		147	0.003	603	379	2835	247
sBLISS_GSM4322063_Enterocyte_High_M1.bed	GSE145598	1468856	0.054	n/a*	n/a*	n/a*	n/a*
sBLISS_GSM4322066_Enterocyte_High_M2.bed		1227510	0.045	n/a*	n/a*	n/a*	n/a*
sBLISS 63 peaks extended 133bp upstream & 133bp downstream that overlapped sBLISS 66 extended 133bp upstream/downstream	GSE145598 (1bp breaks not extended)	489945	4.800	267	267	267	267

*sBLISS peaks are only one nucleotide wide.

Supplementary Table 8-6: RegioneR permutation test results for samples not shown in the heatmap of Figure 4-10.
ST= spermatid, RY= alternating purine-pyrimidine, mESC = mouse embryonic stem cell. Shown in red text are non-significant associations. Highlighted in green are results with a significant positive association and shown in pink are results with a significant negative association. The value shown in blue is still significant but P <0.05 as opposed to <=0.01.

	File 1	File 2	Z-score	P-value
Topoisomerase consensus	ST DSB locations	Topoisomerase consensus (RNYNCNGYNGKTNYNY) top 1000 motifs from FIMO	-3.6	0.001
	ST DSB locations	ALL Topoisomerase consensus (RNYNCNGYNGKTNYNY) from FIMO	–24.4	0.001
complex repeats	ST DSB locations	complex repeats (RYRYRYRYRYRYRYRYRY) top 1000 from FIMO (positive strand only)- all were GCGCGCGCGCGCGCGCGC repeats	N/A	NS
	ST DSB locations	RYRYRYRYRYRYRYRYRY (ALL motifs from FIMO, positive strand only)	29.1	0.001
	ST DSB locations	(RY)$_{26}$ motifs (same mean length as mm10 CA repeats) from FIMO (top 1000, positive strand only)	2.2	0.045
	ST DSB locations	(RY)$_{26}$ motifs (same mean length as mm10 CA repeats) from FIMO (positive strand only)	35.8	0.001
	ST DSB locations	(RY)$_{26}$ FIMO motifs (positive strand only) Not overlapping CA repeats	33.1	0.001
	ST DSB locations	GSM3003548_Mouse_all_w15_th-1_minus.hits.max.PDS.w50.35 (G-quadruplex experimental)	-32.5	0.001
Histone marks	ST DSB locations	SRR1596612-BRD4	39.6	0.001
	ST DSB locations	SRR1596613-H3K9me3	47.7	0.001
	ST DSB locations	SRR1596614-H3K9ac	40.8	0.001
	ST DSB locations	SRR1596615-H4K5ac	-33.6	0.001
	ST DSB locations	SRR1596616-H4K8ac	-8	0.001
	ST DSB locations	SRR1596617-H4K12ac	-18.3	0.001
	ST DSB locations	SRR1596618-H4K16ac	N/A	NS
	ST DSB locations	SRR1596619-H4Kac	-4.5	0.001
	ST DSB locations	SRR625513-H3K4me3 rep1	-43.1	0.001
	ST DSB locations	SRR625514-H3K4me3 rep2	-42.5	0.001
	ST DSB locations	SRR625515-H3K27me3 rep1	-38.8	0.001
	ST DSB locations	SRR625516-H3K27me3 rep2	-61.6	0.001
	ST DSB locations	SRR350906- Kac	4.9	0.001
	ST DSB locations	SRR350907-Kcr	-3.1	0.001
	ST DSB locations	SRR948800-H3K4me3	-58.7	0.001
	ST DSB locations	SRR948805-H3K27me3	-67.8	0.001
	ST DSB locations	SRR948811- H3K27ac	-27.4	0.001
	ST DSB locations	SRR948807-H3K3me1	-83.8	0.001
	ST DSB locations	SRR948814-5hMC	-86.7	0.001
	ST DSB locations	SRR948819.20- H2AZ	18	0.001

	File 1	File 2	Z-score	P-value
ChromHMM states	ST DSB locations	E1	-33.6	0.001
	ST DSB locations	E2	-37.4	0.001
	ST DSB locations	E3	43	0.001
	ST DSB locations	E4	279	0.001
	ST DSB locations	E5	-18.4	0.001
	ST DSB locations	E6	-29.5	0.001
	ST DSB locations	E7	-35.6	0.001
	ST DSB locations	E8	-46.5	0.001
	ST DSB locations	E9	-53.6	0.001
	ST DSB locations	E10	-54.4	0.001
	ST DSB locations	E11	-90.7	0.001
	ST DSB locations	E12	363.2	0.001
	ST DSB locations	E13	-44.5	0.001
	ST DSB locations	E14	-69.7	0.001
	ST DSB locations	E15	-58.1	0.001
	ST DSB locations	E16	-103.1	0.001
Miscellaneous	ST DSB locations not overlapping CA repeats	mESC ZCAN4	629.2	0.001
	ST DSB locations overlapping CA repeats	mESC ZSCAN4	389.6	0.001
	predicted Z-DNA not overlapping STR overlapping ST DSB locations	mESC ZSCAN4	356.2	0.001
	GSM4175885_GFP-Zscan_GFP_ChIP	DRR124329-H3K9me3	1131	0.001

Supplementary Table 8-7: RegioneR correlations between Z- DNA and retained histones.
Shown in green are results with a significant positive association and shown in pink are results with a significant negative association. The value shown in blue is still significant but P=0.021 as opposed to 0.001.

File 1	File 2	Z-score	P-value
predicted Z-DNA	DRR124323-H3-C#1	56.3	0.001
predicted Z-DNA	DRR124324-H3-C#2	61.1	0.001
predicted Z-DNA	DRR124328-H3K4me3#1	33.5	0.001
predicted Z-DNA	DRR124333-H3K4me3#2	81.6	0.001
predicted Z-DNA	DRR124330-H3-N#1	**-2.0**	**0.021**
predicted Z-DNA	DRR124335-H3-N#2	-34.4	0.001
predicted Z-DNA	DRR124329-H3K9me3#1	42.5	0.001
predicted Z-DNA	DRR124334-H3K9me3#2	87.8	0.001

Supplementary Table 8-8: RegioneR correlations with top 1% of OD data.
moderate = one mutation (*Gpx5*), severe = two mutations (*Gpx5* and *SnGPx4*).
STR=short tandem repeat, EXP=experimental.
The value shown in blue is still significant, but the P-value is 0.009 as opposed to 0.001.

File 1	File 2	Z-score	P-value
moderate OD top 1%	Spermatid DSB locations	27.1	0.001
moderate OD top 1%	mESC ZSCAN4	48.1	0.001
moderate OD top 1%	predicted Z-DNA non-B DB no overlap STR	16.2	0.001
moderate OD top 1%	STR non-B DB no overlap predicted Z-DNA	23.3	0.001
moderate OD top 1%	predicted G-quadruplex non-B DB	18.2	0.001
severe OD top 1%	Spermatid DSB locations	27.4	0.001
severe OD top 1%	mESC ZSCAN4	33.9	0.001
severe OD top 1%	predicted Z-DNA non-B DB no overlap STR	**3.2**	**0.009**
severe OD top 1%	STR non-B DB no overlap predicted Z-DNA	14.7	0.001
severe OD top 1%	predicted G-quadruplex non-B DB	11.7	0.001
moderate OD	predicted Z-DNA Not overlapping STR- No overlap spermatid DSB locations	10.7	0.001
moderate OD	STR Not overlapping predicted Z-DNA-No overlap spermatid DSB locations	16.9	0.001
moderate OD	predicted G-quadruplex No overlap spermatid DSB locations	17.6	0.001
moderate OD	G-quadruplex_EXP No overlap spermatid DSB locations	11	0.001
severe OD	predicted Z-DNA Not overlapping STR- No overlap spermatid DSB locations	**1.8**	**0.047**
severe OD	STR Not overlapping predicted Z-DNA-No overlap spermatid DSB locations	10	0.001
severe OD	predicted G-quadruplex No overlap spermatid DSB locations	11.3	0.001
severe OD	G-quadruplex_EXP No overlap spermatid DSB locations	4.3	0.001

Supplementary Table 8-9: Genes within the ROI on **chromosome 6** that are differentially expressed (P<=0.05) between WT-Het and Het-Hom (independent of the direction). The row shown in pink represents a gene that is downregulated in a WT-Het and a Het-Hom comparison. The GIM status refers to confident GIM from the Bhutani dataset (185) Gene names shown in bold are genes with missense mutations.

Chr 6 Gene start (bp)	Chr 6 Gene end (bp)	Gene stable ID	Gene name	Full gene name	Gene variants	GIM	Direction of change
3372257	3399572	ENSMUSG00000047735	***Samd9l***	Sterile alpha motif domain-containing protein 9-like	Upstream /downstream/ synonymous gene variants & **Missense**	No	WT<Het<Hom
4674350	4747207	ENSMUSG00000004631	*Sgce*	Epsilon-sarcoglycan	Upstream gene variant	No	WT<Het<Hom
4807055	4816329	ENSMUSG00000093570	*Gm20714*	predicted gene	upstream/downstream/splice region variant	Yes	WT<Het<Hom
5168090	5193946	ENSMUSG00000002588	*Pon1*	Serum paraoxonase/arylesterase 1	downstream	No	WT>Het>Hom
5264147	5298455	ENSMUSG00000032667	*Pon2*	Serum paraoxonase/arylesterase 2	Upstream/downstream & synonymous	No	WT<Het<Hom
5383386	5433022	ENSMUSG00000042607	*Asb4*	Ankyrin repeat and SOCS box protein 4	Upstream/downstream & synonymous	yes	WT<Het>Hom
5963909	5977393	ENSMUSG00000085416	*1700019G24Rik*	RIKEN cDNA 1700019G24 gene	upstream/downstream	No	WT<Het<Hom
8259450	8597480	ENSMUSG00000107705	*Gm45062*	predicted gene	upstream/downstream	No	WT<Het<Hom

Chr 6 Gene start (bp)	Chr 6 Gene end (bp)	Gene stable ID	Gene name	Full gene name	Gene variants	GIM	Direction of change
8509600	8597548	ENSMUSG00000029638	*Glcci1*	Glucocorticoid-induced transcript 1 protein.	Upstream/downstream/splice region variant	yes	WT<Het<Hom
10804828	1.1E+07	*ENSMUSG00000108249*	*Gm43960*	predicted gene	upstream/downstream	No	**WT>Het<Hom**
14713976	1.5E+07	ENSMUSG00000042717	***Ppp1r3a***	Protein phosphatase 1 regulatory subunit 3A	Upstream/downstream synonymous/ **Missense**	yes	WT<Het<Hom
14901348	1.5E+07	ENSMUSG00000029563	*Foxp2*	Forkhead box protein P2	Upstream/downstream & synonymous	No	WT<Het<Hom
17197750	1.7E+07	ENSMUSG00000085171	*D830026I12Rik*	RIKEN cDNA D830026I12 gene	upstream/downstream	No	WT>Het>Hom
17206822	1.7E+07	ENSMUSG00000085264	*Gm15581*	predicted gene	upstream/downstream	yes	WT>Het>Hom
17692932	1.8E+07	ENSMUSG00000029534	*St7*	Suppressor of tumorigenicity 7 protein.	upstream/downstream/splice region variant	yes	WT<Het<Hom
17834584	1.8E+07	ENSMUSG00000097700	*Gm26738*	predicted gene	upstream	No	WT<Het<Hom
18842933	1.9E+07	ENSMUSG00000106874	*Gm20186*	predicted gene	upstream/downstream	No	WT<Het<Hom

Chr 6 Gene start (bp)	**Chr 6 Gene end (bp)**	**Gene stable ID**	**Gene name**	**Full gene name**	**Gene variants**	**GIM**	**Direction of change**
24528143	2.5E+07	ENSMUSG00000029685	***Asb15***	Ankyrin repeat and SOCS box protein 15	Upstream/downstream/synonymous/splice region variant & **Missense**	No	WT<Het>Hom

Supplementary Table 8-10: Genes within the ROI on **chromosome 16** that are differentially expressed (P<=0.05) between WT-Het and Het-Hom (independent of the direction). The rows shown in pink represent genes that are downregulated in a WT-Het and a Het-Hom comparison. The GIM status refers to confident GIM from the Bhutani dataset (185). Gene names shown in bold are genes with missense or stop-gained mutations.

chr16 Gene start (bp)	Chr16 Gene end (bp)	Gene stable ID	Gene name	Full gene name	Gene Variants	GIM	Direction of change
13799563	13804752	ENSMUSG00000065968	*Ifitm7*	Interferon-induced transmembrane protein 7.	upstream/ downstream	No	WT>Het>Hom
15369941	15413637	ENSMUSG00000022674	*Ube2v2*	Ubiquitin-conjugating enzyme E2 variant 2	upstream/ downstream/ splice region variant	yes	**WT>Het<Hom**
15455730	15660099	ENSMUSG00000022672	***Prkdc***	DNA-dependent protein kinase	upstream/ downstream/ splice region variant/synonymous/ **missense**	yes	WT<Het<Hom
16031182	16090576	ENSMUSG00000041957	*Pkp2*	Plakophilin 2.	synonymous	No	WT>Het>Hom
16167067	16176875	ENSMUSG00000115923	*Gm49521*	predicted gene	upstream/ downstream	No	**WT>Het<Hom**
16234781	16418413	ENSMUSG00000022788	***Fgd4***	FYVE, RhoGEF and PH domain-containing protein 4;	upstream/ downstream /synonymous/ **missense**	yes	WT<Het<Hom
16490092	16494269	ENSMUSG00000048101	*Or7a40*	Olfactory receptor	not found in SnpEff data	No	WT<Het<Hom
17437218	17462692	ENSMUSG00000005899	*Smpd4*	Sphingomyelin phosphodiesterase 4	upstream/ downstream/ synonymous	No	WT<Het>Hom
17650985	17652872	ENSMUSG00000116652	*B830017H08 Rik*	family with sequence similarity 246 member A	upstream/ downstream/ synonymous	No	WT>Het>Hom
17870724	17889496	ENSMUSG00000003531	***Dgcr6***	DiGeorge syndrome critical region gene 6	upstream/ downstream/ synonymous/ **stop-gained**	No	WT<Het<Hom
19916292	19947971	ENSMUSG00000062901	*Klhl24*	Kelch-like protein 24	upstream/ downstream/ synonymous	No	WT<Het<Hom
4412577	4442788	ENSMUSG00000014303	*Glis2*	Zinc finger protein GLIS2.	upstream/ downstream	No	WT<Het>Hom

chr16 Gene start (bp)	Chr16 Gene end (bp)	Gene stable ID	Gene name	Full gene name	Gene Variants	GIM	Direction of change
21023505	21042055	ENSMUSG00000005958	*Ephb3*	Ephrin type-B receptor 3	upstream/ downstream/ synonymous	No	**WT>Het<Hom**
4653280	4685550	ENSMUSG00000022518	*4930562C15 Rik*	RIKEN cDNA 4930562C15 gene	upstream/ downstream /synonymous	No	WT<Het>Hom
4804722	4815716	ENSMUSG00000022542	*Septin12*	Septin 12; Belongs to the TRAFAC class TrmE-Era-EngA-EngB-Septin-like GTPase superfamily	upstream/ downstream/ synonymous	No	WT<Het>Hom
4826594	4831417	ENSMUSG00000022540	*Rogdi*	Protein rogdi homolog; Belongs to the rogdi family.	upstream/ downstream	yes	WT<Het>Hom
21772567	21814413	ENSMUSG00000044626	***Liph***	Lipase member H	upstream/ downstream/ splice region /synonymous/ **missense**	No	WT>Het>Hom
10175812	10213354	ENSMUSG00000039179	***Tekt5***	Tektin-5; May be a structural component of the sperm flagellum.	upstream/ downstream/ synonymous /**missense**	No	WT<Het>Hom
10229812	10242292	ENSMUSG00000022503	***Nubp1***	Cytosolic Fe-S cluster assembly factor NUBP1	downstream/ upstream /splice region variant/ synonymous/ **missense**	No	WT>Het>Hom
10238421	10265226	ENSMUSG00000050908	*Tvp23a*	Golgi apparatus membrane protein TVP23 homolog.	upstream/downstream	No	WT>Het>Hom
24348144	24349112	ENSMUSG00000116875	*Morf4l1-ps1*	mortality factor 4 like 1, pseudogene 1	upstream/downstream	No	WT<Het>Hom
26400454	26548867	ENSMUSG00000022514	***Il1rap***	Interleukin-1 receptor accessory protein	upstream/downstream/ synonymous/ **missense**	No	WT<Het<Hom

chr16 Gene start (bp)	Chr16 Gene end (bp)	Gene stable ID	Gene name	Full gene name	Gene Variants	GIM	Direction of change
12927548	12968481	ENSMUSG00000022545	*Ercc4*	DNA repair endonuclease XPF	upstream/downstream	No	WT<Het>Hom
13172238	13286619	ENSMUSG00000087526	*Gm15738*	predicted gene	upstream/downstream	No	**WT>Het<Hom**
13636709	13653315	ENSMUSG00000022681	*Ntan1*	Protein N-terminal asparagine amidohydrolase	upstream/downstream	No	WT<Het<Hom
4782090	4796826	ENSMUSG00000022543	*Dnaaf8*	dynein axonemal assembly factor 8	upstream/downstream	No	WT>Het>Hom
26492212	26574275	ENSMUSG00000092545	*Gm20319*	predicted gene	upstream/downstream	No	WT<Het<Hom
23941449	24201140	ENSMUSG00000115852	*Gm52969*	predicted gene	upstream/downstream/ splice region variant	No	WT<Het<Hom
10605800	10606524	ENSMUSG00000043050	*Tnp2*	Nuclear transition protein 2	upstream/downstream/ synonymous	No	**WT>Het<Hom**
10608369	10608778	ENSMUSG00000050058	*Prm3*	Protamine-3	upstream/downstream/ synonymous	No	**WT>Het<Hom**
10609244	10613998	ENSMUSG00000038015	***Prm2***	Protamine-2	upstream/downstream/ synonymous/ **missense**	yes	**WT>Het<Hom**
16956928	16965093	ENSMUSG00000022768	***Ccdc116***	Coiled-coil domain-containing protein 116.	upstream/downstream/ **missense**	No	WT<Het>Hom
16962485	16978565	ENSMUSG00000041774	*Ydjc*	CaRobohydrate deacetylase	upstream/downstream/ synonymous	yes	WT<Het>Hom
17026467	17031846	ENSMUSG00000071636	***Rimbp3***	RIMS-binding protein 3	upstream/downstream/ synonymous/ **missense**	No	WT<Het>Hom
17098215	17224178	ENSMUSG00000041720	*Pi4ka*	Phosphatidylinositol 4-kinase alpha	upstream/downstream/ synonymous	No	WT<Het>Hom

Supplementary Table 8-11: Genes containing SNPs with high or moderate effect on chr6 and chr 16 that are genoinformative markers (GIM) according to Bhutani *et al* 2021 (185). There were also other genes with high and moderate effect SNPs that did not overlap the confident GIM genes from the Bhutani *et al* 2021 paper (185) (Supplementary Table 8-12).

Gene name / chr	Function	Number of SNPs	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)*	Comments
Col28a1 (chr6)	Collagen, type XXVIII, alpha 1. Predicted to be an extracellular matrix structural constituent, involved in extracellular matrix organization and to act upstream of or within cell adhesion and negative regulation of peptidase activity. Located in basement membrane.	2 missense	No	0.45	
Cped1 (chr6)	cadherin-like and PC-esterase domain containing 1.	4 missense	No	0.47	
Ica1 (chr6)	islet cell autoantigen 1. Predicted to enable membrane curvature sensor activity and protein domain specific binding activity. Predicted to be involved in regulation of protein-containing complex assembly and regulation of transport. Predicted to act upstream of or within neurotransmitter transport. Located in cytosol and synaptic vesicle membrane.	1 missense	No	6.70	ICA69 KO does not impact fertility. He *et al*, (2015) (367)
Ppp1r3a (chr6)	protein phosphatase 1, regulatory subunit 3A. Predicted to enable glycogen binding activity; protein phosphatase 1 binding activity; and protein serine/threonine phosphatase activity. Predicted to be involved in regulation of glycogen biosynthetic process. Predicted to act upstream of or within glycogen metabolic process. Predicted to be located in membrane.	5 missense	No	0.032	
Rnf148 (chr6)	ring finger protein 148. Predicted to enable ubiquitin protein ligase activity. Predicted to be involved in ubiquitin-dependent protein catabolic process. Predicted to be integral component of membrane. Predicted to be active in Golgi apparatus; endoplasmic reticulum; and late endosome.	1 missense	Yes	86.60*	Fecundity remained largely unaffected in *Rnf148* knockout Nozawa et al 2022 (368)
Umad1 (chr6)	UBAP1-MVB12-associated (UMA) domain containing 1	1 missense	No	2.50	

Gene name / chr	Function	Number of SNPs	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)*	Comments
Wasl (chr6)	Actin nucleation-promoting factor WASL Regulates actin polymerization by stimulating the actin-nucleating activity of the Arp2/3 complex. Involved in various processes, such as mitosis and cytokinesis, via its role in the regulation of actin polymerization. Together with CDC42, involved in the extension and maintenance of the formation of thin, actin-rich surface projections (filopodia). In addition to its role in the cytoplasm, also plays a role in the nucleus by regulating gene transcription, probably by promoting nuclear actin polymerization.	2 missense	No	2.50	
Adcy9 (chr16)	adenylate cyclase 9-Enables adenylate cyclase activity. Acts upstream of or within adenylate cyclase-activating G protein-coupled receptor signalling pathway; cAMP biosynthetic process; and in utero embryonic development. Located in membrane.	2 missense	No	1.461	
Ccdc116 (chr16)	coiled-coil domain containing 116. Predicted to be located in cytoplasm and cytoskeleton. Predicted to be active in centrosome.	2 missense mutations	No	40.75	
Fgd4 (chr16)	FYVE, RhoGEF and PH domain containing 4. Member of the FYVE, RhoGEF and PH domain containing (FGD) family. The encoded protein is a Cdc42-specific guanine nucleotide exchange factor (GEF) that plays an essential role in regulating the actin cytoskeleton and cell morphology. Alternative splicing results in multiple transcript variants.	4 missense mutations	No	0.17	
Hic2 (chr16)	hypermethylated in cancer 2. Predicted to enable DNA-binding transcription factor activity, RNA polymerase II-specific; RNA polymerase II cis-regulatory region sequence-specific DNA binding activity; and protein C-terminus binding activity. Predicted to be involved in regulation of transcription by RNA polymerase II. Predicted to be located in nucleoplasm and active in nucleus.	2 missense mutations	No	1.78	
Il1rap (chr16)	Interleukin 1 receptor accessory protein. Enables interleukin-1 receptor activity; interleukin-33 receptor activity; and protein tyrosine kinase binding activity. Involved in several processes, including interleukin-33-	1 missense mutation	No	0.23	

Gene name / chr	Function	Number of SNPs	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)*	Comments
	mediated signalling pathway; positive regulation of cytokine production; and regulation of synapse assembly. Acts upstream of or within cytokine-mediated signalling pathway; positive regulation of dendrite development; and positive regulation of synapse assembly. Located in glutamatergic synapse.				
Kng1 (chr 16)	kininogen 1. Predicted to enable cysteine-type endopeptidase inhibitor activity; protease binding activity; and signalling receptor binding activity. Predicted to be involved in several processes, including antimicrobial humoral immune response mediated by antimicrobial peptide; negative regulation of blood coagulation; and negative regulation of endopeptidase activity. Located in collagen-containing extracellular matrix.	1 missense mutation	No	1.81*	
Olfr19 (chr16)	olfactory receptor family 7 subfamily A member 40. M12; Olfr19; MTPCR15; MOR140-1. Olfactory receptors interact with odorant molecules in the nose, to initiate a neuronal response that triggers the perception of a smell.	1 missense mutation	No	2.03*	
Parn (chr 16)	poly(A)-specific ribonuclease (deadenylation nuclease). Predicted to enable several functions, including cation binding activity; poly(A)-specific ribonuclease activity; and protein kinase binding activity. Acts upstream of or within nuclear-transcribed mRNA poly(A) tail shortening. Predicted to be located in nuclear speck and nucleolus. Predicted to be active in cytoplasm.	2 missense mutations	No	7.20	
Pla2g10 (chr 16)	phospholipase A2, group X. This gene encodes a member of the phospholipase A2 family of lipolytic enzymes that hydrolyses glycerophospholipids to produce free fatty acids and lysophospholipids. The encoded protein undergoes proteolytic processing to generate a calcium-dependent enzyme that plays pivotal roles in the liberation of arachidonic acid from membrane phospholipids leading to the production of various inflammatory lipid	2 missense mutations	No	58.50	*Pla2g10* is involved in the acrosome reaction and has a role in fertility. Nahed *et al* 2022 (357).

Gene name / chr	Function	Number of SNPs	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)*	Comments
	mediators, such as prostaglandins. Alternative splicing results in multiple transcript variants encoding different isoforms.				
Ppl (chr16)	periplakin. Predicted to enable structural molecule activity. Predicted to be involved in intermediate filament cytoskeleton organization. Predicted to act upstream of or within keratinization. Located in cytoplasm and plasma membrane.	4 missense mutations	-	0.21	
Ppp1r3a (chr6)	protein phosphatase 1, regulatory subunit 3A, predicted to enable glycogen binding activity; protein phosphatase 1 binding activity; and protein serine/threonine phosphatase activity. Predicted to be involved in regulation of glycogen biosynthetic process. Predicted to act upstream of or within glycogen metabolic process. Predicted to be located in membrane.	4 missense	No	0.03	
Prkdc (chr 16)	protein kinase, DNA activated, catalytic polypeptide. Enables DNA-dependent protein kinase activity; double-stranded DNA binding activity; and enzyme binding activity. Acts upstream of or within several processes, including DNA metabolic process; ectopic germ cell programmed cell death; and immune system development. Located in nucleus.	2 missense mutations	No	0.55	
Prm2 (chr16)	Protamine 2-Protamines substitute for histones in the chromatin of sperm during the haploid phase of spermatogenesis and are the major DNA-binding proteins in the nucleus of sperm in many vertebrates. They package the sperm DNA into a highly condensed complex in a volume less than 5% of a somatic cell nucleus. Protamine 2 is synthesized as a precursor and then cleaved to give rise to a family of protamine 2 peptides.	1 missense mutation	Yes	5910.46	

Gene name / chr	Function	Number of SNPs	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)*	Comments
Spidr (chr16)	scaffolding protein involved in DNA repair. Predicted to be involved in cellular response to camptothecin; cellular response to hydroxyurea; and regulation of double-strand break repair. Predicted to act upstream of or within DNA recombination and DNA repair. Predicted to be located in nucleus. Predicted to be active in nuclear chromosome and nucleoplasm.	2 missense mutations	No	1.22	
Tmem191c (chr16)	Transmembrane protein 191- Predicted to be located in membrane. Predicted to be integral component of membrane.	1 missense mutation	No	91.9	
Tmem41a (chr16)	Transmembrane protein 41a-Predicted to be located in membrane. Predicted to be integral component of membrane.	1 missense mutation	No	3.79	
Txndc11 (chr16)	thioredoxin domain containing 11-Predicted to be located in endoplasmic reticulum and membrane	1 missense mutation	No	12.08	
Ubn1 (chr16)	ubinuclein 1. Predicted to enable DNA binding activity. Predicted to be involved in DNA replication-independent chromatin assembly. Located in bicellular tight junction.	1 missense mutation	No	7.11	
Vasn (chr16)	Vasorin. Predicted to enable transforming growth factor beta binding activity. Acts upstream of or within cellular response to hypoxia and cellular response to redox state. Located in mitochondrion and plasma membrane.	3 missense mutations	No	6.09	
Ypel1 (chr16)	yippee like 1. Predicted to enable metal ion binding activity. Predicted to be located in nucleus.	1 missense mutation	No	54.80	
Zc3h7a (chr16)	zinc finger CCCH type containing 7 A. Predicted to enable miRNA binding activity. Predicted to be involved in production of miRNAs involved in gene silencing by miRNA.	1 missense mutation	No	5.20	
Zdhhc8 (chr 16)	zinc finger, DHHC domain containing 8. Predicted to enable palmitoyl transferase activity. Acts upstream of or within locomotory behaviour. Located in mitochondrion.	1 missense mutation	-	5.20	

*RPKM values were obtained from Evo Devo mammalian organs Kaessmann lab app Cardoso-Moreira *et al* 2019 (346).

Supplementary Table 8-12: Genes containing SNPs with high or moderate effect on chr6 and chr 16 that are NOT confident genoinformative markers (GIM) according to the Bhutani et al 2021 (185).
RPKM values were obtained from Evo Devo mammalian organs Kaessmann lab app Cardoso-Moreira *et al* 2019 (346).
Gene expression across mammalian organ development.

Gene name / chr (not confident GIM)	**Function**	**SNP Number**	**Testis specific expression** (Evo-Devo app)	**Testis RPKM (NCBI)**	**Comments**
Asb15 (chr6)	ankyrin repeat and SOCS box-containing 15. Predicted to be involved in intracellular signal transduction.	1 missense	No	4.40	Testis-enriched Asb15 is not required for spermatogenesis and male fertility in mice Wu *et al* 2022 (369)
Dync1i1 (chr6)	Enables dynein light chain binding activity. Predicted to be involved in vesicle transport along microtubule. Predicted to be located in several cellular components, including axon cytoplasm; kinetochore; and spindle pole.	1 missense	No	2.68	
Hyal6 (chr6)	hyaluronoglucosaminidase 6. Predicted to enable hyaluronongluosaminidase activity. Predicted to be involved in hyaluronan catabolic process. Predicted to be active in cytoplasmic vesicle.	2 missense	Yes	6.10	Knockout, does not affect sperm parameters, Bang *et al* 2022 (349)
Lsm8 (chr6)	LSM8 homolog, U6 small nuclear RNA associated. Predicted to enable RNA binding activity. Predicted to be involved in mRNA splicing, via spliceosome. Predicted to act upstream of or within RNA splicing and mRNA processing. Predicted to be located in nucleus. Predicted to be part of Lsm2-8 complex; U2-type precatalytic spliceosome; and spliceosomal snRNP complex.	2 missense	No	7.86	
Pot1a (chr6)	protection of telomeres 1A. Enables single-stranded telomeric DNA binding activity. Acts upstream of or within chromosome organization. Located in chromosome, telomeric region.	2 missense	No	0.79	

Gene name / chr (not confident GIM)	Function	SNP Number	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)	Comments
Samd9l (chr6)	sterile alpha motif domain containing 9-like, acts upstream of or within several processes, including common myeloid progenitor cell proliferation; endosomal vesicle fusion; and hematopoietic or lymphoid organ development. Located in early endosome.	2 missense	No	0.21	
Slc13a1 (chr6)	solute carrier family 13 (sodium/sulphate symporters), member 1. Enables secondary active sulphate transmembrane transporter activity. Acts upstream of or within sulphate transport. Predicted to be integral component of plasma membrane.	2 missense	No	0.01*	
Slc25a13 (chr6)	solute carrier family 25 (mitochondrial carrier, adenine nucleotide translocator), member 13. Enables L-glutamate transmembrane transporter activity. Acts upstream of or within aspartate transmembrane transport and malate-aspartate shuttle. Located in mitochondrion.	1 missense	No	0.51	
Spam1 (chr6)	sperm adhesion molecule 1. Enables hyaluronoglucosaminidase activity. Acts upstream of or within single fertilization. Located in acrosomal vesicle; external side of plasma membrane; and membrane raft.	1 missense	No	9.70	Zheng *et al* 2001 (356), indicate that the Spam1 protein expression does appear to be compartmentalised, but Bhutani *et al* 2021 (185) do not identify the Spam1 gene as a GIM
Vwde (chr6)	von Willebrand factor D and EGF domains. Predicted to be located in extracellular region.	3 missense	No	0.00	

Gene name / chr (not confident GIM)	Function	SNP Number	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)	Comments
Wnt2 (chr6)	wingless-type MMTV integration site family, member 2. Predicted to be located in collagen-containing extracellular matrix and cytoplasm. Predicted to be extrinsic component of external side of plasma membrane.	1 missense	No	0.46	
A930003A15Rik (chr16)	RIKEN cDNA A930003A15 gene-Is expressed in cerebellum; retina inner nuclear layer; and retina outer nuclear layer	1 missense	No	0.04	
Alg3 (chr16)	asparagine-linked glycosylation 3 (alpha-1,3-mannosyltransferase)-Predicted to enable dol-P-Man:Man(5)GlcNAc(2)-PP-Dol alpha-1,3-mannosyltransferase activity. Predicted to be involved in protein glycosylation. Predicted to be located in and active in endoplasmic reticulum membrane.	1 missense	No	1.58	
Anks3 (chr16)	ankyrin repeat and sterile alpha motif domain containing 3-Located in cilium and cytoplasm	3 missense	No	4.84	
Atf7ip2 (chr16)	activating transcription factor 7 interacting protein 2-Predicted to enable transcription coregulator activity. Predicted to be involved in positive regulation of DNA methylation-dependent heterochromatin assembly and regulation of transcription, DNA-templated. Predicted to be part of transcription regulator complex. Predicted to be active in nucleus.	1 missense	No	1.20	
B3gnt5 (chr16)	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 5-Predicted to enable beta-galactosyl-N-acetylglucosaminylgalactosylglucosyl-ceramide beta-1,3-acetylglucosaminyltransferase activity. Predicted to be located in Golgi apparatus and membrane. Predicted to be integral component of membrane. Predicted to be active in Golgi membrane.	2 missense	No	0.03*	

Gene name / chr (not confident GIM)	Function	SNP Number	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)	Comments
Chrd (chr16)	chordin-Enables heparin binding activity and syndecan binding activity. Acts upstream of or within several processes, including gastrulation; negative regulation of BMP signalling pathway; and negative regulation of osteoblast differentiation. Located in extracellular space.	1 missense	No	0.63	
Cldn16 (chr16)	claudin 16-This gene encodes a member of the claudin family. Claudins are integral membrane proteins and components of tight junction strands.	2 missense	No	0*	
Crygs (chr16)	crystallin, gamma S-A structural constituent of eye lens. Acts upstream of or within lens development in camera-type eye and morphogenesis of an epithelium	3 missense	No	0.18	
Dgcr6 (chr16)	DiGeorge syndrome critical region gene 6-This gene encodes a protein that is similar to the gonadal protein in Drosophila (fruit fly).	1 stop gained variant	No	66.47	
Dnm1l (chr16)	dynamin 1-like-This gene encodes a member of the dynamin family. The encoded protein is localized to the cytoplasm and mitochondrial membrane, is involved in mitochondrial and peroxisomal division, and is essential for mitochondrial fission. Alternative splicing results in multiple transcript variants. A related pseudogene has been identified on chromosome 2.	1 missense	No	9.76	
Ece2 (chr16)	endothelin converting enzyme 2-Predicted to enable metalloendopeptidase activity. Predicted to be involved in protein processing. Predicted to act upstream of or within peptide hormone processing. Predicted to be located in cytoplasmic vesicle membrane and trans-Golgi network. Predicted to be active in plasma membrane.	1 missense	No	0.47	

Gene name / chr (not confident GIM)	Function	SNP Number	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)	Comments
Eef1akmt4 (chr16)	enoyl-Coenzyme A, hydratase/3-hydroxyacyl Coenzyme A dehydrogenase-Predicted to enable several functions, including dodecenoyl-CoA delta-isomerase activity; enoyl-CoA hydratase activity; and long-chain-3-hydroxyacyl-CoA dehydrogenase activity. Acts upstream of or within fatty acid beta-oxidation. Located in mitochondrion.	1 missense	Not in database	-	
Etv5 (chr16)	Ets variant 5-Predicted to enable DNA-binding transcription activator activity, RNA polymerase II-specific and RNA polymerase II transcription regulatory region sequence-specific DNA binding activity. Predicted to be located in nucleoplasm. Predicted to be active in nucleus.	1 missense	No	3.71	
Gm49333 (chr16)	Eef1akmt4-endothelin converting enzyme 2 readthrough-This locus represents naturally occurring read through transcription between the adjacent genes eukaryotic translation elongation factor 1 alpha lysine specific methyltransferase 4 and endothelin converting enzyme 2.	2 missense	Not in database	-	
Igll1 (chr16)	immunoglobulin lambda-like polypeptide 1-Predicted to enable antigen binding activity and immunoglobulin receptor binding activity. Predicted to be located in endoplasmic reticulum and extracellular region. Predicted to be part of immunoglobulin complex, circulating. Predicted to be active in external side of plasma membrane.	1 missense	No	0.14*	
Kng2 (chr16)	kininogen 2-Predicted to enable cysteine-type endopeptidase inhibitor activity; protease binding activity; and signalling receptor binding activity. Predicted to be involved in several processes, including antimicrobial humoral immune response, negative regulation of blood coagulation; and negative regulation of endopeptidase activity. Located in collagen-containing extracellular matrix	3 missense	No	0.41*	

Gene name / chr (not confident GIM)	Function	SNP Number	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)	Comments
Liph (chr16)	lipase, member H-Predicted to enable heparin binding activity; lipoprotein lipase activity; and phospholipase activity. Predicted to be involved in fatty acid biosynthetic process and triglyceride catabolic process. Predicted to act upstream of or within lipid metabolic process. Predicted to be located in extracellular region and plasma membrane. Predicted to be active in extracellular space.	1 missense	No	0.47	
Mcm4 (chr16)	Mini-chromosome maintenance complex component 4-Enables single-stranded DNA binding activity. Contributes to DNA helicase activity. Acts upstream of or within DNA unwinding involved in DNA replication. Predicted to be located in nucleoplasm. Predicted to be part of MCM complex. Predicted to be active in nucleus.	1 missense	No	1.43	
Mettl22 (chr16)	methyltransferase 22, Kin17 lysine-Predicted to enable heat shock protein binding activity and protein methyltransferase activity. Predicted to be involved in protein methylation. Predicted to act upstream of or within methylation. Predicted to be located in nucleolus and nucleoplasm. Predicted to be part of protein-containing complex. Predicted to be active in nucleus.	1 missense	No	31.32	
Mgrn1 (chr16)	mahogunin, ring finger 1-Enables ubiquitin protein ligase activity. Involved in negative regulation of smoothened signalling pathway. Acts upstream of or within protein polyubiquitination. Predicted to be located in endoplasmic reticulum. Predicted to be active in early endosome; nucleus; and plasma membrane.	1 missense	No	8.66	

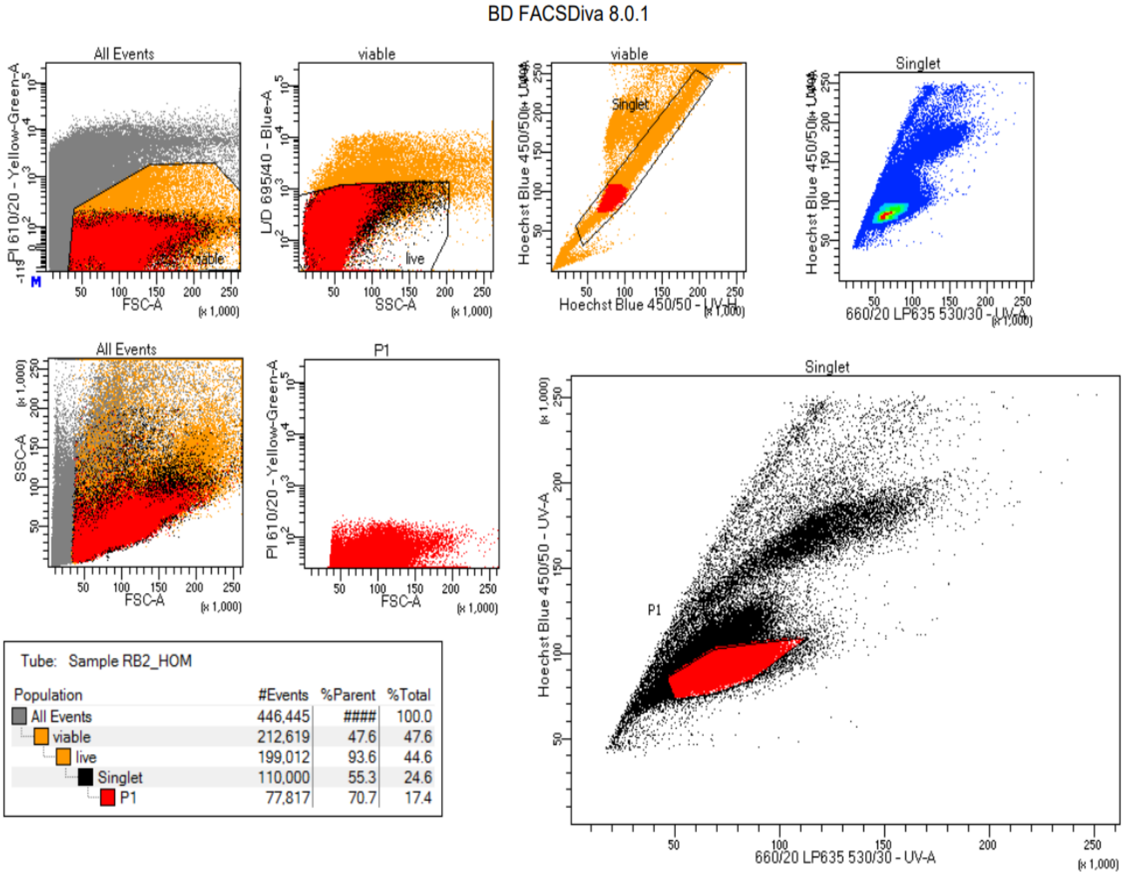
Gene name / chr (not confident GIM)	Function	SNP Number	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)	Comments
Nubp1 (chr16)	nucleotide binding protein 1-Predicted to enable iron-sulphur cluster binding activity. Acts upstream of or within centrosome localization; negative regulation of centrosome duplication; and protein localization to cell cortex. Predicted to be located in plasma membrane. Predicted to be active in cytosol.	2 missense	No	7.07	
Olfr164 (chr16)	olfactory receptor family 2 subfamily M member 12-Olfactory receptors interact with odorant molecules in the nose, to initiate a neuronal response that triggers the perception of a smell.	1 missense	Yes (e16.5)	0*	
P3h2 (chr6)	prolyl 3-hydroxylase 2-Predicted to enable procollagen-proline 3-dioxygenase activity. Predicted to be involved in collagen metabolic process; negative regulation of cell population proliferation; and peptidyl-proline hydroxylation. Located in basement membrane.	3 missense	No	0.26	
Prodh (chr16)	proline dehydrogenase-Predicted to enable FAD binding activity; amino acid binding activity; and proline dehydrogenase activity. Predicted to be involved in positive regulation of cell death and proline catabolic process to glutamate. Predicted to act upstream of or within proline metabolic process. Located in mitochondrion.	1 missense	No	1.19	
Rimbp3 (chr16)	RIMS binding protein 3. Predicted to enable benzodiazepine receptor binding activity. Involved in fertilization and spermatid development. Located in nucleus. Colocalizes with manchette.	2 missense	No	150.12*	Zhou *et al* 2009 (370) Targeted deletion of the RIM-BP3 gene resulted in male infertility owing to abnormal sperm heads, which are characterized by a

Gene name / chr (not confident GIM)	Function	SNP Number	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)	Comments
					deformed nucleus and a detached acrosome. Consistent with its role in morphogenesis
Rtp2 (chr16)	Receptor transporter protein 2-Predicted to enable olfactory receptor binding activity. Predicted to be involved in detection of chemical stimulus involved in sensory perception of bitter taste; protein insertion into membrane; and protein targeting to membrane. Predicted to: be located in plasma membrane, an integral component of membrane and active in cell surface.	1 missense	No	0.08	
Rtp4 (chr16)	Receptor transporter protein 4-Predicted to enable olfactory receptor binding activity. Predicted to be involved in defence response to virus; detection of chemical stimulus involved in sensory perception of bitter taste; and establishment of protein localization to membrane. Predicted to: be located in membrane, an integral component of membrane and active in cytoplasm.	1 missense	No	0.09	
Sec14l5 (chr16)	SEC14-like lipid binding 5-Is expressed in brain.	1 missense	Not in database	0.04	
Shisa9 (chr16)	shisa family member 9- Located in glutamatergic synapse and Is an integral component of postsynaptic density membrane.	1 missense	No	0.04	
Slx4 (chr16)	SLX4 structure-specific endonuclease subunit homolog-This gene encodes a protein containing a BTB (POZ) domain that comprises a subunit of structure-specific endonucleases. The encoded protein aids in the resolution of DNA secondary structures that arise during the processes of DNA repair and recombination. Knock out of this gene in mouse recapitulates the phenotype of the human disease Fanconi anaemia,	7 missense	No	20.84	KO of Slx4 can cause abnormal spermatogenesis Crossan *et al* 2011 ((340)

Gene name / chr (not confident GIM)	Function	SNP Number	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)	Comments
	including blood cytopenia and susceptibility to genomic instability.				
Snap29 (chr16)	synaptosomal-associated protein 29-Predicted to enable SNAP receptor activity and syntaxin binding activity. Predicted to be involved in several processes, including autophagosome membrane docking; regulation of synaptic vesicle cycle; and synaptic vesicle exocytosis. Predicted to act upstream of or within autophagy; cell projection organization; and protein transport. Located in autophagosome.	1 missense	No	4.01	
Srl (chr16)	sarcalumenin-Predicted to enable GTP binding activity. Acts upstream of or within response to muscle activity. Predicted to be located in membrane and sarcoplasmic reticulum. Predicted to be active in cytoplasm; intracellular membrane-bounded organelle; and plasma membrane.	2 missense	No	0.71	
Tbccd1 (chr16)	TBCC domain containing 1. Predicted to be involved in several processes, including maintenance of Golgi location; maintenance of centrosome location; and regulation of cell shape. Predicted to be located in cytoplasm and cytoskeleton and active in spindle pole centrosome.	3 missense	No	4.42	
Tbx1 (chr16)	T-box 1-Enables DNA-binding transcription activator activity, RNA polymerase II-specific and RNA polymerase II intronic transcription regulatory region sequence-specific DNA binding activity. Involved in several processes, including animal organ morphogenesis; positive regulation of tongue muscle cell differentiation; and regulation of transcription by RNA polymerase II. Acts upstream of or within several processes, including animal organ development; regulation of	2 missense	No	1.57	

Gene name / chr (not confident GIM)	Function	SNP Number	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)	Comments
	transcription, DNA-templated; and vasculature development. Located in nucleus.				
Tekt5 (chr16)	Predicted to be involved in cilium assembly and cilium movement involved in cell motility. Located in sperm flagellum.	1 missense mutation	No	92.5	Implicated in sperm mobility. Cao *et al* 2011 (350). This is not listed as a GIM of any kind in the Bhutani data (185).
Tmem186 (chr16)	Transmembrane protein 186-Located in mitochondrion.	1 missense	No	5.70	
Tmem207 (chr16)	Transmembrane protein 207-predicted to be located in membrane and an integral component of membrane.	1 missense	Not in database	-	
Top3b (chr16)	Topoisomerase (DNA) III beta-Predicted to enable DNA topoisomerase activity. Acts upstream of or within chromosome segregation. Located in condensed chromosome.	1 missense	No	13.52	
Tprg (chr16)	Transformation related protein 63 regulated 1-Located in cytoplasm.	1 missense	No	0.01	
Txnrd2 (chr16)	thioredoxin reductase 2-The protein encoded by this gene belongs to the pyridine nucleotide-disulfide oxidoreductase family, and is a member of the thioredoxin (Trx) system. TrxRs are selenocysteine-containing flavoenzymes, which reduce thioredoxins, with a key role in redox homoeostasis. This gene	2 missense	No	1.89	

Gene name / chr (not confident GIM)	Function	SNP Number	Testis specific expression (Evo-Devo app)	Testis RPKM (NCBI)	Comments
	encodes a mitochondrial form important for scavenging reactive oxygen species in mitochondria.				
Vwa5b2 (chr16)	von Willebrand factor A domain containing 5B2.	1 missense	No	0*	
Zfp174 (chr16)	Zinc finger protein 174-Predicted to enable DNA-binding transcription factor activity, RNA polymerase II-specific; RNA polymerase II cis-regulatory region sequence-specific DNA binding activity; and protein homodimerization activity. Predicted to: be involved in negative regulation of transcription by RNA polymerase II, to be located in several cellular components, including actin cytoskeleton; cytosol; and nucleoplasm.	1 missense	No	0.77	



Tube: Sample RB2_HOM

Population	#Events	%Parent	%Total
All Events	446,445	####	100.0
viable	212,619	47.6	47.6
live	199,012	93.6	44.6
Singlet	110,000	55.3	24.6
P1	77,817	70.7	17.4

Supplementary Figure 8-11: Spermatid flow sorting gating strategy for the FACS Aria for dissociated testis stained with Hoechst 33342 and propidium iodide.

The cells within the red gate are the round spermatids that were sorted. They represented 17.4% of the total dissociated testis population in this example.

Supplementary Table 8-13: chr6 significant ASE scores per SNP and genes with significant ASE scores when performing per gene ASE analysis.

Gene ID	Gene name	chr	SNP start	Ref sum	Total count	ASE score per SNP	corrected P-value (per SNP) <=0.05	Dominant allele per SNP	SNP type*	No. of SNPs per gene	Ref sum (Gene)	Total count (Gene)	ASE score per gene	corrected P-value (per Gene) <=0.05	Dominant expressed allele (whole gene)
ENSMUSG00000093570	Gm20714	6	4816230	130	133	0.98	5.58E-32	WT	SR	3	43	45	0.96	5.17E-10	WT
ENSMUSG00000032667	Pon2	6	5264834	209	289	0.72	4.18E-12	WT	NTA	82					
			5264857	216	290	0.74	8.08E-15	WT	NTA						
			5264900	225	310	0.73	2.38E-13	WT	NTA						
			5265054	151	213	0.71	1.20E-07	WT	NTA						
			5267010	138	199	0.69	4.84E-06	WT	Sy						
ENSMUSG00000029757	Dync1i1	6	5843838	1	39	0.97	2.10E-08	Fu	NTA	781					
ENSMUSG00000093482	Gm20619	6	6315225	35	44	0.80	5.12E-03	WT	Nc	225					
			6359120	11	44	0.75	4.03E-02	Fu	NTA						
ENSMUSG00000029752	Asns	6	7677239	654	1542	0.58	3.36E-07	Fu	Sy,Nc	108					
			7677323	610	1430	0.57	3.10E-06	Fu	Sy,Nc						
			7685362	530	1217	0.56	4.88E-04	Fu	Sy						
			7693121	241	663	0.64	3.64E-10	Fu	NTA						
ENSMUSG00000107394	1700012J15Rik	6	7767855	11	44	0.75	4.03E-02	Fu	NTA	26					
			7767868	11	44	0.75	4.03E-02	Fu	NTA						
			7777352	374	582	0.64	9.92E-10	WT	Nc						
			7777403	149	400	0.63	3.24E-05	Fu	Nc						
			7777521	111	316	0.65	1.23E-05	Fu	Nc						
ENSMUSG00000042460	C1galt1	6	7847141	15	78	0.81	3.72E-06	Fu	Nc	135					
			7871280	38	52	0.73	3.83E-02	WT	NTA						
ENSMUSG00000068794	Col28a1	6	8049478	11	11	1.00	3.16E-02	WT	NTA	815					
			8049488	11	11	1.00	3.16E-02	WT	NTA						
			8050055	15	15	1.00	3.10E-03	WT	NTA						
			8050080	15	15	1.00	3.10E-03	WT	NTA						
			8050089	15	15	1.00	3.10E-03	WT	NTA						
			8050113	11	11	1.00	3.16E-02	WT	NTA						
			8057959	18	20	0.90	1.59E-02	WT	NTA						
			8058131	21	22	0.95	6.82E-04	WT	NTA						
ENSMUSG00000029638	Glcci1	6	8593267	484	581	0.83	1.62E-59	WT	Nc	191					
			8593359	463	535	0.87	1.41E-67	WT	Nc						
			8594804	41	54	0.76	7.81E-03	WT	Nc						
ENSMUSG00000107705	Gm45062	6	8593267	484	581	0.83	1.62E-59	WT	-	648					
			8593359	463	535	0.87	1.41E-67	WT	Nc						
			8594804	41	54	0.76	7.81E-03	WT	Nc						

Gene ID	Gene name	chr	SNP start	Ref sum	Total count	ASE score per SNP	corrected P-value (per SNP) <=0.05	Dominant allele per SNP	SNP type*	No. of SNPs per gene	Ref sum (Gene)	Total count (Gene)	ASE score per gene	corrected P-value (per Gene) <=0.05	Dominant expressed allele (whole gene)
ENSMUSG00000062995	Ica1	6	8656371	362	966	0.63	1.52E-12	Fu	Mis	403					
			8659420	178	273	0.65	4.65E-05	WT	NTA						
			8662266	105	153	0.69	3.17E-04	WT	NTA						
			8662360	89	126	0.71	2.79E-04	WT	NTA						
			8662483	68	97	0.70	4.56E-03	WT	NTA						
			8662647	95	145	0.66	1.01E-02	WT	NTA						
			8662748	120	175	0.69	7.70E-05	WT	NTA						
			8663036	119	180	0.66	1.09E-03	WT	NTA						
			8663202	96	138	0.70	3.33E-04	WT	NTA						
			8666655	103	149	0.69	2.42E-04	WT	NTA						
			8667207	90	123	0.73	2.34E-05	WT	NTA						
			8667382	105	138	0.76	8.04E-08	WT	NTA						
			8667448	90	118	0.76	9.86E-07	WT	NTA						
			8668306	60	88	0.68	2.99E-02	WT	NTA						
			8668308	58	86	0.67	4.98E-02	WT	NTA						
			8668331	57	83	0.69	3.08E-02	WT	NTA						
			8668488	74	105	0.70	1.80E-03	WT	NTA						
			8669950	72	108	0.67	2.46E-02	WT	NTA						
			8670299	76	102	0.75	5.98E-05	WT	NTA						
			8670386	64	86	0.74	4.31E-04	WT	NTA						
			8670666	56	76	0.74	2.32E-03	WT	NTA						
			8737416	32	36	0.89	1.41E-04	WT	NTA						
			8737460	31	35	0.89	2.43E-04	WT	NTA						
			8749741	1006	2167	0.54	3.16E-02	Fu	Sy, Nc						
ENSMUSG00000101894	1700016P04Rik	6	13413422	1171	2803	0.58	9.66E-16	Fu	Nc	38					
			13413435	1215	2897	0.58	1.26E-15	Fu	Nc						
			13415832	1231	3057	0.60	2.41E-24	Fu	Nc						
			13415988	41	53	0.77	4.03E-03	WT	Nc						
ENSMUSG00000029552	Tes	6	17105690	1	18	0.94	6.53E-03	Fu	NTA	101					
ENSMUSG00000054556	Gm4876	6	17105690	1	18	0.94	6.53E-03	Fu	NTA	381					
ENSMUSG00000085171	D830026I12Rik	6	17208950	0	19	1.00	2.60E-04	Fu	Nc	60					
			17209011	0	13	1.00	1.02E-02	Fu	Nc						
ENSMUSG00000085264	Gm15581	6	17208950	0	19	1.00	2.60E-04	Fu	Nc	494					
			17209011	0	13	1.00	1.02E-02	Fu	Nc						
ENSMUSG00000029534	St7	6	17733339	0	11	1.00	3.16E-02	Fu	NTA	608					
			17734348	0	12	1.00	1.81E-02	Fu	NTA						
ENSMUSG00000010796	Asz1	6	18094299	0	12	1.00	1.81E-02	Fu	NTA	21					
			18095529	2	20	0.90	1.59E-02	Fu	NTA						
			18098149	1	19	0.95	3.77E-03	Fu	NTA						
			18098826	0	17	1.00	9.08E-04	Fu	NTA						
			18099974	0	11	1.00	3.16E-02	Fu	NTA						
ENSMUSG00000023089	Ndufa5	6	24527627	129	155	0.83	3.93E-15	WT	NTA	5	52	65	0.809	2.90E-05	WT
			24527665	128	160	0.80	1.96E-12	WT	NTA						
ENSMUSG00000029685	Asb15	6	24558509	269	462	0.58	1.81E-02	WT	SR,Sy	65					
ENSMUSG00000029679	Hyal6	6	24733307	110	152	0.72	3.28E-06	WT	NTA	88					
			24733398	302	481	0.63	2.34E-06	WT	NTA						
			24734158	256	414	0.62	1.26E-04	WT	Sy						
			24734713	292	478	0.61	1.08E-04	WT	Sy						
			24734881	208	342	0.61	3.71E-03	WT	Sy						
			24734931	229	373	0.61	7.79E-04	WT	Mis						
			24743365	317	531	0.60	5.73E-04	WT	Mis						
			24743457	321	537	0.60	4.45E-04	WT	Sy						
			24743940	207	343	0.60	6.73E-03	WT	NTA						
			24743944	202	342	0.59	3.16E-02	WT	NTA						
ENSMUSG00000029682	Spam1	6	24796021	528	855	0.62	1.10E-09	WT	Nc	23	178	290	0.612	7.36E-03	WT
			24796215	545	863	0.63	2.35E-12	WT	Sy,Nc						
			24796457	597	978	0.61	8.87E-10	WT	Mis						
			24796707	523	906	0.58	2.59E-04	WT	Sy,Nc						
			24800369	476	741	0.64	1.88E-12	WT	Sy,Nc						
			24800607	395	635	0.62	1.07E-07	WT	Mis						
			24800840	468	768	0.61	1.83E-07	WT	NTA						
			24800890	276	446	0.62	4.78E-05	WT	NTA						
			24800990	173	279	0.62	3.60E-03	WT	NTA						

Supplementary Table 8-14: chr16 significant ASE scores per SNP and genes with significant ASE scores when performing per gene ASE analysis.

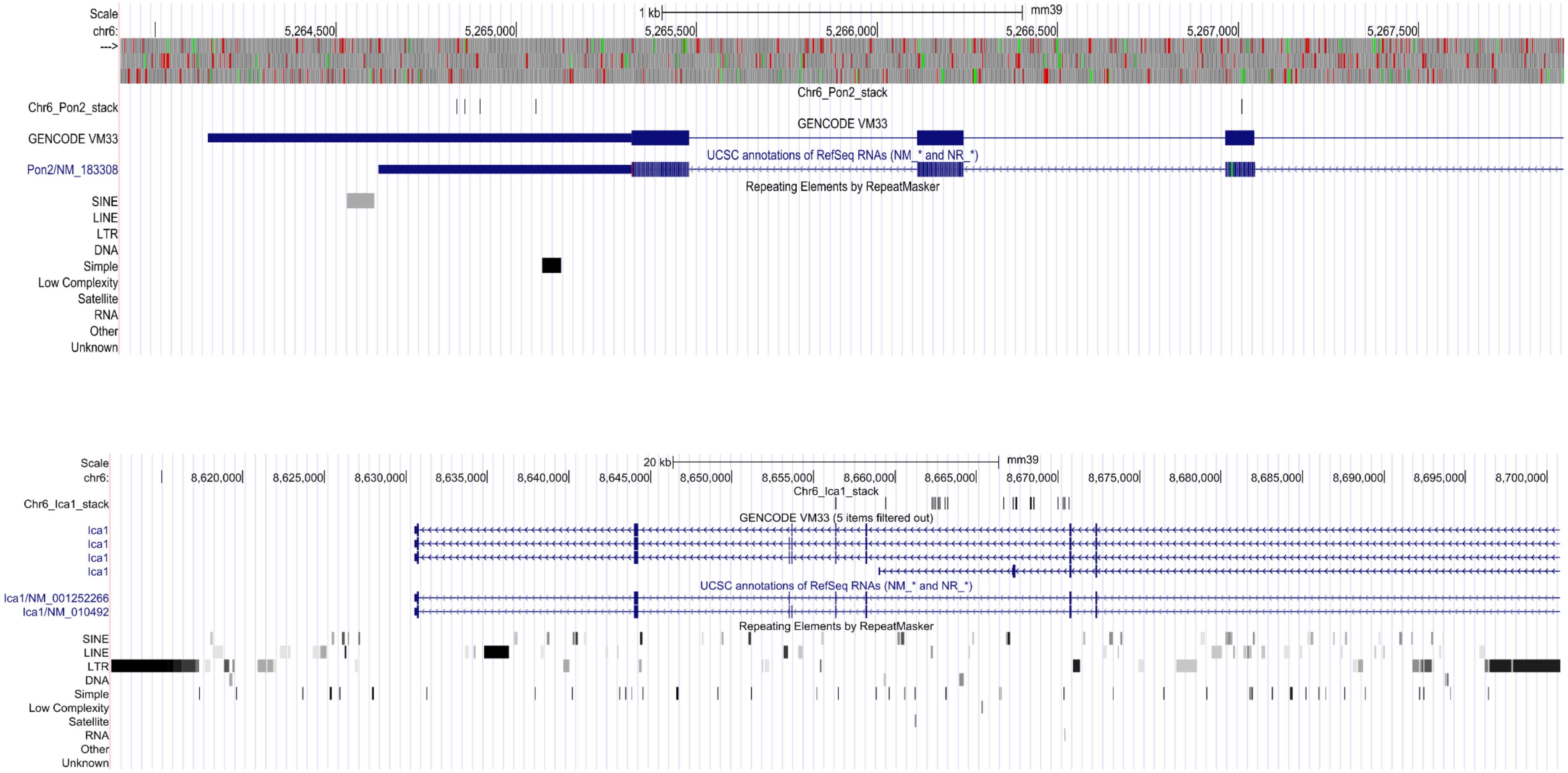
Gene ID	Gene name	chr	SNP start	Ref sum	Total count	ASE score per SNP	corrected P-value (per SNP) <=0.05	Dominant allele per SNP	SNP type*	No. of SNPs per gene	Ref sum (Gene)	Total count (Gene)	ASE score per gene	corrected P-value (per Gene) <=0.05	Dominant expressed allele (whole gene)
ENSMUSG00000005982	Naa60	16	3690269	79	122	0.65	4.5E-02	WT	NTA	56					
ENSMUSG00000039789	Zfp597	16	3690269	79	122	0.65	4.5E-02	WT	Nc	43					
ENSMUSG00000005983	1700037C18Rik	16	3725122	70	206	0.66	3.4E-04	Fu	NTA	13					
			3725916	145	221	0.66	2.7E-04	WT	NTA						
ENSMUSG00000093575	Gm20695	16	3725122	70	206	0.66	3.4E-04	Fu	NTA	43					
			3725916	145	221	0.66	2.7E-04	WT	NTA						
ENSMUSG00000039738	Slx4	16	3798548	352	579	0.61	2.0E-05	WT	Mis,Nc	72					
			3798732	354	594	0.60	2.3E-04	WT	Sy,Nc						
			3803139	311	521	0.60	6.9E-04	WT	Mis,Nc						
			3803531	246	405	0.61	1.1E-03	WT	Sy,Nc						
			3803613	271	438	0.62	6.1E-05	WT	Mis,Nc						
			3804535	230	387	0.59	1.0E-02	WT	Sy,Nc						
			3804901	361	534	0.68	7.8E-14	WT	Mis,Nc						
			3808765	236	387	0.61	1.1E-03	WT	Sy,Nc						
			3812747	253	397	0.64	4.9E-06	WT	Mis						
			3813726	244	360	0.68	2.2E-09	WT	Mis						
			3818822	268	459	0.58	1.5E-02	WT	Mis						
ENSMUSG00000014301	Pam16	16	4434333	12	12	1.00	1.8E-02	WT	Nc	11					
ENSMUSG00000014303	Glis2	16	4434333	12	12	1.00	1.8E-02	WT	Nc	54					
ENSMUSG00000022518	4930562C15Rik	16	4679496	8	43	0.81	2.2E-03	Fu	NTA	26					
ENSMUSG00000022515	Anks3	16	4771889	0	16	1.00	1.7E-03	Fu	NTA	51					
ENSMUSG00000022543	Dnaaf8	16	4782352	7	46	0.85	1.4E-04	Fu	Nc	41	40	221	0.8170	4.55E-20	Fu
			4783716	3	26	0.88	4.3E-03	Fu	Nc						
			4783812	4	24	0.83	4.8E-02	Fu	Nc						
			4783817	4	24	0.83	4.8E-02	Fu	Nc						
			4783973	1	21	0.95	1.2E-03	Fu	Nc						
			4794042	247	1407	0.82	1.14E-137	Fu	Nc						
			4794085	257	1475	0.83	1.09E-145	Fu	Nc						
			4795886	128	894	0.86	1.14E-107	Fu	Nc						
			4795917	123	809	0.85	6.77E-92	Fu	Nc						
			4795919	128	814	0.84	2.22E-89	Fu	Nc						
			4796112	178	1262	0.86	8.45E-155	Fu	Nc						
			4796626	193	1445	0.87	4.67E-186	Fu	Nc						
ENSMUSG00000022540	Rogdi	16	4830519	93	116	0.80	5.64E-09	WT	NTA	3	51	75	0.6875	4.66E-02	WT
ENSMUSG00000106967	Gm42477	16	4830519	93	116	0.80	5.64E-09	WT	NTA	2	52	69	0.7518	2.12E-03	WT

Gene ID	Gene name	chr	SNP start	Ref sum	Total count	ASE score per SNP	corrected P-value (per SNP) <=0.05	Dominant allele per SNP	SNP type*	No. of SNPs per gene	Ref sum (Gene)	Total count (Gene)	ASE score per gene	corrected P-value (per Gene) <=0.05	Dominant expressed allele (whole gene)
ENSMUSG00000008658	Rbfox1	16	5750541	38	47	0.81	1.4E-03	WT	NTA	10200					
			5752787	47	59	0.80	3.4E-04	WT	NTA						
			5752834	43	53	0.81	3.7E-04	WT	NTA						
			5752891	39	48	0.81	9.1E-04	WT	NTA						
			5752930	28	28	1.00	8.33E-07	WT	NTA						
			5752948	27	28	0.96	1.90E-05	WT	NTA						
			5764702	18	72	0.75	1.5E-03	Fu	NTA						
			5764705	18	77	0.77	2.2E-04	Fu	NTA						
			5764707	18	77	0.77	2.2E-04	Fu	NTA						
			5764800	17	81	0.79	1.28E-05	Fu	NTA						
			5764809	19	79	0.76	2.8E-04	Fu	NTA						
			5764837	28	91	0.69	1.3E-02	Fu	NTA						
			5765073	195	274	0.71	3.18E-10	WT	NTA						
			5767304	0	59	1.00	1.08E-15	Fu	NTA						
			5767393	0	46	1.00	6.23E-12	Fu	NTA						
			5767424	0	44	1.00	2.36E-11	Fu	NTA						
			5767498	0	105	1.00	3.48E-29	Fu	NTA						
			5768331	11	11	1.00	3.2E-02	WT	NTA						
			5893750	36	36	1.00	4.60E-09	WT	NTA						
			5895990	27	28	0.96	1.90E-05	WT	NTA						
			5895997	27	28	0.96	1.90E-05	WT	NTA						
			5896003	26	27	0.96	3.48E-05	WT	NTA						
			5896474	72	78	0.92	4.56E-13	WT	NTA						
			5896530	89	96	0.93	1.07E-16	WT	NTA						
			5896589	84	93	0.90	5.91E-14	WT	NTA						
			5901686	11	11	1.00	3.2E-02	WT	NTA						
			5901977	30	31	0.97	3.00E-06	WT	NTA						
			5904556	92	92	1.00	2.31E-25	WT	NTA						
			5906955	42	44	0.95	1.67E-08	WT	NTA						
			5906991	43	45	0.96	8.90E-09	WT	NTA						
			5907057	28	29	0.97	1.03E-05	WT	NTA						
			5907060	28	29	0.97	1.03E-05	WT	NTA						
			7104808	11	11	1.00	3.2E-02	WT	NTA						
			7105369	13	13	1.00	1.0E-02	WT	NTA						
			7105380	13	13	1.00	1.0E-02	WT	NTA						
			7105410	13	13	1.00	1.0E-02	WT	NTA						
			7105524	16	16	1.00	1.7E-03	WT	NTA						
			7106297	13	13	1.00	1.0E-02	WT	NTA						
			7106372	15	15	1.00	3.1E-03	WT	NTA						
			7106555	21	21	1.00	7.38E-05	WT	NTA						
			7106559	20	20	1.00	1.4E-04	WT	NTA						
			7106575	20	20	1.00	1.4E-04	WT	NTA						
ENSMUSG00000022711	Pmm2	16	8455864	22	26	0.85	1.9E-02	WT	NTA	53					
ENSMUSG00000022710	Usp7	16	8586492	13	13	1.00	1.0E-02	WT	NTA	169					
			8587020	12	12	1.00	1.8E-02	WT	NTA						
			8587275	11	11	1.00	3.2E-02	WT	NTA						
			8587436	12	12	1.00	1.8E-02	WT	NTA						
			8587868	11	11	1.00	3.2E-02	WT	NTA						
			8587870	11	11	1.00	3.2E-02	WT	NTA						
			8589541	13	13	1.00	1.0E-02	WT	NTA						
ENSMUSG00000039200	Atf7ip2	16	10010867	130	209	0.62	1.9E-02	WT	Nc	72					
			10022469	186	260	0.72	4.79E-10	WT	SR, Sy						
ENSMUSG00000115943	Gm49455	16	10010867	130	209	0.62	1.9E-02	WT	Nc	57					
			10022469	186	260	0.72	4.79E-10	WT	SR,Sy,Nc						

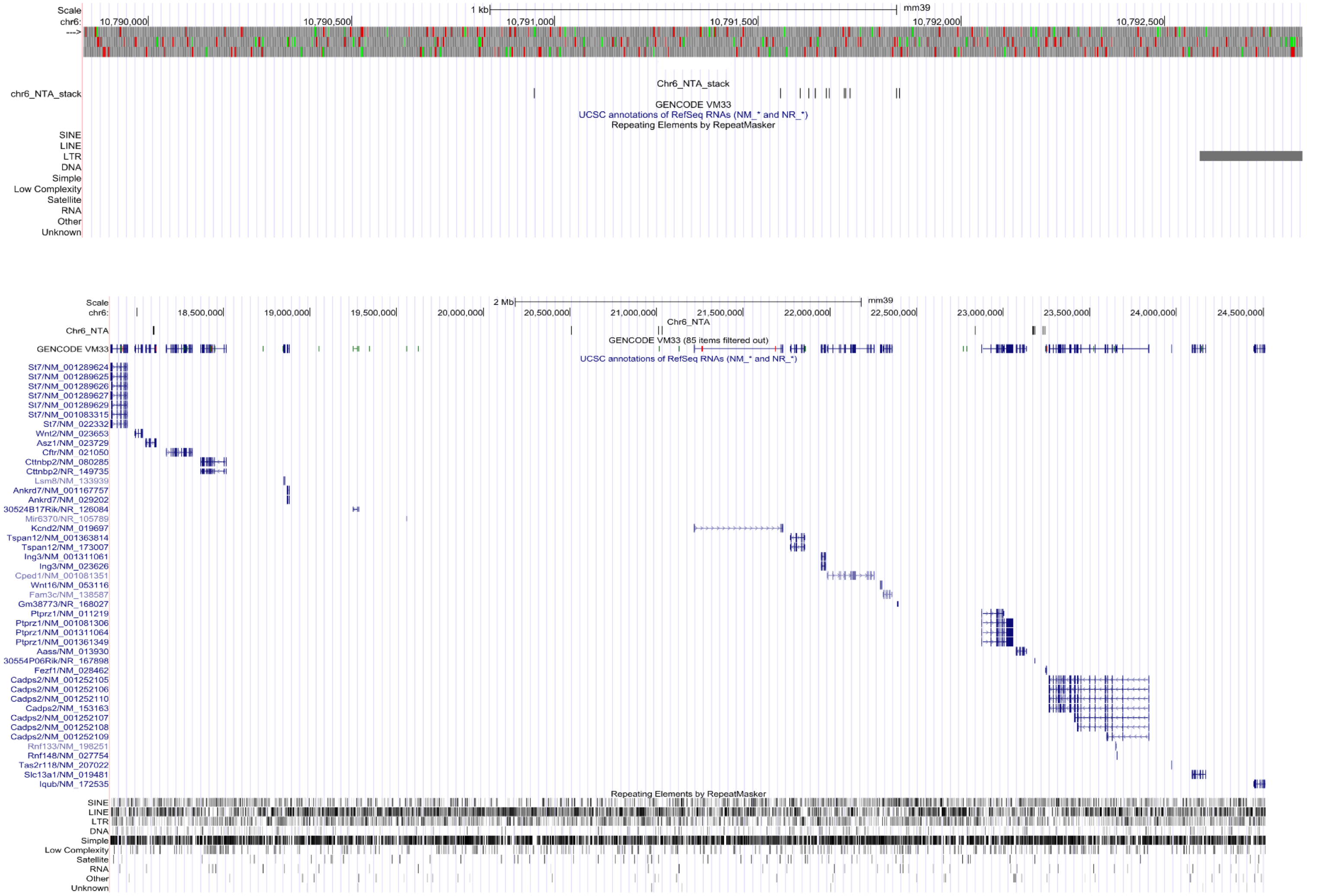
Gene ID	Gene name	chr	SNP start	Ref sum	Total count	ASE score per SNP	corrected P-value (per SNP) <=0.05	Dominant allele per SNP	SNP type*	No. of SNPs per gene	Ref sum (Gene)	Total count (Gene)	ASE score per gene	corrected P-value (per Gene) <=0.05	Dominant expressed allele (whole gene)
ENSMUSG00000039179	Tekt5	16	10175903	614	984	0.62	1.62E-12	WT	Sy	132					
			10176003	789	1171	0.67	3.02E-30	WT	Mis						
			10200995	72	91	0.79	2.06E-06	WT	NTA						
			10201396	72	102	0.71	2.1E-03	WT	NTA						
			10210683	4	29	0.86	5.0E-03	Fu	NTA						
ENSMUSG00000022503	Nubp1	16	10231587	94	377	0.75	1.75E-20	Fu	Sy,Nc	70					
			10235527	96	361	0.73	6.70E-17	Fu	SR,Nc						
			10239497	87	264	0.67	3.21E-06	Fu	Sy,Nc						
			10239540	93	277	0.66	4.81E-06	Fu	Mis, Nc						
			10241129	86	269	0.68	3.93E-07	Fu	Mis, Nc						
ENSMUSG00000050908	Tvp23a	16	10239497	87	264	0.67	3.21E-06	Fu	Sy,Nc	105					
			10239540	93	277	0.66	4.81E-06	Fu	Mis,Nc						
			10241129	86	269	0.68	3.93E-07	Fu	Mis,Nc						
ENSMUSG00000043050	Tnp2	16	10606175	36334	####	0.51	2.83E-07	Fu	Sy	1	36334	74297	0.51	3.44E-07	Fu
ENSMUSG00000050058	Prm3	16	10608574	1523	3661	0.58	1.23E-21	Fu	Sy	2	1360	3270	0.584	2.07E-19	Fu
			10608598	1197	2879	0.58	5.21E-17	Fu	Sy						
ENSMUSG00000038015	Prm2	16	10609683	32436	####	0.53	2.05E-34	WT	Sy	33	2154	4027	0.535	8.45E-04	WT
			10609736	34156	####	0.53	7.12E-54	WT	Mis						
			10609988	100	293	0.66	5.73E-06	Fu	NTA						
			10610237	53	53	1.00	6.02E-14	WT	NTA						
			10610951	342	519	0.66	7.62E-11	WT	NTA						
			10610996	287	456	0.63	3.60E-06	WT	NTA						
			10611099	172	284	0.61	1.7E-02	WT	NTA						
			10611195	188	297	0.63	3.6E-04	WT	NTA						
			10611304	196	291	0.67	3.84E-07	WT	NTA						
			10611892	154	224	0.69	2.11E-06	WT	NTA						
			10611926	155	215	0.72	1.07E-08	WT	NTA						
			10611951	160	252	0.63	1.3E-03	WT	NTA						
			10612013	161	170	0.95	3.24E-34	WT	NTA						
			10612118	167	177	0.94	6.34E-35	WT	NTA						
			10612121	170	180	0.94	9.73E-36	WT	NTA						
			10612155	158	166	0.95	2.37E-34	WT	NTA						
			10613571	77	77	1.00	5.86E-21	WT	NTA						
			10613788	236	303	0.78	1.36E-20	WT	NTA						
			10613865	244	347	0.70	5.53E-12	WT	NTA						
			10613874	243	327	0.74	1.50E-16	WT	5P						
ENSMUSG00000022501	Prm1	16	10614856	31	107	0.71	9.5E-04	Fu	NTA	38					
			10614922	52	168	0.69	6.90E-05	Fu	NTA						
			10615009	73	216	0.66	1.6E-04	Fu	NTA						
			10615032	73	223	0.67	2.35E-05	Fu	NTA						
			10615043	71	225	0.68	3.24E-06	Fu	NTA						
			10616549	106	287	0.63	7.0E-04	Fu	NTA						
			10617859	108	290	0.63	9.8E-04	Fu	NTA						
			10618035	124	310	0.60	1.9E-02	Fu	NTA						
			10618036	124	311	0.60	1.7E-02	Fu	NTA						
			10619201	71	191	0.63	1.8E-02	Fu	NTA						
			10619847	86	226	0.62	1.6E-02	Fu	NTA						
			10620211	98	250	0.61	2.7E-02	Fu	NTA						
			10620516	101	259	0.61	1.8E-02	Fu	NTA						
ENSMUSG00000116038	Gm46563	16	10616549	106	287	0.63	7.0E-04	Fu	NTA	44					
			10617859	108	290	0.63	9.8E-04	Fu	NTA						
			10618035	124	310	0.60	1.9E-02	Fu	NTA						
			10618036	124	311	0.60	1.7E-02	Fu	NTA						
			10619201	71	191	0.63	1.8E-02	Fu	NTA						
			10619847	86	226	0.62	1.6E-02	Fu	NTA						
			10620211	98	250	0.61	2.7E-02	Fu	NTA						
			10620516	101	259	0.61	1.8E-02	Fu	NTA						
ENSMUSG00000022500	Litaf	16	10873133	18	21	0.86	4.7E-02	WT	NTA	258					
			10873849	29	36	0.81	1.3E-02	WT	NTA						
			10873909	31	37	0.84	2.2E-03	WT	NTA						
			10873922	32	38	0.84	1.4E-03	WT	NTA						
			10874099	32	42	0.76	3.2E-02	WT	NTA						
			10874396	16	18	0.89	4.2E-02	WT	NTA						

Gene ID	Gene name	chr	SNP start	Ref sum	Total count	ASE score per SNP	corrected P-value (per SNP) <=0.05	Dominant allele per SNP	SNP type*	No. of SNPs per gene	Ref sum (Gene)	Total count (Gene)	ASE score per gene	corrected P-value (per Gene) <=0.05	Dominant expressed allele (whole gene)
ENSMUSG00000037972	Snn	16	10890284	147	221	0.67	7.95E-05	WT	Sy	28					
			10890459	158	230	0.69	1.56E-06	WT	NTA						
			10891092	84	99	0.85	1.56E-10	WT	NTA						
			10892131	46	58	0.79	5.2E-04	WT	NTA						
ENSMUSG00000116520	4930414F18Rik	16	12567458	33	43	0.77	2.2E-02	WT	NTA	652					
			12567945	29	32	0.91	1.8E-04	WT	NTA						
ENSMUSG00000109857	Gm53058	16	13281579	386	565	0.68	5.84E-16	WT	Nc	29					
ENSMUSG00000087526	Gm15738	16	13281579	386	565	0.68	5.84E-16	WT	Nc	94					
ENSMUSG00000022680	Pdxdc1	16	13651888	36	37	0.97	7.32E-08	WT	NTA	86					
			13651912	33	34	0.97	4.72E-07	WT	NTA						
			13652248	851	1006	0.85	4.23E-113	WT	Nc						
			13653808	10	431	0.98	1.12E-106	Fu	Nc						
			13658237	970	1553	0.62	3.01E-20	WT	Sy, Nc						
ENSMUSG00000022681	Ntan1	16	13651888	36	37	0.97	7.32E-08	WT	NTA	17	56	65	0.8584	9.17E-07	WT
			13651912	33	34	0.97	4.72E-07	WT	NTA						
			13652248	851	1006	0.85	4.23E-113	WT	Nc						
ENSMUSG00000065968	Ifitm7	16	13799746	931	1455	0.64	4.70E-24	WT	Nc	14					
ENSMUSG00000022674	Ube2v2	16	15413487	5	33	0.85	3.3E-03	Fu	Nc	64					
			15413493	5	33	0.85	3.3E-03	Fu	5p, Nc						
ENSMUSG00000022673	Mcm4	16	15448199	2	23	0.91	3.3E-03	Fu	Sy	54					
ENSMUSG00000022672	Prkdc	16	15587718	19	20	0.95	2.1E-03	WT	Sy	507					
			15594928	24	25	0.96	1.2E-04	WT	Sy						
			15608319	23	26	0.88	4.3E-03	WT	Sy						
			15617769	20	22	0.91	5.6E-03	WT	Mis						
ENSMUSG00000041957	Pkp2	16	16058429	4	41	0.90	9.56E-06	Fu	Sy,Nc	6					
ENSMUSG00000022789	Dnm1l	16	16162692	4	48	0.92	1.90E-07	Fu	NTA	206					
			16162711	8	55	0.85	7.60E-06	Fu	NTA						
ENSMUSG00000116096	4933404G15Rik	16	16528043	318	766	0.58	2.1E-04	Fu	Nc	55					
			16531059	31	34	0.91	6.12E-05	WT	NTA						
ENSMUSG00000022783	Spag6l	16	16571329	462	1199	0.61	4.68E-13	Fu	NTA	172					
			16571518	696	1559	0.55	1.5E-03	Fu	NTA						
			16581005	813	1824	0.55	2.7E-04	Fu	Sy						
			16646983	10	41	0.76	4.6E-02	Fu	Nc						
ENSMUSG00000022773	Ypel1	16	16904301	70	186	0.62	3.2E-02	Fu	NTA	16					
ENSMUSG00000049916	2610318N02Rik	16	16931361	3114	3978	0.78	3.28E-291	WT	Sy	19	174	228	0.7606	3.32E-13	WT
		16	16933013	9	39	0.77	3.4E-02	Fu	Sy						
ENSMUSG00000116658	Gm49580	16	17847034	76	114	0.67	1.8E-02	WT	Nc	206					
			17847056	67	100	0.67	3.1E-02	WT	Nc						
ENSMUSG00000003526	Prodh	16	17887951	385	523	0.74	3.00E-25	WT	SG	51					
ENSMUSG00000003531	Dgcr6	16	17887951	385	523	0.74	3.00E-25	WT	SG	37					
ENSMUSG00000009097	Tbx1	16	18405368	18	20	0.90	1.6E-02	WT	Mis	6					
ENSMUSG00000062901	Klhl24	16	19926326	240	381	0.63	3.71E-05	WT	Sy	47					
ENSMUSG00000115293	Eef1ece2	16	20440282	16	18	0.89	4.2E-02	WT	NTA	43					
ENSMUSG00000044626	Liph	16	21773837	9	53	0.83	9.36E-05	Fu	NTA	126					
			21795017	8	119	0.93	1.18E-21	Fu	Sy, Nc						
			21800193	14	87	0.84	1.28E-08	Fu	Mis, SR						
			21802735	15	96	0.84	7.06E-10	Fu	Sy, Nc						
ENSMUSG00000022855	Senp2	16	21854585	29	34	0.85	2.1E-03	WT	Nc	17					
ENSMUSG00000043870	Gm5809	16	22049199	20	20	1.00	1.4E-04	WT	Nc	3					
			22049282	0	25	1.00	5.69E-06	Fu	Nc						
			22049570	12	12	1.00	1.8E-02	WT	Nc						
ENSMUSG00000004460	Dnajb11	16	22684281	165	268	0.62	8.1E-03	WT	Sy	235					
			22688176	205	330	0.62	7.7E-04	WT	Sy						
			22688182	212	331	0.64	3.04E-05	WT	Sy						
			22688194	218	347	0.63	1.5E-04	WT	Sy						
			22690622	251	393	0.64	4.16E-06	WT	NTA						
			22690907	133	202	0.66	5.0E-04	WT	NTA						
			22690908	133	202	0.66	5.0E-04	WT	NTA						
			22690994	88	128	0.69	1.5E-03	WT	NTA						
			22691007	85	120	0.71	3.8E-04	WT	NTA						

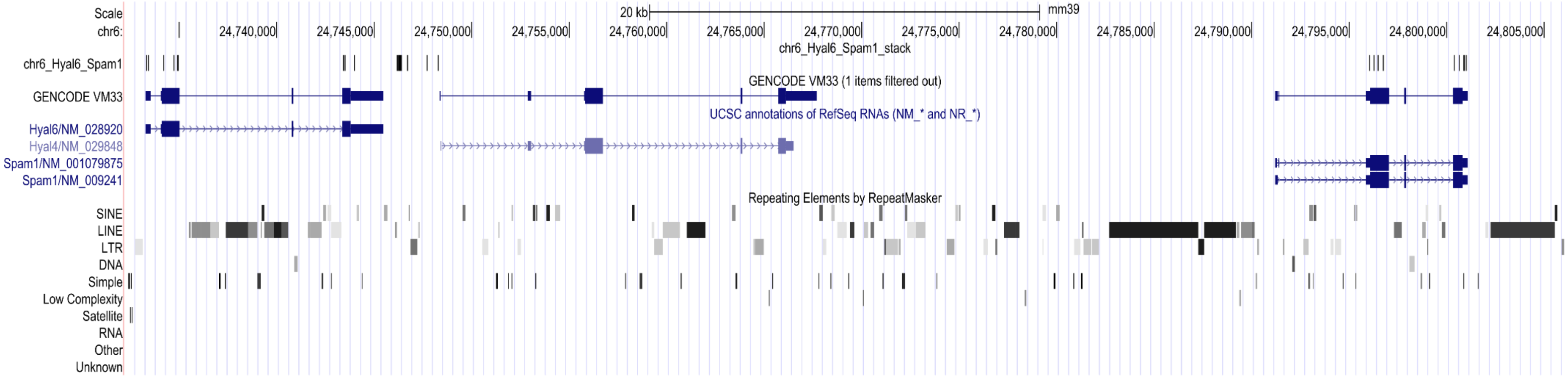
Gene ID	Gene name	chr	SNP start	Ref sum	Total count	ASE score per SNP	corrected P-value (per SNP) <=0.05	Dominant allele per SNP	SNP type*	No. of SNPs per gene	Ref sum (Gene)	Total count (Gene)	ASE score per gene	corrected P-value (per Gene) <=0.05	Dominant expressed allele (whole gene)
ENSMUSG00000060459	Kng2	16	22808170	11	11	1.00	3.2E-02	WT	NTA	101					
			22808469	12	12	1.00	1.8E-02	WT	NTA						
			22808518	18	19	0.95	3.8E-03	WT	NTA						
ENSMUSG00000115869	Gm31814	16	24012103	18	20	0.90	1.6E-02	WT	Nc	270					
			24012843	20	21	0.95	1.2E-03	WT	NTA						
			24014616	26	32	0.81	1.9E-02	WT	NTA						
			24014950	28	33	0.85	3.3E-03	WT	NTA						
			24016374	27	31	0.87	1.9E-03	WT	NTA						
			24016417	30	33	0.91	1.1E-04	WT	NTA						
			24016436	33	36	0.92	1.99E-05	WT	NTA						
			24017336	16	18	0.89	4.2E-02	WT	NTA						
			24022256	35	40	0.88	1.1E-04	WT	NTA						
			24022934	16	18	0.89	4.2E-02	WT	Nc						
			24023094	18	20	0.90	1.6E-02	WT	Nc						
			24023160	17	19	0.89	2.6E-02	WT	Nc						
			24025404	59	66	0.89	3.82E-09	WT	Nc						
			24070022	0	11	1.00	3.2E-02	Fu	NTA						
			24070140	0	18	1.00	4.9E-04	Fu	NTA						
			24070239	0	18	1.00	4.9E-04	Fu	NTA						
			24070294	0	16	1.00	1.7E-03	Fu	NTA						
			24070531	1	27	0.96	3.48E-05	Fu	NTA						
			24070560	1	28	0.96	1.90E-05	Fu	NTA						
			24070622	1	16	0.94	1.9E-02	Fu	NTA						
			24072325	0	12	1.00	1.8E-02	Fu	NTA						
			24072398	0	18	1.00	4.9E-04	Fu	NTA						
			24072421	0	18	1.00	4.9E-04	Fu	NTA						
			24072453	0	21	1.00	7.38E-05	Fu	NTA						
			24072591	0	14	1.00	5.6E-03	Fu	NTA						
			24072598	0	14	1.00	5.6E-03	Fu	NTA						
			24072670	0	13	1.00	1.0E-02	Fu	NTA						
			24072676	0	12	1.00	1.8E-02	Fu	NTA						
			24072804	0	12	1.00	1.8E-02	Fu	NTA						
			24073470	0	11	1.00	3.2E-02	Fu	NTA						
			24075033	0	21	1.00	7.38E-05	Fu	NTA						
			24076438	0	20	1.00	1.4E-04	Fu	NTA						
			24081709	0	14	1.00	5.6E-03	Fu	NTA						
ENSMUSG00000115852	Gm52969	16	24012103	18	20	0.90	1.6E-02	WT	Nc	606					
			24012843	20	21	0.95	1.2E-03	WT	NTA						
			24014616	26	32	0.81	1.9E-02	WT	NTA						
			24014950	28	33	0.85	3.3E-03	WT	NTA						
			24016374	27	31	0.87	1.9E-03	WT	NTA						
			24016417	30	33	0.91	1.1E-04	WT	NTA						
			24016436	33	36	0.92	1.99E-05	WT	NTA						
			24017336	16	18	0.89	4.2E-02	WT	NTA						
			24022256	35	40	0.88	1.1E-04	WT	Nc						
			24022934	16	18	0.89	4.2E-02	WT	Nc						
			24023094	18	20	0.90	1.6E-02	WT	Nc						
			24023160	17	19	0.89	2.6E-02	WT	Nc						
			24025404	59	66	0.89	3.82E-09	WT	Nc						
			24070022	0	11	1.00	3.2E-02	Fu	NTA						
			24070140	0	18	1.00	4.9E-04	Fu	NTA						
			24070239	0	18	1.00	4.9E-04	Fu	NTA						
			24070294	0	16	1.00	1.7E-03	Fu	NTA						
			24070531	1	27	0.96	3.48E-05	Fu	NTA						
			24070560	1	28	0.96	1.90E-05	Fu	NTA						
			24070622	1	16	0.94	1.9E-02	Fu	NTA						
			24072325	0	12	1.00	1.8E-02	Fu	NTA						
			24072398	0	18	1.00	4.9E-04	Fu	NTA						
			24072421	0	18	1.00	4.9E-04	Fu	NTA						
			24072453	0	21	1.00	7.38E-05	Fu	NTA						
			24072591	0	14	1.00	5.6E-03	Fu	NTA						
			24072598	0	14	1.00	5.6E-03	Fu	NTA						
			24072670	0	13	1.00	1.0E-02	Fu	NTA						
			24072676	0	12	1.00	1.8E-02	Fu	NTA						
			24072804	0	12	1.00	1.8E-02	Fu	NTA						
			24073470	0	11	1.00	3.2E-02	Fu	NTA						
			24075033	0	21	1.00	7.38E-05	Fu	NTA						
			24076438	0	20	1.00	1.4E-04	Fu	NTA						
			24081709	0	14	1.00	5.6E-03	Fu	NTA						
			24088617	0	19	1.00	2.6E-04	Fu	NTA						
			24120961	0	11	1.00	3.2E-02	Fu	NTA						
			24134547	0	20	1.00	1.4E-04	Fu	NTA						
			24134658	0	27	1.00	1.60E-06	Fu	NTA						
			24170515	1	20	0.95	2.1E-03	Fu	NTA						
ENSMUSG00000092545	Gm20319	16	26532871	36	38	0.95	6.19E-07	WT	Nc	252					
			26545898	32	37	0.86	4.9E-04	WT	Nc						
			26546386	35	42	0.83	9.1E-04	WT	Nc						
ENSMUSG00000022514	Il1rap	16	26532871	36	38	0.95	6.19E-07	WT	Nc	292					
			26545898	32	37	0.86	4.9E-04	WT	Nc						
			26546386	35	42	0.83	9.1E-04	WT	Nc						



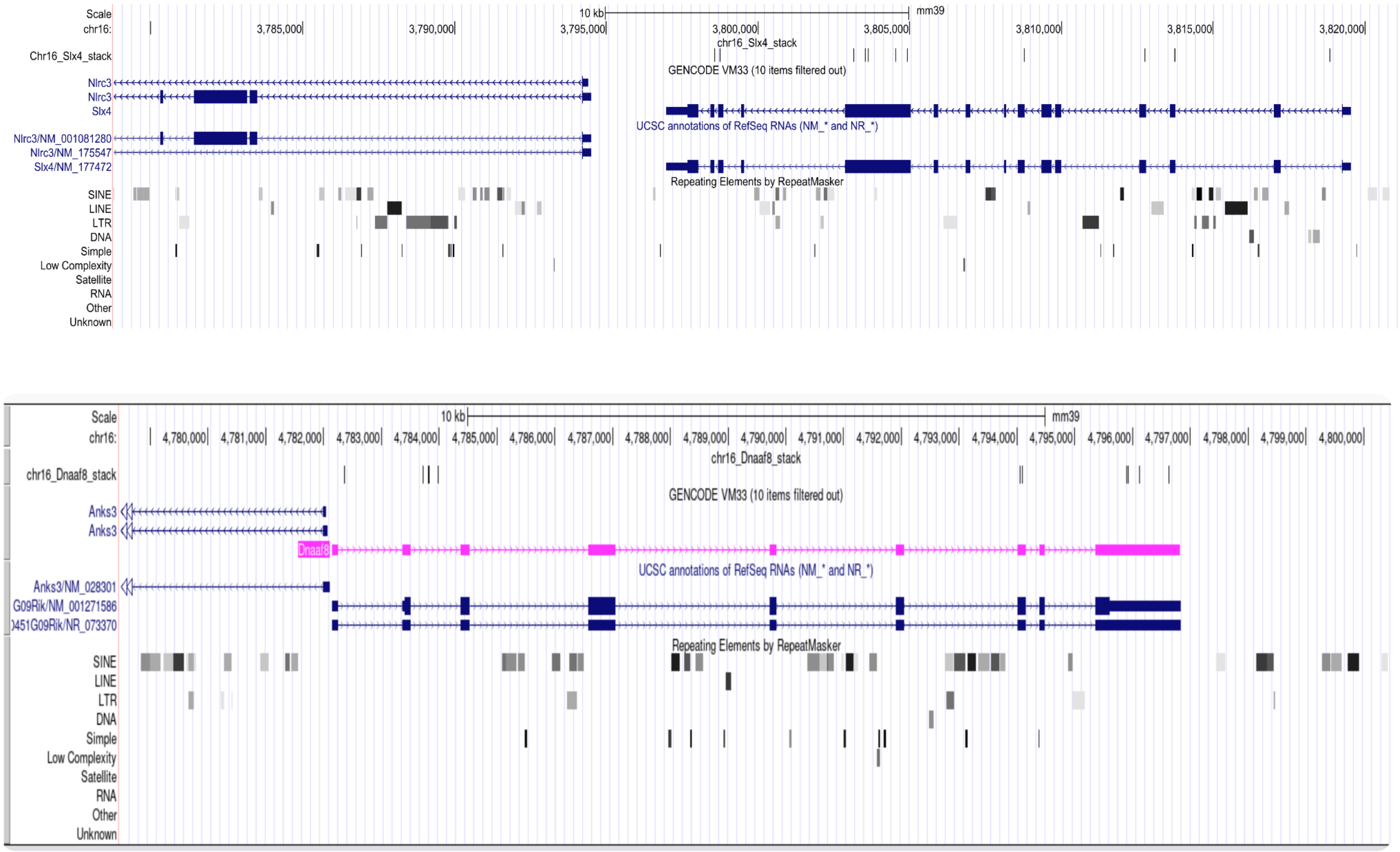
Supplementary Figure 8-12: UCSC genome browser plots of ASE chr6 ROI *Pon2* and *Ica1* stacks related to Figure 5-5.



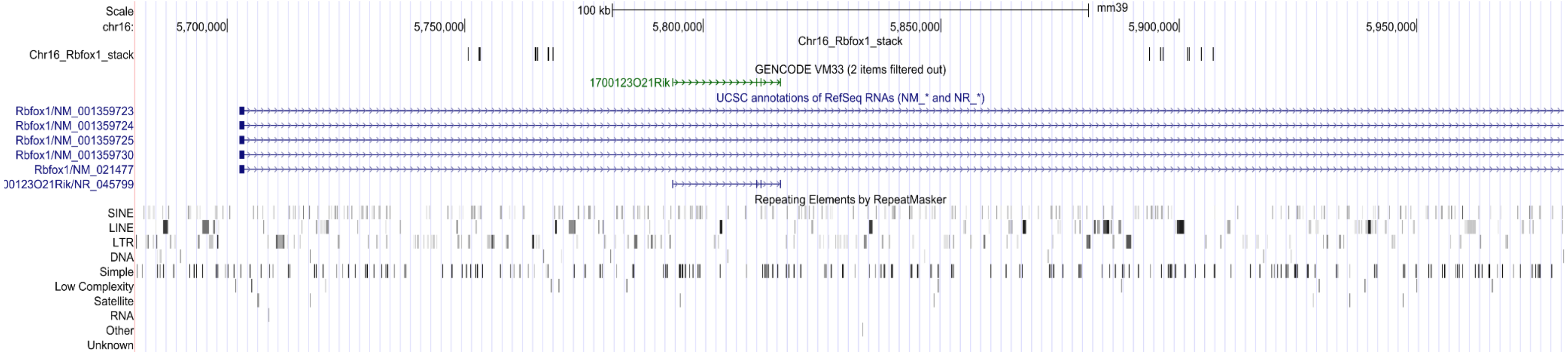
Supplementary Figure 8-13: UCSC genome browser plots of chr6 ASE stacks not annotated regions (NTA) left and right, related to Figure 5-5.

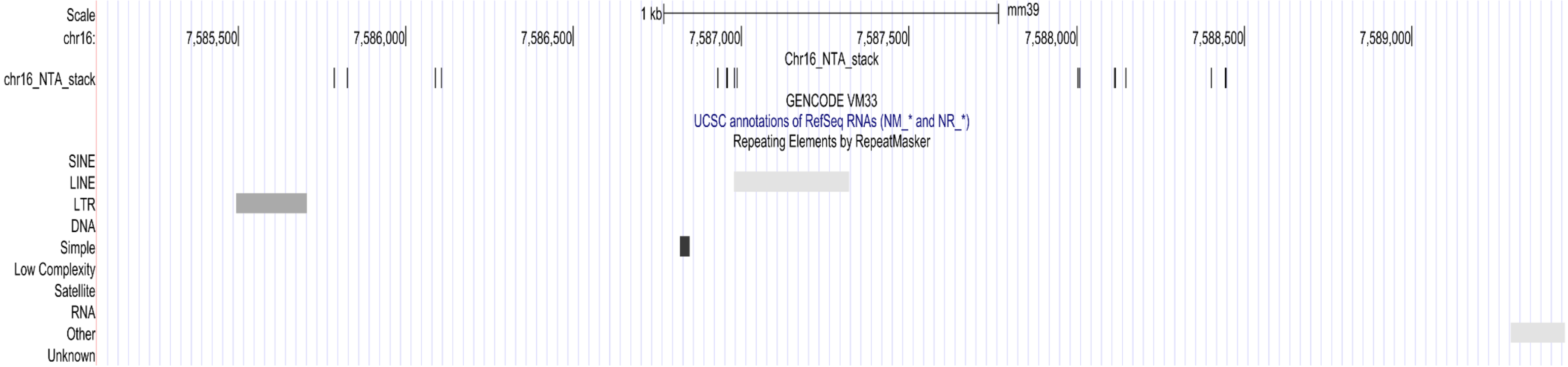


Supplementary Figure 8-14: UCSC genome browser plots of chr16 *Hyal6* and *Spam1* ASE stacks, related to Figure 5-5.

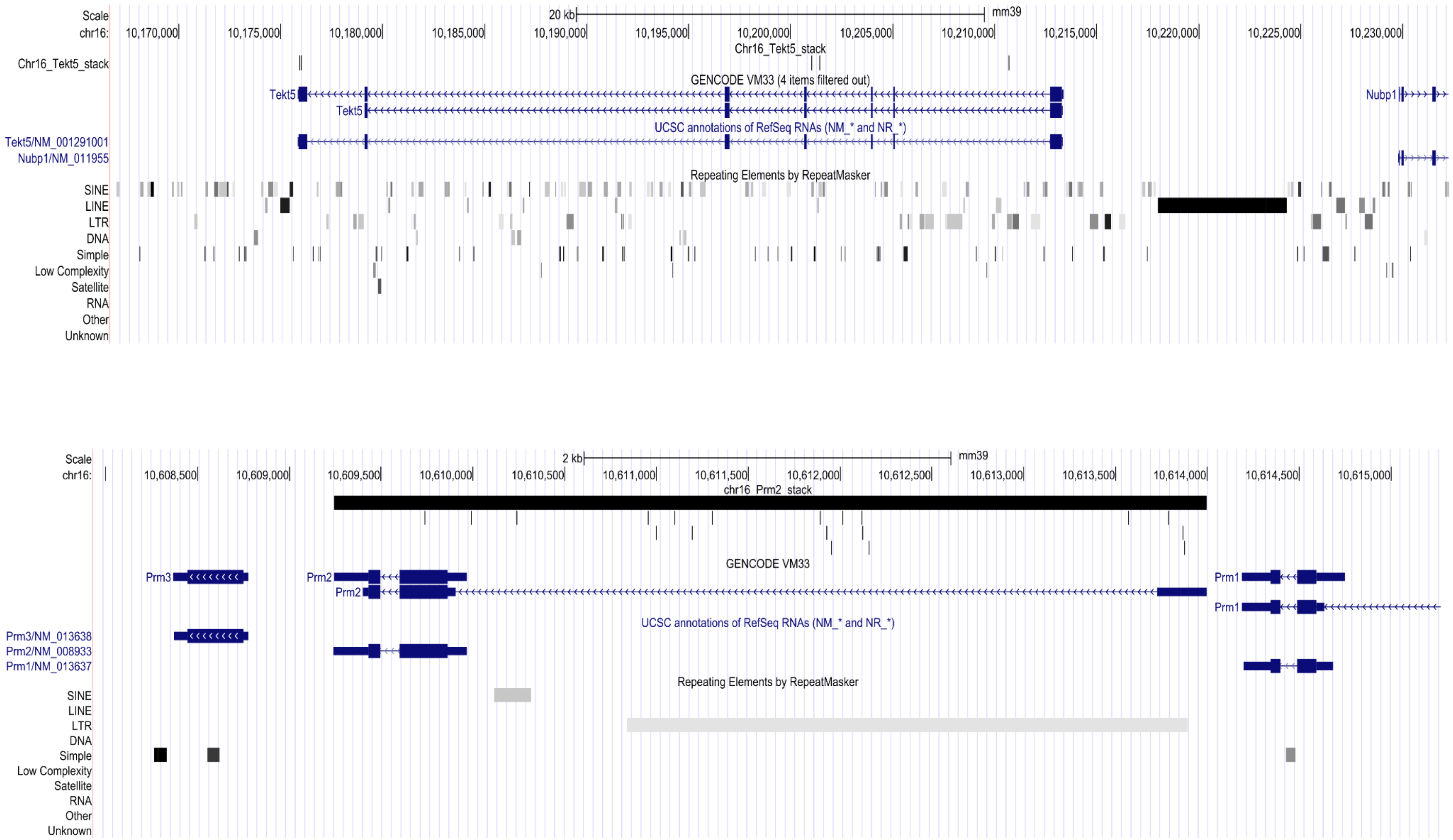


Supplementary Figure 8-15: UCSC genome browser plots of chr16 *Slx4* and *Dnaaf8* ASE stacks, related to Figure 5-5.

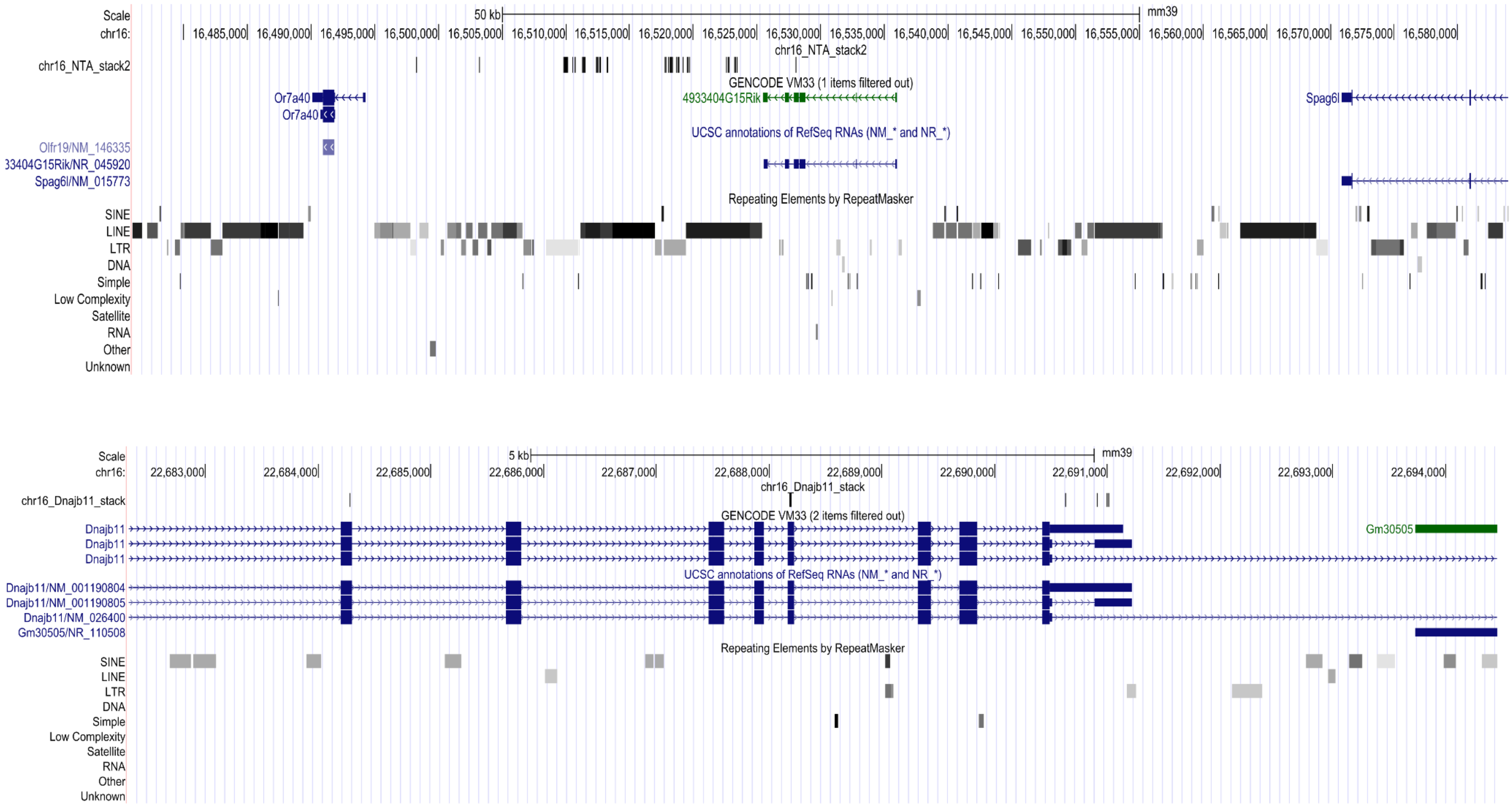




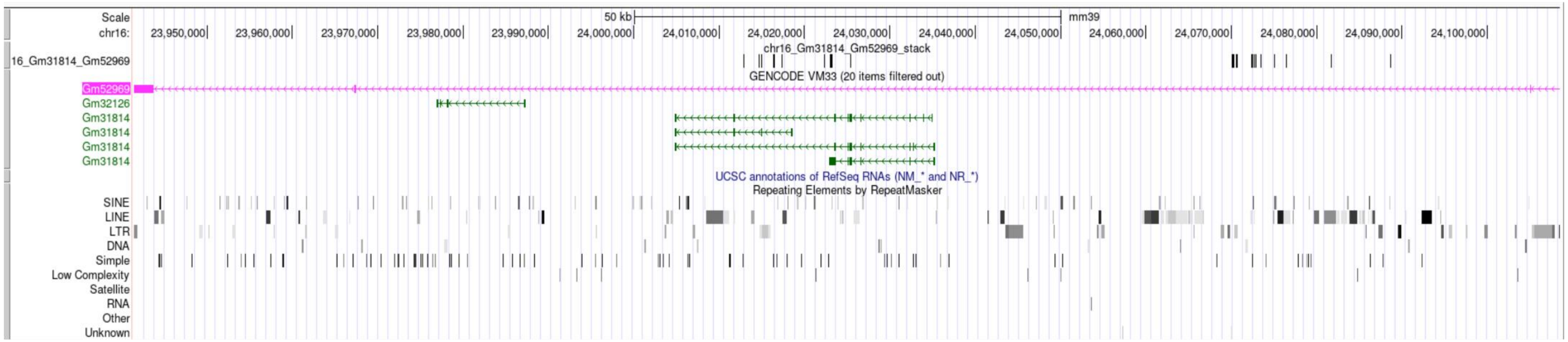
Supplementary Figure 8-16: UCSC genome browser plots of chr16 *Rbfox1* and a not-annotated ASE stack (NTA), related to Figure 5-5.



Supplementary Figure 8-17: UCSC genome browser plots of chr16, *Tekt5* and *Prm2* ASE stacks, related to Figure 5-5.



Supplementary Figure 8-18: UCSC genome browser plots of chr16, Not-annotated region (NTA) stack2 and Dnajb11 ASE stack, related to Figure 5-5.



Supplementary Figure 8-19: UCSC genome browser plots of chr16, *Gm52969* and *Gm31814*, related to Figure 5-5.