The use of novel bioinformatic approaches to understand the sex-linked vs autosomal contributions to mouse sperm development.

Eleanor Mary Watson

A thesis submitted for the degree of Doctor of Philosophy in Life Sciences

School of Life Sciences

University of Essex

January 2023

Impact of COVID-19

The COVID-19 pandemic unfortunately impacted upon this thesis in quite a significant way. The research for this thesis commenced October 2019, and in March 2020 the University suspended all face-to-face research. The original plan for this thesis was to be focusing on the imaging of different sperm cells and quantifying these into different categories that could potentially be linked to lifestyle factors. With the closure of the laboratories so early on in the project, only a small number of samples had been imaged due to project planning and training needs. Because it was unknown how long the labs would be closed for, it was decided to turn the project almost entirely on its head and focus mainly on bioinformatic analysis of sperm cell genomic data. This meant re-working the project aims, learning a whole new skillset and migrating to long-term remote study.

Acknowledgements

There are several people without whom this PhD would not have been possible. I would first and foremost like to thank my supervisor Ben Skinner, who took a chance on me as his very first PhD student and who has provided me with so much patience and guidance throughout the entirety of my research. Our lab group might have been small, but we're mighty.

Secondly, I would like to thank our wider research group over at Kent, specifically Darren Griffin and Peter Ellis, with whom we have collaborated on multiple projects, alongside drinking multiple pints and eating many a Friday night curry.

I am not sure whether I should be more grateful for starting the course, giving me the opportunity to meet Livy, or to Livy for keeping me sane throughout the course, but regardless I have met a friend for life. Thanks, Liv, for inviting me to lunch on our very first week where we had no idea what on Earth we were meant to be doing. Thanks for still inviting me for lunch now, when we still have no idea what on Earth we are meant to be doing – I love winging life with you.

Jo, Nora, Annabel and Lowri you have been more important than you'll ever know. Jo, I am so grateful to you for all our wonderful, albeit manic, dog walks. They brighten my week. Dr Nora, for being my very own PhD guiding light, my sounding board when I have any questions/moans/gripes, and always down for a gin or two. Lowri and Annabel, you don't know how much I look forward to our weekends together. They reset my batteries every single time and I hope we're still partying in trashy clubs together when we're 80. I love all four of you to the moon and back.

To my dad, for providing coffee breaks and beers, and for always supporting me in my dreams. To my mum, for giving me the best advice and always cheering me on from the side-lines, not just through my PhD but right through school and uni as well.

Finally, to my very own little family, Marcus, Nelly and Marnie. I thank my lucky stars for you guys every day. You don't always make my life easier (looking at you, puppies) but my goodness do you make it better. Marcus, you have been my absolute rock, from the evening I found out I had been accepted for the PhD position, to dropping me off on my very first day, to suffering all the highs and lows with me, the endless tears and stresses. 2023 is the year I become Dr Watson, but also the year I become Mrs Grainger, and I don't know which I'm more excited for.

Contents

List of Figures 13

List of Tables 18

Abbreviations 21

Abstract 24

1 Introduction 25

1.1 Importance of Sperm Development Research 26

1.1.1 Sperm development in agriculture 26

1.1.2 Current Status of Fertility in Humans 27

1.1.2.1 Current Fertility Options in Humans 29

1.2 General Sperm Biology 30

1.2.1 Sperm Structure 30

1.2.2 Sperm Production 31

1.3 Importance of sperm nuclear morphology 36

1.3.1 Advancements in quantifying sperm nuclear morphologies 36

1.3.2 Imaging techniques used in morphology analysis 39

1.4 Skinner's Nuclear Morphology Analysis Software 41

1.4.1 Nuclear morphology analysis methods 43

1.4.2 Methods of Edge Detection 45

1.4.3 Texture Analysis 48

1.5 Data Processing Methods 49

1.5.1 UMAP vs tSNE Dimensionality Reduction 49

1.6 Mouse hybrids and knockouts used in this study 53

1.6.1 Mouse subspecies 53

1.6.2 Male mouse hybrids show asymmetric sterility 53

1.6.3 Yq-del mice are mice with part of their Y chromosome long arm knocked out, and they show abnormal sperm nuclear morphology 55

1.6.4 Y-introgression is the process of combining the Y-chromosome from one subspecies onto the X-chromosome and autosomal background of another subspecies 57

1.6.5 *Sly* and *Slx* are sex-linked ampliconic master regulators in mouse 58

1.7 Mouse Sperm Development 60

1.7.1 Stages of Meiosis 62

1.7.1.1 Meiosis I 62

1.7.1.2 Meiosis II 63

1.7.1.3 Prophase I of Meiosis 63

1.7.2 Gene expression in spermatogenesis 64

1.7.3 Chromosome Synapsis and Transcriptional Silencing 68

1.7.4 Transcriptional Silencing 69

1.7.5 Gene Expression at Different Stages of Spermatogenesis 71

1.8 Genetic Analysis 71

1.8.1 Next Generation Genome Sequencing 71

1.8.1.1 RNA-Sequencing 72

1.8.1.2 Paired-end vs single-ended sequencing 73

1.8.1.3 Single-Cell Sequencing 73

1.8.2 RNA Sequencing Analysis Methods 74

1.8.2.1 FASTQC 75

1.8.2.2 GTF annotation files 76

1.8.2.3 STAR read aligner 77

1.8.2.4 featureCounts annotation tool 78

1.8.3 Differential Gene Expression Methods 78

1.8.3.1 SCDE is a differential gene expression analysis package designed for single cell data 79

1.8.4 Analysis of gene expression data 80

1.8.5 Motif analysis 80

1.8.6 Micro-RNA binding sites 83

1.8.7 Cis-limited genes 84

1.8.7.1 Mechanisms of Cis-Limitation 84

1.8.7.2 Known cis-limited genes 86

1.9 Rationale for this thesis 86

1.9.1 Overall aim of this research 88

1.9.2 Specific chapter aims 88

2 Materials and methods 89

2.1 Morphometric analysis of mice sperm nuclei 89

2.1.1 Mice sperm samples used 90

2.1.2 Mice sperm parameters assessed 92

2.1.3 Preparation of sperm samples for imaging 92

2.1.4 Microscopes and cameras used 93

2.1.5 Microscopy of the sperm samples 93

2.1.6 Checking concordance of cameras 95

2.1.7 Analysis of images with Nuclear Morphology Analysis 97

2.2 RNA Sequencing Analysis 98

2.2.1 Downloading FASTQ Files 98

2.2.2 FASTQC Scores and Trimming 99

2.2.3 BedTools was used to create a custom gene transfer format file (GTF file) that contains information about introns, as well as exons 100

2.2.4 STAR Analysis 101

2.2.5 Feature Counts 103

2.2.6 Analysing output in R to classify cells as X or Y bearing 103

2.3 Differential gene expression analysis 105

2.3.1 SCDE was used to generate a list of differentially expressed genes for X and Y bearing spermatids 105

2.2.2 Gene enrichment analysis pathways allow analysis of collective gene functions 107

2.2.3 The STREME package in Meme Suite (Bailey, 2021) was used for common motif identification in differentially expressed genes 107

2.2.4 Identifying the function of the enriched motifs using Tomtom 109

2.2.5 Identification of common microRNA binding sites using seedVicious (Marco, 2018) 109

2.2.6 Known cis-limited genes were analysed for common motifs 110

3 The use of high throughput nuclear morphology software to explore the impact of XY chromosomal incompatibilities and X-autosomal incompatibilities upon sperm nuclear morphologies and sterility across mouse strains 111

3.1 Introduction 111

3.1.1 Specific aims 112

3.2 Methodology 112

3.3 Results 115

3.3.1 Auto-exposure times were sufficient for robust analysis, both microscopes gave concordant results, and non-parametric tests were used 115

3.3.2 Significant differences in morphological parameters were seen across strains 118

3.3.3 Control hybrids provide a good substitute for the wild types whilst removing any negative morphological effect of inbreeding 120

3.3.4 F1 hybrids show distinct morphological abnormalities, similar to those seen in Yq-del sperm 123

3.3.5 Y-introgression hybrids are more similar in shape to the control hybrids, whilst F1 hybrids differ 126

3.3.6 Adding a Y-introgression rescue to the F1-hybrids does not negate the morphological impact 128

3.3.7 Control hybrids with Y-introgressions showed the same morphologies as Y-introgressions on a wild-type background 130

3.3.8 Genetic background is important in how X-Y or X-autosomal mismatch affects male reproductive phenotypes 131

3.3.9 There is a strong correlation between morphology of sperm and the sperm quality assessment parameters routinely used in clinic 133

3.4 Discussion 137

4 Developing an RNA Sequencing Analysis Pipeline which can Genotype Spermatids as either X Bearing or Y Bearing 141

4.1 Introduction 141

4.1.1 Rationale for Genotyping Spermatids and the need to identify cis-limited genes 142

4.1.2 The Importance of Introns for Genotyping 144

4.1.3 Specific aims 145

4.2 Materials and Methods 146

4.3 Results 146

4.3.1 Initial Results with SmartSeq2 Chen *et al* Data 146

4.3.1.1 STAR Mapping 146

4.3.1.2 FeatureCounts Analysis 147

4.3.1.3 Analysis of Number of Hits to Introns and Exons 149

4.3.1.4 Round Spermatids 157

4.3.1.5 Initial Conclusions from Chen Data 158

4.3.2 Results with 10X data 158

4.3.2.1 10X Pipeline Tweaks 159

4.3.2.2 Initial Dimensionality Reduction Results 159

4.3.3 Refining the Pipeline 162

4.3.3.1 Strandedness Corrections 162

4.3.3.2 Exon Spanning Regions 163

4.3.3.3 Revised GTF and Hermann Data 166

4.3.5 Final SmartSeq Results with full Bhutani dataset 173

4.4 Conclusion 178

5 Determining candidate cis-limited genes involved in sex ratio skew and sterility in mice, from differential gene expression analysis, gene ontology, enrichment analysis and genoinformativity 180

5.1 Introduction 180

5.1.1 Rationale 180

5.1.2 The novel genotyping pipeline allowed identification of 250 spermatids from SmartSeq2 data 181

5.1.3 The genotyping pipeline gave a roughly equal split of X bearing and Y bearing spermatids in the SmartSeq2 dataset 183

5.1.4 Specific Aims 184

5.2 Methods 184

5.3 Results 187

5.3.1 SCDE analysis produced lists of genes upregulated in both the X bearing and the Y bearing cells 187

5.3.2 Analysis of the autosomal differentially expressed genes 189

5.3.2.1 Gene ontology analysis of the autosomal DEGs 190

5.3.2.2 The presence of motifs in the autosomal DEGs 195

5.3.2.3 Motif analysis of just the long non-coding genes 200

5.3.2.4 There were no differences in microRNA binding sites between the autosomal protein coding DEGs from the X bearing vs Y bearing spermatids.....201

5.3.3 Key transcription factors were identified in common motif analysis of both the overall autosomal gene lists and the long non-coding genes 204

5.3.3.1 GO analysis of the putative transcription factors in X-bearing spermatids showed common functional pathways 205

5.3.3.2 Refining the candidate gene list 206

5.3.4 Analysis of the sex-linked differentially expressed genes 209

5.3.4.1 GO analysis of the sex-linked DEGs.....209

5.3.5 Refining the candidate cis-limited gene list 211

5.3.4.1 The candidate X-linked cis-limited genes are expressed to higher levels at the early-mid spermatogenesis stage.....213

5.3.4.2 The candidate cis-limited genes share common gene ontologies including chromatin binding 214

5.3.4.3 The levels of each candidate gene were plotted to show the change in expression over time, in both the X and Y bearing spermatids.....215

5.3.4.4 Expression levels of the candidate genes in each individual cell shows the distribution between the X bearing and Y bearing spermatids 217

5.3.4.5 comparing known cis-limited genes with the novel candidate cis-limited gene list 217

5.3.6 Biological significance of the candidate genes 218

5.3.6.1 *Gata1* 218

5.3.6.2 *Taf1*.....219

5.3.6.3 *Astx6* 220

5.3.6.4 *Platr21* 220

5.3.6.5 *Serpina7* 221

5.3.7 Gene signatures from the SmartSeq2 data cannot easily be used to genotype 10X data 222

5.4 Future Uses 228

6 General Discussion 230

6.1 Differences between hybrids and knockouts 230

6.2 Scope for improvements and limitations of the pipeline 232

6.3 Candidate cis-limited genes and their functions 234

6.4 Applications of the candidate cis-limited gene list 234

6.5 Applying this to 10X data 235

7 Conclusion and Evaluation 237

8 References 239

9 Appendices 263

9.1 Appendix 1: the cell IDs of the spermatids labelled as X or Y bearing from the Bhutani *et al.* dataset (Bhutani *et al.*, 2021) 263

9.2 Appendix 2: the motifs found in the differentially expressed genes and their candidate transcription factor targets 264

9.3 Appendix 3: Published peer-reviewed paper 271

9.4 Appendix 4: Published peer-reviewed paper 304

9.5 Appendix 5: Working title and co-authors for paper in preparation 358

List of Figures

Figure 1: Process of spermiogenesis..... 32

Figure 2: Image of a DAPI stained PWK mouse sperm nucleus..... 33

Figure 3: Comparison of consensus nuclei of human sperm from a cohort of fertile vs a cohort of infertile men..... 42

Figure 4: Distribution of telomeres and chromatin in human fertile and infertile cohorts..... 43

Figure 5: Representation of Laplacian edge detetction..... 45

Figure 6: Diagram of edge detection thresholds..... 47

Figure 7: Comparison of edge detection method..... 48

Figure 8: Comparison of fuzzy and not fuzzy connections in UMAP..... 52

Figure 9: Images of wild-type *mus* mouse sperm and Yqdel mouse with 2/3 deletion..... 52

Figure 10: Basic statistics for FASTQ files generated from FASTQC..... 75

Figure 11: Potential mechanisms of cis-limitation..... 85

Figure 12: A representation of the bounding height and bounding width of a nucleus..... 97

Figure 13: Basic workflow of RNA sequencing analysis..... 104

Figure 14: Dimensionality reduction plot showing comparison of samples imaged on the two different microscopes..... 115

Figure 15: The effect of exposure time on the area of 100 sperm nuclei..... 116

Figure 16: Basic morphologoical statistics for all strains of mouse imaged..... 120

Figure 17: Basic morphological statistics for the wild-type strains and the control crosses..... 121

Figure 18: Images of sperm..... 123

Figure 19: Basic morphological statistics for the control crosses and F1 hybrids .. 124

Figure 20: UMAP dimensionality reduction plot coloured by strain, for the control crosses and F1 hybrids 126

Figure 21: UMAP dimensionality reduction plot coloured by strain, for the control crosses, F1 hybrids, and wild-types with Y-introgression 127

Figure 22: UMAP dimensionality reduction plot coloured by strain, for the control crosses, F1 hybrids, wild-types with Y-introgression, and F1 hybrids with Y-introgression 129

Figure 23: UMAP dimensionality reduction plot coloured by strain, for the control crosses, F1 hybrids, wild-types with Y-introgression, F1 hybrids with Y-introgression, and control crosses with Y-introgression 130

Figure 24: Basic morphological statistics for the hybrid cross types and the introgression lines compared to the controls 132

Figure 25: Basic sperm quality assessment parameters for each cross type, data collected by collaborators at the University of Montana. 135

Figure 26: The relationship between bounding height and the sperm quality phenotype assessment parameters 136

Figure 27: The relationship between bounding width and the sperm quality phenotype assessment parameters 137

Figure 28: The percentage of each BAM file from the Chen *et al* dataset that was successfully annotated by FeatureCounts 147

Figure 29: The percentage of each BAM file from the Chen *et al* dataset that was successfully annotated by FeatureCounts, by developmental stage of the cells 148

Figure 30: The number of reads from the Chen *et al* dataset that map to introns as a percentage of all genomic features. 150

Figure 31: The number of hits from the Chen *et al* dataset to introns and exons, for each chromosome.....152

Figure 32: The hits to introns and exons for each gene in the Chen *et al* dataset, the hits to introns and exons excluding any of the genes that have 0 hits to introns or exons, and the hits to introns and exons excluding outlier gene ENSMUSG00000106106.....153

Figure 33: The number of genes from Chen *et al* dataset that had reads hitting to introns or exons.....154

Figure 34: The number of hits to introns and exons from the Chen *et al* dataset for each chromosome, excluding gene ENSMUSG00000106106.....155

Figure 35: The log2 ratios of the number of hits to introns in the X and Y chromosomes in each individual cell in the Chen *et al* dataset.....157

Figure 36: Comparison of the novel genotyping pipeline and dimensionality reduction tecniques used in this study vs the alternative pipeline and dimensionality reduction techniques used at University of Kent.....160

Figure 37: UMAPs produced using different versions of the Hermann *et al* data...161

Figure 38: Representation of exon spanning regions.....164

Figure 39: UMAPs produced using different versions of the Hermann *et al* data, using a refined pipeline.....167

Figure 40: UMAP dimensionality reduction performed in R on the Hermann *et al* data, coloured by marker gene expression levels.....168

Figure 41: UMAPs produced from the exonic read counts of the Hermann *et al* data, produced by the refined novel pipeline, and coloured by spermatid classification from this study vs the original study.....169

Figure 42: UMAP of Hermann *et al* dataset spermatids, coloured by levels of ZPBP (early spermatids), TNP1 (mid spermatids) and PRM2 (late spermatids) 170

Figure 43: Alternative genotyping calculations used to try and separate Hermann *et al* cells into X and Y bearing. 171

Figure 44: Intron:exon ratio for the X chromosome compared to the intron:exon ratio for the autosomes for the Hermann *et al* data 172

Figure 45: UMAPs of exonic read counts from the Bhutani *et al* dataset, coloured by normalised expression levels of 20 key genes. 173

Figure 46: UMAP of the full Bhutani *et al* dataset ,and the spermatids only 174

Figure 47: UMAP dimensionality reduction on just the spermatids, coloured by gene. Showing cluster one and two are the early spermatids, cluster three and four are the mid spermatids and clusters five and six are the late spermatids. 175

Figure 48: Genotyping calculation performed on the Bhutani *et al* spermatids 176

Figure 49: Intronic/exonic ratio for the X chromosome plotted against intronic/exonic ratio for the autosomes, for each different cluster in the Bhutani *et al* dataset. 177

Figure 50: Process by which *Sly/Slx* might cause differences between X and Y bearing spermatids 181

Figure 51: UMAP of all cells present in Bhutani *et al* dataset (323 cells), where 250 were classified as spermatids 181

Figure 52: UMAP of all spermatids present in Bhutani *et al* dataset, classified from early through to late spermatids. 182

Figure 53: Distribution of putative X bearing and putative Y bearing spermatids from the Bhutani *et al* dataset on the UMAP 183

Figure 54: Sample output chart of SCDE 187

Figure 55: Volcano plot showing the genes upregulated in X bearing spermatids and the genes upregulated in Y-bearing spermatids 188

Figure 56: Frequency distribution of the number of significantly enriched genes mapping to each of the chromosomes in the X and Y bearing spermatids 189

Figure 57: STRING analysis of the significantly enriched autosomal genes in X-bearing spermatids..... 190

Figure 58: STRING analysis of the significantly enriched autosomal genes in Y-bearing spermatids..... 193

Figure 59: Absolute read counts of the X and Y bearing spermatids for key genes across spermatid developmental stage 206

Figure 60: The log2 of the read counts of each potential transcription factor in X bearing and Y bearing spermatids. 207

Figure 61: Potential mechanisms of regulatory pathways that differentiate between X-bearing and Y-bearing spermatids 208

Figure 62: The read counts per cell of the candidate X-linked genes coloured onto the UMAP of the spermatids..... 213

Figure 63: Expression level of each candidate cis-limited gene for cells at different stages of development, by genotype. 215

Figure 64: Expression level of the revised candidate cis-limited genes for cells at different stages of development, by genotype..... 216

Figure 65: Expression levels per cell of the identified candidate cis-limited genes, coloured by X bearing or Y bearing spermatid 217

Figure 66: STRING analysis pathway of proteins encoded by the candidate X-linked cis-limited genes and their interactions. 218

List of Tables

Table 1: Comparison between t-SNE and UMAP methods for dimensionality reduction 51

Table 2: Copy number estimates for *Sly/Slx* in *mus* and *dom* mice (Larson *et al*., 2022). 59

Table 3: Description of the GTF file fields..... 76

Table 4: Symbols used in STREME motifs and their meanings..... 82

Table 5: Mice strains used in study..... 90

Table 6: Number of nuclei imaged per strain, number of samples per strain and location of imaging 94

Table 7: The p-values for the Shapiro-Wilk test of normality for each nuclear morphological parameter, per strain 118

Table 8: Median values for sperm nuclei morphological parameters, by strain, for the controls..... 123

Table 9: Median values for morphological parameters, by strain, for the controls and F1 hybrids..... 125

Table 10: Cell types and their corresponding developmental stages from the Chen *et al* dataset 149

Table 11: Number of hits to introns and exons, per chromosome, in the Chen *et al* dataset. 151

Table 12: Number of hits to introns and exons per chromosome in the Chen *et al* dataset, excluding gene ENSMUSG00000106106..... 156

Table 13: Adaptive thresholds for genotyping each cluster of spermatids..... 177

Table 14: The number of cells in each cluster, which represents developmental stages..... 184

Table 15: The biological processes, molecular functions and cellular components of proteins encoded by the autosomal genes upregulated in X bearing spermatids ... 192

Table 16: The biological processes, molecular functions and cellular components of proteins encoded by the autosomal genes upregulated in Y bearing spermatids ... 194

Table 17: Motif logos and percentage of genes containing the motif for motifs found in the genes upregulated in X-bearing spermatids ... 196

Table 18: Motif logos and percentage of genes containing the motif for motifs found in the genes upregulated in Y-bearing spermatids ... 196

Table 19: The novel motifs found in the genes upregulated in the X-bearing spermatids, and potential transcription factors that may bind these motifs ... 198

Table 20: The percentage of the DEGs upregulated in X and Y bearing spermatids with at least one miRNA binding site ... 202

Table 21: miRNA binding sites found in the genes differentially upregulated in the X bearing spermatids ... 203

Table 22: miRNA binding sites found in the genes differentially upregulated in the Y bearing spermatids ... 204

Table 23: Sex-linked genes upregulated in X-bearing and Y-bearing spermatids, and the roles of those genes ... 210

Table 24: Functional annotation chart of the protein domains found in the sex-linked genes upregulated in X-bearing and Y-bearing spermatids ... 211

Table 25: Functions of the sex-linked differentially expressed genes for X and Y bearing spermatids, and the sex-linked transcription factors that may bind the autosomal DEGs ... 211

Table 26: The number of X-bearing and Y-bearing cells falling into each batch which represents developmental stage ... 223

Table 27: Functional annotation chart for the autosomal genes upregulated in late-stage X-bearing and late-stage Y-bearing spermatids....226

Table 28: Ampliconic copy numbers of S*lx, Slxl1, Sstx, Sly, Ssty1* and *Ssty2* in *mus* and *dom* subspecies....231

Table S1: The cell IDs of the spermatids labelled as X bearing and the spermatids labelled as Y bearing....263

Table S2: The novel motifs found in the genes upregulated in the Y-bearing spermatids, and potential transcription factors that may bind these motifs....264

Table S3: Motif logos and percentage of genes containing the motif for motifs found in the transcription factors that may bind the genes upregulated in X-bearing spermatids....265

Table S4: Motif logos and percentage of genes containing the motif for motifs found in the transcription factors that may bind the genes upregulated in Y-bearing spermatids....266

Table S5: The novel motifs found in the candidate transcription factors that bind genes upregulated in X bearing spermatids, and potential second-level transcription factors that may bind these motifs....267

Table S6: The novel motifs found in the candidate transcription factors that bind genes upregulated in Y bearing spermatids, and potential second-level transcription factors that may bind these motifs....270

Abbreviations

ANOVA – analysis of variance

ART – assisted reproductive technology

BAM – binary alignment and map

BASH – Bourne Again Shell

BED – Browser Extensible Data

CASA – computer assisted sperm analysis

CASMA – computer assisted sperm morphometry analysis

cDNA – complimentary deoxyribonucleic acid

CDS -coding sequence

CERES – University of Essex's high performance computing cluster

DAPI – 4′,6-diamidino-2-phenylindole

DAVID – Database for Annotation, Visualization and Integrated Discovery

DEG – differentially expressed gene

DNA – Deoxyribonucleic acid

DSB – double stranded break

EBI – European Bioinformatics Institute

FDR – false discovery rate

FISH – fluorescence in situ hybridisation

GFP – green fluorescence protein

GIM – genoinformative marker

GLCM – grey-level occurrence matrix

GO – gene ontology

GTF – general transfer format

hn-RNA – heterogeneous nuclear ribonucleic acid

HPC – high performance cluster

ICSI – intracytoplasmic sperm injection

IVF in vitro fertilisation

KEGG – Kyoto Encyclopaedia of Genes and Genomes

LED – light emitting diode

MEME – multiple EM for motif elicitation

miRNA – micro-ribonucleic acid

MMP – maximal mappable prefixes

mRNA – messenger ribonucleic acid

MSCI – meiotic sex chromosome inactivation

NCBI – National Center for Biotechnology Information

PCA – princial components analysis

PCR – polymerase chain reaction

PFA – paraformaldehyde

RNA – ribonucleic acid

SAM – sequence alignment map

SCDE – single cell differential expression analysis tool

SNOM – scanning near-field optical microscopy

SRA – sequencing read archive

SRR – sequencing run

STAR – spliced transcripts alignment to a reference

STREME – simple, thorough, rapid, enriched motif elicitation

TIFF – tag image file format

tSNE – t-distributed stochastic neighbour embedding

UMAP – uniform manifold projection algorithm

UTR – untranslated regions

ZR – Zahn-Roskies

Abstract

Globally, populations are ageing, and birth rates are falling. This presents a variety of issues, including food shortages and a lack of working-age humans to sustain the older generations. This presents a global requirement for research into human and animal fertility: animal research to help with global food security allowing a larger population to be fed more effectively, and human research to help tackle the declining fertility rates to ensure that the population continues to grow at a level that can support the larger number of elderly people.

The area of fertility research that needs to be investigated most is that of sperm development. There is traditionally a lot more female-factor research, leaving a knowledge gap for male-factor research. Sperm development in mice has been shown to be directly linked to successful fertilisation outcomes: mouse sperm with perturbed development do not produce viable embryos. In some cases, this has been shown to be a sex-linked effect, whereby female mouse embryos are produced, but male mouse embryos are not. This study also shows that autosomal genes are critical for proper mouse sperm development, and so differential gene expression analysis methods are used to determine which genes have a greater impact: autosomal, or sex-linked, and to determine which specific genes may be the most implicated in poor mouse sperm development. The list of candidate genes found in this study to potentially be involved in poor mouse sperm development provides possible targets for gene therapy to improve fertilisation outcomes, and orthologues of these genes can be investigated in agricultural species such as pigs and cows, as well as humans.

1 Introduction

The introduction to this project, the use of novel bioinformatic approaches to understand the sex-linked vs autosomal contributions to mouse sperm development, consists of a structured literature review of the current knowledge surrounding the research that will be carried out. In this study, morphological and genomic approaches are used to study mouse sperm quality, which has significant future impact for both agriculture and healthcare.

This literature review starts by looking at the wider picture: the current status of fertility in both humans and animals, why we should care about improving it, and the current role of sperm and genetics research in agriculture and healthcare. This review will then look at current imaging analysis strategies, and sequencing techniques, as these areas of analysis are the main approaches used in this research project to assess mouse sperm quality and development.

Sperm development research for improving agriculture is of growing importance in a world in which food sustainability is vital, and sperm development research is also crucial for clinical reasons, in a world where human fertility levels are rapidly falling, and the current major approaches to treating infertility are lacking in efficiency. In vitro fertilisation (IVF) is costly, invasive, and has a low success rate. The high cost of current artificial reproductive technologies (ART) acts as a financial gatekeeper; IVF is not accessible to everyone, and a greater understanding of sperm development will help on the pathway to making ART accessible to all. Sperm development research can also be used for optimising agricultural yields. The aim of this review is to fully justify the need for research into sperm development.

The research questions that this thesis is aiming to answer are:

- “How is mouse sperm morphology currently measured, and how can it be better measured?”
- “How does mouse sperm morphology correlate with mouse sperm quality?”
- “Which are the key genes implicated in normal and abnormal morphology of mouse sperm, what are their functions, and can we target them in future?”
- “How can we apply these findings to other species, in order to increase agricultural yield, and human fertility levels?”

1.1 Importance of Sperm Development Research

Sperm is a fundamental part of life as we know it. Without sperm, there would be no sexual reproduction, and without sexual reproduction, humanity could not continue procreating and there would be no production of animals for food to sustain the human race. However, what happens when sperm exist, but are of declining quality? What are the implications and hallmarks of poor sperm, and why should this be investigated?

1.1.1 Sperm development in agriculture

Global demand for animal and plant food products is increasing at an ever-growing pace. Food sustainability has become one of the most important worldwide challenges, and more innovative methods are needed to ensure that the world’s population will not go hungry. Genomics is a tool that can be used to help this: the most desirable traits in species of animals can be identified and the genes leading to these phenotypes (Aryee *et al.*, 2021). With robust analysis of the morphology of sperm, as well as being able to analyse the sub-nuclear localisations of organelles

within the sperm, there is the potential to categorise sperm abnormalities, find treatments for different abnormalities, and use this to screen sperm from livestock in order to make farming methods more efficient, for example, allowing the breeding of transgenic pigs by screening the sperm to ensure it is the transgene-containing sperm (Bou *et al.*, 2014). Previous studies have shown that the mouse Y chromosome has an abnormally high level of gene amplification affecting genes expressed in spermatids, attributed to continuous conflict between X and Y chromosomes for control of sex ratio. Partial deletions of the Yq (long arm of the Y chromosome) have been shown to lead to sperm head malformation as well as DNA damage and a distorted sex ratio skew in favour of females. In some animal species, production of female offspring is more valuable than production of males in terms of agricultural efficacy. This is an example of how furthering knowledge of the sub-nuclear arrangement of sperm could prove invaluable in terms of sex selection in livestock for farming (Rathje *et al.*, 2019).

In addition to this, cryopreservation of animal sperm in agriculture is a widely used tool, especially bull sperm, whereby semen from some of the most high-quality breeding bulls is harvested, cryopreserved and then shipped worldwide to inseminate cows globally. Bettering techniques used to study sperm is necessary to improve the viability of sperm after thawing to ensure food security.

1.1.2 Current Status of Fertility in Humans

Human sperm morphology is notoriously inferior compared with many other species (Kumar and Singh, 2015), and it has long been accepted that abnormal sperm morphology is linked to subfertility in human males – however there is dispute over

what constitutes ‘abnormal’ sperm morphology, and there is not yet a reliable automated method of obtaining morphometric human sperm data, such as nuclear size and shape. There have been many studies attributing abnormal genomic arrangements to male-factor infertility, but further work is needed to determine the role that size, density, texture and shape abnormalities in sperm nuclei plays in male-factor infertility (Ioannou and Griffin, 2011).

Some definitions to mention here are infertility and sub-fertility. Definitions vary between sources, but for the purpose of this study, infertility is the inability to conceive a child within one year of starting to try and conceive, and sub-fertility is used for similar but slightly less severe cases; whereby couples struggle more than the average person to conceive, but who are maybe not clinically defined as fully infertile. When mouse sperm quality is assessed in this review, ‘sterile’ is used. For the purpose of this project sterility is synonymous with infertility – whereby offspring cannot be conceived.

The exact causes of rising global sub-fertility in humans remain unknown, but there is speculation there may be a link between mobile phone/Wi-Fi usage and sub-fertility; though there is only limited evidence to support this statement (Yildirim *et al.*, 2015). Alongside technological changes, diet trends and fashion (such as tighter clothing) may also play a role. High protein diets are considered ‘optimal’ for many men with an interest in ‘health and fitness’ due to the effect of muscular gain, but some rodent and avian studies have shown this can negatively impact upon sperm (Murugesan and Prakash, 2016).

Human infertility is a complex health challenge that affects rising numbers of couples around the world. Not only is it disappointing for the couple, it can cause social

problems due to the importance society places on becoming a parent (Yazdani, Kazemi and Ureizi-Samani, 2016), and financial problems due to counselling that may be required for the couple, as well as the hugely expensive assisted reproductive therapies that are currently available which have low success rates.

There is also the argument to say that lower fertility levels lead to an ageing population, which is associated with many negative economic consequences; fewer working age people who are less capable of supporting the larger number of societal dependents. The combination of these effects can place massive strain on health care providers, as well as governments, and because of this, infertility can be considered a major global problem to tackle.

Infertility is estimated to affect over 50 million couples worldwide (Ray *et al*., 2017). In a 2016 paper, Shenk *et al* described the link between human fertility and the demographic transition, suggesting that there is evidence to support the idea of fertility decisions being affected by access to resources, access to technology, risk of mortality, and social status. (Shenk, Kaplan and Hooper, 2016). The study offers hypotheses for the observed levels of fertility decline; one proposed reason being that financial investment in 'status' means less financial availability to support a larger family, therefore leading to fewer children.

1.1.2.1 Current Fertility Options in Humans

An overwhelming feature of the current literature is the fact that human fertility treatments are heavily female-orientated. In 2016, it was estimated there were five million humans alive who were conceived using assisted reproductive techniques (ART) – such as IVF (in vitro fertilisation) or ICSI (intracytoplasmic sperm injection).

This shows that ARTs can be used extremely successfully, however in almost half of all cases, ARTs fail to achieve a successful pregnancy (Ray *et al.*, 2017). It is also described as being a 'mainly palliative' process – with the outcome being to create a pregnancy, rather than to understand what the specific problem within the couple is (Ray *et al.*, 2017). One of the key clinical disadvantages to IVF and ICSI procedures is that they often result in multiple pregnancies, with the multiple pregnancy rate being around 30% higher in IVF/ICSI conceptions than after natural conception. This brings higher risk deliveries, lower birth weights and greater risks to the mother of complications such as pre-eclampsia and gestational diabetes, haemorrhage and premature delivery (Elster *et al.*, 2000). Single embryo transfer is now becoming more commonplace, rather than multiple embryo transfer, in assisted reproductive techniques, which has been shown to reduce multiple pregnancy rate(Gerris, 2005). However, it is likely that this in turn reduces the overall success rate of ARTs, and not all patients are eligible to undergo single embryo transfer. It is clear that current assisted reproductive techniques are still sub-optimal in the battle against fertility problems. This demonstrates the clear need for further research into more efficient and safer future methods. A focus on healthier sperm may help in the process of selecting sperm for use in ARTs, which would mean single embryo transfer may become more successful.

1.2 General Sperm Biology

1.2.1 Sperm Structure

Sperm cells consist of the sperm head, mid-piece (body) and the tail. The head contains the nucleus, and the nucleus takes up most of the sperm head. Sperm

nucleus shape therefore influences sperm head shape, and is the main factor considered in this study. The nucleus contains the chromosomes (in humans, this is 23 and in mice, 21), packed tightly as chromatin. The sperm head also contains the acrosomal region, which is essential for the acrosome reaction, whereby the sperm penetrates the outer layer of the egg in order to fuse with the egg plasma membrane (Gu *et al.*, 2019).

The midpiece of the sperm cell is roughly the same size as the head, and contains mitochondria, very tightly packed, in order to provide the vast amounts of energy required for the cell to swim.

The tail of the sperm cell is the largest part, making up around 80% of total length of the sperm. It is long and thin and is made up of various parts. The connecting piece attaches the tail to the midpiece of the sperm. The principal piece and end piece of the sperm cell are required to provide the propulsion motion required for movement of the sperm.

1.2.2 Sperm Production

Spermatogenesis is the formation of sperm cells from the germinal epithelium of males, where four functional spermatozoa are produced. The diploid germ cells from the germinal epithelium become the four haploid gametes. It occurs in the seminiferous tubules in the testes.

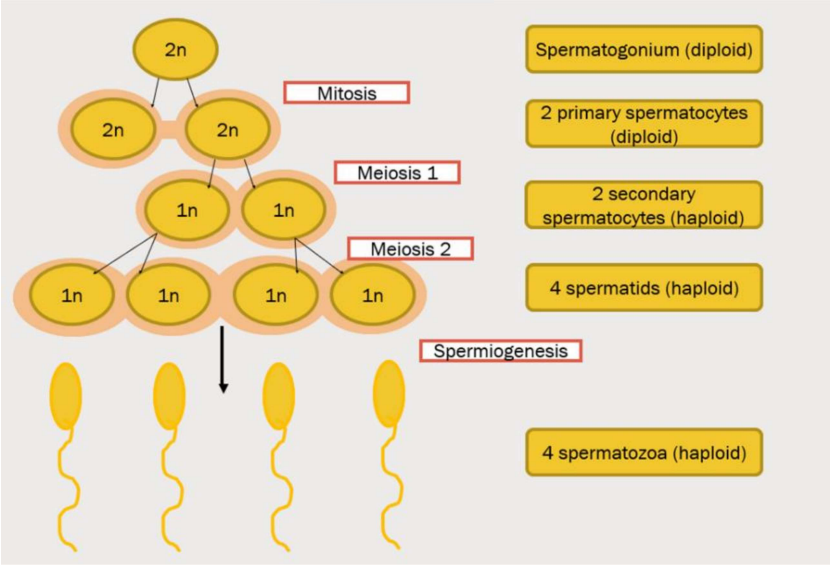
The first stage of spermatogenesis is the mitosis of spermatogonia, where cells divide, and form primary spermatocytes(Panawala, 2017). This is followed by meiosis 1, to form secondary spermatocytes, and then meiosis 2 to form spermatids. The spermatids then mature in order to produce motile spermatozoa. This process

produces four functional spermatozoa from one spermatogonium. The main phases involved in spermatogenesis are multiplication, growth, maturation and differentiation.

Spermiogenesis is the final stage of spermatogenesis, the process whereby the spermatids mature and differentiate into spermatozoa.

The sperm cells develop in a syncytium, which allows the sharing of gene transcripts between haploid spermatids, until spermiogenesis where the cytoplasmic bridges are broken.

Figure 1

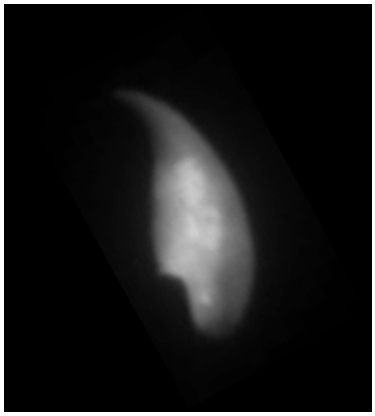


The process of sperm development, from a single diploid spermatogonium into four haploid spermatozoa

1.2.3 Sperm Morphology in Various Animals

Sperm morphology varies between different species: mature sperm are the most diverse shaped animal cell type (Kopania *et al.*, 2022). Genes involved in spermatogenesis evolve rapidly, and there is intense selection pressure on sperm function. Since the sperm nucleus makes up almost the entirety of the sperm head, the nuclear shape is a large determinant of overall sperm shape. Pigs and humans are examples of species which have spatulate shaped sperm nuclei, whilst rodent sperm nuclei have a falciform hooked shape with two axes of symmetry. The function of the hook in unknown: there are multiple theories including that the hooks allow trains of sperm to form and swim to the egg more successfully, but none of these theories have been proven to be correct. DAPI is a stain used to stain the nucleus of a cell, so can be used to visualise the sperm nuclei, and therefore overall sperm head shape.

Figure 2



DAPI stained PWK mouse sperm nucleus

Because of the specific axes and regions in sperm nuclei, in theory, the locations of FISH (fluorescence in-situ hybridisation) signals can be localised, mapped and compared between samples of the same species. It has been shown that human males with fertility problems tend to have more sporadic chromosome territories than healthy human males (Finch *et al.*, 2008); the position of chromosomes in sperm is an important parameter to research. Manually locating the position, however, is both subjective and takes a long time, and therefore there is a limit to how many nuclei

can be analysed with manual location of chromosomal position. Sequencing techniques, for example, Hi-C, can produce 3D maps of the chromatin structure of nuclei (Battulin *et al.*, 2015), but it is also necessary to use 3D imaging techniques to analyse the chromatin structure, such as by using microscopy. Nuclear morphology analysis software (Skinner *et al.*, 2019a) allows for automatic landmark detection which can greatly assist the process of locating chromatin inside sperm nuclei.

In humans, male-factor infertility is believed to be multifactorial in the majority of cases, caused by both genetic factors and external factors. The majority of human male sub-fertility cases are due to a low sperm count, however, many of the remaining cases are caused by qualitative defects such as abnormal morphology, preventing the sperm from being able to fertilise the egg as it should. There have been several common morphological defects that have been defined and characterised, and during this process several gene defects have been identified as the cause for poor morphology. The most severe cases of male infertility are more likely to be predominantly of genetic origin; nine human genes have so far been directly linked with qualitative sperm defects (Ray *et al.*, 2017) and there are many more that have not yet been identified, in both humans and other species.

It has been found that the most severe shape abnormalities are caused by single gene defects leading to sperm head shape abnormalities (Ray *et al.*, 2017). For example, a mutation in aurora kinase C (AURKC) leading to a truncated transcript means the transcript gets degraded. This leads to the production of large-headed and polyploid spermatozoa, leading to sub-fertility. Sperm heads are responsible for spermiogenesis, the acrosome reaction and flagellum biogenesis – therefore any abnormality in the sperm head could have extreme consequences for the sperm's function. This provides a strong rationale for studying sperm nucleus morphology.

Following insemination, sperm must be transported to the egg and subsequently fertilise it, in order to successfully create an embryo and lead to a pregnancy. There is still much research to be done to elucidate precisely how sperm shape influences transport and fertilisation. As described previously, it is known that the sperm heads are primarily involved in the acrosome reaction, and it is also known that the flagella influence transport (García-Vázquez *et al.*, 2016). Sperm encounter uterine fluid, which, in many species including mouse, is one of the first 'barriers' to fertilisation – the uterine fluid can be toxic to sperm in the absence of the correct proteins in seminal fluid. Also, within the uterus, sperm come into contact with a variety of cells, which may or may not be able to distinguish between normal and abnormal morphologies. It is possible that if cells in the uterus cannot interact with morphologically abnormal sperm, this may be one of the pathways hindering abnormally shaped sperm from achieving fertilisation.

Furthermore, it has been shown that some damaged sperm gets eliminated by phagocytosis in the uterus. It is entirely possible that phagocytes can detect abnormal morphologies, as well as damaged sperm, or even that phagocytes mistake abnormally shaped sperm as damaged. The uterus is the more hostile environment for the sperm; once in the oviducts, the environment is much better at promoting sperm survival. However, even in the oviduct there is evidence to suggest that morphologically normal sperm, and sperm with normal chromatin distributions and densities, are favoured, with studies in pigs showing preferential binding of pig oviduct epithelial cells to morphologically normal boar sperm, over morphologically 'abnormal' sperm. This binding to oviduct epithelial cells facilitates sperm transport along the oviduct (García-Vázquez *et al.*, 2015). Although it is only morphologically normal sperm that are able to bind to epithelia in the oviduct, there is still further

selection that occurs in the female genital tract; not all sperm that bind end up fertilising – known as “cryptic female selection” (Birkhead, 1998).

1.3 Importance of sperm nuclear morphology

1.3.1 Advancements in quantifying sperm nuclear morphologies

Traditional measures of general sperm morphology have disadvantages; measurements of the organisation and morphology of the nuclei in sperm is complex, since sperm nuclei are normally asymmetric in shape across the majority of species. Therefore, historically there have been difficulties in rapidly analysing morphologies of these asymmetric shapes, with many methods being manual and subject to investigator bias. Finch *et al.* used shell analysis, which was an automated approach (Finch *et al.*, 2008), but Skinner *et al.* developed a novel approach to automate the process using a nuclear morphology analysis tool which could be used in conjunction with fluorescence in-situ hybridisation (FISH) signals to locate chromosome position (Skinner *et al.*, 2019a). The analysis software uses automatic landmark detection to locate equivalent structures in each nucleus, and then the FISH signals are warped into a common shape to show the composite signal of the entire cell population.

Back in the 1990s, it was hypothesised that a stain could be used in order to assess various parameters of sperm in clinical tests. Eggert-Kruse et al describe how Acridine Orange was used in the screening process of sperm in 103 males from sub-fertile couples. (Eggert-Kruse *et al.*, 1996) However, the stain was not shown to produce any useful results; with the results unrelated to any of the important parameters of sperm analysis, such as motility, morphology, and pH. The staining

patterns observed were also not related to sperm function, the medical history of the patient, and there was no difference in males with explained vs unexplained fertility, or with those who later went on to become fertile. This meant the acridine orange stain test was deemed not clinically useful as a screening tool for sperm analysis (Eggert-Kruse *et al*., 1996). Other methods for clinical sperm screening in ARTs were required. The 'swim-up method' was first described in 1984 (Mahadevan and Baker, 1984) and is the oldest, most common method for sperm separation and screening. It is deemed to be an "easy, reliable and effective" method (Jameel, 2008). However, there are a few reasons for why the swim-up method is not perfect.

Clinically, the 'swim-up method', an assay for sperm motility, is the method used to select what is considered to be the ideal sperm for taking part in assisted reproductive procedures (Xie *et al*., 2010). However, the swim-up method is simply a tool for selecting which sperm are fastest and best at swimming, rather than a measure of any other parameters, such as shape, ability to be attracted to the egg (chemotaxis), chromosomal arrangements and many others. Xie *et al*. (2010) designed a novel method to monitor the chemotaxis of sperm as well as sperm motility, by developing a microdevice composed of a channel that mimics the mammalian female reproductive tract. This was the first device able to monitor both chemotaxis and motility. However, even this device – an improvement on traditional screening techniques due to its ability to monitor chemotaxis – is not able to give any information on the shape of the sperm (Xie *et al*., 2010).

In addition to the fact that the swim-up method does not give as much information as is required for significant improvements in sperm health, it could also be argued that it may potentially cause harm to the sperm during the centrifugation step of the process. One study tested this, by using semen samples from patients undergoing

ARTs and splitting each sample into two parts; using one part for standard swim-up analysis (involving centrifugation) and using the other for swim-up analysis modified to not include centrifugation. Whilst in this particular study, there was not a significant difference in DNA damage of sperm from the centrifugation vs non-centrifugation methods, there was still less DNA fragmentation seen in the non-centrifuged samples. The study concluded that the centrifugation did not cause significant damage (Younglai *et al.*, 2001). However, this was just one, fairly small (only 16 sperm samples) study and it is possible that the higher levels of DNA fragmentation observed could be doing more damage than this study suggests. This is another reason for looking for alternatives to traditional screening methods.

The diversity in sperm cell morphologies is hypothesised to mirror the diversity in sperm function (Maroto-Morales *et al.*, 2016). However, how diversity in sperm came about during speciation, and the different roles in sperm function that diversity plays is unknown. Selective pressure forces during adaptation to different environments have played a role in influencing sperm morphology, length and sperm count. Computer assisted sperm analysis (CASA), fluorescence probes, flow cytometry and chromatin assessment have facilitated less subjective and more efficient analysis of sperm. Computer assisted sperm morphometric analysis (CASA-morph) has been used to demonstrate the relationships between sperm morphology and male fertility, and sperm freezing ability. CASA-morph has become the most popular technique to assess sperm morphology, due to the increased reliability and reduced subjectivity. However, there are still areas of variation such as sample preparation, fixation method, staining, microscopes used, and the technician. These are all areas still to be resolved; the assessment of sperm even with CASA-morph is not ideal, with results varying across different laboratories. Higher resolution optics and imaging

analysis programs are required to analyse specific sperm structures, such as the nucleus.

1.3.2 Imaging techniques used in morphology analysis

As described previously, there are many possible ways in which sperm morphology could be analysed and assessed, and this project aims to make use of a tool that has taken the most effective methods and put them together to automate and speed up the screening and analysis process.

The paper “the application of scanning near field optical imaging to the study of human sperm morphology” (Andolfi *et al.*, 2015) focuses on scanning near-field optical microscopy (SNOM) as a method of visualising the shape and structure of sperm components with a higher resolution than conventional optical microscopy. The study found that SNOM could analyse morphology in both healthy and ‘abnormal’ sperm, and that head shape, midpiece shape and tail shape, as well as cytoplasmic structures, the acrosome, the nucleus and the mitochondria, were able to be observed, without requiring staining or invasive protocol that may be damaging to the sperm. The study cites that several morphological studies previously occurred in order to define ‘normal’ shape of sperm but argues that the methods of characterising ‘normal’ are left open to interpretation by the technician analysing, as well as the staining procedures that are used.

Conventional optical microscopy is cited as the “most common technique to assess sperm morphology”(Andolfi *et al.*, 2015) – likely due to the easy sample preparation, but it has limited resolution. Their paper describes how in the 1980/90s, computer-

assisted sperm morphometry analysis (CASMA) was developed to use alongside optical microscopy (Moruzzi *et al.*, 1988), and this was later improved upon by using fluorescence microscopy rather than optical microscopy. The computerised process allowed quantification and classification of human sperm morphology in a reduced time, where sperm heads were imaged through a microscope and measured on an image analysing system, to measure stain content, length, width, and perimeter. The study describes that, more recently, scanning and transmission electron microscopies have been used to analyse the morphology as sperm, which provide higher resolutions, but which require complex preparation.

Atomic force microscopy has more recently allowed higher resolution analysis without the need for complex preparation, and can be used with fixed or living cells (Kumar *et al.*, 2005). Scanning near-field optical microscopy (SNOM) works similarly to atomic force microscopy, with a probe scanning the sample surface. Andolfi et al describe the many advantages than SNOM has over other techniques – the main ones being the fact that the process is non-destructive and inexpensive (Andolfi *et al.*, 2015). However, a limitation to SNOM as an imaging technique is the fact that the cells must still be purified, fixed and dried. Whilst this does not cause cell damage which could lead to inaccurate results, it still means that there is, albeit significantly less than with other techniques, a reasonable amount of preparation to be done. This is a limitation that the nuclear morphology analysis tool (Skinner, *et al.*, 2019) has overcome.

Another disadvantage to past methods of analysing sperm head morphology with computer-aided image analysis is the fact that these methods rely on area, perimeter, length and width measurements which typically reflect size rather than shape and do not therefore give enough information regarding the variation in sperm

head shape (Ostermeier *et al.*, 2001). However, Ostermeier *et al.* (2001) used the Fourier method to classify bovine sperm shape and found it to be a repeatable and accurate method, which differs from any other previous methods in that it produces continuous data; more effective at explaining shape distributions, and it also an objective method of describing shape rather than previous methods that were subjective in terms of the human analyst deciding 'normal/abnormal'. Fourier harmonic shape analysis can evaluate perimeter curvature of shapes and has been used successfully in sand grains, murine sperm and epithelial cells, amongst others.

However, Fourier analysis is deemed to have a complicated mathematical background, and therefore can be considered difficult to interpret and conceptualise for non-mathematicians, which is why it is still not a widely used method. However, the objectivity of the Fourier method provides scope for the future identification of "novel sperm head shapes, distribution characteristics, or a combination of these, that affect male fertility" (Ostermeier *et al.*, 2001).

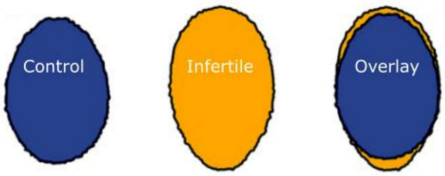
Another limitation of Fourier analysis, described by Skinner *et al.* (2019), is that it relies on smooth curves of shapes, rather than sharp points, which can prove a problem with large scale sperm image analysis due to the presence of sharp points in the outlines of sperm nuclei, especially in mice, which have hook shaped nuclei. It is clear to see that each novel method improves upon previous imaging techniques, finding solutions to some previous problems, but even the most recent techniques are not yet perfect.

1.4 Skinner's Nuclear Morphology Analysis Software

Novel software for analysing nuclear morphology was developed (Skinner *et al.*, 2019; Skinner, 2022) as a new and alternative method of analysing sperm nucleus

shape in rodents, specifically mice, and the software has been used to show functional differences in the swimming ability of mouse sperm linked to morphology of the mouse nuclei (Rathje *et al.*, 2019). Whilst this novel software has been optimally designed to make the link between morphology and functionality in rodents and pigs, it was also used in humans at the very start of this project (Turner *et al.*, 2021).

Figure 3

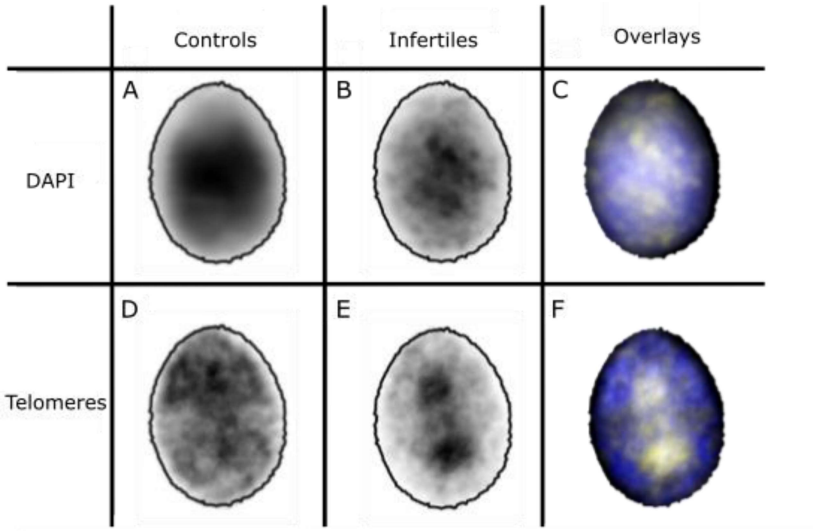


Consensus nuclei of human sperm from a cohort of fertile and infertile men from a fertility clinic, showing subtle yet detectable differences

As described earlier, there is a clear need to further investigate the morphology and function link in humans, not just livestock and rodents, due to the growing problem of human subfertility. Subtle differences in human sperm nuclear morphology (Figure 3) can be detected using the tool (Turner *et al.*, 2021) even though the tool is not optimised for humans, which provides great scope for manipulating the software to apply to a wider range of organisms.

Not only can the tool be used to demonstrate human sperm morphological abnormalities, but it was also used for showing genomic arrangements, and showed there is a correlation between the two (Turner *et al.*, 2021), as shown in Figure 4.

Figure 4



Telomere and DAPI (chromatin) distributions in fertile and infertile cohorts, showing distinct differences in chromatin and telomere distribution between fertile and infertile human males. The overlays show the fertile chromatin distributions in blue and the infertile chromatin distributions in yellow

1.4.1 Nuclear morphology analysis methods

Nuclear Morphology Analysis utilises many different techniques described here to produce a novel automated method of screening, imaging and detecting sperm nuclei from a sample. The software is freely available as an Image-J plugin, and it will automatically detect and analyse nuclei that have been captured as images from a microscope.

The software starts with pre-processing of the image. For this, the Kuwahara filter is used (Skinner *et al.*, 2019) which is a non-linear filter that removes background noise but retains the edges of the image. Because the edges are preserved, it is useful for feature extraction and segmentation and is often used in medical imaging. This helps with flattening of the centromeres (Skinner *et al.*, 2019).

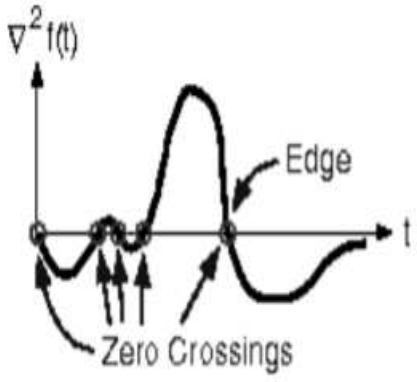
Objects are detected using the Canny edge detection algorithm (Skinner *et al.*, 2019), which is more effective than fixed-threshold detection. All nuclei are aligned, and landmarks are identified on each nucleus, giving rise to a linear trace, which is the angle profile. This angle profile is generated using a modification of the Zahn-Roskies transform (ZR). The window size is 0.05 of the perimeter of the nucleus, so from each starting position there are 20 semi-landmarks before returning to the start. However, multiple overlapping possible ZR profiles are produced, due to the presence of many more than 20 points around the perimeter. The combination of the overlapping ZR profiles gives rise to the angle profile trace produced by the software, giving rise to a much higher resolution of shape, especially in the more detailed areas of the sperm nucleus. Once each angle profile has been generated from each nucleus, they are all aligned against each other, and a median angle profile generated. The median profile is split into different segments for ease of analysis, at local minima and maxima, which allows landmarks at convex and concave areas of the nuclei to be automatically defined, which correspond to corners in the object outline. This allows the landmark locations in each individual nucleus to be identified, by the best fit of the median nucleus profile to the individual profile at each landmark.

Edge detection is a vital part of image processing, the first step involved in detecting objects within images, and there are a number of methods for doing so. Edge detection typically relies on a gradient vector, detecting a difference in intensities between pixels. In general, the changes in pixel intensities define the edges of an object. However, there are some cases where background noise interferes and ‘edges’ can be detected that don’t exist. This is something to be taken into consideration when developing an automated screening tool, which ideally will remove any background noise.

1.4.2 Methods of Edge Detection

There are 2 general categories of edge detection – gradient detection or Laplacian. The Laplacian detection finds 'zero crossings' in the second derivative of the image, whilst the gradient method finds the maximum and minimum pixel magnitudes in the first derivative of the image. Compared with first derivative gradient methods such as the Sobel operator method, the Laplacian second derivative method tends to produce better edge localisation (Wang, 2007).

Figure 5



Laplacian edge detection, showing the second derivative of the image, where the second derivative crosses zero is considered to be the edges. Source: (Shrivakshan and Chandrasekar, 2012).

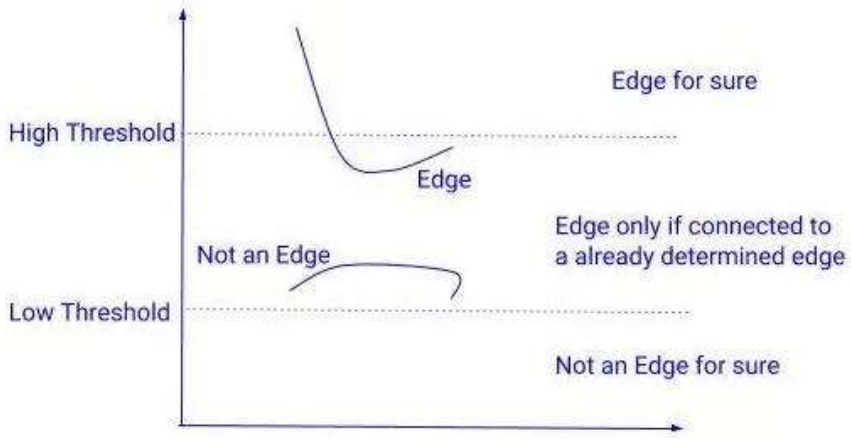
Classical methods of edge detection include the Sobel and Prewitt methods, which have the overarching advantage of being very simple. However, they are very sensitive to noise, which leaves them inaccurate. Zero crossing methods are good for detecting edges and orientations and are fairly simple. Zero crossing methods include Laplacian, which are also very sensitive to noise.

An edge pixel has a corresponding gradient vector, where the pixel intensity increases in the direction of the gradient. The greater the magnitude of the gradient, the stronger the edge. If there is no edge, there will be no pixel intensity change and the gradient vector will be 0. In general, due to background noise and other factors, such as fluctuations in pixel intensity within an object, a threshold is set – so if the gradient vector is greater than that threshold, then it is decided that the gradient vector corresponds to an edge. When a first derivative is at a maximum, the second derivative is zero. This is the basis of Laplacian methods – where the second derivative crosses zero, an edge is detected (Gupta and Ghosh Mazumdar, 2013). The Sobel operator method uses 2D spatial gradient measurements and finds the gradient magnitude at each point in an image. Each gradient estimate is the sum of a pair of orthogonal vectors (Gupta and Ghosh Mazumdar, 2013).

Gaussian filters are often used in combination with edge detection, in order to ‘smooth out’ the image objects, i.e., in order to reduce noise in the image, a Gaussian filter filters out small structures and isolated points. This does come with the disadvantage of spreading the edges, making the edges less defined and less easy to detect. Normally, a Gaussian filter is applied, and then Laplacian edge detection is applied.

However, the Canny edge detection (a gradient detection method) can also be applied after a Gaussian filter (Shrivakshan and Chandrasekar, 2012). Skinner *et al* decided to use the Canny edge detection in the nuclear morphology analysis software, due to its sensitivity and accuracy (Skinner *et al.*, 2019). The Canny edge detection used by Skinner *et al.* was first described by John Canny in 1986 (Canny, 1986). There are three main steps to the process: smoothing the image with a Gaussian filter, computing the gradient magnitude and orientation, and then applying non-maxima suppression to the gradient magnitude (Shrivakshan and Chandrasekar, 2012). The double thresholding algorithm is used to detect and link the edges. An 'edge' is defined as a point where the strength is locally maximum, in the direction of the gradient, and is based on a low and high threshold. If the strength is above the high threshold, it is an edge for sure, between the low and high threshold it may be an edge if connected to an already determined edge. If the strength of the gradient doesn't even meet the low threshold, then that part is not an edge.

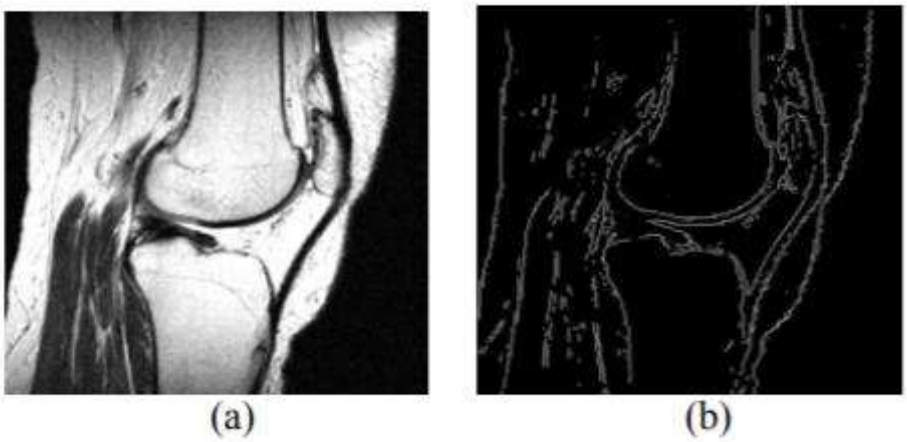
Figure 6



Edge detection thresholds, taken from https://cv-tricks.com/opencv-dnn/edge-detection-hed/, accessed 19/05/2020

This is a more accurate method of edge detection than traditional thresholding, because of the noise reduction. The main disadvantage of Canny edge detection is the relatively lengthy time it takes due to the complex computation required (Shrivakshan and Chandrasekar, 2012). Shrivakshan *et al.* believed that Canny edge detection, whilst probably the most expensive, time consuming and complex, was the more effective option of all edge detection methods they reviewed, which included Laplacian of the Gaussian, Sobel, Prewill and Robert's (Shrivakshan and Chandrasekar, 2012).

Figure 7



Edge detection comparisons, image (a) edges of an MRI scan detected with Sobel edge detection and (b) with Canny. Shows significant noise reduction in the Canny edge detection. Source: (Othman et al., 2009)

1.4.3 Texture Analysis

Texture is useful as a spatial feature used in the identification of objects in an image. There are a number of methods of texture analysis, with grey-level cooccurrence matrix (GLCM) being one of the more popular methods (Haralick, Dinstein and Shanmugam, 1973). However, this does not always give satisfactory results; the matrix depends on spatial relationships of grey levels as well as regional intensity

background variation in the image. A texture unit is the smallest complete unit which best characterises local texture in a given pixel and its neighbourhood of eight pixels (Wang, 1990). The local texture information for a pixel can be obtained by a neighbourhood of 3x3 pixels – this is the smallest complete unit, the texture unit. A texture spectrum describes the distribution of all the texture units within an image. A real texture image is generally composed of 2 parts: the texture elements and background. Texture can be better envisaged by humans if there are more texture elements relative to background (Wang, 1990).

1.5 Data Processing Methods

One of the main hurdles to overcome when analysing the data produced from the analysis software is the fact that the data is very high-dimensional – cell nuclei are described by numerous variables, and pixel intensity vectors in images tend to have thousands of dimensions. Interpreting high-dimensional data is not possible to humans without some method of dimensionality reduction – dimensionality reduction techniques allow the high-dimensional dataset to be displayed on an interpretable scatter graph. To simplify high dimensional data, it is assumed that the data of interest lie on an embedded non-linear manifold within the higher-dimensional space.

1.5.1 UMAP vs tSNE Dimensionality Reduction

Principle component analysis (PCA) was one of the traditional methods of dimensionality reduction. Two more recent methods are T-distributed stochastic neighbour embedding (t-SNE) and uniform manifold approximation and projection

(UMAP, which use graph layout algorithms to arrange high dimensional data into low dimensional space (Van Der Maaten and Hinton, 2008).

t-SNE is an algorithm based on four equations: the Gaussian probability equation, perplexity equation, student t-distribution and the Kullback-Leibler divergence loss function. t-SNE leads points that are close in high dimensions to get even closer in lower dimensions, and points far in high dimensions to get even further in low dimensions.

Stochastic neighbour embedding (SNE) is the basis for t-SNE. SNE works by using an algorithm to convert the high-dimensional distances between datapoints into 'conditional probabilities' that represent similarities of datapoints to each other. The algorithm computes the probability that pairs of datapoints in the high-dimensional space are related, and then chooses low-dimensional embeddings which produce a similar distribution – it can change slightly every time you run the algorithm. t-SNE is only able to preserve the local structure of a dataset, due to the nature of the Kullback-Leibler divergence loss function, which is an exponential function and therefore when the distance between data points in the higher dimensional data is large, this is projected onto the lower dimensional data and could give a resulting output of anywhere between 0 and infinity; as the larger distances are amplified in lower dimensions, the global structure won't necessarily be preserved. Therefore, a big weakness of t-SNE is the fact that the dimensionality is reduced based on local properties of datapoints, meaning that in datasets with high intrinsic dimensionality may not yield successful t-SNE results in terms of showing clusters within datasets.

In contrast, UMAP is a different algorithm for dimensionality reduction which is able to preserve more of the global structure, therefore leading to clearer visualisation of

clustering. It is also arguably a faster process and can be applied to much larger datasets than t-SNE (McInnes, Healy and Melville, 2018). In the paper on the nuclear morphology analysis software, Skinner *et al* use t-SNE dimensionality reduction techniques to present the data (Skinner *et al.*, 2019). In this analysis, UMAP has been chosen as the preferred dimensionality reduction technique, simply because of the advantages over t-SNE, specifically the speed advantage and the fact that it preserves local and global data structure. The speed advantage that UMAP has over t-SNE, as detailed in Table 1 means that, when experimenting with the different parameters to be used within the algorithm, the algorithm can be run a number of times in a sensible time frame to find the right parameters for the analysis to be done.

Table 1: Comparison between t-SNE and UMAP for dimensionality reduction

t-SNE	UMAP
Performs well on smaller datasets, but poorly on larger datasets	Performs well on most sizes of datasets
Can only embed into two or three dimensions	Can embed into a higher number of dimensions
Slow; 45 minutes for a 70,000-point dataset	Fast, under three minutes for a 70,000-point dataset
Doesn't well preserve global structure, proximity of clusters has no meaning	Preserves global structure, so clusters presented closer together are more similar

UMAP works by graphically representing the high-dimensional data and then creating a low-dimensional graph that is as similar as possible to the high-dimensional data. UMAP works by attributing each point a radius extending outwards, which connects with overlapping radii. The size of the radius is therefore important; if the radius is too small, the clusters will be too small and very isolated,

but if the radius is too large, many points will be connected without distinct clusters. However, a feature of UMAP is that the radius gets 'fuzzy' after the first connection, decreasing the likelihood of connection to other points, as the radius becomes larger. This builds up what is known as a 'fuzzy simplicial complex'.

Figure 8

Difference between fuzzy and not fuzzy connections in UMAP, adapted from https://umap-learn.readthedocs.io/en/latest/how_umap_works.html, accessed 18/05/2020

There are two important parameters used in UMAP to allow the balance between local and global structure to be controlled. They are n_neighbours – the approximate number of neighbours used to construct the high dimensional graph; where low values will focus more on local structure, and higher values will show global structure more accurately at the expense of finer detail of the local structure. The other parameter is min_distance – the minimum distance between the points in low dimensional space, with lower values leading to tightly packed clusters and larger values leading to more loose clusters.

1.6 Mouse hybrids and knockouts used in this study

The sperm development specifically studied in this study is that of the sperm from various mice subspecies, including their hybrid crosses and some knockouts.

1.6.1 Mouse subspecies

The mouse subspecies investigated in this study derive from two recently diverged mouse subspecies, *Mus musculus domesticus* (hereby referred to as *dom)* and *Mus musculus* (hereby referred to as *mus). Mus musculus spretus* (hereby referred to as *spr)* samples are imaged but not used for analysis. These subspecies diverged approximately 0.35-0.5 million years ago (Geraldes *et al.*, 2011) and the two subspecies show subtle differences in their mouse sperm morphologies. *Mus* subspecies (CZE and PWK) originated in Eastern Europe: Bratislava, Slovakia, and Lhotka and Czech Republic, respectively. *dom* subspecies (WSB and LEWES) originated in the USA: from Eastern Shore, Maryland and Lewes, Delaware, respectively. *Mus* and *dom* form a small hybrid zone in central Europe (JanouŠek *et al.*, 2012).

1.6.2 Male mouse hybrids show asymmetric sterility

Ellis *et al* (2011) demonstrated that creating hybrids between a strain of subspecies *M. musculus* (PWK) and of subspecies *M. musculus domesticus* (LEWES) led to some clear patterns in sex-ratio skew (Ellis, Bacon and Affara, 2011). Male hybrid sterility has since been observed in further F1 crosses of house mouse strains *mus* and *dom*, as a result of disruption to gametogenesis at various developmental stages, from mitosis, to meiosis, to post meiotic differentiation of cells (Larson *et al.*, 2022). Hybrid sterility is typically complex, due to the interaction of multiple alleles

which are often polymorphic, which means that the extent of gametogenic disruption can vary, both within and between strains. The classic model of hybrid incompatibilities is the Dobzhansky-Muller Incompatibilities (Muller and HJ, 1942; Dobzhansky, 2015), whereby multiple interacting loci diverge in early emerging species, which do not then function correctly when brought together in hybridisation. Mutations that accumulate on genes e.g. AABB in a diverging population gain mutations leading to AaBB and AABb. a and b are mutually incompatible and because this a-b interaction is not present in the pure species, hybrid incompatibility between the diverging populations evolves. Whilst this model is a simple model of hybrid sterility it provides a basic explanation for how hybrid incompatibilities emerge, but there are many other factors at play.

It has long been established that closely related subspecies of mice easily hybridise in natural populations, but they show asymmetric sterility. Crosses between *mus* females and *dom* males produce sterile hybrid males (Turner, Schwahn and Harr, 2012), however with the reciprocal cross, *dom* females and *mus* males, hybrid male sterility depends upon the individual genotypes. *mus* and *dom* mouse subspecies diverged between 350 and 500 million years ago, and it is well known that female hybrids are more fertile than male hybrids (Suzuki and Nachman, 2015), which fits in line with patterns of sterility seen with partial Y chromosome deletion (Ellis, Bacon and Affara, 2011) affecting Y-bearing sperm more severely.

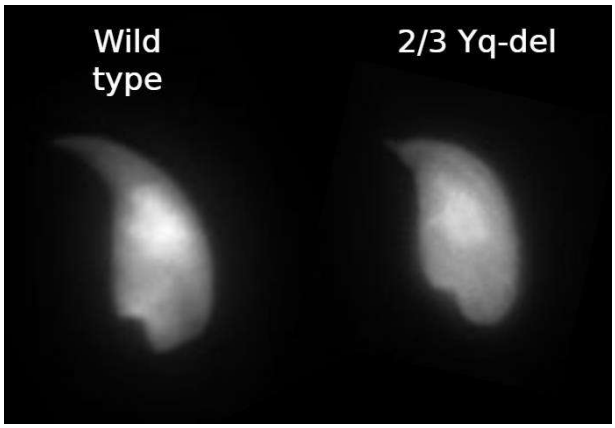
Many of the loci involved in these hybrid incompatibilities have been previously identified. It is known that the hybrid sterility in mouse is associated with the overexpression of sex chromosome genes, and post-meiotic X chromosome overexpression in F1 hybrids, leads to sex-ratio skew and sterility as seen in Yq-deletion (Larson *et al.*, 2022). Hybrid sterility occurs as a result of disruption to

gametogenesis at various developmental stages, from mitosis, to meiosis, to post meiotic differentiation of cells (Larson *et al.*, 2022). It is known that *Prdm9*-mediated incompatibilities between autosomes and X chromosomes can cause disruptions to meiotic sex chromosome inactivation (Good *et al.*, 2010; Bhattacharyya *et al.*, 2013). *Prdm9*is a zinc finger protein, which is essential for meiosis and is not found in somatic cells. It binds to specific locations of DNA to methylate histone H3 at lysines 4 and 36 (Paigen and Petkov, 2018). This helps to determine the site of recombination. *Prdm9* is also involved in programmed double strand breaks (DSBs), that allow genetic exchange between chromosomes.

1.6.3 Yq-del mice are mice with part of their Y chromosome long arm knocked out, and they show abnormal sperm nuclear morphology

The research that this study builds on was a story started a long time ago. Back in 2005, it was found that deletions on the mouse long arm of the Y chromosome, Yq, led to the loss of mouse Yq-linked transcripts, and the subsequent upregulation of several genes in the spermatid cells of these Yq-del males (Ellis *et al.*, 2005). Sperm nuclear morphology of these Yq-del males was shown to be abnormal, leading to one of two

Figure 9



Wild type mus mouse sperm vs Yqdel mouse with 2/3 deletion

distinct phenotypes: either a shortened acrosomal region with a blunt hook, or increased hook curvature (Figure 9).

Ampliconic genes are present in very large copy numbers on the sex chromosomes. During spermatogenesis, meiotic sex chromosome inactivation (MSCI) causes sex chromosome silencing, and sex-linked ampliconic genes are expressed during post meiotic spermatogenesis during the time when sex chromosomes are repressed. It is proposed that the high copy numbers are a mechanism to overcome the repressive effect, and copy numbers on the sex chromosomes are driven up in conflict with each other (Ellis, Bacon and Affara, 2011). Different subspecies have evolved to have different copy numbers of these ampliconic genes. *Sly* is known to be the most highly transcribed of the known ampliconic Yq genes and is found in very high copy numbers on the Y chromosome. Loss of *Sly,* as in the Yq-del mice, causes a copy number imbalance with the X-linked *Slx* gene family. *Sly* is a master regulatory transcriber, loss of *Sly* leads to the upregulation of many other genes.

It is demonstrated that knocking out different proportions of the Yq chromosome led to different magnitude of morphological effects. Smaller knockouts of the Yq led to a sex ratio skew in the offspring, in favour of females. However, larger knockouts led to total sterility of the mouse. This shows that Y-bearing sperm are selectively affected by the knockout of *Sly* and the subsequent upregulation of other genes to a greater degree than X-bearing sperm. It is likely that the upregulation of some genes as a result of Yq-deletion directly contributes to the abnormal Yq-del sperm phenotypes seen; *Mgclh* is homologous to genes involved in human and mouse sperm head development, and CK2 is also known to regulate sperm head morphology (Ellis *et al.*, 2005). The upregulated genes seen in Yq-del males are predominantly X- and Y-linked, whereas the autosomal homologues are not upregulated in any way.

The sex-ratio skew towards female offspring was shown using an X-linked GFP transgene (Rathje *et al.*, 2019); this skew was seen in embryos mid-gestation as well as at the 2-stage blastocyst after natural mating. This shows that it was not a post-fertilisation effect, but rather something that occurs pre-fertilisation or during fertilisation. X- and Y-bearing sperm in the males with a Yq deletion had differential motility and morphology but were equally able to penetrate the cumulus and fertilize the egg once at the site of fertilization, meaning that altered motility was likely to be the cause of the sex ratio skew, and the sterility (Rathje *et al.*, 2019). IVF was shown to abolish this sex-ratio skew; leading the study to conclude that it was the transport to the egg that was impacted in Yq-deletion sperm, rather than sperm-egg interactions.

1.6.4 Y-introgression is the process of combining the Y-chromosome from one subspecies onto the X-chromosome and autosomal background of another subspecies

Introgression is also known as introgressive hybridization (Anderson, 1949), and it is the integration of alleles from one species into the gene pool of another species (Harrison and Larson, 2014). This normally occurs through hybridisation and then repeated backcrossing (Twyford and Ennos, 2011). A key feature of introgression of the Y chromosome onto an autosomal background in mouse is that, as discussed above, mice have large ampliconic regions of DNA on their sex chromosomes (Bachtrog, 2014), and different species of mouse have differing sizes of ampliconic region. Therefore, Y-introgression in mouse leads to a mismatch in ampliconic copy numbers between the X and Y chromosome – the X chromosome will have

ampliconic copy numbers specific to one species, and the Y chromosome will have ampliconic copy numbers specific to another species. Y-introgression via repeated backcrossing can be used for experimental purposes to cause an X-Y mismatch, which can address the question “does X-Y mismatch cause abnormal sperm morphology, as in Yq-del?”. The mice used in this study underwent ten generations of backcrosses in reciprocal directions to generate mice with a *Mus musculus domesticus* (*dom*) genetic background and *Mus musculus musculus* (*mus*) Y chromosome (dom^{musY}) and mice with a *mus* genetic background and *dom* Y chromosome (mus^{domY}). This was further expanded to a number of more complex crosses, including F1 hybrids, as summarised in Table 2.

1.6.5 *Sly* and *Slx* are sex-linked ampliconic master regulators in mouse

It is well established that mammalian sex chromosomes undergo evolutionary pressures that do not act upon autosomes, and these pressures affect both structure and function of the sex chromosomes. This arises because the sex chromosomes are only present in one copy in males, and so the impact of mutation and genetic drift is greater on the sex chromosomes than the autosomes, where two copies are present. Mutations on one copy of an autosome are therefore able to be masked by the other copy of the autosome, whereas a mutation on the X or Y chromosome in males will not be masked.

One of these evolutionary pressures has led to gene amplification on sex chromosomes, which is seen to a greater extent in mouse than human. This has caused the presence of large amplicons on the sex chromosomes in mouse Ellis,

Bacon and Affara, 2011). *Slx* is a sex-linked ampliconic gene which originated from *Sycp3* – a single copy autosomal mouse gene, which was transposed onto the X chromosome, first becoming *Slx1* and then *Slx.* Eventually, the presence of this ampliconic gene on the X chromosome led to a copy emerging on the Y chromosome, *Sly* (Kruger *et al.*, 2019). Because gene amplification on sex chromosomes is more pronounced in mice than in humans, there are no genetic equivalents of these high copy number ampliconic genes in humans, however it is possible that some of the downstream genetic mechanisms implicated in this study may have human equivalents, which would be an area of further research for the medical side of fertility. Male mice with deletions on Yq show sex-ratio skew in their offspring in favour of females, which in turn leads to the upregulation of X-linked genes in spermatids. This skew is facilitated by X-linked spermatid-expressed genes favouring their own transmission, in contrast to genes on the Yq which suppress X transcription to restore the natural 50:50 sex ratio. It is this effect that has led to such massive expansion of copy number for these ampliconic genes over many years: a genetic 'arms-race' to try to keep the natural sex ratio. *Sly* is known to be a key gene on Yq, since selective knockout of *Sly* leads to the same effects as a partial Yq deletion - including the sperm head shape abnormalities and increased X chromosome transmission. This suggests that the main gene causing these effects is indeed *Sly. Sly* inhibits transcription from sex chromosomes in spermatids; the accumulation of amplicons on the mouse sex chromosomes counteracts this repressive effect of *Sly* in order to maintain expression levels of the sex-linked spermatid genes (Ellis, Bacon and Affara, 2011).

The co-amplification of sex-linked ampliconic genes has occurred to different extents in different species of mouse, leading to divergent copy numbers of *Slx/Sly* in the

mus and *dom* mice strains. A recent study estimated the copy numbers of *Slx/Sly* by two different methods, Mosdepth and AmpliCoNE (Larson *et al.*, 2022) to show that *mus* mice had significantly higher *Sly* and *Slx* copy numbers than *dom*. Their results are summarised below.

Table 2: copy number estimates for *Sly/Slx* in *mus* and *dom* mice (Larson et al., 2022).

		Mosdepth		AmpliCoNE	
		Sly	***Slx***	***Sly***	***Slx***
musculus	PWK	192	48	213	50
	CZEII	206	51	217	62
domesticus	LEWES	152	16	134	15
	WSB	155	13	127	13

Mismatch in copy number levels of *Sly* in *mus* and *dom* is a proposed explanation for asymmetric male sterility seen in crosses between LEWES (*dom)* and PWK (*mus)* subspecies, since *Sly* levels in *mus* are too high to be counteracted by the ampliconic regions from *dom*. (Ellis, Bacon and Affara, 2011) Pairing of a LEWES Y chromosome with a PWK X chromosome has been shown to lead to X overexpression in spermatids, causing sterility, which is hypothesised to be because of the reduced *Sly* copy number from the LEWES Y chromosome. Further work is needed to elucidate the extent of the role that *Sly/Slx* mismatch plays in hybrid sterility.

1.7 Mouse Sperm Development

The first stage of spermatogenesis is the mitosis of spermatogonia, where differentiated spermatogonia divide, and form primary (preleptotene) spermatocytes. (Panawala, 2017) This is followed by meiosis 1, to form secondary spermatocytes

(meiotic cells), and then meiosis 2 to form round spermatids. The spermatids then mature in order to produce motile spermatozoa.

Spermiogenesis is the final stage of spermatogenesis, whereby the spermatids mature and differentiate into spermatozoa.

Mammalian spermatogenesis is a complex process controlled by transcriptional and translational mechanisms, leading to temporal and stage-specific gene expression patterns (Yu *et al.*, 2003). During spermatogenesis, there is a series of transcriptional silencing followed by selective reactivation of specific genes. Understanding the times at which certain genes become activated or inactivated, at which stages during spermatogenesis, can help improve the understanding of which genes are switched on and switched off in mRNA, and its pre-cursor, hn-RNA.

Whilst the majority of cells in the body contain the same genome, not all genes are required in each cell; genes are expressed when needed to specify unique features in different cells. Gene expression profiles can be defined by using reverse transcription-polymerase chain reactions, RNase protection assays, or Northern blot analysis. Gene expression profiles compare patterns of gene expression in different tissues, at different developmental stages, and can be used to compare healthy and disease states. Yu *et al.* describe how traditional methods of gene expression profiling stated above only focus on a few genes at a time, and cDNA microarray technology - hybridisation of entire cDNA populations to nucleic acid arrays - is a much more promising high-throughput analysis, allowing rapid detection of gene expression profiles of hundreds/thousands of genes at one time (Yu *et al.*, 2003). Yu *et al.* describe how cDNA microarray technology was used to identify

spermatogenesis-related, and differentially expressed genes among six spermatogenic cell types isolated from mouse testis (Yu *et al.*, 2003).

The basic stages of spermatogenesis are mitosis, meiosis I, meiosis II and spermiogenesis, and the stages involved in meiosis are described below.

1.7.1 Stages of Meiosis

Meiosis produces gametes, yielding daughter cells with exactly half as many chromosomes as the starting cell. Meiosis occurs in a two-stage division process. Homologous pairs separate in the first round of cell division, meiosis I, and then sister chromatids separate during the second round, meiosis II.

1.7.1.1 Meiosis I

Before meiosis I, cells go through interphase, which has a G_1 growth phase, all the chromosomes are copied during the S phase, and then prepares for division during G_2. Following interphase, prophase I can occur. Each chromosome aligns with its homologous partner, so the 2 chromosomes match up, and crossing over of the chromosomes occurs, while the pair of homologues are held together by the synaptonemal complex. Chiasmata occur from the crossing over of homologous chromosomes. These are cross-shaped structures that link the homologues together and keep them connected after the synaptonemal complex breaks down. Because of this, each homologous pair requires at least one, but it is common for homologous pairs to have up to 25 chiasmata. Crossing over occurs completely randomly, leading to unique combinations of alleles. Following crossing over, the spindle begins to move chromosomes towards the centre of the cell, and each chromosome attaches to microtubules from just one pole of the spindle. The two homologues of a

pair therefore bind to microtubules from opposite poles; this process is metaphase I. The orientation of the homologous pairs during metaphase I is random, further increasing the number of possibilities of different gametes that could be formed. During anaphase I, the homologues are pulled apart and move to opposite poles. The sister chromatids of each chromosome remain attached to each other. In telophase I, cytokinesis occurs; chromosomes arrive at opposite poles of the cell and two haploid daughter cells are formed.

1.7.1.2 Meiosis II

Meiosis II is a simpler process, since DNA is not copied. In meiosis II, it is the sister chromatids that separate, rather than the homologous pairs. During prophase II, the chromosomes condense, and the centrosomes move apart, the spindle forms and microtubules begin to separate the sister chromatids. In metaphase II, the chromosomes line up along the metaphase plate and in anaphase II, the sister chromatids are separated. Telophase II occurs, and nuclear membranes are formed around each set of chromosomes. Cytokinesis splits the chromosomes into new cells, forming the final products – four haploid cells in which each chromosome has a single chromatid.

1.7.1.3 Prophase I of Meiosis

Prophase I is the longest stage of meiosis, when homologous chromosomes are pairing up and exchanging genetic information. Because it is such a long phase, it is divided into sub-stages. Leptotene is the first stage of prophase I – whereby individual chromosomes become individualised to form visible strands in the nucleus.

Leptotene is a very short sub-stage and lateral elements of the synaptonemal complex start to assemble. Leptotene is followed by zygotene, which occurs as the chromosomes line up into homologous chromosome pairs. Homologous chromosomes start to come together, and the assembly of the central element of the synaptonemal complex occurs. Pachytene follows, whereby homologous combination – crossing over – occurs, and chiasmata form. Diplotene is the sub-stage where the synaptonemal complex degrades and the homologous chromosomes begin to separate, but chiasmata will remain on chromosomes until anaphase I and so the chromosomes will not properly separate until then. Diplotene is the stage at which human oocytes are arrested in until ovulation, all developing oocytes in a fetal human develop until the diplotene stage and then remain in that stage until puberty, when the oocytes will continue the rest of meiosis ready for potential fertilisation.

Following diplotene, diakinesis occurs, which is where the chromosomes further condense and the four parts of the tetrads of the chromosomes are visible for the first time, and the meiotic spindle begins to form. In this stage, chiasmata are clearly visible. The last stage is meiotic spindle formation.

1.7.2 Gene expression in spermatogenesis

Transcription is an ongoing process throughout spermatogonia production, meiosis and early spermiogenesis, and in mice, transcription rate increases after meiosis. However, mRNA in sperm cells must be synthesised and stored prior to translation, because global transcription in sperm cells stops several days before the completion of spermiogenesis. It is essential that the correct pattern of gene expression occurs

during spermatogenesis, and disruptions to the pattern of gene expression have been shown to lead to the arrest of germ cell differentiation, a process leading to infertility (Acharya *et al.*, 2006). Several studies have shown critical cis elements in the untranslated regions (UTRs) of mRNA relating to correct temporal gene expression.

Gene regulation during transcription involves activation of transcription at the correct development stage as well as the prevention of transcription in the incorrect cell types. Cis-elements recruit transcriptional activators or repressors in the appropriate cell types, and germ cell-specific transcriptional activators, such as CREM are able to activate a number of differentiation markers during the early stages of spermiogenesis (Acharya *et al.*, 2006).

An example of the need for transcriptional regulation is the mRNA for the SP-10 mouse gene, which codes for an acrosomal protein, and must only be synthesised in the round spermatid stage of development, a process regulated by the proximal promoter of SP-10 which contains cis-regulatory elements, essential for both positive and negative regulation (Acharya *et al.*, 2006).

In mammalian spermatids, chromosome condensation is initiated by transition proteins, and completed by protamines (Braun, 1998). The mRNA encoding transition proteins and protamines needs to be carefully controlled, to ensure the chromosomes condense at the correct time – the genes are transcribed a lot earlier than they are translated. For example, two protamine genes *Prm-1* and *Prm-2* are transcribed in step seven round spermatids, but the mRNAs are not translated until around step 13 elongating spermatids (Braun, 1998). However, other mRNAs need to be actively translated during this time. Braun (1998) described how the *Prm1* 3’

untranslated region (UTR) contains redundant translational repression elements, and separate sequences for translational activation, and elucidated the need for further study regulation of UTRs to determine if they contain similar or different cis-acting regulatory elements (Braun, 1998).

It is generally considered that preleptotene spermatocytes are the starting point of entry into meiosis, however, it should be mentioned that type B spermatogonia play a critical role in terms of preparing for entering meiosis, so between type A spermatogonia and type B spermatogonia, transcriptional activity of many genes including cyclins, growth factors, transcription factors and oncogenes are upregulated. 20 genes were also found to be expressed in differentiated spermatogonia which decreased to near background level by meiosis – so these genes may not function during the later stages of spermatogenesis, or other germ cell specific genes are expressed to compensate for the lack of presence of inactivated genes.

The G2/mitotic-specific cyclin A1 and G1-specific cyclin C are both expressed consistently at a high level in all germ cells. The S/G2/M-specific cyclin F, G1/S-specific cyclin D1, and G1/S-specific cyclin E1 are only expressed in spermatogonia and are highly up-regulated from type A to type B spermatogonia. Cyclin D3 (of the G1/S-specific cyclin family) is undetectable in spermatogonia but is expressed strongly in spermatocytes and spermatids. Gene E2F, which is related to the function of cyclin D1 and cyclin E1, which have regulatory roles in the G1/ S-phase transition of the cell cycle, is also up-regulated from type A to type B spermatogonia, which could reflect the increase in the proliferative capacity of type B spermatogonia compared to type A.

It may be that E2F is a transcriptional activator of both cyclin E1 and cyclin D1, which cause growth and proliferation of type B spermatogonia.

Numerous growth factors and receptors play essential roles in spermatogenesis, and this is already well-understood. Yu *et al.* (2003) showed that the expression of transforming growth factor b1 (*Tgfb1*) was undetectable in primitive type A spermatogonia and only weakly expressed in type B spermatogonia, whilst no *Tgfb2* expression was seen in either type A spermatogonia or type B spermatogonia. Other previous studies have found that *Tgfb2* is solely expressed in late spermatogenic phases.

Another study (Eddy, 1998) showed that targeted inactivation of genes for several *Tgfb* superfamily members expressed in the testis did not cause primary defects in spermatogenesis, suggesting that they may not have a key role in spermatogenesis; this remains to be clarified.

Yu *et al.* (2003) showed IGF-1 was expressed at a low level in primitive type A and type B spermatogonia, whereas IGF-2 had a strong expression signal in both, however IGF-1 receptor (IGFR-I) was expressed strongly in the two cell types, whereas IGF-2 receptor (IGFR-II) was undetectable (Yu *et al.*, 2003). The study suggests that IGF-2 plays a more important role than IGF-1 during spermatogenesis. As well as IGF-2, some adapters and death receptor-associated proteins, such as STAM and FAN, were strongly up-regulated in type B spermatogonia, supporting the idea that apoptosis plays an important role in establishing the adult population of spermatogonia. DNA polymerase d and DNA polymerase a were also expressed in type B spermatogonia; these participate in synthesis of DNA during the early phase of meiosis and play roles in DNA replication of spermatogonia.

When the process of meiosis begins, with spermatogenic cells differentiating into preleptotene spermatocyte stage, most gene expression levels are maintained from preleptotene to pachytene spermatocytes, except three genes which are up-regulated: *Six2, Sim2,* and *Osm. Six2* and *Sim2* behave as transcriptional repressors in mammals, so may cause gene transcription inactivation in spermatids (Yu *et al.*, 2003).

The number of expressed genes and their expression levels decreases markedly from pachytene spermatocytes to round and elongating spermatids; histone-to-protamine transition and the resulting chromatin compaction in condensing spermatids mean sperm gene transcription becomes virtually inactive. Yu *et al* (2003) showed that seven genes (*Fcgr2b, Lim2, Madh1, Hsp86-1, Dad1, Gdnf,* and *Mapk14*) were up-regulated as pachytene spermatocytes differentiated to round spermatids and that three genes were up-regulated during the transition from round spermatids to elongating spermatids; genes that may be involved in the induction of spermiogenesis (Yu *et al.*, 2003).

1.7.3 Chromosome Synapsis and Transcriptional Silencing

During prophase I of meiosis, double-strand break and repair of DNA occurs, by topoisomerase-like enzyme *Spo11*. Chromosomal regions that do not have a pairing partner, specifically X and Y chromosomes, show delayed DSB repair, and also undergo transcriptional silencing. Inagaki, Schoenmakers and Baarends(2010) reviewed meiosis-specific aspects of double strand break repair and meiotic silencing.

In many species, the homologous chromosome pairing during prophase I requires formation and repair of the meiotic-specific DNA double strand breaks. In somatic cells, DSBs can occur due to DNA damage and so they must be repaired before mitosis to maintain the integrity of the genome. However, this is even more important in germ cells, due to the fact that the combination of the two parental genomes forms the new individual. In mice, the formation and repair of meiotic DSBs is essential for correct pairing, synapsis and segregation of the homologous chromosomes during the rest of meiosis I. In some species, such as *D. melanogaster*, meiotic DSBs are absent in males (Inagaki, Schoenmakers and Baarends, 2010).

As described above, chromosome pairing begins during leptotene of prophase I, progresses during zygotene and completes during pachytene. Following repair of meiotic DSBs in mouse and human, synapsis occurs. Knocking out genes involved in synapsis has shown that chromosomes are able to align but won't synapse and meiotic DSB repair is then stalled. Many cells that have meiotic DSB repair stalled end up undergoing apoptosis during pachytene. It has been shown that there is a sex-difference in severity of the effect of knockout of synapsis genes, with males experiencing more severe effects (Inagaki, Schoenmakers and Baarends, 2010). It could be that some of the downstream genes that *Sly* impacts upon could be involved in synapsis.

1.7.4 Transcriptional Silencing

Sex chromosomes exhibit unique behaviour during prophase 1 of meiosis in spermatogenesis. Sex chromosomes lack a homologous pairing partner, so they undergo gradual condensation until the middle of the pachytene stage, which is

when the XY body is fully formed (Homolka *et al.*, 2007). The condensed chromatin in the XY body undergoes transcriptional repression, via the process of meiotic sex chromosome inactivation (MSCI). MSCI begins in the late zygotene and occurs by BRCA1-directed ATR kinase phosphorylating histone H2AX. The arrest of mouse spermatogenesis in prophase I is a frequent cause of male sterility – due to mutations or deletions of functional genes or chromosomal rearrangements. Homolka *et al* (2007) investigated a link between the incomplete synapsis of chromosomes in sterile male autosomal rearrangements, and X chromosome inactivation in male meiosis. This was investigated by analysing the abnormal chromosome synapsis with BRCA1 and gamma H2AX markers, as well as analysing global transcriptome changes with microarrays. It was shown that a significant transcriptional down-regulation of genes in the unsynapsed autosomal chromatin in pre-mid-pachytene spermatocytes was followed by a partial breakdown of transcriptional silencing of the X chromosome during the mid-late-pachytene stage (Homolka *et al.*, 2007).

It has been shown that mice lacking ubiquitin ligase UBR2 are infertile due to arrest of meiosis at prophase I (An *et al.*, 2010). UBR2 localises to meiotic chromatin regions and mediates transcriptional silencing by ubiquitination of histone H2A. UBR2-deficient spermatocytes are unable to induce the ubiquitination of H2A during meiosis, impairing these spermatocytes in chromosome-wide transcriptional silencing of genes. It is suggested that insufficient UBR2-dependent histone H2A ubiquitination triggers a checkpoint system during pachytene of prophase I (An *et al.*, 2010).

1.7.5 Gene Expression at Different Stages of Spermatogenesis

This study relies on using data from round spermatid stages only. In order to determine which of the cells in each dataset are true round spermatids, it is essential to be able to identify which genes are switched on at which points in the development of sperm cells. The expression levels of these marker genes in each cell can then be assessed to determine which cells are at which developmental stage.

It was determined that a number of marker genes based upon the literature were to be used to qualify developmental stage (Cocquet *et al.*, 2012; Jan *et al.*, 2017). These genes were *Vim, Cd14, Vwf, Acta2, Dlk1, Dazl, Magea4, Utf1, Id4, Fgfr3, Kit, Dmrt1, Dmrtb1, Stra8, Sycp3, Spo11, Mlh3, Zpbp, Tnp1* and *Prm2*.

Early round spermatids should have high levels of *Zpbp* but low levels of *Magea4*, *Dazl, Spo11, Mlh3* and *Sycp3*. Late round spermatids through to late elongating spermatids should have high levels of *Tnp1*. Elongating through to condensing spermatids should have high levels of *Prm2*. Analysing the gene expression levels of these key genes allows identification of which cells are round spermatids and which are not.

1.8 Genetic Analysis

1.8.1 Next Generation Genome Sequencing

Dideoxynucleotide sequencing of DNA was first described in 1977 and has steadily undergone a transformation into a highly specialised and technical large-scale process, where the costs have dramatically fallen, making it a much more accessible

analysis method in the field of genetics. Next-generation sequencing has become readily available and has revolutionised genomics (Mardis, 2008).

One of the major advantages of next-generation sequencers is the speed at which the sequencers can process reads; millions of sequence reads at once, rather than just 96 at a time. In addition to this, next generation sequence reads are produced from libraries of fragments, which have not undergone the conventional methods of *E. coli*-based amplification which leads to a number of cloning biases. Next generation sequencers can also be used for paired-end reads, rather than just single-end reads.

The three main methods of next-generation sequencing are Roche, Illumina and SOLiD. Roche sequences fewer megabases per run but is much faster and more expensive per megabase to run. It uses an emulsion PCR amplification with pyrosequencing. Illumina uses bridge amplification and polymerase-based sequencing-by-synthesis; it is the mid-range method in terms of cost and timings. SOLiD is the cheapest to run but also the slowest. It also uses an emulsion PCR based amplification with ligation-based sequencing (Mardis, 2008).

1.8.1.1 RNA-Sequencing

RNA sequencing methods have been developed to allow for studying RNA structure, translation and for looking at single cell gene expression. RNA sequencing has been around for many years now (Emrich *et al.*, 2007) and is most commonly used for analysing differential gene expression. Techniques advance often, but the key stages of RNA sequencing haven't changed since the very first applications. RNA is extracted, mRNA is enriched, or ribosomal RNA is depleted, cDNA is synthesised

and then an adaptor ligated sequencing library is prepared. The sequencing library is sequenced, up to a read depth of 30 million reads per sample, and then the sequencing reads are aligned computationally to a transcriptome (Stark, Grzelak and Hadfield, 2019).

1.8.1.2 Paired-end vs single-ended sequencing

During sequencing, the number of base pairs read is an important consideration: longer reads can provide more information however it is normally more expensive. Two types of sequencing are paired end and single ended. In paired end reading, the sequencer reads fragments from one end of the fragment and stops at the specified read length, and then another round of reading occurs starting at the opposite end of the fragment. Paired end reading is much more effective than single end reading, especially when used for the assembly of repetitive regions. However, it is not always necessary, and cost must be balanced with the need for this higher degree of accuracy.

1.8.1.3 Single-Cell Sequencing

Single-cell RNA sequencing was introduced in 2009 (Wang *et al.*, 2021) and allows the RNA of each individual cell to be analysed, allowing comparisons of different cell types to show the compositions of tissues and organs. It also allows the detection of rare cell types. Two frequently used single cell RNA sequencing methods are a droplet-based method, 10X Genomics Chromium approach (hereon referred to as 10X) and the plate-based, full-length method, SmartSeq2. SmartSeq2 samples to a much higher read depth, detecting many more genes in every cell, including low

abundance and alternatively spliced transcripts. SmartSeq2 also captures higher levels of mitochondrial genes. In comparison, 10X data shows more gene dropout, especially with genes with lower expression levels. 10X techniques can be used to cover a much larger number of cells than SmartSeq2, however, which means that rare cell types can be detected with 10X sequencing.

SmartSeq2 is based on microtiter plates, and mRNA is separated, and reverse transcribed for every cell. Conversely, 10X is droplet-based, and allows genome-wide expression profiling of thousands of cells, all at the same time. Unique molecular identifiers are used to label individual cells.

Both sequencing types are popular and have their own unique advantages and disadvantages depending on what they would be used for. Some studies would benefit from the larger scale analysis of 10X to a smaller read depth, while other studies would favour the higher read depth of SmartSeq2, even though the number of cells being able to be studied would be smaller. Because of this, large amounts of data exist that have been derived from each different sequencing type, and the analysis of single cell data in this study should be applied to both sequencing methods, if possible.

1.8.2 RNA Sequencing Analysis Methods

RNA sequencing analysis begins with downloading the sequencing reads used in single-cell RNA sequencing studies from the SRA of the NCBI database. Single cell RNA sequence reads are available as a list of SRR identifiers, which can be converted to FASTQ files. FASTQ files are text files which encode the sequence data from read sequencing. Paired end reads will have 2 FASTQ files per read.

1.8.2.1 FASTQC

FASTQC gives an overview of basic quality control metrics for next generation sequencing data. The output is a html file that can be viewed in a web browser, with

Figure 10

Measure	Value
Filename	SRR6342093_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	99012114
Sequences flagged as poor quality	0
Sequence length	151
%GC	50

Basic statistics for FASTQ file SRR6342093_1.fastq, from the Chen dataset. Basic statistics show the filename, type of file, the type of sequencing, total number of sequences, number of poor-quality sequences, average sequence length and the GC percentage.

a report section for each FASTQC module. Basic statistics are first printed, as demonstrated in Figure 10, which include the filename, file type, the number of sequences in the file and the number flagged as poor quality.

The FASTQC modules are: Per Base Sequence Quality, Per Sequence Quality Scores, Per Base Sequence Content, Per Sequence GC Content, Per Base N Content, Sequence Duplication Levels, Overrepresented Sequences, Adapter Content, Kmer content. Graphs and list data are presented for each module as well as a ‘pass’, ‘warn’ or ‘fail’ flag, but these flags need to be assessed with caution as they are fine tune for good quality whole genome DNA sequencing, so for RNA sequencing or targeted sequence capture they may not be reliable. It is recommended that ‘warn’ and ‘fail’ flags simply mean that particular care must be

taken to consider the results in the context of the sample and type of sequencing that was run. In many cases, RNA sequencing is perfectly good quality but will produce warn or fail flags, as detailed below, since the flag parameters are based upon whole genome DNA sequence – rather than RNA sequencing. Therefore, although important to take the time to acknowledge the FASTQC output, it was not of any concern that some of the parameters produced warnings or fails.

1.8.2.2 GTF annotation files

General Transfer Format (GTF) annotation files are tab-delimited text files that store gene structure information. Each GTF file contains nine fields, explained below.

Table 3: GTF file fields

Identifier	Description
Seqname	The name of either the chromosome or the scaffold ID
Source	The source that generated the feature e.g., the program, database or project
Feature	The type of feature: e.g., gene, transcript, exon
Start	The start position of the feature (starting from 1)
End	The end position of the feature (starting from 2)
Score	Floating point score
Strand	The strand direction: either forward (+) or reverse (-)
Frame	'0' would indicate that the first base of the feature in the first base of the codon, whilst '1' would show the second base is the first base of the codon, and '2' would show that the third base is the first base of the codon
Attribute	Additional information about the feature as a semi-colon separated list

GTF files are used to annotate genomes. The seqname from column one needs to match the file being annotated. Common types of features include gene, CDS, mRNA, exon, 5' UTR, 3' UTR, rRNA, tRNA, ncRNA, tmRNA, transcript, promoter and repeat region. Feature types not standardly included are introns and proteins.

1.8.2.3 STAR read aligner

The STAR Aligner tool helps determine from where on the genome reads have originated, and it is specifically designed in order to address many of the complexities of using RNA-sequencing data, by accounting for spliced alignments. As is a consideration with FASTQC, RNA-sequencing does have some distinct differences to DNA-sequencing and STAR is specialised to work with RNA sequencing data. STAR has a very high accuracy but is memory-intensive – for this reason enough threads from the HPC cluster needed to be assigned for it to run in an appropriate time. The mapping occurs in two stages – seed searching and then clustering, stitching, and scoring.

Seed searching is where, for every read that to be aligned, STAR searches for the longest sequence that exactly matches one or more locations on the reference genome. These sequences are called the Maximal Mappable Prefixes (MMPs). The first MMP mapped to the genome is called seed1, with each ‘seed’ being a part of the read mapped separately. Following mapping of seed1, STAR will search again, amongst only the unmapped portion of the read, to find the next longest sequence to exactly match the reference genome, which becomes seed2.

Next, the clustering, stitching, and scoring occurs. In this process, the separate seeds are ‘stitched’ together to create a complete read. The seeds are first clustered based on proximity to a set of ‘anchor’ seeds (seeds that are not mapped to multiple positions). They are stitched together based on the best alignment for the read – a score is given based on mismatches and gaps in the sequencing.

Prior to any mapping, the first stage of STAR analysis is generation of the index genome. After generation of the reference genome comes the mapping stage. Each

subsequent time STAR is ran using the same genome, only the mapping stage needs to be run – the reference genome only needs to be generated once per genome.

1.8.2.4 featureCounts annotation tool

In RNA sequencing, the term "reads" refers to sequences from the sequencing instrument, and the "hits" are the mappings of a read to a specific genomic region e.g., intron, gene, exon. A read "hits" some defined region, or in the case of multi-mappers, several regions. The number of hits shows how many reads are associated with a transcript.

FeatureCounts is a programme that is part of the Subread package. The subread package comprises multiple programs for processing next-generation sequencing data. The featureCounts program assigns mapped reads to genomic features such as genes, exons, promoters etc, and is suitable for genomic DNA-seq and RNA-seq reads. It uses BAM files as an input, which are generated directly from the STAR aligner package (Liao, Smyth and Shi, 2014).

The rationale for using the featureCounts package is that the sequencing reads analysed need to be annotated to genomic features of interest – i.e., does the read map to a gene/a transcript/an exon, or in this case an intron. The output of featureCounts read summarisation is a count table, where the number of reads assigned to each feature in each library is given.

1.8.3 Differential Gene Expression Methods

There are a variety of packages that will perform differential gene expression analyses. Differential expression analysis quantifies the difference in expression

levels of genes between experimental groups, using normalised read count data and statistical testing to decide if the observed differences are likely due to chance or if they are significant.

Some tools can only perform pairwise comparisons, but many can perform multiple comparisons.

1.8.3.1 SCDE is a differential gene expression analysis package designed for single cell data

The package SCDE (Kharchenko, Silberstein and Scadden, 2014) is optimised for the differential gene expression analysis of single cell data, and is chosen for use in this study. The package allows detection of differentially expressed genes in a way that is more tolerant of the higher levels of background noise seen in single cell data; comparing individual cells typically results in much more variability than seen in bulk RNA-sequencing analysis. Typical measures of differentially expressed genes often include genes that have been driven by outliers and 'drop-out' events, both of which are seen to high degrees in single cell data. The drop-out is highly informative if it can be accounted for in the expression analysis. Overall drop-out rates are consistently higher in some cells that others, and dropouts are more frequent in genes that are expressed to a lower magnitude. This means that where there is drop out of a gene observed in cells with a very high expression magnitude, the effect seen is much more likely represent true expression differences between the cells (Kharchenko, Silberstein and Scadden, 2014).

1.8.4 Analysis of gene expression data

KEGG (Kyoto Encyclopaedia of Genes and Genomes) is a collection of databases for allowing gene function analysis (Kanehisa and Goto, 2000). KEGG pathway maps are produced by the input of gene lists, to show the genes upregulated in different cells, and show the number of genes from the list which are implicated in various biological processes, coloured by their p-values for how significantly those genes are associated with that process.

KEGG plots list gene functions and have dots next to the functions. The larger the dot next to the stated biological function, the more genes associated with this function. The redder the dot, the more significantly these genes are associated with the function. String analysis can be used to identify common pathways in the proteins encoded by the genes. String also provides gene ontology information from several databases.

DAVID (Dennis *et al.*, 2003) is a functional annotation tool which uses Kappa statistic scores to determine relationships between gene lists based on a novel fuzzy clustering algorithm which groups genes that are classed as similar, heterogeneous or redundant from various resources into annotation categories.

1.8.5 Motif analysis

Searching for common motifs in gene sequences has many applications. In this analysis it is used to identify potential common transcription factor binding sites. This can be done using the package STREME. STREME stands for sensitive, thorough, rapid, enriched motif elicitation, and searches for motifs that are enriched in input sequences (Bailey, 2021). STREME is optimised to work on 50 sequences or more.

The genes significant in the X-bearing spermatids had their promoter regions extracted and STREME was used to discover motifs enriched in the X-bearing cells. STREME takes a set of primary sequences as an input, and can also take a set of control sequences, although in this study, control sequences were not used. In the absence of a control sequence, STREME creates its own control from random shuffling of the primary sequences inputted. STREME will use either the Fisher's exact test or the Binomial test in order to test for if there is significant enrichment of motifs in the input sequences compared to the control sequence. If the primary and control sequences have the same average length, Fisher's exact is used, otherwise Binomial testing is the test of choice (Bailey and Grant, 2021).

The outputs of STREME include an HTML file, where results are shown in order of decreasing statistical significance. There is a sequence logo of each motif that was shown to be enriched, the name of the file containing said motif, the ID of the motif, number of times it occurs in primary sequences – true positives – as well as false positives, the number of times the motif appears in the control sequences. STREME gives each motif a name, made from a consensus sequence. The output files include the motif, the logo, the reverse complement logo, positional distribution and matches per sequence.

The sequence logo of the motif uses letter stacks based on equations by Schneider and Stephens (Schneider and Stephens, 1990), and the positional distribution refers to the positional distribution of best match to the motif in all the sequences with a match.

Some of the letters in the motif are not the typical ‘A, C, T, G’, this is because sometimes the letter is not exactly known. Table 4 shows the symbol letters used and their potential meanings.

Table 4: symbols used in STREME motifs and what they mean

Symbol in motif name	Name	Matches
N	Any base	A C G T
X	Any base	A C G T
V	Not T	A C G
H	Not G	A C T
D	Not C	A G T
B	Not A	C G T
M	Amino	A C
R	Purine	A G
W	Weak	A T
S	Strong	C G
Y	Pyrimidine	C T
K	Keto	G T

Motifs are short recurring patterns in gene sequences, and they have biological functions that include being promoter regions, or enhancer regions. Motifs recruit regulatory proteins and transcription factors to their target gene. Therefore, genes with common motifs likely have common regulators. Elucidating what these regulators are may be key in determining candidate cis-limited genes. STREME identifies novel motifs in a set of given genes, but does not give any further information about the motifs; what binds them, what they have in common with other motifs. The package Tomtom (Gupta *et al.*, 2007), part of MEME suite, can be used

to compare novel motifs against a database of known motifs, to identify potential transcription factors that may bind the motifs.

1.8.6 Micro-RNA binding sites

microRNAs are short non-coding RNA made up of 18-23 nucleotide molecules (Godden and Immler, 2022) that regulate genes post-transcriptionally. They inhibit translation and destabilise mRNA. Most of the functions of miRNAs are unknown but they are predicted to regulate up to 30% of all protein coding genes (Wienholds and Plasterk, 2005). Mouse models that do not produce miRNAs have shown developmental defects, suggesting that many miRNAs are essential for differentiation and maintenance of tissues.

miRNAs expression is most likely to be affected by environmental stressors. Altered expression of miRNAs as a result of environmental stressors is known to lead to post-transcriptional gene expression changes, as well as potential degradation or mRNA and therefore downregulated gene expression. The developing spermatids from each individual are exposed to the same environmental stressors, as they develop syncytially. Regardless of whether they are X bearing or Y bearing, there should not be a significant difference in the microRNA binding sites. This means that looking at the sample as a whole, there should be no overall difference in miRNA binding sites found in the protein coding genes upregulated in X bearing spermatids, and the protein coding genes upregulated in Y bearing spermatids.

1.8.7 Cis-limited genes

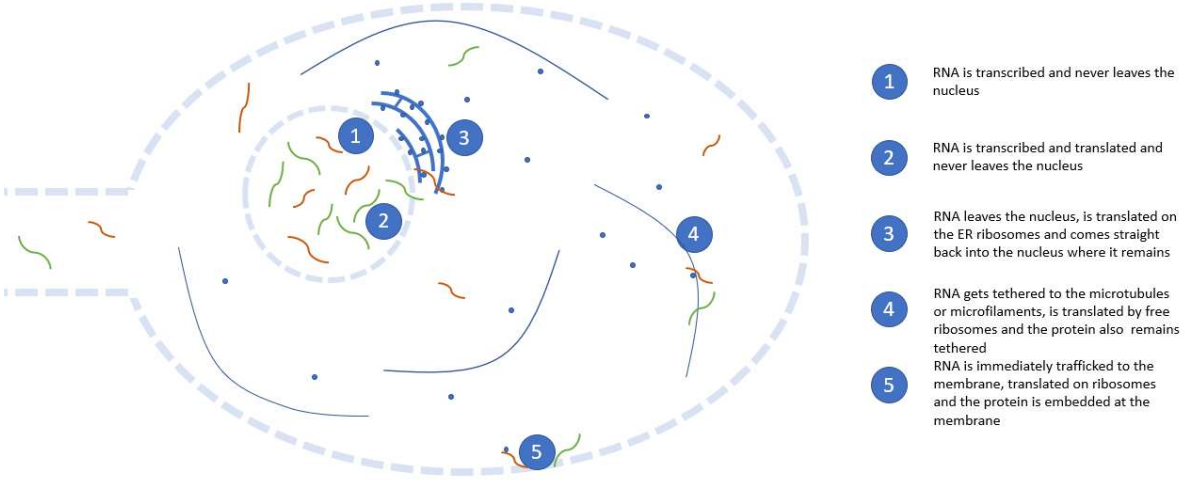
The majority of gene transcripts are able to be shared across the cytoplasmic bridges of sperm cells developing in syncytia, until the spermatid cells undergo spermiogenesis. At this stage, transcripts can no longer be shared but it means that transcripts originating in one spermatid may have migrated to an alternative spermatid and that is where it will remain to be translated into protein. Some genes, however, evade this process. Cis-limitation of genes describes the phenomenon by which gene transcripts are not shared between developing spermatids: they remain tethered in the cell they originated from by some mechanism. Cis-limited genes are important to this study because any genes that lead to functional differences between X and Y bearing sperm must be cis-limited. This is because, if a functional difference is seen in all X bearing sperm vs all Y bearing sperm, there is a protein or gene leading to this functional difference and if that gene transcript had been able to be shared across cytoplasmic bridges, there should be no difference between the X or the Y bearing sperm. Any differences seen must be caused by gene transcripts that are able to remain either in the X or the Y bearing sperm cells only – i.e., the genes are cis-limited. Cis-limitation does not necessarily mean that the genes cannot act in trans within the same cell. *Slx/Slxl1/Sly* likely act in trans and regulate autosomal factors.

1.8.7.1 Mechanisms of Cis-Limitation

There are a number of ways in which genes may be cis limited. For genes translated early in spermatogenesis, for the gene to be fully cis-limited, both the gene transcript and resulting protein need to remain compartmentalised and unable to be shared

across cytoplasmic bridges. It is proposed that mRNAs may bind and remained tethered to the endoplasmic reticulum, or to cytoskeletal proteins, and the resultant proteins may be directed straight to the plasma membrane after translation. However, for genes where the mRNA is not translated until after the cytoplasmic bridges have separated in spermatogenesis, just the mRNA needs to remain compartmentalised, so these transcripts likely remain within the nucleus. For this reason, it is important that any transcripts that are not long-noncoding mRNA are considered to potentially bind common micro RNAs or transcription factors, as they may be able to leave the nucleus.

Figure 11



Potential mechanisms of cis-limitation

1.8.7.2 Known cis-limited genes

There are two already known cis-limited genes: *Smok-2b* and *Spam-1. Smok-2b* encodes a sperm motility kinase 2B, which enables protein kinase activity and acts within protein phosphorylation. Mouse SMOK-2B sequences are coded for by genes located on mouse chromosome 17, consisting of three exons.

SPAM-1 is sperm adhesion molecule 1, a glycosyl phosphatidylinositol-linked sperm membrane protein that is dually expressed in testis and epididymis (Chen *et al.*, 2006). SPAM-1 is secreted in all regions of the epididymis in mammals, and SPAM-1 uptake significantly increases cumulus penetration, acting therefore as a marker of sperm maturation. SPAM-1 has many functions in the process of fertilisation and is found on mouse chromosome 6 (Martin-DeLeon, 2006). SPAM-1 transcripts are expressed only in the testis in spermatids, but the RNA and protein are present in both the male and female reproductive tracts in mice.

Both *Smok-2b* and *Spam-1* are known cis-limited genes, evading sharing across cytoplasmic bridges in sperm development. They are both autosomal genes, so under normal circumstances do not lead to differing sperm cell phenotypes; whereas the candidate cis-limited genes investigated in this study that are either controlled by or are on the sex chromosomes can plausibly lead to the functional sex differences seen in Yq-del mice.

1.9 Rationale for this thesis

In mammalian spermatogenesis, male gametes are produced at a natural ratio of 50:50 X-bearing sperm to Y-bearing sperm. X bearing vs Y bearing sperm have different uses and implications: for example, in the farming industry it would be

beneficial to have the ability preselect female calves for dairy or male calves for beef. Sex-selection in humans could also have profound effects with approved roles in medicine in the case of sex-linked diseases. However, being able to sort sperm requires detectable differences between the X and the Y bearing sperm, which is where the problem lies. Y-bearing sperm have a lower DNA content due to the size of the Y chromosome in comparison to the X, and currently the only proven sperm sorting method is to use a DNA-binding dye such as Hoechst (Ellis *et al.*, 2011). This is however a slow and expensive procedure and can cause considerable damage to the cells. There is ever-growing interest in being able to more efficiently and easily separate X bearing and Y bearing sperm for the future of agriculture and medicine.

A better understanding of the genes involved in producing healthy sperm is also a key driver of this research. As detailed in this introduction, human fertility is rapidly declining which is a large and ever-growing public health problem that will need to be addressed. In addition to this, understanding chromosomal compatibilities/incompatibilities implicated in cross breeds is just as important as being able to select for sex in agriculture. Cross breeding in farming can be used in order to select for the most favourable qualities of an animal, but if the crossbreeds then produce sterile offspring, this is not a sustainable food supply. Better understanding the genes involved in hybrid crosses is crucial for the future of agriculture where crosses of subspecies can be created that produce animals with the most favourable characteristics for sustainable food production, but that can also produce healthy, good quality sperm.

1.9.1 Overall aim of this research

The thesis focuses on the contribution of sex-linked vs autosomal genes to mouse sperm development. The project starts by looking at hybrid crosses of mouse subspecies and determining whether X-linked and autosomal incompatibilities between different subspecies, or X-linked and Y-linked chromosomal incompatibilities cause morphological, and functional, detriment to the sperm. This project then uses a novel pipeline to genotype cells as X or Y bearing based upon RNA sequencing data, which allows for differential gene expression analysis of the X vs Y bearing cells to gain a better understanding of the genes that are responsible for functional differences seen between X and Y bearing sperm.

1.9.2 Specific chapter aims

Given the rationale for the thesis, the project outline and specific aims for each chapter are:

Chapter 2: Materials and methods

Chapter 3: to assess the morphological differences of sperm and the sperm quality parameters between different hybrid mouse crosses

Chapter 4: to develop a pipeline to enable us to genotype sperm cells as X or Y bearing

Chapter 5: to perform differential expression analysis on RNA sequencing data from spermatids that are X bearing and Y bearing, to search for candidate cis-limited genes

Chapter 6: general discussion

Chapter 7: overall conclusion

2 Materials and methods

This thesis uses both published genomic datasets for genomic analysis as well as sperm samples that were imaged during the study for morphological analysis. All animal procedures for collecting the sperm samples imaged by the Skinner lab and collaborators were in accordance with the United Kingdom Animal Scientific Procedures Act 1986 and the University of Montana Institute for Animal Care and Use Committee (protocol 002-13) and were subject to local ethical review.

2.1 Morphometric analysis of mice sperm nuclei

Chapter three focuses on the morphometric analysis of mice sperm samples. The methods used in this chapter are described below.

2.1.1 Mice sperm samples used

Sperm samples were obtained from the Jeff Good species survey. All samples were collected between 2015 and 2017 by Erica Larson, Vanessa Stewart and Jeff Cloutier, University of Montana, and were fixed in 2% PFA and Methanol-Acetic Acid. See Table 5 for the list of strains used.

Table 5: mice strains used in study

Genotype	Strain	X	Y	Autosomal	Match/Mismatch
PWK	PWK	***Mus*** (PWK)	***Mus*** (PWK)	***Mus*** (PWK)	Match XY, match X/autosomal
CZE	CZE	***Mus*** (CZE)	***Mus*** (CZE)	***Mus*** (CZE)	Match XY, match X/autosomal
LEWES	LEWES	***Dom*** (LEWES)	***Dom*** (LEWES)	***Dom*** (LEWES)	Match XY, match X/autosomal
WSB	WSB	***Dom*** (WSB)	***Dom*** (WSB)	***Dom*** (WSB)	Match XY, match X/autosomal
STF	STF	***Spr*** STF	***Spr*** (STF)	***Spr*** (STF)	Match XY, match X/autosomal
SFM	SFM	***Spr*** SFM	***Spr*** (SFM)	***Spr*** (SFM)	Match XY, match X/autosomal
CC.PP	CZExPWK	***Mus*** (CZE)	***Mus*** (PWK)	***Mus*** (CZExPWK)	Match XY, match X/autosomal
WW.LL	WSBxLEWES	***Dom*** (WSB)	***Dom*** (LEWES)	***Dom*** (WSBxLEWES)	Match XY, match X/autosomal
PP.LL	PWKxLEWES	***Mus*** (PWK)	***Dom*** (LEWES)	***MusXDom*** *hybrid* (PWKxLEWES)	Mismatch XY, mismatch X/autosomal
PP.WW	PWKxWSB	***Mus*** (PWK)	***Dom*** (WSB)	***MusXDom*** *hybrid* (PWKxWSB)	Mismatch XY, mismatch X/autosomal
CC.WW	CZExWSB	***Mus*** (CZE)	***Dom*** (WSB)	***MusXDom*** *hybrid* (CZExWSB)	Mismatch XY, mismatch X/autosomal
LL.PP	LEWESxPWK	***Dom*** (LEWES)	***Mus*** (PWK)	***DomXMus*** *hybrid* (LEWESxPWK)	Mismatch XY, mismatch X/autosomal
WW.PP	WSBxPWK	***Dom*** (WSB)	***Mus*** (PWK)	***DomXMus*** *hybrid* (WSBxPWK)	Mismatch XY, mismatch X/autosomal
PP.LY	$PWK^{Y\text{-}LEWES}$	***Mus*** (PWK)	***Dom*** (LEWES)	***Mus*** (PWK)	Mismatch X/Y, match X/autosomal
WW.CY	$WSB^{Y\text{-}CZE}$	***Dom*** (WSB)	***Mus*** (CZE)	***Dom*** (WSB)	Mismatch X/Y, match X/autosomal
LL.PY	$LEWES^{Y\text{-}PWK}$	***Dom*** (LEWES)	***Mus*** (PWK)	***Dom*** (LEWES)	Mismatch X/Y, match X/autosomal
PP.LL.PY	$PWKx(LEWES^{Y\text{-}PWK})$	***Mus*** (PWK)	***Mus*** (PWK)	***MusXDom*** *hybrid* (PWKxLEWES)	Match X/Y, mismatch X/autosomal
LL.PP.LY	$LEWESx(PWK^{Y\text{-}LEWES})$	***Dom*** (LEWES)	***Dom*** (LEWES)	***DomXMus*** *hybrid* (LEWESxPWK)	Match X/Y, mismatch X/autosomal
CC.PP.LY	$CZEx(PWK^{Y\text{-}LEWES})$	***Mus*** (CZE)	***Dom*** (LEWES)	***Mus*** (CZExPWK)	Mismatch XY, match X/autosomal
PP.CC.WY	$PWKx(CZE^{Y\text{-}WSB})$	***Mus*** (PWK)	***Dom*** (WSB)	***Mus*** (PWKxCZE)	Mismatch XY, match X/autosomal
WW.LL.PY	$WSBx(LEWES^{Y\text{-}PWK})$	***Dom*** (WSB)	***Mus*** (PWK)	***Dom*** (WSBxLEWES)	Mismatch XY, match X/autosomal
LL.WW.CY	$LEWESx(WSB^{Y\text{-}CZE})$	***Dom*** (LEWES)	***Mus*** (CZE)	***Dom*** (LEWESxWSB)	Mismatch XY, match X/autosomal

These mice represented all combinations of X-Y match/mismatch and X-autosomal match/mismatch. The six parental samples, PWK, CZE, LEWES, WSB, STF and

SFM are of three different subspecies: *musculus*, *domesticus* and *spretus*. The *spretus* subspecies is notoriously abnormal in shape and has low fertility, so for this study, the *spretus* samples have been disregarded, and research focused on the *musculus* and *domesticus* samples. The CC.PP and WW.LL hybrids were bred as a control; to reduce the impact of inbreeding that would occur from crossing two mice of the same strain on reproductive phenotypes, but to allow a *musculus* and *domesticus* control for comparison.

F1 hybrids with a mismatched X and Y were bred, for both reciprocals of the cross. It has already been documented and described that these are largely infertile (Larson *et al.*, 2022). These F1 hybrids are compared to the *musculus* or *domesticus* controls, to show the difference in morphology between non-hybrids and hybrids. In addition, hybrids between a *musculus* and *domesticus* with a *musculus* Y-chromosome introgression and hybrids between a *domesticus* and *musculus* with a *domesticus* Y-chromosome introgression were bred. These were bred to investigate whether adding back the correct Y-chromosome had any effect upon the hybrid – to determine whether it is the X and Y mismatch or the X and autosomal mismatch, or both, that cause morphological impacts on nuclei.

Crosses were also made where the autosomal background was a hybrid of two strains, but of the same subspecies; so either fully *musculus* or *domesticus*, but the Y-chromosome introgression did not match the X chromosome and autosomal background. Here, the X and autosomal background matched but the X and Y chromosomes did not. This allowed for comparison with the hybrids that had a Y chromosome introgression where the X and autosomal background was mismatched, as well as the X and Y background (see Table 5).

2.1.2 Mice sperm parameters assessed

Reproductive phenotypes assessed included testis weight, sperm count, sperm motility and seminal vesicle weight. These were assessed at the University of Montana by collaborators. Testis weight and sperm count tend to be very closely linked in mice and are deemed to be a better measure of fertility than other factors such as seminal vesicle weight which varies in response to testosterone levels. All males with their reproductive phenotypes assessed were sacrificed at or after 60 days old. A Makler counting chamber and light microscope were used for the sperm motility assessment and sperm counts. Full details of the reproductive phenotype analysis protocol can be found in the materials and methods of a study carried out by collaborators in 2007 (Good, Handel and Nachman, 2008).

2.1.3 Preparation of sperm samples for imaging

Approximately half of the samples had been previously imaged at the University of Kent by Claudia Rathje. For the remaining samples, 4µl of the sperm was taken and dropped onto a slide with 16µl of DAPI solution. This was covered with a 22x50mm coverslip and then stored in a slide box in the fridge, to prevent photobleaching of the DAPI stain. The DAPI takes some time to develop, so slides were prepared at least one day in advance of microscopy, to ensure the nuclei were stained to a suitable depth.

2.1.4 Microscopes and cameras used

The microscope used for the imaging was the Nikon Widefield Ti-E microscope, using the DAPI filter and the 60x magnification lens, set to 1.5x zoom, leading to an overall 90x magnification. The widefield was used on its optical setting, without the DAPI filter, at a lower magnification in order to find the sperm cells on the slide, and once the cells were identified, the magnification was increased and the DAPI filter switched on. Samples were imaged and the images ran through Nuclear Morphology Analysis software (Skinner, 2022).

2.1.5 Microscopy of the sperm samples

From the University of Kent, 41 different samples from 15 different strains were analysed, and run through the software. All images had a resolution of 15.2 pixels per micron. Another 36 different samples, from seven different strains were imaged at Essex with a resolution of 19.8 pixels per micron. Finally, seven more samples from two different strains needed imaging. These samples were stored at the University of Kent, so imaging was finished at the University of Kent on their microscope, with 15.2 pixels per micron.

Claudia Rathje imaged 15 different strains, 41 samples in all, and Eleanor Watson imaged 43 samples of nine different strains. There were 22 different strains overall, summarised in Table 6, which doubled up on the imaging of WW.LL, PP.LL and LL.PP strains. Each sample contained sperm from a mixture of between two and three different animals, which means that for every genotype, at least four and up to 48 biological replicates were used for analysis. Hierarchical clustering was performed on a sample of cells that were from the same mouse strain, but that had been imaged both by Eleanor Watson and Claudia Rathje. These cells clustered in

the same hierarchical cluster, showing that there was no discernible difference in the imaging techniques or imaging equipment that would skew the results. This eliminated any observer bias.

Table 6: number of nuclei imaged per strain, number of samples per strain and location of imaging

Genotype	Strain	Original Number of Cells	Curated Number of cells	Number of samples	Aged between (days)	Imaged At
PWK	PWK	651	623	2	172-249	Kent
CZE	CZE	907	872	3	73-120	Kent
LEWES	LEWES	735	680	3	63-131	Kent
WSB	WSB	740	692	3	67-102	Kent
STF	STF	747	725	3	65-149	Kent
SFM	SFM	893	803	3	79-84	Kent
CC.PP	CZExPWK	641	391	6	68-102	Essex
WW.LL	WSBxLEWES	581	467	2	66-106	Essex/Kent
PP.LL	PWKxLEWES	1369	870	7	65-192	Essex/Kent
PP.WW	PWKxWSB	782	722	3	70-106	Kent
CC.WW	CZExWSB	583	533	3	79-80	Kent
LL.PP	LEWESxPWK	1978	1583	9	74-171	Essex/Kent
WW.PP	WSBxPWK	817	807	3	70-119	Kent
PP.LY	$PWK^{Y\text{-}LEWES}$	646	638	3	65-171	Kent
WW.CY	$WSB^{Y\text{-}CZE}$	719	672	3	113-225	Kent
LL.PY	$LEWES^{Y\text{-}PWK}$	965	907	3	74-171	Kent
PP.LL.PY	$PWKx(LEWES^{Y\text{-}PWK})$	1696	847	12	92-178	Essex
LL.PP.LY	$LEWESx(PWK^{Y\text{-}LEWES})$	1081	650	6	86-133	Essex
CC.PP.LY	$CZEx(PWK^{Y\text{-}LEWES})$	575	401	3	64-69	Kent
PP.CC.WY	$PWKx(CZE^{Y\text{-}WSB})$	625	545	3	62-63	Kent
WW.LL.PY	$WSBx(LEWES^{Y\text{-}PWK})$	702	443	4	60-84	Kent
LL.WW.CY	$LEWESx(WSB^{Y\text{-}CZE})$	469	414	3	63-69	Kent

2.1.6 Checking concordance of cameras

Before the main study commenced, two of the mouse strains were used to practice and develop microscopy skills. This was vital to ensure concordance between the equipment available at the University of Essex, and the equipment used for imaging previously at the University of Kent. The resolution on the microscope was 0.09 microns per pixel (11.11 pixels per micron), and the aim was to image approximately 500 nuclei from each sample (1000 nuclei in total), allowing for the fact that during this training period, some nuclei would not be detected by the software due to being slightly out of focus, not stained well enough, and general less-than-perfect technique (as this was still the early stages of the research). The final count after running through the software was 822 nuclei detected from the 2 samples, ready for the development of a downstream analysis pipeline.

The images of these nuclei were converted from their original format exported by the microscope camera (JP2) into TIFF format, using a batch convert command in ImageJ. In addition to this, it was decided that the resolution on the images was too low, and the images would be best rescaled to maximise resolution. The pixels of the images were manually doubled with the creation of a plug-in using ImageJ. The plugin rescaled all the images in a folder to have double the pixels (yielding 22.22 pixels per micron). Following this step, the images were ready to be run through the nuclear morphology analysis software, where all the nuclear statistics could be easily exported as comma separated files for further downstream analysis in R (Team, 2018). At this stage, some initial comparisons were run with the images from the University of Kent. All images taken at the University of Kent had a resolution of 15.2 pixels per micron, and so this was something that needed to be considered when running images through the software, adjusting the scale accordingly.

The camera on the microscope at Essex was later upgraded and an LED light source was added. The new camera had a resolution of 50.44nm/pixel, giving image resolution of 19.826 pixels per micron, and the LED light source allowed for better detection of the DAPI. These modifications meant that the ImageJ plugin to double the resolution of the images was no longer required during analysis, and the only image manipulation required was the conversion of files from JP2 into TIFF format.

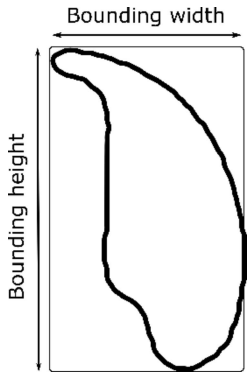
As described earlier, validation was needed that images captured at Essex on the Nikon Widefield, were in concordance with those images captured of the same strains at the University of Kent. The way in which concordance of results was checked for was by plotting the nuclei imaged from a single strain (PP.LL) but that had been imaged on the two different microscopes, onto a dimensionality reduction plot and colouring this by which microscope was used. The spread of colours across the plot would show whether or not the images taken from the 2 different cameras were clustering differently.

In addition to ensuring results were comparable across microscopes, optimum exposure times determined in the initial analysis at Essex needed to be checked against the updated camera at Essex, knowing the new camera was very different to the original camera used. The effect of automatic camera exposure time versus fixed exposure times on samples was tested. Auto-exposure had been used initially, and this was checked to ensure it was still appropriate. Approximately 100 nuclei were imaged each at 200ms, 350ms, 500ms and the auto exposure setting. All statistics were exported, and charts plotted of each parameter to assess the differences between exposure times, and determine the optimal exposure for imaging.

2.1.7 Analysis of images with Nuclear Morphology Analysis

All images taken from the University of Essex and the University of Kent were pooled for analysis. All images were run through Nuclear Morphology Analysis (Skinner, 2022), and then the nuclei were curated to ensure that only those correctly detected made the final sample. The software was developed to analyse nuclear shape (Skinner *et al.,* 2019). It uses a detects objects within images from microscopy using the Canny edge detection algorithm. The software then orientates and aligns images, and uses a modification of the Zahn-Roskies transformation of the nucleus outlines generates angle profiles for every nucleus detected. The angles are measured at the interior angle between one point and its two neighbours at every 0.05 of the perimeter of the nucleus, but at every point around the nucleus perimeter, which leads to multiple overlapping traces. This creates a set of angles which are then mapped onto a graph to generate a trace, the angle profile. The angle profiles are generated for each nucleus detected, and then a median trace is generated based on each individual trace. This allows the median trace to be projected back onto individual profiles, to identify landmark positions in individual traces, such as the hook, and the tail attachment point. Local minima and local maxima on the angle profile correspond to corners within the object.

Figure 12



Demonstrates what is referred to as the bounding height and bounding width of a nucleus

R Studio was used to compare various parameters including area, perimeter, bounding height and bounding width, and hook length – the typical characteristic component of a mouse sperm nucleus. Figure 12 shows a description of bounding height and bounding width of a nucleus.

Nuclear Morphology Analysis provided comma separated files containing all data statistics, and this was imported into R Studio. Using the 'dplyr' package, just the information required for this analysis was selected. In this case, a number of morphological parameters were chosen including length of hook and width of nucleus. The 'ggplot' package was used to produce different violin plots and box plots to show the distribution of the data. ggplot allows the titles, axes labels, legends and sizes of the plots to be manipulated easily within one function.

2.2 RNA Sequencing Analysis

Chapter 4 and 5 use previously published single cell genomic data to determine differences in gene expression in X bearing vs Y bearing sperm cells. The methods used for downloading and analysing the single cell data, and the methods for distinguishing X vs Y bearing sperm cells are described here.

2.2.1 Downloading FASTQ Files

EMBL's European Bioinformatics Institute (EBI) stores genomic data from published studies. Genomic data for each cell are given as SRR identifiers, which can be used to download FASTQ files. For this download stage, the fastq-dump command was used on CERES, the University of Essex's high performance computing cluster.

Fastq-dump requires specification of a directory for the output files as well as each input file, in this case the SRR codes from the published studies taken from EBI.

The list of SRRs were provided as an input to a download wrapper, produced in conjunction with a download script. This wrapper script allowed for iteration over each SRR in order to download the FASTQ file for each SRR given, meaning many FASTQ files for many cells can be downloaded simultaneously. To ensure easy debugging at any stage, a log was kept of each stage of the pipeline, which was inbuilt to each script. The script was written to search for FASTQ files before downloading them, and to log this stage (FASTQ file present or not), to ensure that downloads were not duplicated – saving time and storage.

2.2.2 FASTQC Scores and Trimming

Following the download of the FASTQ files, FASTQC was used to obtain quality scores for the reads. Following this, a trimming tool was run on the FASTQ files, in order to remove any adapters and low-quality sequences. Since the sequence reads for all published datasets used had already been trimmed, when running the trimming tool there was no difference seen in the length or quality of the reads before vs after trimming. However, the trimming tool was still run, to ensure that all adaptors were definitely removed, and in future this could be an important step because not all datasets will be pre-trimmed. The trimming tool used was trim-galore (Krueger, 2015) and the quality read was assessed using FASTQC. In this instance, no adapters were removed during the trimming as they were all pre-trimmed.

For the FASTQC analysis, it was essential to specify the output directory, the FASTQ files to input, and the flag “—nogroup” was used to prevent grouping of

bases past read length of 50bp, so that the reports produced show data for every base in the read, rather than grouping them together for quality assessment.

Trim-galore (Krueger, 2015) is a tool which expects four arguments to be specified: the FASTQ file for read 1, trimming directory for output, report directory for FASTQC, and FASTQ file for read 2 (this is optional and for paired end reads only - which were used in the Chen and Bhutani datasets). This trimmed the FASTQ files and then ran FASTQC analysis on them, which allowed comparisons of the FASTQC from before and after trimming with trim galore.

2.2.3 BedTools was used to create a custom gene transfer format file (GTF file) that contains information about introns, as well as exons

FASTQ files then needed to be aligned to the mouse genome. Reads were aligned using a tool called STAR (Dobin *et al.*, 2013). However, before running STAR, it was essential to optimise the pipeline to allow for analysing intronic data, which was an important part of this analysis. Downstream analysis would include using the featureCounts programme on data to count the introns and exons within the sequencing data. For this, an annotation file that specifies both intronic and exonic regions was required. Standard GTF files do not have introns specified within them, so modification of GTF file was needed to produce a file with introns listed as a feature.

This was achieved firstly by downloading the standard mouse GTF file from Ensembl (Yates *et al.*, 2020) version GRCm39.109, and then by creating a BED file of all of the introns, using Bedtools (Quinlan and Hall, 2010). This was accomplished by

subtracting the exonic regions from the genic regions, both of which are labelled in the freely available GTF annotation file. The subtractBed utility was the tool to achieve this. The output given was a GTF annotation file with genes, transcripts and exons labelled, and a BED file which gave chromosome number and coordinates of the introns, calculated from the exonic and genic regions. This BED file containing intronic data was used to add in intronic regions to the original, standard GTF file.

An R script (Team, 2018) was written to loop over each row of the GTF file, and to compare with each intron from the BED file, to see if there was a match – if the intron coordinates were between the coordinates of the GTF file, then the GTF file was labelled with the feature as 'intron'. This process was sped up by only comparing the files which had matching chromosomes, and skipping over the rows in the GTF annotation file labelled as 'genes', as otherwise the process would have been very time intensive.

Finally, the two GTF files, the original from Ensembl, and the one that was created with the introns labelled, were merged and sorted in BASH using the 'cat' and 'sort' commands. This final GTF file that was created had all introns and exons labelled, and was ready for use with the Subread package featureCounts command (Liao, Smyth and Shi, 2014) to identify all gene features.

2.2.4 STAR Analysis

With the annotation GTF file optimised to include intronic data, alignment could commence using STAR (Dobin *et al.*, 2013); a 2-stage process that first requires the generation of a reference genome using annotation files (in this case, the GTF file that had just been created). The STAR documentation recommends using either

Ensembl or Gencode – in this case, Ensembl was used for the genome sequence files and for the annotation file that was then adapted for use with intronic data (Yates *et al.*, 2020)

The six basic options in order to generate a STAR genome index are the number of threads, the run-mode, the genome directory path, the path to the FASTQ files, the path to the annotation files and the overhang read length. The number of threads defines how many threads are used for genome generation and needs to be set to be equal to the number of available cores on the server node; eight was chosen for this analysis – sufficient to allow for fast enough analysis. The genomeGenerate option for run mode was selected, to run the genome index generation. The overhang read length was 149, which is equal to the length of the reads subtract 1. The "–limitGenomeGenerateRAM" flag was also used to specify the required memory for the task.

For the mapping, the basic options required are the number of threads, the path to the genome directory and the path to the sequence reads. For paired ends, two different comma separated lists were provided, for read one and read two. The other advanced options used were the out-file name prefix, the out-file temporary directory, and the output type, specified as BAM, sorted by coordinate. Default (standard) output SAM attributes were used, and the "–outSAMunmapped" flag specified as 'within', meaning that unmapped reads were outputted within the main SAM file. The "-outFilterScoreMinOverLread" was set as 0.3 (the alignment will only output if its score is higher than or equal to this value, normalised to the sum of read lengths in paired end reads). "-outFilterMatchNminOverLread" was also set as 0.3 (the alignment will only output if the number of matched bases is higher than or equal to this value, normalised to sum of read lengths in paired end reads).

2.2.5 Feature Counts

As well as the BAM files, a list of genomic features is required as an input to the program. As described above, a GTF (gene transfer format) file was created, which contained all of the features. featureCounts has some inbuilt GTF annotation files, but they do not label intronic regions, and the standard annotation files available from Ensembl also do not contain intronic regions.

There are a number of arguments used by the featureCounts program, whereby one can specify the GTF files, and the BAM input files as well as the features to be counted. A wrapper script and several other scripts containing the loop of files to submit to the wrapper were created. The dataset was analysed as a whole, as well as on a cell-by-cell basis, and by developmental stage. Different wrapper scripts were used to look for exons, introns, or all genomic features.

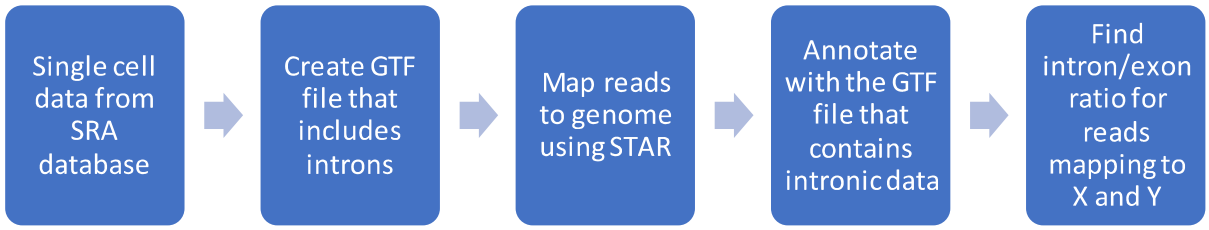
2.2.6 Analysing output in R to classify cells as X or Y bearing

The featureCounts comma separated output files with Gene-ID, chromosome number, feature type and the coordinates of the read were analysed in RStudio (Team, 2018).

The initial analysis of the Chen *et al* dataset (Chen *et al*., 2018) was to understand which cell types were mapping better than others, to find the right downstream tools for analysis and to check that the outputs were in line with what would be expected from the literature.

When intron and exon matrices were generated from the two larger datasets (Hermann *et al*;Bhutani *et al*), the main downstream analysis in R was to classify the cells as X bearing or Y bearing.

Figure 13



Basic workflow of analysis

With the exonic and intronic read counts data, a calculation was needed that could normalise the intronic reads per gene but that could compare the number of intronic X transcripts to the rest of the cell transcripts. Comparing X intronic transcripts to Y intronic transcripts was the obvious choice for genotyping as X bearing vs Y bearing, however the Y chromosome is so much smaller than the X that this may not give reliable information. Therefore, one calculation that was used was comparing the intronic/exonic read ratio for the reads to the X chromosome with the intronic/exonic read ratio for the hits to the Y chromosome: [log2(Xintronic) + log2(Yexonic) – log2(Xexonic) – log2(Yintronic)].

Another one, looking at the autosomes rather than the very small Y chromosome was comparing the intronic/exonic read ratio for the reads to the X chromosome with the intronic/exonic read ratio for the hits to the autosomes: [log2(Xintronic) + log2(Autosomal-exonic) – log2(Xexonic) – log2(Autosomal-intronic)].

The exonic read counts normalised the intronic reads, and the log normalising meant that even a small difference seen in ratios accounts for a very large difference in number of X transcripts per cell. The cells with the higher ratios therefore would have exponentially higher numbers of X transcripts compared to the cells with the lower ratios and could therefore be considered to be X bearing.

2.3 Differential gene expression analysis

Once cells were able to be classified as X or Y bearing, differential gene expression analysis was run on the two groups in order to compare gene signatures. This is the focus of chapter 5, and the methods are described below.

2.3.1 SCDE was used to generate a list of differentially expressed genes for X and Y bearing spermatids

SCDE is used for differential gene expression analysis. It takes the direct output counts matrix from featureCounts and the first step is to clean the data using the command ‘clean.counts’. Here, the minimum number of genes detected in a cell are defined, the minimum number of reads per gene are defined and the minimum number of cells a gene must be seen in are defined. This ensures that any extreme outlier genes or cells are filtered out. Next, an error model is fit to the data; the cells

are compared to a subset of genes with consistent expression, and then a Poisson-NB mixture model is fitted. The individual cell error models are used to test for the differential expression between the two defined groups, in this case X and Y. One of the most important steps in SCDE analysis is the batch correction. The need for batch correction arises as a result of genomic data being produced in batches, due to practicalities of collecting the data. Batch effects need to be addressed by normalising the data, and this is a built-in function of the SCDE package.

The SCDE package allows filtering of cells and genes based on certain parameters. The parameters specified here were a minimum of 1800 cells detected in a cell, at least ten reads per gene detected and at least 20 cells for each gene to be detected in. 250 genes were given as an input: 132 X-bearing and 118 Y-bearing.

Dimensionality reduction and hierarchal clustering of the exonic read counts was used to split the cells into groups, from the earliest stage spermatids through to the latest, condensing spermatids. These groupings were the batches used in SCDE. This was important, since some genes are expressed at different times throughout development. The clusters from dimensionality reduction represent different developmental stages.

This differential gene expression analysis of the entire sample of spermatids gave genes upregulated and genes downregulated. The upregulated genes were the genes that were differentially expressed in X bearing spermatids, and the downregulated genes were differentially expressed in Y bearing spermatids. These differentially expressed genes (DEGs) could be separated into sex-linked and autosomal. Significantly differentially expressed genes were classed as those with a log fold change (log2fc) > 2.0 and unadjusted $p<0.05$. If the adjusted p-value were to

be used this would be more conservative as there would be fewer false positives in the results. However, this analysis was only looking at genes that met very specific criteria. For this reason, the unadjusted p-values could be used, since any false positives would be filtered out along the pipeline.

2.2.2 Gene enrichment analysis pathways allow analysis of collective gene functions

The gene lists upregulated in X bearing and Y bearing spermatids are given as inputs to gene expression analysis pathways. KEGG, STRING and David are all used to investigate common pathways in the functions of the upregulated genes.

2.2.3 The STREME package in Meme Suite (Bailey, 2021) was used for common motif identification in differentially expressed genes

The autosomal genes that were differentially expressed in X-bearing and Y-bearing spermatids were split into two groups, the long-noncoding RNAs and the protein coding RNAs. Long non-coding RNAs were of particular interest in searching for common regulatory regions. First, motif analysis was run on the entire gene list and then just the long non-coding genes. Motifs found in the DEGs are potential transcription factor binding sites. Long non-coding RNAs are widely expressed with key gene regulation roles, so finding motifs in these may shed some light on which other key genes they regulate, which could lead to functional differences in X bearing and Y bearing spermatids. However, there are likely to be many layers of

mechanisms involved in regulation of these differentially expressed genes, and so the motif analysis was also applied to the whole gene list, including the protein coding genes. For the motif analysis of the long non-coding genes, MEME was used as opposed to STREME, because there were fewer than 50 input sequences (Bailey and Elkan, 1994). Any transcription factors identified in the long non-coding genes are more likely to be involved in the cis-limited pathway.

The autosomal genes shown to be enriched in X-bearing cells were the first to be ran through the STREME algorithm, and then the autosomal genes enriched in the Y-bearing cells were run through. Only the promoter regions of the genes were used as an input, as this is where transcription factors bind. Transcription factor binding sites can be located very close to the transcription start site in gene promoter regions or can be found more distantly. The promoter region lies 5' to the gene sequence, so the inputs used for motif discovery was 1000 base pairs upstream of the transcriptional start site.

Common motifs were identified in the gene sequences of differentially expressed genes. Motifs could represent transcription factor binding sites or microRNA binding sites. An important consideration here was that, if the sequence input was the entire gene sequence, motifs discovered would be likely to be random repeat regions as well as potential transcription factor binding sites. Another important consideration is that these motifs may be miRNA binding sites, although one would expect microRNA binding sites to not differ between X bearing and Y bearing spermatids. Almost every single gene in the genome is controlled by transcription factors, with approximately 10-20% of animal protein coding genes bound by miRNAs (Martinez and Walhout, 2009).

2.2.4 Identifying the function of the enriched motifs using Tomtom

The enriched novel motifs identified in STREME were run through Tomtom to compare to known transcription factor binding motifs. Tomtom showed which transcription factors are likely to bind to the identified motifs, so that potential transcription factors that bind the differentially expressed genes in X and Y bearing cells could be identified.

2.2.5 Identification of common microRNA binding sites using seedVicious (Marco, 2018)

The steps involved in identifying common microRNA binding sites are: 1) identification of miRNA binding sites in each gene, 2) compiling a list of all miRNAs identified, 3) for each miRNA, counting how many genes they may bind. There is currently not a program that could do this all in one, but seedVicious (Marco, 2018) can identify miRNA binding sites in single genes from its default list of miRNAs, and can do pairwise comparisons of genes that contain these binding sites. SeedVicious finds canonical microRNA sites as well as sites that are very close to being target sites, which allows for analysis of population variation in microRNAs. The software is written in Perl and runs on 64-bit Unix computers, but the program can also be run on a dedicated web server.

The genes used in this analysis were the same as for the common transcription factor binding motif analysis; the autosomal differentially expressed genes, excluding those that were long non-coding, and the analysis was ran twice, once for the genes

upregulated in X-bearing cells, and once for the genes upregulated in the Y-bearing cells.

Standard SeedVicious output files show the microRNA, the transcript, the position of the last nucleotide of the target site in the transcript, and the type of target (Marco, 2018). This standard output file was generated for all genes in the differentially expressed gene lists, and a basic R script looped over the output file to total up the number of genes out of the input gene list that contained at least one binding site for the specified miRNA.

2.2.6 Known cis-limited genes were analysed for common motifs

Smok2b and *Spam1* were analysed common motifs. *Smok2b* and *Spam1* are the two already-known cis-limited genes, and common motifs could be used to identify shared transcription factor binding sites with the candidate gene list. The Meme suite of tools was used, but since there were only two sequences here for analysis, MEME, rather than STREME was used (Bailey and Elkan, 1994). The maximum motif width was set to the same as that of the default value in STREME, 15. STREME stops searching for motifs when too many consecutive, non-significant motifs are found, three unless otherwise specified. For MEME, parameters were specified to stop searching when the last motif has an E value of greater than 0.05, or when 1000 motifs have been found.

3 The use of high throughput nuclear morphology software to explore the impact of XY chromosomal incompatibilities and X-autosomal incompatibilities upon sperm nuclear morphologies and sterility across mouse strains

3.1 Introduction

This chapter makes use of the novel software Nuclear Morphology Analysis (Skinner, 2022) for analysis of sperm from a number of mouse cross types, to assess a number of morphometric parameters and investigate a link between sperm morphology, and the genetic basis of fertility. The automated software not only allows fast analysis of large amounts of data, but it provides quantitative statistics for further downstream analysis, rather than qualitative data. This allows the detection of small differences in morphology that would otherwise be undetectable if relying on the naked eye to assess. The software allows the analysis of thousands of nuclei in minutes, which is key to allowing data collection on a large enough scale to demonstrate how subtle morphological changes have large impacts on fertility.

This study investigates mouse sperm phenotypes, working in conjunction with collaborators at the University of Kent and University of Montana.

The early stages of imaging allowed development and honing of several skills that were critical for the more in-depth analysis. For example, running the morphometric software on the images required identification of clusters of shapes within the images from dimensionality reduction analysis (Van Der Maaten and Hinton, 2008), and the preliminary imaging allowed for experimentation with different types of dimensionality reduction analysis.

3.1.1 Specific aims

The main questions that this chapter aims to answer are:

(1) are there significant morphological differences seen in the sperm nuclei of mouse strains with X and Y chromosome incompatibilities?

(2) are there significant morphological differences seen in the sperm nuclei of mouse strains with X chromosome and autosomal incompatibilities?

(3) How well do traditional fertility assessment parameters such as sperm count, sperm motility, testis weight and seminal vesicle weight correlate with morphology? i.e., do poor sperm nuclear morphologies indicate poor sperm quality?

3.2 Methodology

Mouse sperm nuclei from each cross were imaged and then run through Nuclear Morphology Analysis (Skinner, 2022). Statistics were exported and analysed in R (Team, 2018). A Shapiro-Wilk test of normality was performed on the four main morphological parameters of each strain; area, bounding height, bounding width and difference from median. This was to determine whether parametric or non-parametric tests were to be used for downstream analysis. Parametric tests are favourable due to their higher statistical power but can only be used on normally distributed data. Although non-parametric tests generally have less statistical power, they become more accurate the larger the sample size. All groups had a large sample size of over 300, so in this case both parametric and non-parametric tests would have sufficient power. Parametric tests that could be used if the data were normally distributed

include paired t-tests and analysis of variance (ANOVA). Non-parametric tests that may be used downstream if data are not normally distributed include Kruskal-Wallis, Wilcoxon-rank, Mann-Witney U, and Friedman tests.

Simple plots comparing the basic statistics (area, body width, aspect ratio, difference from the median and hook length) were plotted, for all 24 strains.

Whilst *musculus* and *domesticus* strains both have a very similar overall shape, there are some subtle differences. Over time, the breeding of pure wild types in labs may have led to a degree of inbreeding which could lead to morphological defects. For this reason, crossing two different *musculus* or two different *domesticus* strains to produce a control hybrid is likely to be favourable, to ensure inbreeding had no confounding effect on morphology. The wild-type *mus* and *dom* were compared with the *musXmus* and *domXdom* control hybrids to ensure the controls were a suitable proxy for the wild-types, and control hybrids were used for further analysis. Wild types were then excluded from further analysis.

Dimensionality reduction and hierarchal clustering were performed on the global population of the samples in order to elucidate the relationship between strains based upon their morphologies. Dimensionality reduction was performed on every nucleus that passed the curation stage of the morphology analysis software, excluding the *spretus* samples as they have such poor morphology that even the wild types are subfertile and therefore were not used for hybridisation in this study. 15,285 nuclei, minus the 1528 *spretus* nuclei, meant that a total of 13,757 nuclei were involved in the dimensionality reduction. Various plots were then generated by plotting just the nuclei belonging to different sample groups, to compare their clusterings.

The dimensionality reduction was performed solely on the angle profiles of the nuclei, so any clustering patterns occurring were due to similarities and differences in angle profiles, rather than any other of the morphological parameters. This was because the angle profiles give the best overall measure of nuclear shape. Both tSNEs and UMAPs were generated, but UMAPs were deemed the best overall analysis method since a UMAP will preserve the global structure of the data much more accurately than a tSNE – this means that the closer two points are on a UMAP, the closer they are to each other. In a tSNE, the local structure is preserved well, leading to very accurate intra-cluster relationships, but two clusters next to each other on the plot may not be the closest in relation to each other in reality. On a UMAP, the further two clusters are on a plot, the more different they are in reality. For the purpose of this analysis, where focus is specifically looking at the relationship between different strains, the fact that UMAPs don't preserve the local structure of the data quite as accurately as a tSNE does not matter. The purpose of this study is elucidating the relationship between different hybrid crosses, and therefore global structure is more important to show the relationships between different strains, rather than within strains.

Because of the non-linear relationship between morphological factors and quality of sperm, to assess the relationship between the morphological parameters and the sperm assessment parameters, the morphological parameters needed to be normalised. Sperm with areas too large are abnormal as are sperm with areas that are too small; it is not the case that the larger the area, the more normal the sperm or vice versa. It is the deviation from the norm, in either direction which is more critical.

In order to normalise the morphological parameters and quantify the level of abnormality, the mean value for the two control crosses was calculated, and then the percentage deviation in either direction from the mean of the two controls was calculated for all other cross types. Spearman's correlation coefficient was used to determine correlation between morphology and sperm assessment parameters.

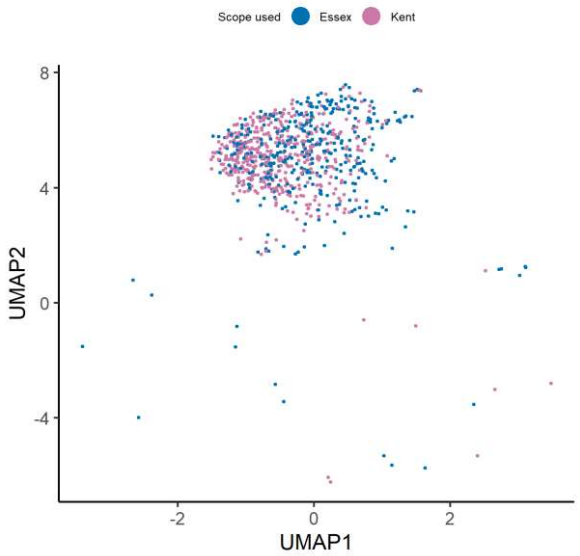
3.3 Results

3.3.1 Auto-exposure times were sufficient for robust analysis, both microscopes gave concordant results, and non-parametric tests were used

When plotting the UMAP of the PP.LL sample which had been imaged at both Kent and Essex, the majority of the points fell into one cluster, and there was an even distribution of the two colours across the cluster, showing that there was no detectable difference in the nuclei detected, despite to the fact that they were imaged on two separate cameras.

This allowed all the nuclei to be assessed together, without the need to take into account where they had been imaged. It was a

Figure 14

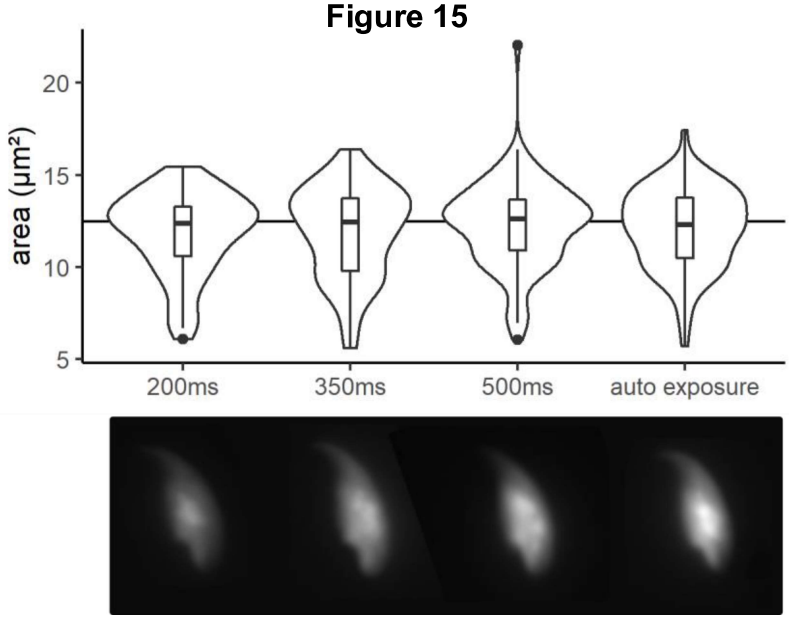


Dimensionality reduction plot of the PP.LL strain coloured by which microscope the nuclei were imaged on

sanity check to ensure that the microscopes used were not skewing the results and that results from either microscope would be comparable against each other.

Figure 15 shows that the nuclear area measured across the range of exposure times (200ms-500ms) remained constant, and the automatic exposure always fell within this range. Automatic exposure also remained consistent with the range of other exposure times, across all parameters. Figure 15 shows the area of the nuclei, with each exposure time and auto exposure being almost identical in terms of median, upper and lower quartiles and spread of values. The median line of all samples was plotted. Kruskal-Wallis tests confirmed that there was no significant differences between the area in square microns and the exposure time used to capture the images, with a test statistic of 1.0655 and a p-value of 0.7854.

Figure 15



Effect of different exposure times on the area of 100 sperm nuclei, with a sample DAPI stained nucleus for each exposure time, and the median line for the median area measured at all exposure times. Even though by eye the 500ms and the auto-exposure time are clearer than the 200ms/350ms nuclei, the area as calculated by Nuclear Morphology Software does not change by any meaningful amount.

From this it was concluded that automatic exposure is sufficient for robust analysis of the samples. Consistency of results across exposure times coupled with convenience of automatic exposure time meant auto-exposure was the chosen parameter.

The p-values for the Shapiro-Wilk tests performed on each parameter are summarised in Table7. Grey boxes indicate a non-significant p-value, which means the data are normally distributed. This is only the case for the bounding width of the CC.PP and the bounding height of the LL.WW.CY. Because of the low p-values across the rest of the data, the strains are not normally distributed, which means downstream statistical tests need to be non-parametric.

Table 7: p-values for the Shapiro-Wilk test of normality for each morphological parameter, per strain. Grey boxes indicate a normal distribution. Difference from median represents the variability within the dataset.

Subspecies	Genotype	Area (μm^2)	Bounding height (μm)	Bounding width (μm)	Difference from median

mus	**CC.PP**	2.624e-06	0.0017	0.3005	<2.2e-16
dom	**WW.LL**	3.282e-14	0.000127	2.217e-13	<2.2e-16
musXdom	**PP.LL**	0.01338	2.679e-09	<2.2e-16	<2.2e-16
	PP.WW	6.852e-11	9.146e-06	6.04e-11	<2.2e-16
	CC.WW	< 2.2e-16	0.007759	4.55e-10	<2.2e-16
domXmus	**LL.PP**	5.106e-16	4.265e-13	<2.2e-16	<2.2e-16
	WW.PP	8.656e-11	1.291e-10	0.006067	<2.2e-16
mus $^{Y\text{-}dom}$	**PP.LY**	2.804e-15	4.95e-08	4.597e-06	<2.2e-16
dom $^{Y\text{-}mus}$	**WW.CY**	8.572e-05	0.0006323	1.67e-06	<2.2e-16
	LL.PY	9.174e-15	<2.2e-16	0.001367	<2.2e-16
musXdom $^{Y\text{-}mus}$	**PP.LL.PY**	0.0005442	0.001638	2.419e-15	<2.2e-16
domX mus $^{Y\text{-}dom}$	**LL.PP.LY**	8.222e-08	9.215e-10	2.256e-06	<2.2e-16
musX mus $^{Y\text{-}dom}$	**CC.PP.LY**	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16
	PP.CC.WY	5.735e-09	0.0008346	1.226e-05	<2.2e-16
domXdom $^{Y\text{-}mus}$	**WW.LL.PY**	6.812e-08	<2.2e-16	<2.2e-16	<2.2e-16
	LL.WW.CY	3.438e-12	0.426	3.079e-16	<2.2e-16

3.3.2 Significant differences in morphological parameters were seen across strains

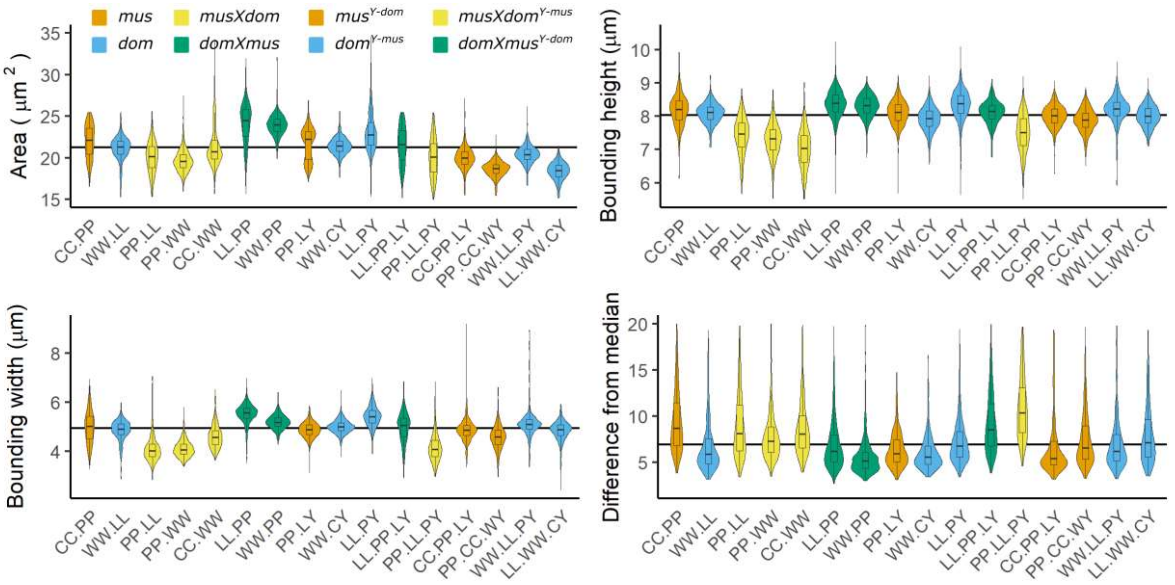
Bounding height is a measure of hook length; the larger the bounding heights, the larger the hook (refer back to Figure 12). In general, anything deviating significantly from the median line on the violin plots show an abnormality. It is well known that spretus strains have poor morphology and fertility; they have much lower bounding heights than the rest of the strains, showing a smaller hook length, and a more squashed nucleus.

The three *musXdom* hybrids (PP.LL, PP.WW and CC.WW) show smaller bounding heights and widths than the median. This shows a squashed morphology with shorter hook. In contrast the reciprocal crosses *domXmus* (LL.PP and WW.PP) have higher than median bounding heights and widths, showing a more elongated hook and more stretched out, larger nucleus.

Bounding heights and bounding widths for each strain are quite closely correlated; in general, the larger bounding height the larger the bounding width of the sample. This

also correlates with area. The F1 hybrids with Y-introgression (PP.LL.PY and LL.PP.LY) show a similar pattern to the F1 hybrids. The *musXdom* with Y introgression (PP.LL.PY) has lower bounding height and width than the median, like the *musXdom* hybrids. The *domXmus* with Y introgression (LL.PP.LY) has higher bounding height and width than the median, like the *domXmus* hybrids. The wild-types with Y introgression (PP.LY, WW.CY and LL.PY) and the control hybrids with Y-introgression (CC.PP.LY, PP.CC.WY, WW.LL.PY and LL.WW.CY) tend to sit closer to the median line than any of the hybrid strains. All strains show a fairly similar distribution of 'difference from median', with the *musXdom* strains (PP.LL, PP.WW and CC.WW)and the *musXdom* with a Y-introgression (PP.LL.PY) having slightly higher differences from median. This shows that these strains show a higher level of variability within their populations.

Figure 16

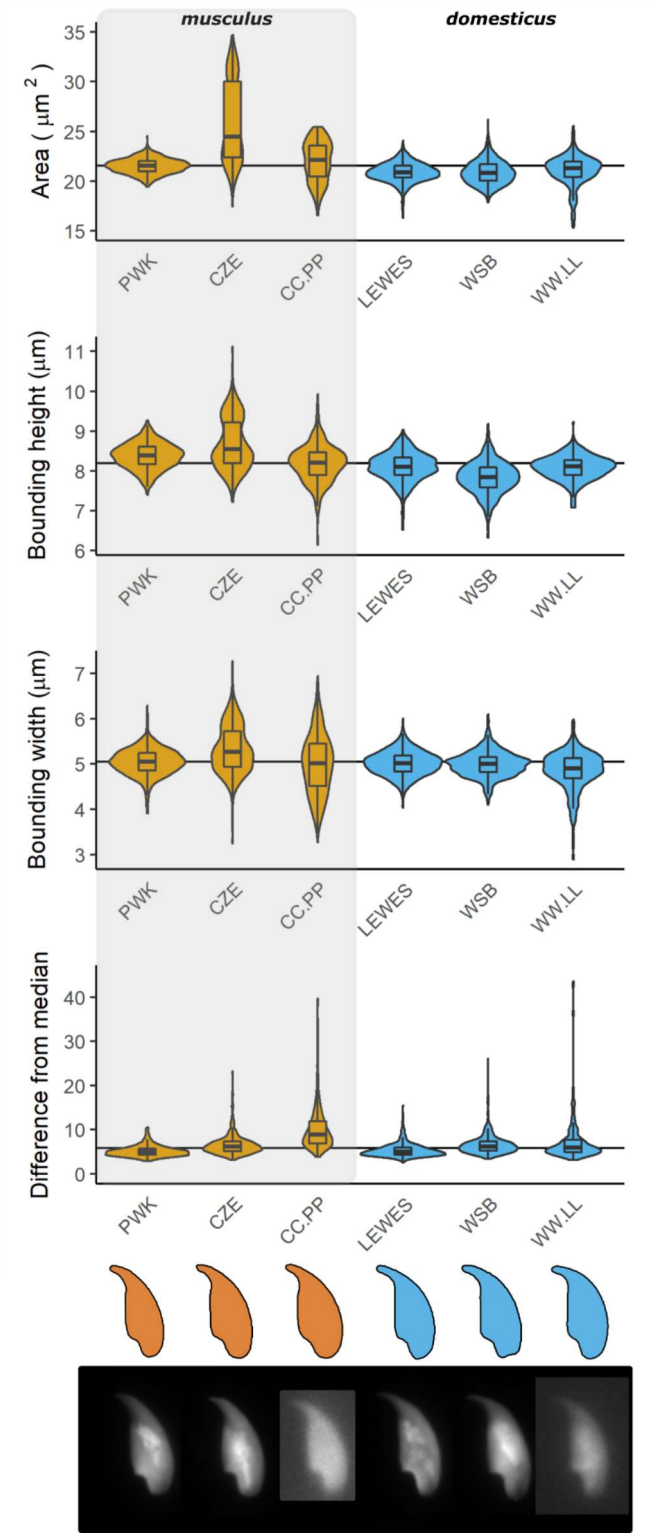


Area (square microns), bounding height (microns), bounding width (microns) and difference from median of all strains imaged. The orange (mus) and blue (dom) show the wild-types and the control crosses, and the wild-types or controls with a Y introgression. The yellow refers to the musXdom hybrids, and the musXdom hybrids with a Y introgression. Green is the domXmus hybrids and domXmus with a Y introgression. Box plots show the median, lower quartile and upper quartile of each strain, as well as the full spread of the data in that sample. Median lines show the median for all strains of each parameter.

3.3.3 Control hybrids provide a good substitute for the wild types whilst removing any negative morphological effect of inbreeding

It was concluded that the control hybridisation of both *mus* and *dom* led to suitable substitutes to the wild-type parental strains for comparative analysis. Whilst the control hybrids showed subtle differences to the wild-types, it is unknown whether the wild-types had any morphological abnormalities due to in-breeding, and the controls upheld the subtle differences between strains. For example, the control *mus* showed slightly larger areas, bounding heights and bounding widths than the control *dom,* as was seen in all the wild types (Figure 17). There was however a little more variation in the control *mus* than the wild type *mus* (Figure 17).

Figure 17



Area in square microns, bounding height (microns), bounding width (in microns) and difference from median of the 2 musculus strains, the F1 control musculus cross, the 2 domesticus strains and the F1 control domesticus cross, with the median line. Box plots show the median, lower quartile and upper quartile of each strain, as well as the full spread of the data in that sample. The musculus have slightly higher median areas, bounding heights and bounding widths than the entire group median. All strains show fairly similar spread about the mean. Consensus nuclei of each strain and DAPI stained images of a nucleus from each strain are shown.

The violin plots show that the control crosses tend to follow the same general trends as the wild-type – i.e., the *mus* tend to be slightly larger than the *dom.*

Table 8 shows the median values for morphological parameters between the wild type *mus* and *dom* strains and the F1 control hybrids. * denotes a significant difference (FDR-corrected Wilcoxon rank sum test $P < 0.05$) between F1 hybrid controls of different subspecies. ɣ denotes a significant difference between the wild type and all samples from the other subspecies. Grey boxes denote a significant difference between a subspecies and the F1 control of the same subspecies.

Table 8 shows that whilst there are in many cases statistical differences between the wild-types and the F1 controls of the same subspecies, there are statistically significant differences for all parameters between the two different F1 controls, and for the most part, each strain is significantly different to all other strains of the opposite subspecies. It can be confidently said that the F1 controls therefore can represent the overall *mus* or *dom* standard control morphologies, as they are different to each other but close to their respective wild-types, the violin charts show clear differences between the *mus* or *dom* subspecies and the F1 controls conform to the same pattern, but the F1 controls remove the negative effect of any inbreeding within the wild-types

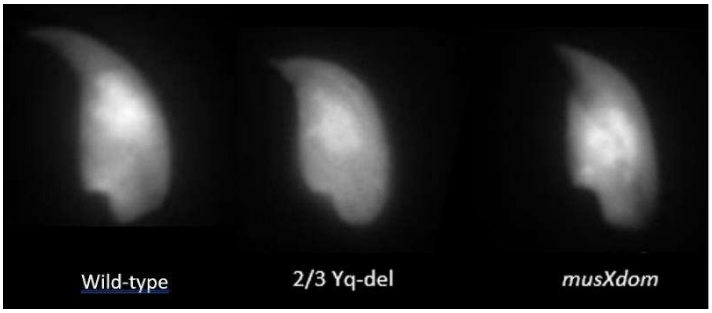
Table 8: median values for morphological parameters by strain.

	musculus			*domesticus*		
	PWK	CZE	CC.PP	LEWES	WSB	WW.LL
Area (square microns)*	21.52 ɣ	24.50 ɣ	22.11 ɣ	20.86 ɣ	20.81 ɣ	21.30 ɣ
Bounding height (microns)*	8.39 ɣ	8.54 ɣ	8.21 ɣ	8.10 ɣ	7.84 ɣ	8.11 ɣ
Bounding width (microns)*	5.05 ɣ	5.26 ɣ	5.02	5.01	5.00	4.90 ɣ
Difference from median*	4.95	6.13	8.88 ɣ	4.97	6.20	5.86

3.3.4 F1 hybrids show distinct morphological abnormalities, similar to those seen in Yq-del sperm

The F1 hybrids showed two classes of abnormality dependent on the cross direction. Whilst pure *dom* tends to have slightly lower area, bounding height and bounding width than pure *mus*, *musXdom* (*mus* female crossed with *dom* male) tends to have much lower area, bounding width and bounding height than the reciprocal hybrid cross *domXmus* (Figure 19). The *musXdom* nuclei showed a similar shape to the 2/3 Yq-del sperm, being much shorter with a blunter hook.

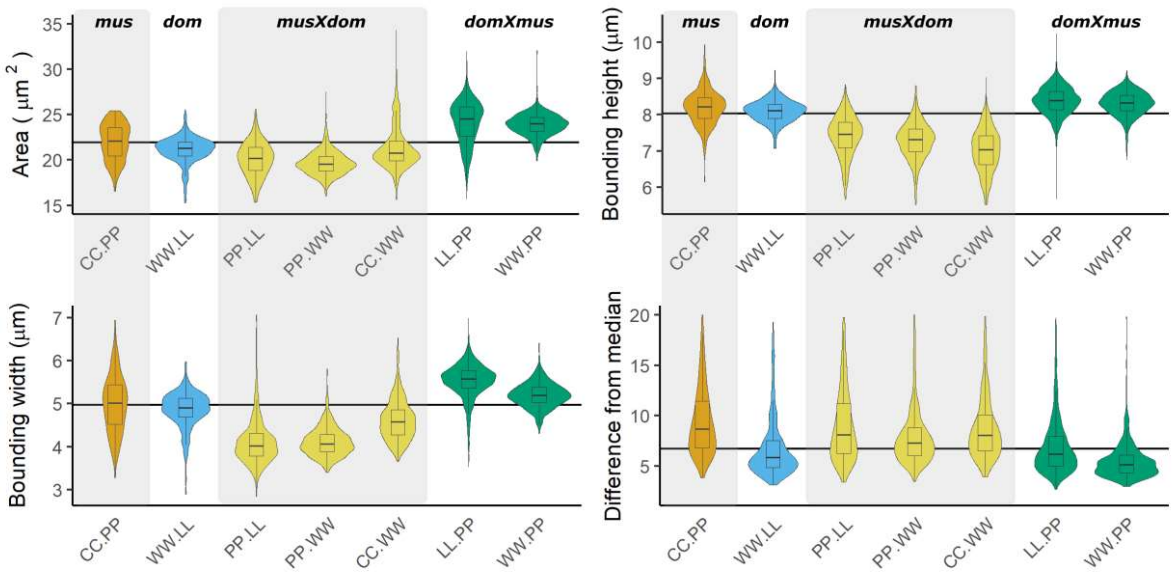
Figure 18



Wild-type *mus* sperm, 2/3 Yq-del MF1 mouse sperm and *musXdom* sperm

There is not a great difference in variabilities, shown by difference from the median (Figure 21), which is to be expected from a large sample. However, the *musXdom* seems to have slightly higher variation within the samples than *domXmus.*

Figure 19



Area (square microns), bounding height (microns), bounding width (microns) and difference from median of the musculus control, the F1 musculusXdomesticus hybrids, the domesticus control and the F1 domesticusXmusculus hybrids, with the median line for each measure shown. Box plots show the median, lower quartile and upper quartile of each strain, as well as the full spread of the data for each sample. The musXdom have slightly lower medians than the entire group median, with the domXmus having slightly higher medians than the group median. The controls are very close to the median line. Both hybrid crosses deviate from the median, in opposite directions. Difference from median doesn't vary much by sample, suggesting all samples have a similar level of variation.

Pairwise Wilcoxon tests were run with FDR corrections, on the median values for morphological parameters in the control hybrids and the F1 hybrids. The results are summarised in Table 9, where the median values are displayed. * indicates the hybrid cross differs from the X background, for example the *musXdom* differs

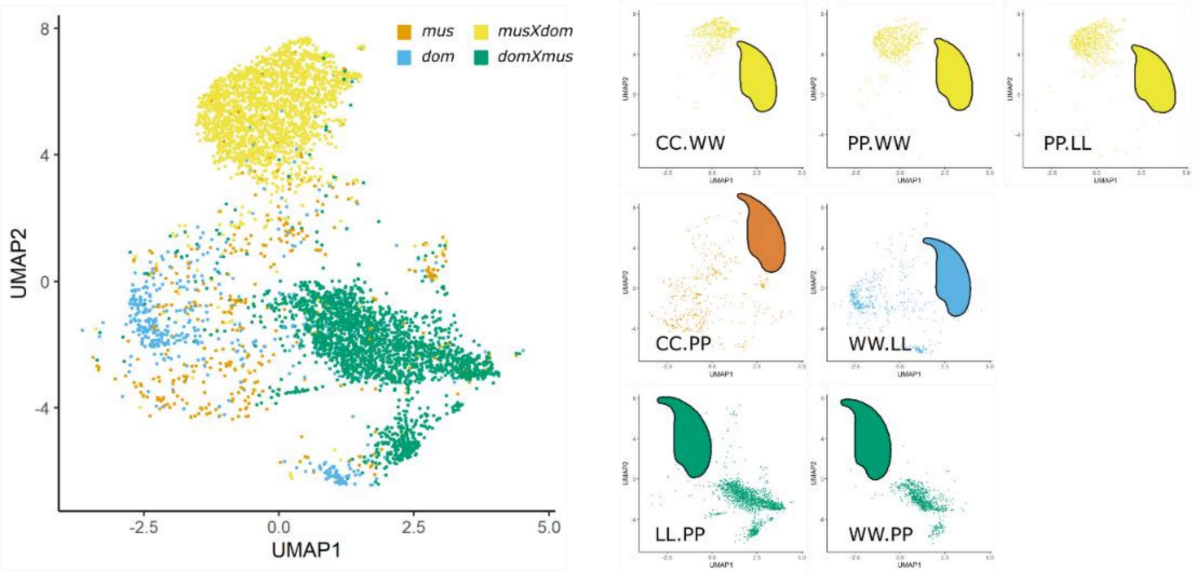
significantly from the *mus.* ɣ indicates the hybrid cross differs from the Y background, for example the *musXdom* differs significantly from the *dom.* Grey boxes indicate the hybrid cross differs significantly from all other reciprocal hybrid crosses.

Table9: median values for morphological parameters

	musculus	***musculusXdomesticus***			***domesticus***	***domesticusXmusculus***	
	CC.PP	PP.LL	PP.WW	CC.WW	WW.LL	LL.PP	WW.PP
Area (square microns)	22.12	20.14 *ɣ	19.55 *ɣ	20.75 *ɣ	21.31	24.52*ɣ	23.98*ɣ
Bounding height (microns)	8.21	7.46*ɣ	7.32*ɣ	7.02*ɣ	8.11	8.39*ɣ	8.32*ɣ
Bounding width (microns)	5.02	4.02*ɣ	4.07*ɣ	4.58*ɣ	4.90	5.58*ɣ	5.19*ɣ
Difference from median	8.88	8.22*ɣ	7.31*ɣ	8.10*ɣ	5.86	6.22*ɣ	5.17*ɣ

The Wilcoxon test with FDR corrections show that each hybrid cross differs significantly from both *mus* and *dom* as well as differing significantly from the reciprocal crosses. To further investigate how nuclear morphologies related to each other, dimensionality reduction plots of the controls and F1 hybrids shows that the controls form a layered band across the middle, with the *dom* sitting above the *mus*. The *musXdom* hybrids then form a tight cluster below the middle band, and the *domXmus* form a layered cluster above the middle band. It shows very clearly the difference in reciprocal crosses from both the parental strains and the opposite cross.

Figure 20



UMAP dimensionality reduction plot coloured by strain, showing the control F1 hybrids across the middle. The interspecific sterile F1 hybrids then diverge off to the top or bottom, with the musXdom at the top and the domXmus at the bottom. Left panel shows subspecies and right-hand side shows where individual genotype within the subspecies lies.

The *domXmus* has a more curved hook and is longer than the controls, as shown in Figure 20, and these form the top cluster, whilst the bottom cluster, *musXdom* is squashed, short and has a blunted hook.

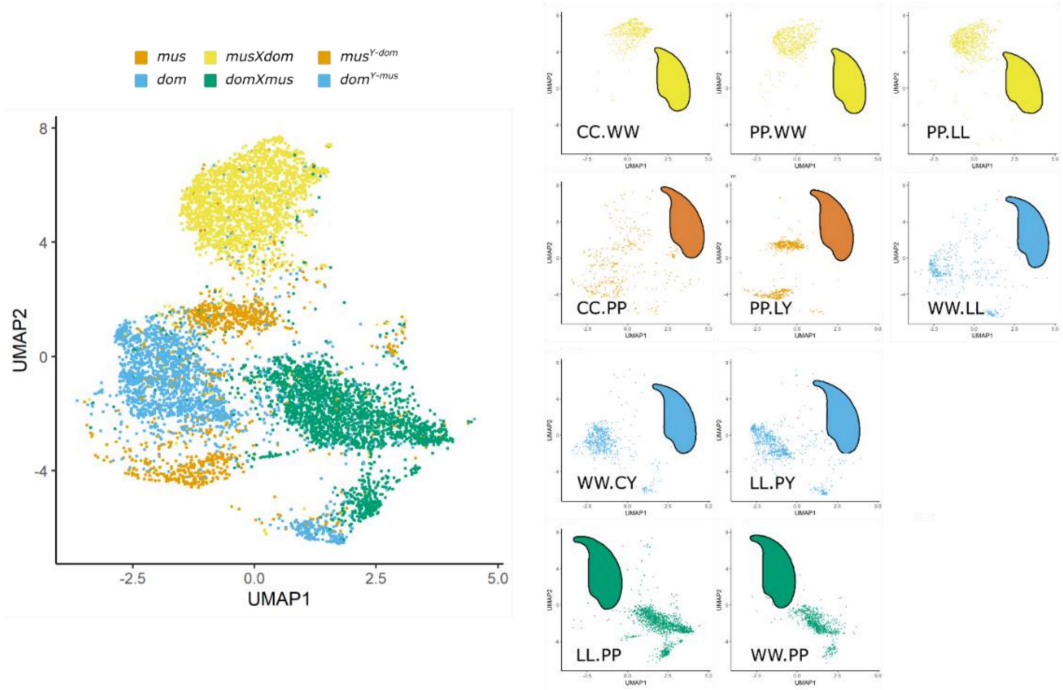
3.3.5 Y-introgression hybrids are more similar in shape to the control hybrids, whilst F1 hybrids differ

The next set of comparisons look at not just the control hybrids and the F1 hybrids, but also *mus* with a Y-introgression of *dom,* and *dom* with a Y-introgression of *mus.* In this case, the controls have X, Y and autosomal background from the same

subspecies, hybrids have the X and Y from different subspecies and the autosomal background is a hybrid of the two, and the introgression line has the X and the autosomal background from the same subspecies, but the Y comes from a different subspecies. Here, the Y-introgressions nicely overlay onto the middle band of the controls, rather than the diverging clusters of the hybrids.

This overlay onto the middle band with the controls shows the Y-introgressions are more morphologically similar to the controls, not the hybrids. Here, mismatch of the Y-chromosome when the X chromosome and autosomal background match causes little to no impact upon morphology. This is in contrast to the hybrids, where the X and Y chromosome mismatch, and so does the X and autosomal backgrounds.

Figure 21



UMAP dimensionality reduction plot coloured by strain, showing the control F1 hybrids as a band across the middle, with the musculus on the left and the domesticus on the right. The interspecific sterile F1 hybrids then diverge off to the top or bottom, with the musXdom at the top and the domXmus at the bottom. The musculus with domesticus Y introgression and domesticus with musculus Y intogression sit round the middle with the control hybrids. Left panel shows subspecies and right-hand side shows where individual genotype within the subspecies lies. Triangles represent a Y-introgression line.

The consensus nuclei of each subspecies highlight this pattern; with the Y-introgression lines showing very similar to their counterparts without introgressions. In the F1 hybrids, X and Y mismatch and so do X and the autosomal background. In the Y-introgression lines, X and Y mismatch but X and the autosomal background match. It can therefore be concluded that it is likely the X and autosomal mismatch that is a greater determinant of morphological abnormality than the X and Y mismatch.

3.3.6 Adding a Y-introgression rescue to the F1-hybrids does not negate the morphological impact

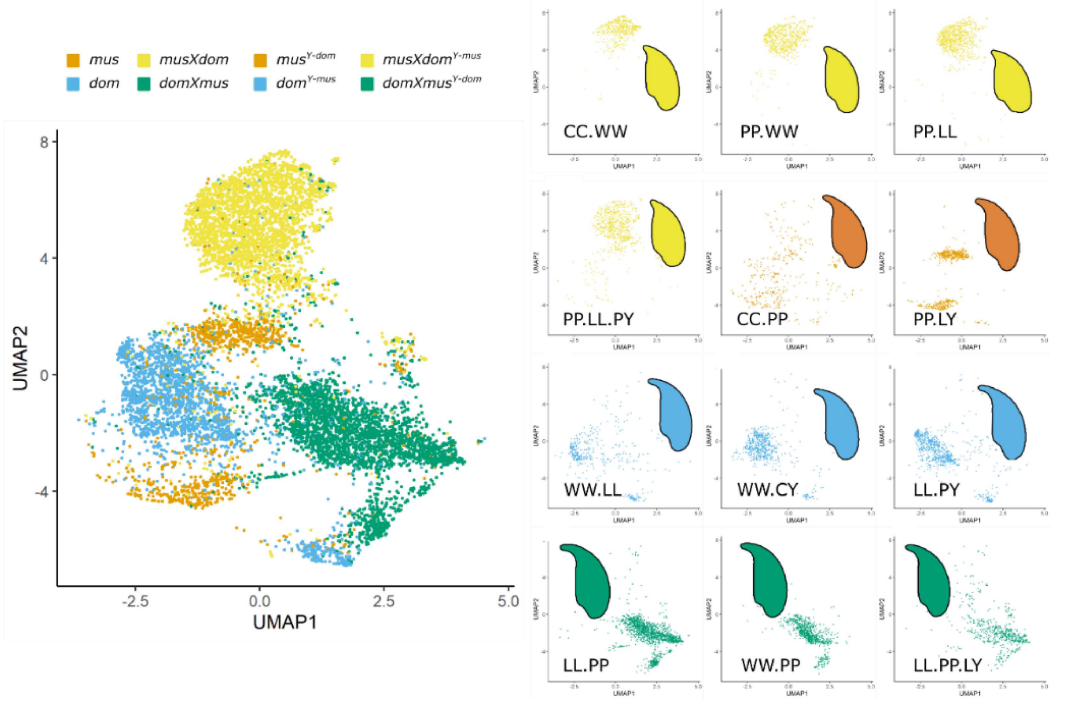
To further test this initial conclusion, and to determine if it is the autosomal and X mismatch that is a stronger determinant of morphological abnormality than the X-Y mismatch, strains that have a hybrid autosomal background but either a *mus* or *dom* X and Y background are added in: these strains were formed of a *mus* mother and a $dom^{Y\text{-}mus}$ father, or a *dom* mother and a $mus^{Y\text{-}dom}$ father. These were known as F1 hybrids with a Y-introgression rescue; to see if adding in the matching Y-chromosome could 'rescue' the abnormality.

Here, strains had an X and autosomal mismatch but an X-Y match, as was the case with the straight Y-intogressions. F1 hybrids with Y-introgression rescue overlay almost perfectly onto the respective divergent clusters of the F1 hybrids- i.e. the morphology of them is almost the same as the F1 hybrids with the same parentals. The Y-rescue makes little to no difference to the morphology. The F1 hybrids with

the Y- introgression rescue are marked in the same colours as the F1 hybrids but with triangle symbols rather than circles.

This overlay of the F1 hybrids with Y-introgression rescue onto the F1 hybrids confirms that X-autosomal mismatch appears to be the cause of the abnormal morphology. In the F1 hybrids, X and Y mismatch, and so do X and the autosomal background. In the Y-introgression rescues, the X and Y match, but the X and autosomal background still mismatch. These two groups (F1 hbrids and hybrids with Y-introgression) show the same morphological abnormalities, showing the X-Y match/mismatch has little to no effect on morphology.

Figure 22

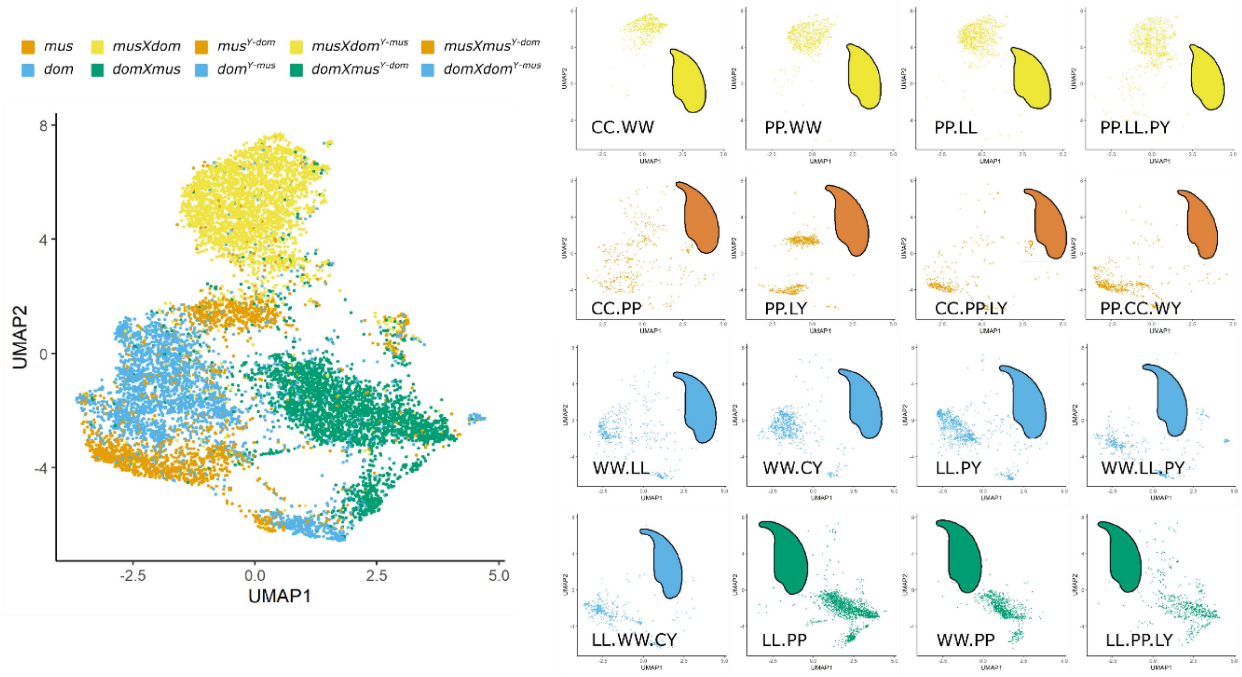


UMAP dimensionality reduction plot coloured by strain, showing the control F1 hybrids as a band across the middle, with the musculus on the left and the domesticus on the right. The interspecific sterile F1 hybrids then diverge off to the top or bottom, with the musXdom at the top and the domXmus at the bottom. The musculus with domesticus Y introgression and domesticus with musculus Y intogression sit round the middle with the control hybrids. The F1 hybrids with Y-introgression rescue almost entirely overlay the interspecific F1 hybrids Left panel shows subspecies and right-hand side shows where individual genotype within the subspecies lies. Triangles represent a Y-introgression line.

3.3.7 Control hybrids with Y-introgressions showed the same morphologies as Y-introgressions on a wild-type background

The final piece of the jigsaw to ensure that it is indeed the X-autosomal mismatch causing the morphological abnormalities was to compare with control hybrids that had a Y-introgression of the opposite subspecies. This meant that these strains had matching X and autosomal backgrounds but a mismatch in X and Y. These strains again sat along the middle band, rather than diverging towards the abnormal clusters. This means that the mismatch of the X and Y had no effect on the morphology of nuclei.

Figure 23



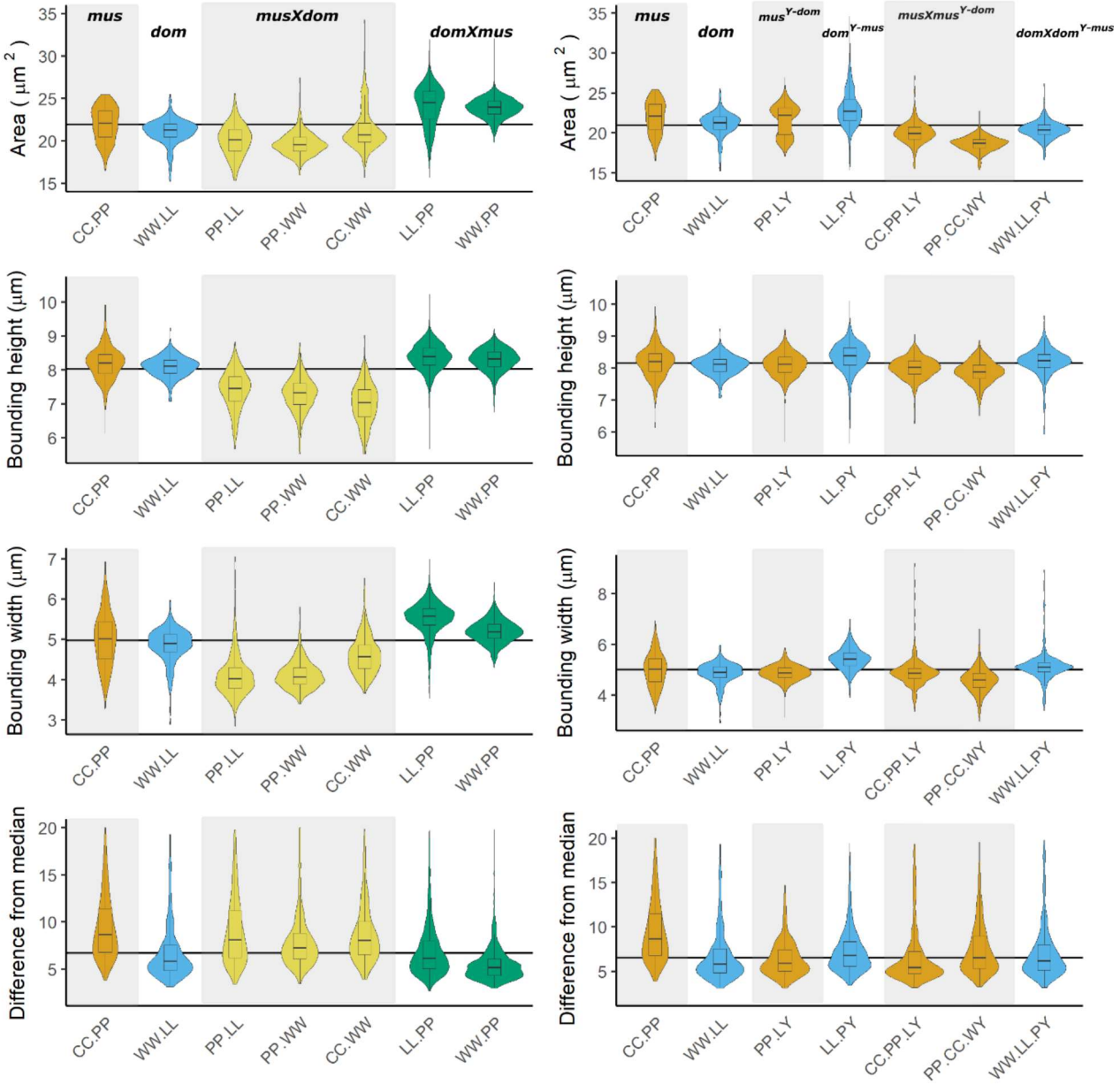
UMAP dimensionality reduction plot coloured by strain, showing the control F1 hybrids as a band across the middle, with the musculus on the left and the domesticus on the right. The interspecific sterile F1 hybrids then diverge off to the top or bottom, with the musXdom at the top and the domXmus at the bottom. The musculus with domesticus Y introgression and domesticus with musculus Y intogression sit round the middle with the control hybrids. The F1 hybrids with Y-introgression rescue almost entirely overlay the interspecific F1 hybrids. The Y-introgression on a control hybrid background stay mostly along the middle band. Left panel shows subspecies and right-hand side shows where individual genotype within the subspecies lies. Triangles represent a Y-introgression line.

3.3.8 Genetic background is important in how X-Y or X-autosomal mismatch affects male reproductive phenotypes

The consensus nuclei showed subtle differences between nuclear morphology of the controls and the Y introgression lines, yet these would be virtually undetectable without automated software to analyse. The Y introgression lines had a matching X and autosomal background but X-Y and X/autosomal incompatibilities: *mus*$^{Y\text{-}dom}$, *musXmus*$^{Y\text{-}dom}$, *dom*$^{Y\text{-}mus}$ and *domXdom*$^{Y\text{-}mus}$. It was shown that genetic background of the X-Y and X-autosomal incompatibility impacted upon the morphological parameters in the same asymmetric pattern as seen in the X-autosomal incompatibilities, but to a lesser extent (Figure 24).

For area, the same trend was seen regardless of the *mus* or *dom* genetic background. The *mus* control had a smaller area than the *mus*$^{Y\text{-}dom}$ ($p=0.0047$) but a larger area than both *musXmus*$^{Y\text{-}dom}$ crosses ($p<2e\text{-}16$). The *dom* control also had a smaller area than the *dom*$^{Y\text{-}mus}$ ($p<2e\text{-}16$) but a larger area than the *domXdom*$^{Y\text{-}mus}$ ($p<2e\text{-}16$). Figure 24 shows the areas of each genotype, showing that the *musXdom* hybrids and the *mus* with Y introgressions had smaller areas than the *domXmus* hybrids and the *dom* with Y introgressions. The difference was much smaller in the Y intogressions.

Figure 24



The area (squared microns), bounding heights (microns), bounding widths (microns) and difference from median of the hybrid cross types and the introgression lines compared to the controls. Median lines are shown for each cohort

For bounding height, the effect on morphology did depend on cross direction. The *mus* control had a larger bounding height than the *mus*$^{Y\text{-}dom}$ (p=0.00185) and the *musXmus*$^{Y\text{-}dom}$ crosses (p=4.8e-11 and p<2e-16). The *dom* control had a smaller

bounding height than the *dom*$^{Y\text{-}mus}$ ($p<2e\text{-}16$) and the *domXdom*$^{Y\text{-}mus}$ ($p=3.0e\text{-}10$). The addition of the mismatched Y chromosome to a *mus* background causes the bounding height of the sperm nuclei to decrease, whereas the addition of the mismatched Y chromosome to a *dom* background causes the bounding height of the sperm nuclei to increase.

For bounding width, the effect on morphology also depended on cross direction. The *mus* control had a larger bounding width than the *mus*$^{Y\text{-}dom}$ ($p=0.0015$) and the *musXmus*$^{Y\text{-}dom}$ crosses ($p=0.0015$ and $p<2e\text{-}16$). The *dom* control had a smaller bounding width than the *dom*$^{Y\text{-}mus}$ ($p<2e\text{-}16$) and the *domXdom*$^{Y\text{-}mus}$ ($p<2e\text{-}16$).

3.3.9 There is a strong correlation between morphology of sperm and the sperm quality assessment parameters routinely used in clinic

Some sperm parameters often used in clinic include relative percent motile sperm and sperm count as measures of fertility. These parameters were in each cross type by collaborators at University of Montana who shared their findings. They also assessed testis weight and seminal vesicle weight, which are parameters that would not routinely be assessed in clinic.

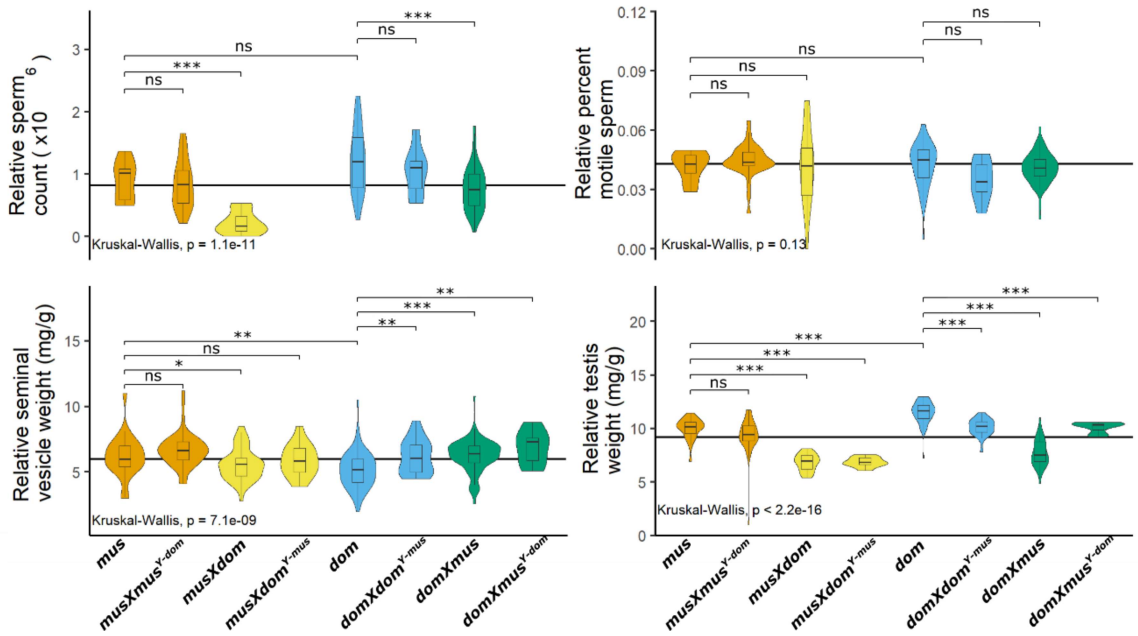
Relative percent motile sperm is mostly consistent across all cross types. However, when relative percent motile sperm is considered in conjunction with sperm count, it is clear that the hybrid crosses and hybrid crosses with Y-introgression have a much lower number of motile sperm cells.

Figure 25 shows *mus* has a relative sperm count of approximately $1x10^6$ and the *dom* has slightly higher than this. The *musXdom* hybrid, with the most severe morphological abnormality has the lowest relative sperm count, below $0.25x10^6$ and the *domXmus* cross also has lower sperm count than control crosses, but not significantly so.

The Y-introgression lines where there is a mismatched Y chromosome on a control cross background with matching X and autosomes have slightly lower relative sperm counts than their respective control cross counterparts, but not as significantly reduced as when the X and Y, and the X and autosomes are mismatched.

The overall percentage of motile sperm per ejaculate remains relatively stable across all crosses, but the hybrid crosses have much lower sperm counts than the control *mus* or *dom.*

Figure 25



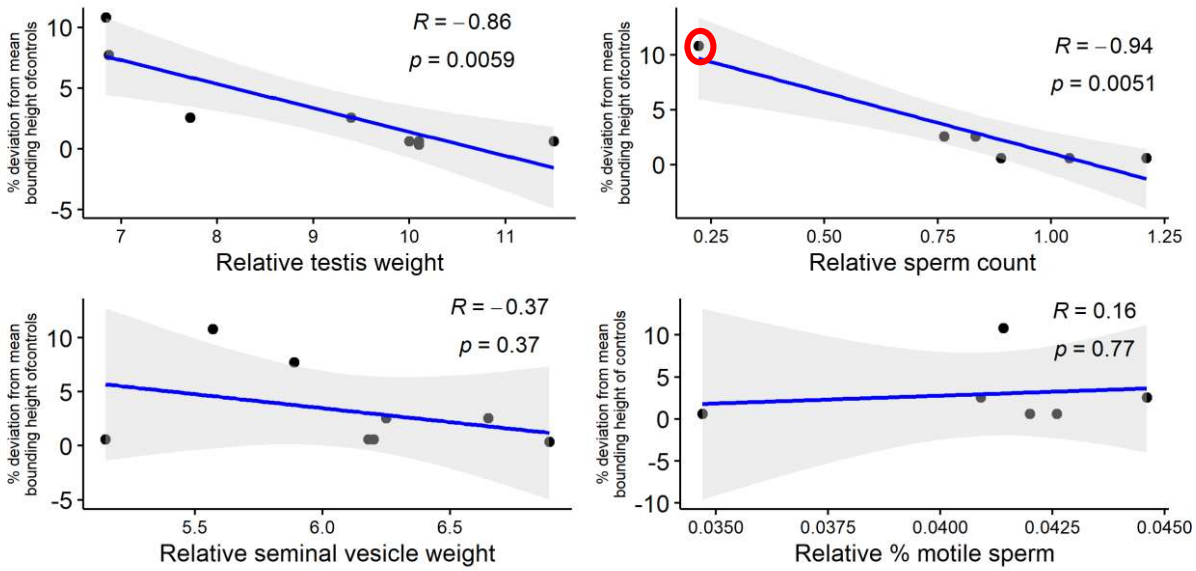
Relative percentage of motile sperm for each cross type, relative sperm count (x10^6), relative seminal vesicle weight and relative testis weight (mg/g) for each cross type, and the median line for each strain. Measurements were not taken for the hybrid crosses with Y introgression (musXdomY-mus and domXmusY-dom). Percent motile sperm data and sperm count data were not collected musXdom$^{Y\text{-}mus}$ or domXmus$^{Y\text{-}dom}$. Sperm count dramatically drops in the musXdom cross, which has the most significant morphological abnormalities. Relative seminal vesicle weight (mg/g) remains consistent across each cross type. Data was collected by collaborators at the University of Montana. Pairwise Wilcoxon test scores were calculated as well as an overall Kruskall-Wallis test score. P-values for pairwise Wilcoxon test scores are indicated with 'ns' for p>0.05, * for p<0.05, ** for p<0.01 and *** for p<0.001.

It appears from the violin charts that testis weights are correlated with morphology; the control crosses with the most 'normal' morphologies of the control crosses have the highest relative testis weights. The most abnormal morphologies, the hybrid crosses, have the lowest relative testis weights, and the Y introgressions appear to fall in the middle for both morphology and testis weight, with the exception of the *domXmus*$^{Y\text{-}dom}$ which has a normal relative testis weight. It must be considered however that testis weight can also be influenced by a number of other factors such

as age of the mouse, and of course will vary between individuals, so could be skewed by one or two individuals.

The percentage deviation from the mean of the controls was plotted as a scatter graph for all parameters. It was shown the further the deviation from the mean of the control strains in bounding height, the further the deviation from the mean of the control strains in relative testis weight, relative sperm count and relative seminal vesicle weight. This relationship was echoed in comparison with bounding width. Bounding height and bounding width did not show a correlation with % motile sperm, which has already been shown to remain almost constant levels in all strains. However, only the relationship between bounding height and testis weight, bounding height and sperm count, and bounding width and testis weight were statistically significant.

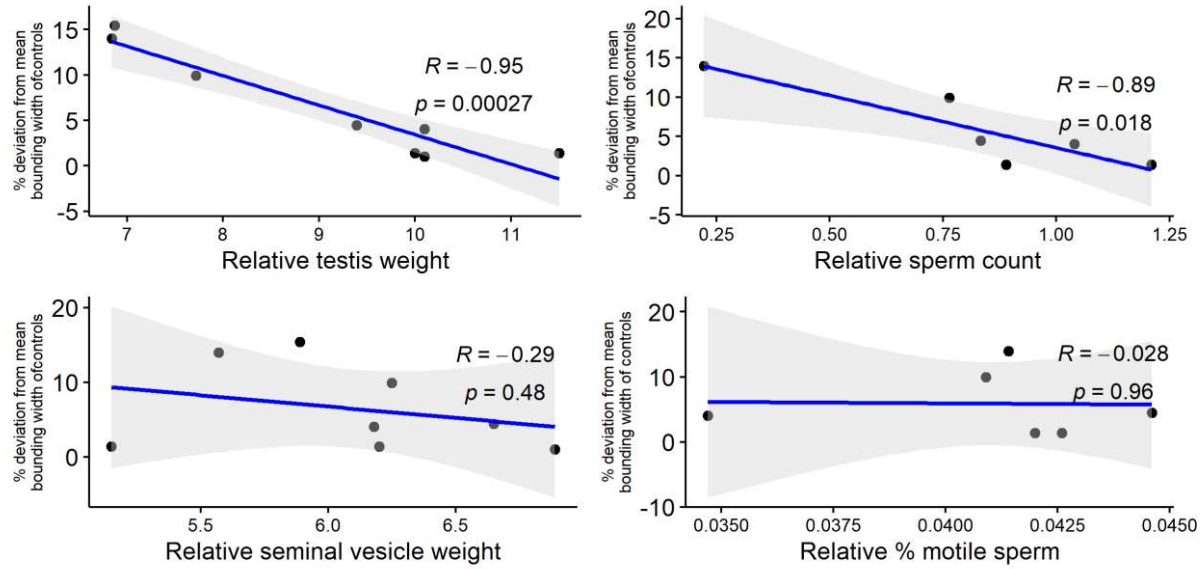
Figure 26



Scatter plots of the relationship between bounding height and the phenotype assessment parameters: relative testis weight, relative sperm count, relative seminal vesicle weight and relative percentage of motile sperm. Correlation coefficient calculated using Pearson's method, the regression line is shown in blue and is smoothed with the Loess method. Relative testis weight, relative sperm count, relative seminal vesicle weight and relative percent motile sperm data were collected by collaborators at University of Montana.

Removing the circled point would remove the association between relative sperm count and bounding height, however the circled point belongs to the genotype with the most abnormal bounding height, suggesting that smaller abnormalities in bounding heights may not affect sperm count too much, but a larger abnormality does.

Figure 27



Scatter plots of the relationship between bounding widths and the phenotype assessment parameters: relative testis weight, relative sperm count, relative seminal vesicle weight and relative percentage of motile sperm. Correlation coefficient calculated using Pearson's method, the regression line is shown in blue and is smoothed with the Loess method. Relative testis weight, relative sperm count, relative seminal vesicle weight and relative percent motile sperm data were collected by collaborators at University of Montana.

3.4 Discussion

This chapter focuses on hybrid mice and Y-introgression strains, comparing the morphology of the hybrid sperm nuclei to the sperm nuclei of Yq-del mice previously determined (Ellis, Bacon and Affara, 2011).

Since the nuclear defects seen in Yq-dels were the same as seen in the *musXdom* hybrids, it had been hypothesised that the mismatch in Sly and Slx copy number between different subspecies of mouse would lead to the same sort of abnormalities seen in Yq-deletion when Sly copy number is reduced to mismatch Slx. However, this study showed that the mismatch in X chromosome subspecies and Y chromosome subspecies has little to no effect on the morphology of sperm; even though there is a significant difference in Slx and Sly copy numbers between species. The results therefore do not support the model of *Slx/Slxl1-Sly* imbalance leading to major disruptions in sperm head morphology in mouse F1 hybrids (Kopania et al., 2022). It is hypothesised that *Slx, Slxl1*, and *Sly* could additionally be involved in epistatic interactions involving not only sex chromosome ampliconic genes, but also autosomal ampliconic genes, and additional autosomal loci. This is consistent with previous work showing that ampliconic gene dosage imbalance also does not lead to patterns of disrupted spermatogenesis gene expression that are common in sterile hybrids (Kopania et al., 2022). It is known that in Yq-del knockout mice, the mismatch in *Sly, Slxl1* and *Slx* copy number leads to the blunted, shortened, squashed morphological change seen in the sperm nuclei, driving the sex ratio skew towards female offspring. It has been shown that this is due to a lack of swimming ability of the Y-bearing spermatids: it was shown that the natural sex ratio can be restored with IVF (Rathje *et al.*, 2019). The study showed that the percentage of motile sperm did not change with the different cross types, rather it was sperm count that was reduced. This suggests that the X-autosomal compatibilities cause a functional impact reducing the number of viable sperm cells. This is in contrast to the X-Y chromosomal incompatibilities in the Yq-del mice which cause a functional impact reducing swimming capability of the Y-bearing spermatids. Whilst the

morphological abnormalities seen are very similar between Yq-del mice and *musXdom* the functional outcomes are not the same. It appears that X/Y mismatch in the Yq-del mice leads to one set of head shape differences and sex ratio skew arising from impacts upon motility, while X/autosomal mismatch in hybrid crosses leads to an alternative set of head shape differences and sterility, not related to motility.

It is known that X-autosomal incompatibilities play a role in hybrid sterility, but it had been hypothesised here that sex-chromosome ampliconic mismatch would play an equally large part, based upon effects seen in Y chromosome ampliconic mismatch in Yq-del mice (Ellis *et al.*, 2005). However, it is now clear that the main driver of post-meiotic X chromosome overexpression in F1 hybrids is unlikely to involve X-Y interactions, based upon the findings in this study. Rather, it is likely that *Sly* and *Slx* are master regulators acting upon autosomal genes, and when the X chromosome and the autosomal backgrounds are mismatched, the autosomal interactions with the amplicons on the X chromosome, such as *Sly* are disrupted.

Slx, Slxl1, and *Sly* are normally only expressed during postmeiotic spermatogenesis, but it has been shown they are aberrantly expressed during early leptotene-zygotene stages of meiosis in sterile F1 hybrids (Larson *et al.*, 2017; Kopania *et al.*, 2022). Whilst leptotene-zygotene stages are earlier cell types than when MSCI occurs, this aberrant expression could potentially arise as a result of related mechanisms to those that result in disrupted MSCI (Skinner *et al.*, 2019; Kopania *et al.*, 2022). It is possible that misexpression of these ampliconic genes during the meiotic stages of spermatogenesis could lead to the misexpression of other genes and contribute to disrupted spermatogenesis in sterile hybrid mice (Kopania *et al.*, 2022).

It is well known that *Sly* and *Slx* are master regulatory genes which acts upon a number of downstream genes, and now it is known that they are likely to regulate autosomal genes as well as sex-linked genes. Since the effects of mismatch and ampliconic imbalance are worse in Y-bearing sperm than X-bearing sperm, any genes involved must be cis-limited; i.e., they evade sharing at the syncytial development stage of sperm and remain tethered in the nucleus as either X or Y linked.

Two cis-limited genes are already known: *Smoc2b* and *Spam1*. These are both autosomal cis-limited genes, so the fact that they are not shared in the syncytia of developing sperm does not lead to differential sperm phenotypes between X and Y bearing sperm. This study showed that it is autosomal genes that are likely to be of the most significance in the morphological changes seen. The changes in Yq-del mice, and likely hybrid mice, selectively impact upon Y bearing sperm to a greater extent than X bearing. Knowing that some genes regulated in this process of sterility must be autosomal, it is of vital importance to understand exactly which genes *Slx* and *Sly* are working downstream on. Whilst it must necessarily be either sex-linked cis-limited genes or sex-linked regulators of cis-limited genes that are causing the sex-ratio skew, this study confirms that the main driver of the sterility is mismatch in autosomal and X-linked genes. Elucidating some autosomal targets of *Sly/Slx* and other potential cis-limited gene regulators will show more about the mechanisms occurring that cause the deformation of sperm nuclei in the hybridisation of two species which have a mismatch in copy number of their ampliconic sex-linked genes *Sly* and *Slx*. This is discussed in the next chapter.

4 Developing an RNA Sequencing Analysis Pipeline which can Genotype Spermatids as either X Bearing or Y Bearing

4.1 Introduction

This study makes use of single cell RNA sequencing data from mouse sperm cells in order to investigate whether there are any genes in single cells that show bias towards either the X or Y chromosome, which could be used in order to genotype the cells into X bearing or Y bearing spermatids. This can be used in the future to investigate which genes are responsible for sperm development defects, since studies have shown that sex linked regulatory genes in conflict with each other can skew sex ratio or even lead to sterility. This study uses data from three previously published single-cell datasets (Chen *et al.*, 2018; Hermann *et al.*, 2018; Bhutani *et al.*, 2021) consisting of sperm cells. Analysis tools in BASH and RStudio such as STAR, Bedtools and genomeFeatures are used in order to assess genes and their features in the single cells.

The previous chapter highlighted the phenotypic consequences of X/autosomal imbalance involved in sterility in hybrid mice, and contrasted this with the phenotypic consequences from X-Y chromosomal imbalances in Yq-del mice. Yq-del mice have a partial deletion on the long arm of the Y chromosome, which produce offspring with a sex ratio skew of 60:40 female: male. This sex ratio skew results from a functional difference arising between Yq-del Y bearing sperm and Yq-del X bearing sperm. The sperm produced from Yq-del mice present with similar nuclear morphological defects to those seen in the hybrid crosses with X and autosomal imbalances. The phenotypes seen in Yq-del mice sperm must be caused by cis-limited genes – genes that evade being shared across cytoplasmic bridges during syncytial sperm

development. In order to answer the next research question “which genes are responsible for the functional differences seen in X bearing and Y bearing spermatids?”, it is crucial to develop a reliable way to characterise spermatids into either X bearing or Y bearing. This chapter focuses on answering the question: “is it possible to genotype X bearing and Y bearing spermatids from a mixed sample of single cell RNA sequencing data”?

4.1.1 Rationale for Genotyping Spermatids and the need to identify cis-limited genes

The previous chapter looks at hybrid crosses of mouse subspecies and Y-introgression between two mouse strains, and how the mismatch in ampliconic copy numbers from different subspecies can affect sperm cells, leading to sterility. It has been shown in Yq-del knockout mice (Ellis, Bacon and Affara, 2011) that *Sly* is a gene involved in a genomic conflict with some X-linked sex-ratio distorter genes. *Sly* is a master regulatory gene, found on mouse Yq chromosome, which acts upon a number of downstream spermatid genes (Ellis, Bacon and Affara, 2011) The large amplicons found in the Y chromosomes of mice have been shown to counteract the effect of *Sly*, maintaining adequate gene expression levels in the developing spermatids. However, when there is either knockout of the ampliconic regions (Rathje *et al.*, 2019) or a mismatch in ampliconic copy numbers in Y-introgressed offspring lines, *Sly* is not repressed as it should be, which leads to overexpression of *Sly. Sly* can therefore act upon some downstream X-linked sex-ratio distorter gene targets (which currently are unidentified). This chapter aims to identify some of these potential targets of *Sly.* It is important from a clinical and agricultural perspective to

be able to identify some of these sex-ratio distorter genes. This would elucidate the genetic pathway of sex-ratio skew or sterility in offspring affected by *Sly* knockout/mismatch. Whilst *Sly* is not present in other organisms, identifying downstream genes that it acts upon may help elucidate some equivalent genes in humans and livestock that may contribute to a similar process of defective sperm formation.

Any genes responsible for the sex-ratio skew patterns highlighted in chapter 3 are likely to be cis-limited: i.e., they evade sharing in sperm syncytial cytoplasm. This is because sperm cells develop in a syncytium, which allows for sharing of RNA transcripts in the cytoplasm between cells.

The knockout of part of the Y chromosome leads to functional differences between X and Y bearing spermatids (Rathje *et al.*, 2019) which means that the genes contributing to this functional difference must remain in either the X bearing or the Y bearing spermatids: they evade being shared in the syncytia. If they did not remain tethered to the X-bearing or Y-bearing spermatids, there would be no difference between X and Y bearing sperm phenotypes, as the transcripts would get shared equally. Cis-limited genes can also be autosomal, but autosomal genes have an equal distribution amongst X and Y bearing spermatids. Therefore, the cis-limited genes that are implicated in the functional difference seen here between X and Y bearing sperm must be sex-linked and remain tethered in the nucleus of the cell.

It is widely accepted that alleles are distributed to offspring with an equal frequency, stated in Mendel's first law, and this means that sperm must be functionally equivalent. Due to the nature of the developing sperm, which form in syncytia, the sharing of gene transcripts means that developing haploid sperm have virtually no

difference in their genetic material. However, there is a class of mammalian genes that are not shared across cytoplasmic bridges, genoinformative markers (GIMs), which can be thought of to be ‘selfish’ leading to natural selection amongst sperm rather than the typical behaviour of Mendel’s first law (Bhutani *et al.*, 2021). Genoinformativity scores from zero to one can be used for each gene to represent the estimated percentage of transcripts retained in cells and not shared across syncytia.

4.1.2 The Importance of Introns for Genotyping

In the developing sperm cell, the nucleus contains hn-RNA, heterogeneous nuclear RNA. The developing sperm cells undergo complex developmental pathways that have been described above, including transcriptional silencing and selective reactivation of genes. hn-RNA is the large pre-mRNA of various nucleotide sequences and is processed in order to be transported out of the cytoplasm and become cytoplasmic mRNAs. This means that hn-RNA contains introns, which have not yet been spliced out. Mature mRNA will not contain introns; they will already have been spliced out. The intronic data is needed to identify the X or Y bearing spermatids. Looking at the intronic reads mapping to different chromosomes helps look solely at the RNA remaining tethered in the nucleus, but the ratio of intronic reads to exonic reads is used to normalise the read counts. If a cell has a higher proportion of intronic hits to the X chromosome, it is very likely to be an X bearing cell, and vice versa if it has a higher proportion mapping to the Y chromosome. This analysis would be impossible without looking at intronic hits, else there would be no way of determining which RNA was still in the nucleus (and therefore not being shared across syncytia and between X and Y bearing cells). This process helps

identify where there is a bias towards intronic hits to either the X or the Y chromosome, which allows genotyping of the cell and determination of whether the cell is an X or Y bearing spermatid.

The mouse genome is readily available from Ensembl (Yates *et al.*, 2020). Several freely available studies have published single cell mouse RNA data that could be used in analysis (Chen *et al.*, 2018; Hermann *et al.*, 2018; Bhutani *et al.*, 2021).

The very first stage of genotyping is to ensure that all cells are haploid – i.e., they are only X bearing or Y bearing. Diploid cells will not be X or Y bearing so they would confound any results and needed to be excluded from analysis. This requires identification of just the spermatids.

It is also important to ensure the removal of any genes that contain no exonic hits – to allow identification of actively transcribing genes in each cell type.

4.1.3 Specific aims

The specific aims of this chapter are:

1) To make use of existing single cell RNA sequence data of sperm cells from both 10X and SmartSeq2 approaches to produce intronic and exonic read count matrices by developing a novel pipeline using existing bioinformatic tools
2) To identify from the single cell data the stages of sperm cell development, so that spermatids can be isolated
3) To use intronic and exonic matrices to identify whether cells are X or Y bearing

4.2 Materials and Methods

This study made use of data from three previous studies. The initial pipeline development was trialled on data from a study by Chen *et al.* (2018) study data, and then applied to the other two datasets (Hermann *et al.*, 2018; Bhutani *et al.*, 2021). A critical consideration was the difference between sequencing methods used in each study. The Chen and Bhutani papers used Smart-seq2 sequencing, whereas the Hermann paper used 10X sequencing. The crucial difference in the protocols is the way RNA is processed to become cDNA. Smart-seq2 captures full-length mRNA, as opposed to 10x that uses a 3′-tag sequencing method. This means that 10X captures many more cells (magnitude of 1000s instead of 100s), but each read may not always be long enough in length to capture any introns – which, as detailed above, are essential to this analysis. The initial pipeline developed was optimised for Smart-Seq2 data but was later modified to attempt to allow it to be used for 10X data. The aim was to have a working pipeline for both types of data, in the hope that the pipeline could genotype data generated by either sequencing method.

4.3 Results

4.3.1 Initial Results with SmartSeq2 Chen *et al* Data

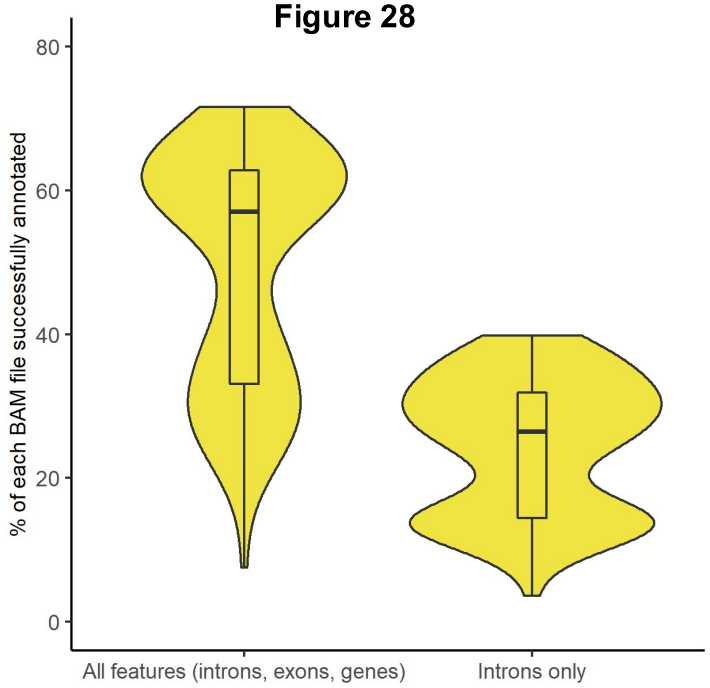
4.3.1.1 STAR Mapping

A high percentage of reads from the Chen *et al.* dataset were uniquely mapped when used as an input to the STAR mapping tool, with the median percentage sitting at around 80% uniquely mapped reads. This was the first check to ensure that a high enough coverage of mapped reads for producing worthwhile data.

4.3.1.2 FeatureCounts Analysis

After establishing a high proportion of successfully mapped reads, the proportion of reads in different genomic features (introns, exons and genes) was calculated. BAM files produced from the STAR aligner were used as an input to the featureCounts tool (Liao, Smyth and Shi, 2014), which annotated the genomic features against the mouse genome annotation file. Figure 28 shows that featureCounts successfully annotated about 60% of each BAM file to the mouse genome annotation, when considering all features, and around 25-30% when annotating introns only.

Figure 28

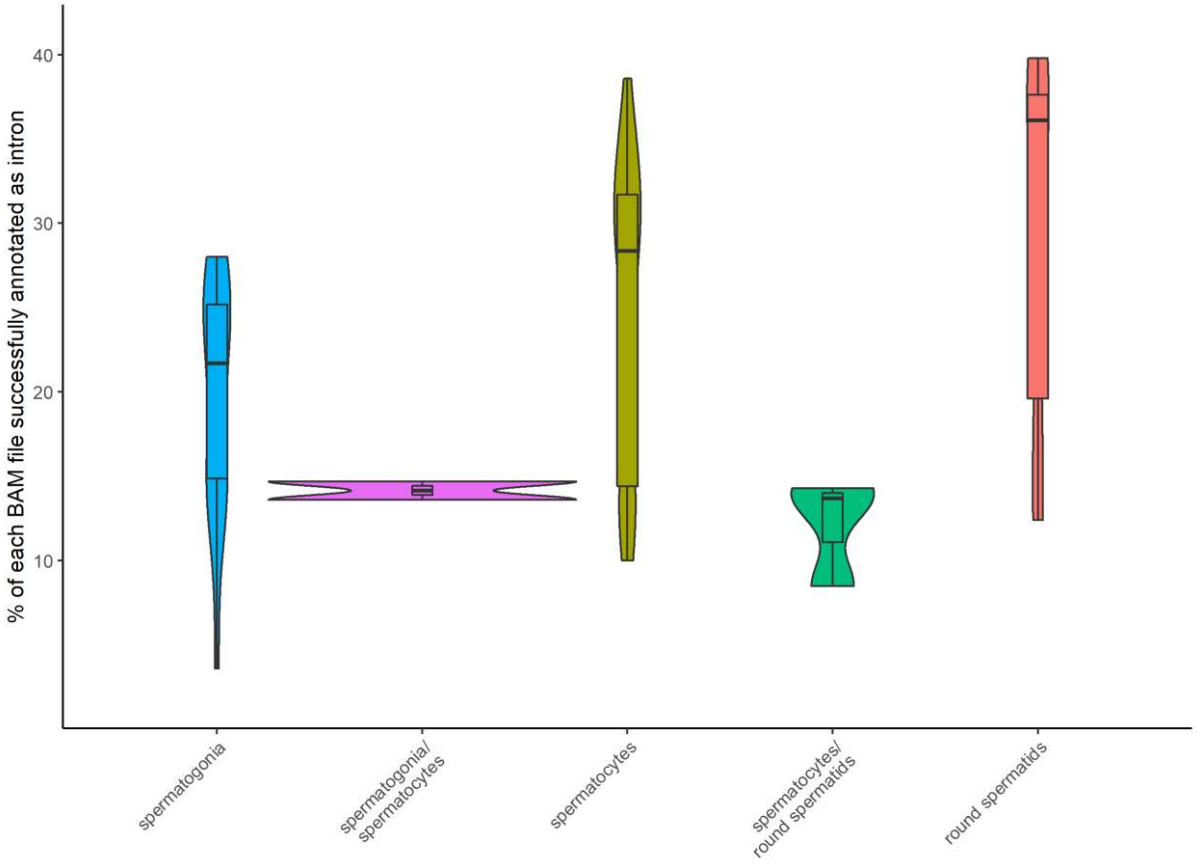


Violin plots to show the percentage of each BAM file successfully annotated by FeatureCounts for all genomic features, and then for intronic regions only

When looking at the annotation for introns only by developmental stage, the mixed ‘spermatogonia and spermatocytes’ and mixed ‘spermatocytes and round

spermatids' seemed to have a much lower percentage of successfully annotated introns than the other developmental stages, as demonstrated in Figure 29.

Figure 29



Violin plots to show the percentage of each BAM file successfully annotated by FeatureCounts for the introns only, by developmental stage of the cells

As discussed in the introduction, developing sperm cells undergo complex processes of RNA silencing and selective gene reactivation, meaning that it is expected that different genes are switched on and off (and therefore mapped) in different developmental stages. To determine if any cell types were outliers compared to the others in terms of successful annotation, annotation was assessed by cell type. The

cell types that were present in the dataset are summarised in Table 10, along with their corresponding developmental stages. It was shown that the pure sample round spermatids had the highest level of successfully annotated features, and analysis would later be restricted to round spermatids.

Table10: cell types and their developmental stages from the Chen dataset

Developmental Stage	Cell Type	Description of Cell Type
Differentiated spermatogonia	A1	Type A1 Spermatogonia
	In	Intermediate Spermatogonia
	BS	S Phase Type B Spermatogonia
	BG2	G2/M Phase Type B Spermatogonia
Preleptotene Spermatocytes	G1	G1 Phase Preleptotene
	ePL	Early S Phase Preleptotene
	mPL	Middle S Phase Preleptotene
	lPL	Late S Phase Preleptotene
Meiotic Cells	L	Leptotene
	Z	Zygotene
	eP	Early Pachytene
	mP	Middle Pachytene
	lP	Late Pachytene
	D	Diplotene
	MI	Metaphase I
	MII	Metaphase II
Round Spermatids	RS2	Steps 1-2 Spermatids
	RS4	Steps 3-4 Spermatids
	RS6	Steps 5-6 Spermatids
	RS8	Steps 7-8 Spermatids

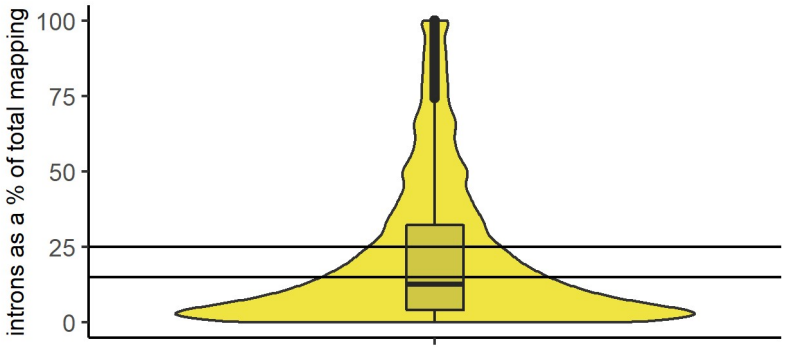
4.3.1.3 Analysis of Number of Hits to Introns and Exons

Given that read mapping demonstrated that both intronic and exonic reads were present in the sequencing data, the relative proportion of intronic reads in each cell were analysed (Figure 30) to determine if there was the expected proportion of intronic reads (La Manno *et al.*, 2018). There was a slightly lower median proportion

of intronic reads in this data than to be expected from the literature: the horizontal lines on Figure 30 show the range of expected proportions of intronic reads.

This is as anticipated; the literature suggests between 15-25% is to be expected (La Manno *et al.*, 2018), as marked on Figure 30. The majority of the cells in this analysis fell just below this range.

Figure 30



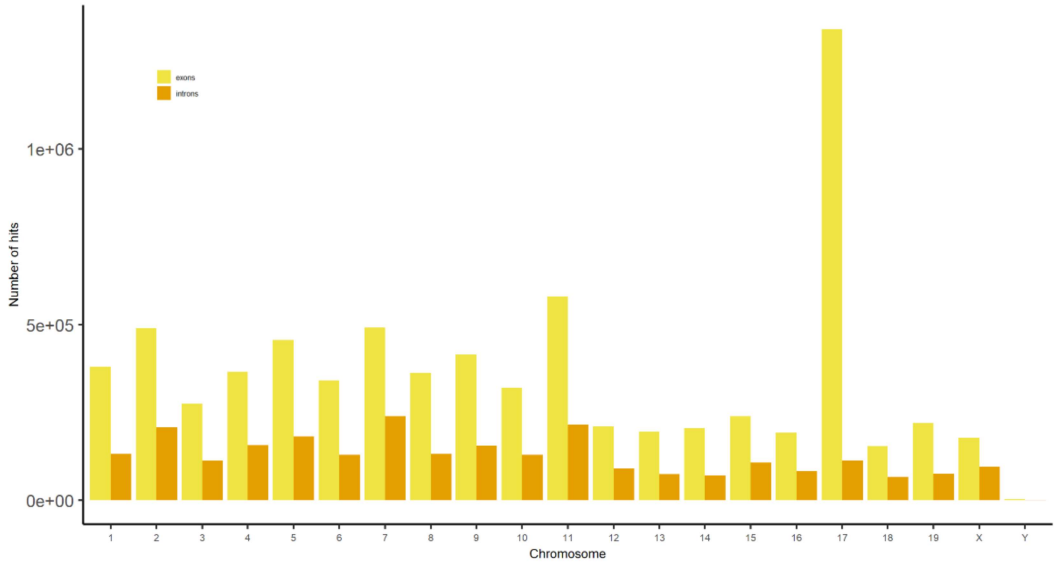
Number of reads mapping to introns as a percentage of mappings all genomic features. The horizontal lines show the expected percentages from the literature (La Manno et al., 2018).

Table 11 summarises the number of hits to both introns and exons on each chromosome for all of the Chen *et al.* dataset cells. Knockout and stem cells were disregarded for this analysis. For most of the chromosomes, except chromosome 17, there were about 2-3 times the number of hits to exons as there were for introns. Chromosome 17 appears a bit of an outlier, with much higher numbers of hits to exons any of the other chromosomes so this was investigated further, and a bar graph summarising the results in Table 11 produced, as seen in Figure 31.

Table11: Number of hits to introns and exons, by chromosome, for all cells in the Chen dataset, excluding SpSC and knockout cells. 95234 reads have hit to introns on chromosome X, and 178552 reads have hit to exons on chromosome X. In contrast, as expected due to the much smaller size of chromosome Y, 1389 reads hit to introns on chromosome Y whereas 3796 reads hit to exons on chromosome Y.

Chromosome	Intron Hits	Exon Hits	Ratio Exons:Introns
1	131845.68	380546.92	2.89
2	208117.07	490268.95	2.36
3	112348.82	275364.03	2.45
4	156875.39	366567.77	2.34
5	181578.55	456506.01	2.51
6	130004.84	341094.39	2.62
7	239437.25	491489.51	2.05
8	132578.74	362779.84	2.74
9	156283.80	414955.03	2.66
10	129729.43	321105.26	2.48
11	215763.22	580293.54	2.69
12	90382.00	211144.30	2.34
13	74666.59	195093.40	2.61
14	70960.95	205803.97	2.90
15	106970.77	293180.93	2.24
16	83631.73	192251.75	2.30
17	114231.02	1341002.86	11.74
18	67700.33	154930.32	2.29
19	76880.28	220096.14	2.86
X	95234.22	178552.87	1.87
Y	1389.99	3796.78	2.73

Figure 31



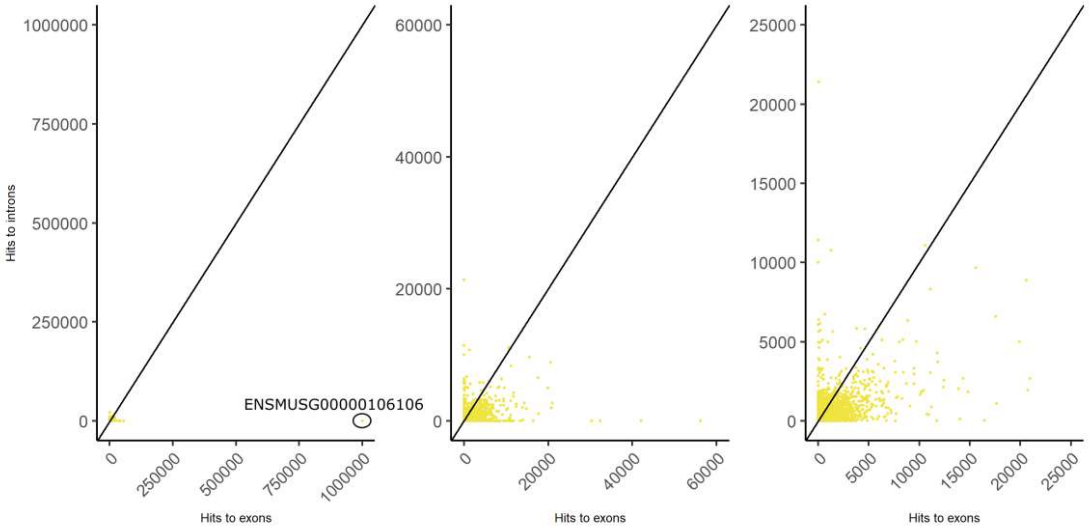
Bar chart showing number of hits to introns (yellow) and exons (orange) for each chromosome. For example, 95234 reads have hit to introns on chromosome X, whereas only 1389 reads have hit to introns on chromosome Y.

To investigate what was causing this anomaly in chromosome 17, as seen in Table 11, some further analysis was run, and scatter plots of every gene with their hits to introns and hits to exons were plotted. This allowed a clear outlier gene to be identified, ENSMUSG00000106106, as shown in Figure 32.

ENSMUSG00000106106, also known as the gene CT010467.1, had over 1,000,000 reads mapping to exons, whereas for all the other genes, the exonic hits were below 125,000. This gene was therefore excluded, as were all the genes that had hits to

exons only or that had hits to introns only. It is important that all genes with no hits to exons or any single exon genes were excluded for genotyping analysis.

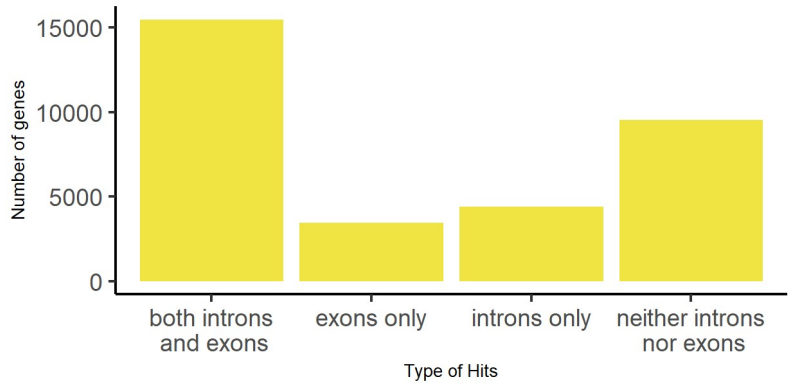
Figure 32



Scatter plot to show hits to introns and exons for each gene. scatter graph to show the number of hits to introns and exons for each gene, excluding any genes that have 0 hits to introns or 0 hits to exons. scatter plot to show hits to introns and exons for each gene, excluding one outlier gene ENSMUSG00000106106

The genotyping analysis will look at the sample of genes that have reads mapping to both the introns and exons on the X and Y chromosomes. Figure 33 shows that the majority of genes have reads mapping to both the introns and exons.

Figure 33

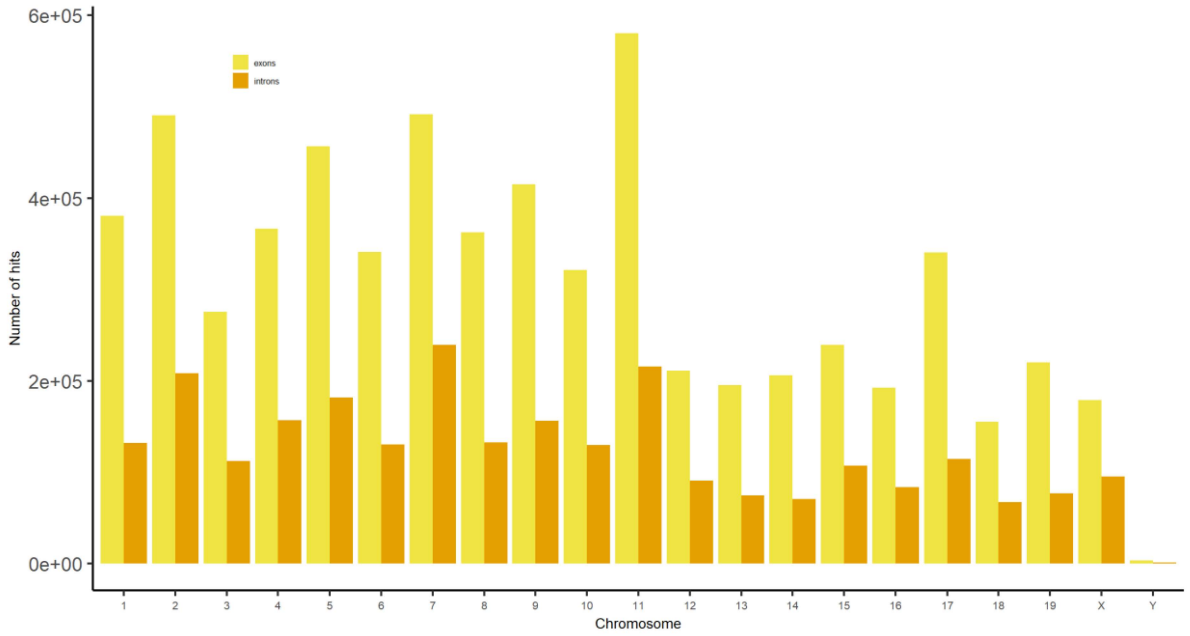


Bar chart to show number of genes that had reads hit to introns/exons within the gene. 15468 had reads hit to both introns and exons, 4397 to introns only, 3459 to exons only and 9513 to neither introns nor exons

A bar chart was plotted again showing the number of hits to introns and exons in each chromosome, but this time excluding gene ENSMUSG00000106106, removing the anomalous result seen in chromosome 17 (Figure 31). Removing this gene made the bar chart much more evenly spread, as shown in Figure 34, with chromosome 17 no longer being an outlier. The raw data is summarised in Table 12.

ENSMUSG00000106106 is an 18S ribosomal RNA gene found on chromosome 17. It is a pseudogene – a non-functional segment of DNA that resembles functional genes. Pseudogenes normally arise as superfluous copies of functional genes, either directly by DNA duplication or indirectly by reverse transcription of an mRNA transcript or by degradation of the original copy with no duplication involved. Pseudogenes are well known to complicate genetic studies, since they are prone to PCR bias or amplification bias, and are sometimes annotated as genes in genome sequences. Because pseudogenes are often misidentified as real genes or exons, confusion can occur in genome analysis. This may explain the anomalous result looking at intron:exon hit ratio in each individual gene.

Figure 34



Bar chart showing number of hits to introns and exons for each chromosome, excluding gene ENSMUSG00000106106

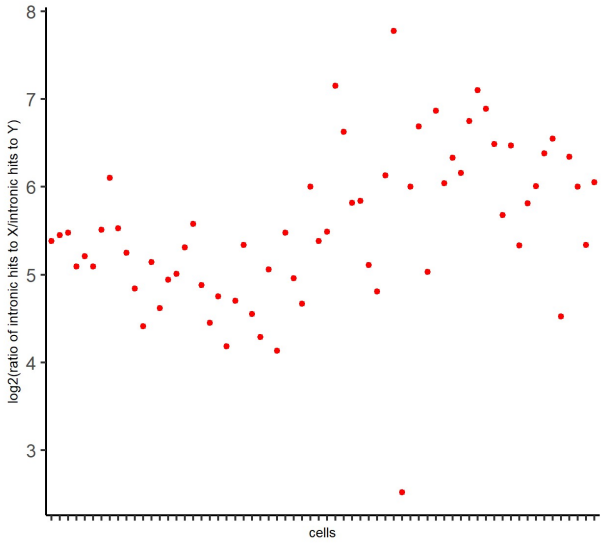
Table12: Number of hits to exons and introns per chromosome, excluding gene ENSMUSG00000106106. The only difference made is to the number of hits to exons on chromosome 17. ENSMUSG00000106106 is a RRNA gene found on chromosome 17.

Chromosome	Intron Hits	Exon Hits	Ratio Exons:Introns
1	131845.68	380546.92	2.89
2	208117.07	490268.95	2.36
3	112348.82	275364.03	2.45
4	156875.39	366567.77	2.34
5	181578.55	456506.01	2.51
6	130004.84	341094.39	2.62
7	239437.25	491489.51	2.05
8	132578.74	362779.84	2.74
9	156283.80	414955.03	2.66
10	129729.43	321105.26	2.48
11	215763.22	580293.54	2.69
12	90382.00	211144.30	2.34
13	74666.59	195093.40	2.61
14	70960.95	205803.97	2.90
15	106970.77	293180.93	2.24
16	83631.73	192251.75	2.30
17	114231.02	340412.88	2.98
18	67700.33	154930.32	2.29
19	76880.28	220096.14	2.86
X	95234.22	178552.87	1.87
Y	1389.99	3796.78	2.73

Next, a method of genotyping these cells needed to be developed. The initial method trialled for genotyping each individual cell was: (1) recording the number of intronic hits to the X chromosome and the number of intronic hits to the Y chromosome. (2) Calculating the log2 ratio of the number of X chromosome hits compared to the number of Y chromosome hits for each cell and plotted onto a graph, Figure 35. The higher the log2 ratio of X chromosome hits to the Y chromosome hits, the more hits to the X compared to hits to the Y. Figure 35 shows that all cells have more hits to the X than the Y, but some cells have higher proportions of X-mapping reads than other cells. It could be hypothesised that cells with proportionally more hits to the X

chromosome are likely to be X-bearing sperm, and those cells with proportionally more hits to the Y chromosome are likely to be Y-bearing sperm. However, due to the fact that the Y chromosome is so much smaller than the X, a method of normalising the read counts was needed, and further refinement of the calculation used for genotyping was needed. Not only this, but the cells needed to be split into spermatids and non-spermatids.

Figure 35



Log2 ratios of the number of hits of genes to introns in the X and Y chromosomes in each individual cell

4.3.1.4 Round Spermatids

It was essential to be focusing purely on genes that are switched on in round spermatids, since meiosis is complete (cells are haploid and therefore either X or Y bearing) in round spermatids, but cytoplasmic bridges still exist and so the cells are still syncytial. In round spermatids, only 5% of the sequence is made up of intronic regions, whereas overall in the full mouse sperm cell population, 35% is intronic.

Isolating just the round spermatids, left very few cells of interest. These were the cells with SRR: SRR6342170, SRR6342171, SRR6342172, SRR6342173, SRR6342174, SRR6342190, SRR6342191, SRR6342192, SRR6342193, SRR6342194. Unfortunately, with just ten cells, it would be impossible to see any clear-cut clustering pattern as the sample size was simply not large enough.

4.3.1.5 Initial Conclusions from Chen Data

At this stage, it was concluded that the pipeline that had been developed worked well for mapping the data and generating intron/exon matrices, which gave the scope to allow for a lot of downstream data analysis to hopefully genotype cells. However, the data available from the Chen *et al.* study simply was not sufficient for the development of a genotyping method – only ten round spermatids were present in the dataset, and as discussed, it is essential that only the round spermatids were used for genotyping as they are haploid. Therefore, no conclusions could be drawn – with such a small sample size, it could be that all of the spermatids were X bearing, or all were Y bearing – it simply could not be determined with such a small sample size.

To overcome this and to refine the pipeline, a dataset with many more cells was used from a study by Hermann *et al (*2018).

4.3.2 Results with 10X data

The novel mapping pipeline was applied to another dataset with a larger number of cells, this time consisting of 10X single cell data. It was crucial to try to apply the novel pipeline to a range of sequencing methods. Some slight tweaks were required

to optimise the initial mapping and annotation stages for 10X data, due to the way 10X data is uploaded to the sequencing read archive database. 10X data is uploaded so that data from multiple cells is stored in a multiplexed format rather than as data from individual cells in individual FASTQs. The Hermann *et al.* 10X dataset had an average read count of 61,000 reads per cell – far less than the sequencing depth would be from SmartSeq2 data.

4.3.2.1 10X Pipeline Tweaks

Multiplexed data is stored as .bam files rather than .fastq files. Instead of using fastq-dump, the .bam files were downloaded from the SRR data access NCBI page, along with a barcode file, which contained cell barcodes to identify which part of the .bam files relate to which cells. 10X's program "bamtofastq" allows conversion of the multiplexed .bam files back to multiplexed FASTQ files, which were all combined into a single file. Umi-tools program 'extract' then allowed the matching of the cell barcodes with reads from the FASTQ files, so that the extracted reads could be mapped using STAR. Then, featureCounts was run, before UMI-tools programme 'count' was used to split gene assignments according to individual cell. This gave an output of Gene ID, cell identifier, read count (for both introns and exons), and chromosome number was matched in using the Gene ID.

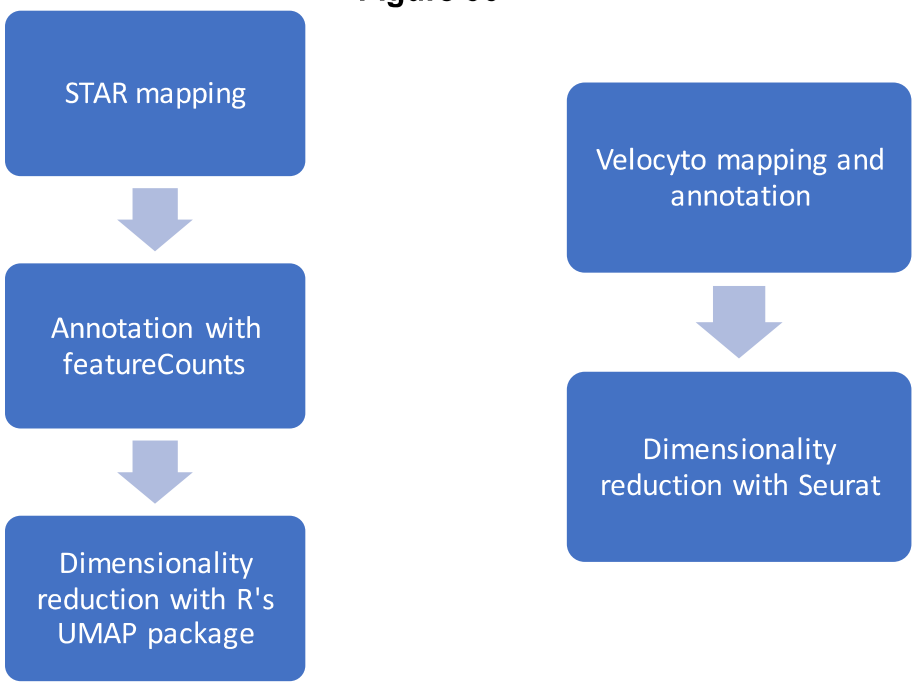
4.3.2.2 Initial Dimensionality Reduction Results

The gene expression levels were log-normalised for ease of interpretation on the dimensionality reduction plots. Cells were filtered and only included if they expressed more than 200 detected genes and genes were only included for analysis if they

were expressed in more than three cells. UMAPs were plotted and split into clusters, then were coloured by the log-normalised gene expression levels. The gene expression levels were used to determine which clusters would be the spermatids.

An important stage of sanity checking here was that the results generated were in concordance with collaborators at the University of Kent, whereby a different already established pipeline was being used to try and produce the same results. The University of Kent pipeline used a package called Velocyto to generate the intron and exon matrices, rather than the novel pipeline used here. Kent then used a programme called Seurat to do downstream analysis and clustering, rather than using R, which was used on the results from the novel pipeline.

Figure 36



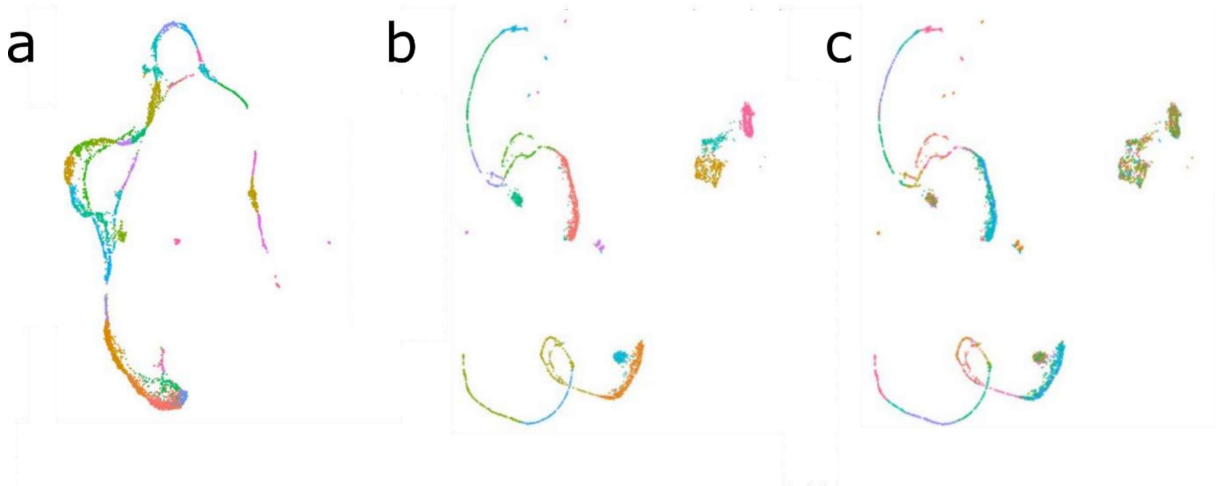
(Left) novel pipeline and dimensionality reduction vs (right) pipeline and dimensionality reduction used at University of Kent

To check concordance of the two methods, the results were compared, first by considering the dimensionality reduction plots. However, when data generated from

the novel pipeline was run through the alternative dimensionality reduction pathway being used at University of Kent (Seurat), a different shape dimensionality reduction plot was produced compared to when the University of Kent were running it, as shown in Figure 36. It is important to note that both pipelines used the same original download data files; the inputs at this stage to Seurat differed only in the initial analysis upon them (novel pipeline vs Velocyto to generate intron/exon matrices).

Figure 37 shows (a) the exonic matrix generated in Velocyto run through Seurat UMAP dimensionality reduction compared with (b) the exonic matrix generated through the novel pathway, but run through the Seurat UMAP dimensionality reduction. Panel (c) shows the exonic matrix generated in Velocyto run through Seurat UMAP dimensionality reduction, but coloured according to the clusters generated by the dimensionality reduction package in R. This shows thatthe clusters generated in Seurat and the clusters generated in R were not concordant.

Figure 37



UMAPs produced using different versions of the data. (a) the exonic read counts produced by Velocyto, run through Seurat dimensionality reduction and coloured by cluster. (b) the exonic read counts produced by the novel pipeline, run through Seurat dimensionality reduction and coloured by cluster. (c) exonic read counts produced by the novel pipeline, run through Seurat dimensionality reduction and coloured by cluster generated from the UMAP dimensionality reduction programme in R.

It is clear to see that somewhere along the pipeline, something had happened to mean that the novel pipeline vs Velocyto pipelines were producing different outputs, as the final dimensionality plots, when run through the same piece of downstream analysis code (Seurat), came out different shapes (panels a and b). Whilst dimensionality reduction leads to slightly different shapes each time as it is based on a seeded algorithm, the general shape should be the same. However, the novel pathway produced two different noodles and some 'blobs', whereas the Velocyto pathway produced one long noodle and a shorter one. The noodle was an expected shape, as it shows the developmental stages of the spermatids, one end being the earliest stage and the other end being the latest stage.

When the hierarchal clusters determined from the dimensionality reduction using UMAP in R were applied to the Seurat plots (panel c), it showed that it was the 'blobs' that were not clustering as they were in the Velocyto pipeline, and led to the need for a final round of pipeline refinements.

4.3.3 Refining the Pipeline

4.3.3.1 Strandedness Corrections

The first parameter that was discussed to be refined within the novel pipeline was the strandedness parameter given to the featureCounts program. The default setting is 'unstranded' but supplying the argument of '1' or '2' meant that only forward or reverse strands were counted respectively. An initial concern was that if the default setting of unstranded was supplied to the program, then read counts could be being over or under-estimated.

In order to test this hypothesis, featureCounts was run for a second time across the indexed BAM files, this time using the forward stranded argument, then the reverse stranded argument and then appending these two outputs together. Manual separation and aggregation of the forward and reverse counts allowed certainty that reads were counted accurately. With the default parameter, reads could somehow have been counted twice by being recognised as being on both forward and reverse strands, or not being counted at all. Once these results were generated, they were compared to the initial featureCounts output to check to see if this was a worthwhile correction. It was shown that the read counts were slightly higher in response to making this strandedness correction, becoming more in line with the read counts given by the Velocyto pipeline. Even though this step involved running featureCounts twice (for forward and reverse strands) and then appending the two files – which is considerably more time consuming, the fact that there was a significant change in results makes it an essential refinement to add to the pipeline. Therefore, in the pipeline, featureCounts was ran twice, once on the forward and once on the reverse strand.

4.3.3.2 Exon Spanning Regions

Introns are non-coding regions located within protein-coding genes and are removed prior to protein formation. Intergenic regions are non-coding regions found in between coding genes. It was important for the purpose of this analysis to ensure that the introns were not counted as part of the spanning exon, and that intergenic DNA is not included as intronic.

Figure 38

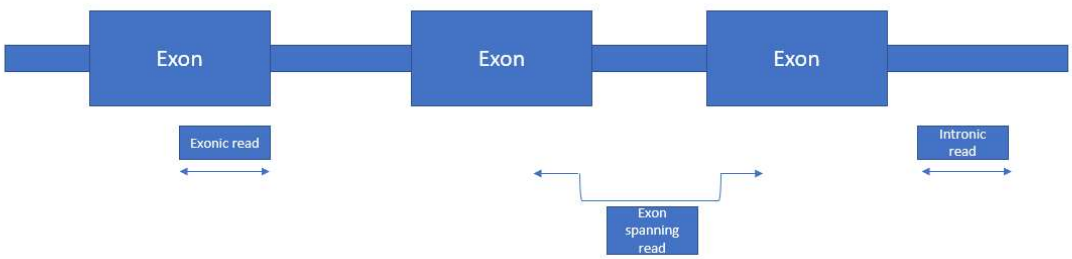


Diagram illustrating exon spanning regions

Reads that were only within an exonic region would be counted as exonic, reads fully within an intronic region would be intronic, but those reads that come from exon-exon boundaries would be exonic reads, however due to the fact they span an intron, they may also be classed as intronic. It was important to check the way featureCounts was classifying these reads as they may lead to a distorted intronic region count. Therefore, the argument "--nonSplitOnly" was given to featureCounts, which would only include the non-split reads in the output. When comparing this parameter with the default, there was zero change in read counts, so both the default parameter and –nonSplitOnly can be used to produce the same values. This means featureCounts was not incorrectly classifying introns within spanning exonic regions, which was an initial concern.

4.3.3.3 Annotation Version

Pipeline development began in spring 2020, just prior to the release of the GRCm39 mouse genome on 24/06/2021. Pipeline development took several months from start to finish, so it wasn't until systematic comparisons of the Velocyto output from University of Kent with the novel pipeline that it came to light that there were some apparent, somewhat interesting differences in the annotation of different features.

Whilst the majority of cells from the Hermann *et al* study clustered into the same clusters when comparing the two pipelines (Figure 37), there were a few small clusters from the Velocyto pipeline that did not match up to the clusters in the novel pipeline. These cells were extracted and the differences in intronic and exonic matrices showed that; whilst across the whole dataset, cells tended to have a slightly higher proportion of reads mapping to introns and exons in the novel pipeline, the mismatching clusters had significantly higher average read counts in matrices generated in the novel pipeline.

The first thing to be investigated here was, were any particular genes causing this skew, and it was found that a certain set of genes were leading to this distortion in read counts. Interestingly, some of the genes implicated were ones that were going to be used as marker genes for identifying spermatids – *Tnp2, Tnp1, Prm2* and *Prm3* were among the 20 highest expressed genes in the mismatched clusters. Here it was important to determine why reads were mapping differently to these genes across the two pipelines, so the inputs to the pipelines were carefully compared.

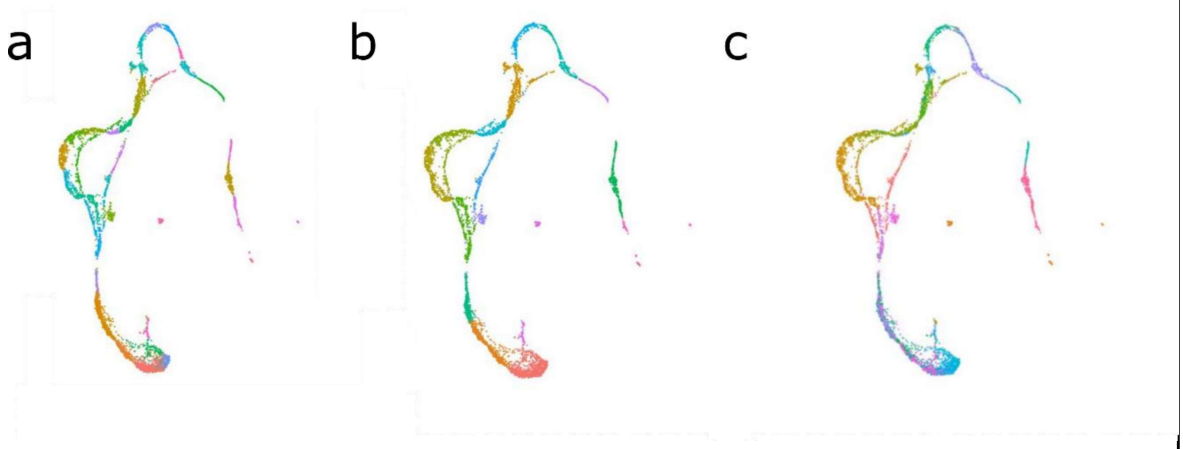
The only difference in inputs to the pipelines were that the Velocyto pipeline used an inbuilt annotation file and automatically determined the intronic regions from this, whereas the novel pipeline was using the annotation file that had been downloaded and had intronic data manually input into it. On side-by-side comparison of the two

annotation files, some of the key genes that had been identified as being annotated differently in the two pipelines showed slightly different coordinates between the two annotation files. At this point, it was determined that it was the use of the out-of-date annotation file that was leading to the differences in the results of the two pipelines. Therefore, the annotation file needed updating to use the latest GRCm39 release to manually add in the intronic data. The annotation file generated here was carefully compared to the annotation file that Velocyto used, to ensure gene coordinates matched up.

4.3.3.3 Revised GTF and Hermann Data

The novel pipeline was re-run, this time using the updated annotation file. The new exon matrices that were generated were put through the Seurat dimensionality reduction pathway, and, using the same seed for the plot creation, the same shape plots were able to be generated as those generated from the data produced in the Velocyto pipeline (Figure 39). This showed both pipelines (Velocyto and the novel pipeline) were generating the same count data, and that the same dimensionality reduction method could be applied to each set of data get the same shaped plots. The plots generated using Seurat were coloured by the cluster assignments generated from the UMAP dimensionality reduction performed in R. The cluster patterns matched up nicely, whilst they are coloured differently (colouring is arbitrary between programmes), the areas which have different colours matched up fairly concordantly. This showed that cells were clustering the same, even when slightly different methods of performing UMAP dimensionality reduction were used.

Figure 39

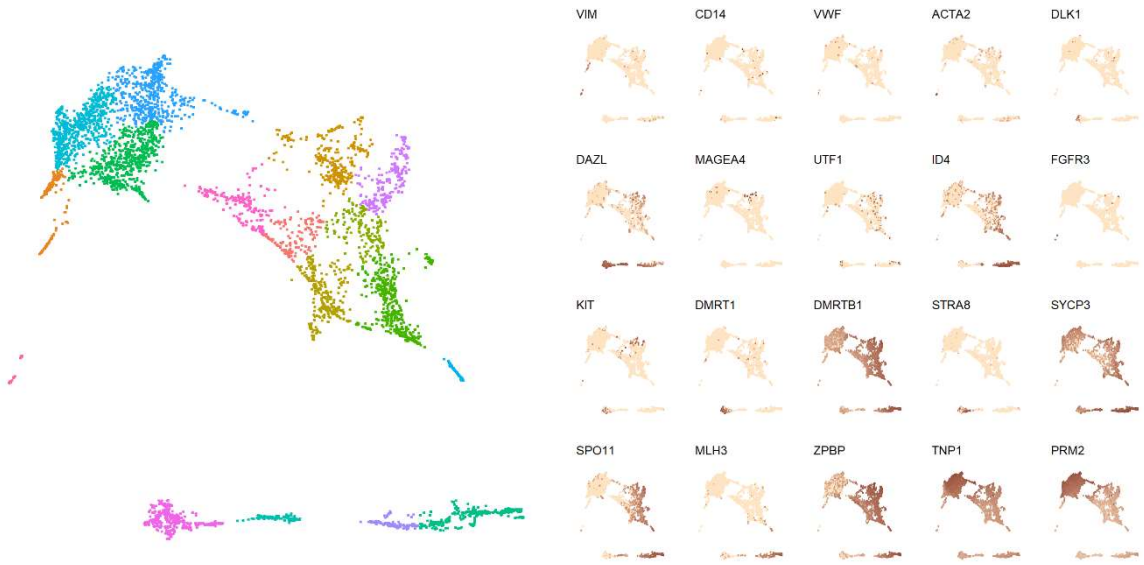


UMAPS produced using different versions of the data. (a) the exonic read counts produced by Velocyto, run through Seurat dimensionality reduction and coloured by cluster. (b) the exonic read counts produced by the refined novel pipeline, run through Seurat dimensionality reduction and coloured by cluster. (c) exonic read counts produced by the refined novel pipeline, run through Seurat dimensionality reduction and coloured by cluster generated from the UMAP dimensionality reduction programme in R.

The exonic count matrix generated from the novel pipeline had dimensionality reduction applied in R, and plots produced were coloured by gene expression levels to visualise gene expression levels at different stages (Figure 40). Gene counts were log-normalised, and cells were filtered and only included if they expressed more than 200 detected genes and genes were only included for analysis if they were expressed in more than three cells. Clusters that had gene expressions matching spermatid classifications were identified as spermatid cells. These clusters were the ones with high levels *of Zpbp, Tnp1* and *Prm2*. Early round spermatids would show low levels of *Magea4, Dazl, Spo11, Mlh3* and *Sycp3*, but higher levels of *Zpbp*. Late round through to elongating spermatids would have high levels of *Tnp1*, and elongating spermatids would have high levels of *Prm2*. It was determined that

clusters 1,3,4,5,6,8,10,11,12,13,14, and 15 were likely to be spermatids, and a separate data-frame was created, consisting of just the putative spermatids.

Figure 40

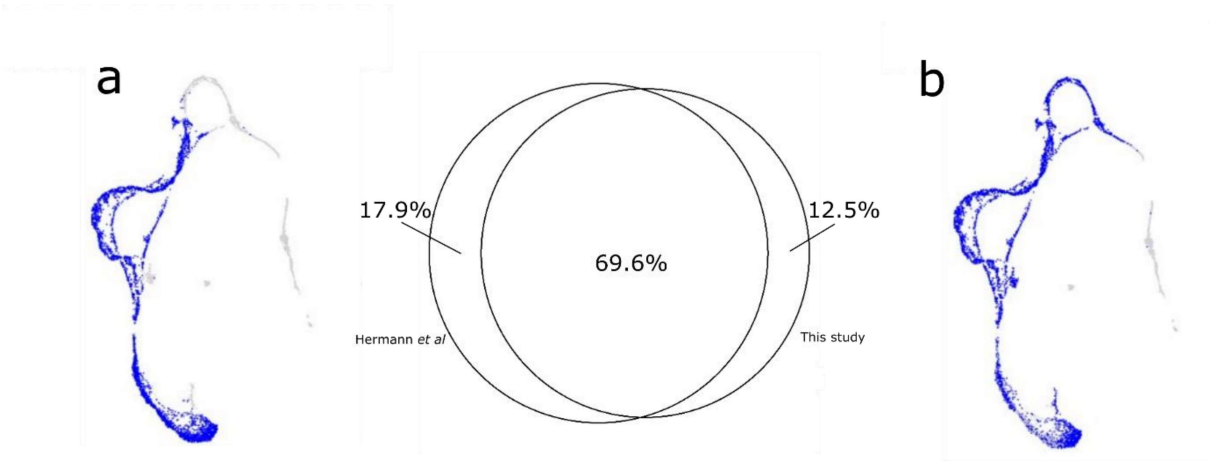


UMAP dimensionality reduction performed in R, run on the refined novel pathway exonic matrix data, coloured by marker gene expression levels

The cells labelled as spermatids from the UMAP clusterings (3819 cells) were compared with the original dataset's classification of spermatids, and the classifications were mostly in agreement (Figure 41). Therefore, the pipeline was working as it should to determine spermatids. For sanity checking, the UMAPs generated in Seurat were coloured, firstly by the spermatid classifications determined here, and secondly by the classifications determined in the original study. These agreed almost entirely – other than slightly more were classified as spermatids here than the original study. It was decided that it did not really matter if some spermatids were misclassified – they should simply drop out at the further stages of analysis as they wouldn't be able to fit into either X or Y bearing spermatid

cluster. These cells can be ignored, as long as there are not too many excess cells that may skew the results.

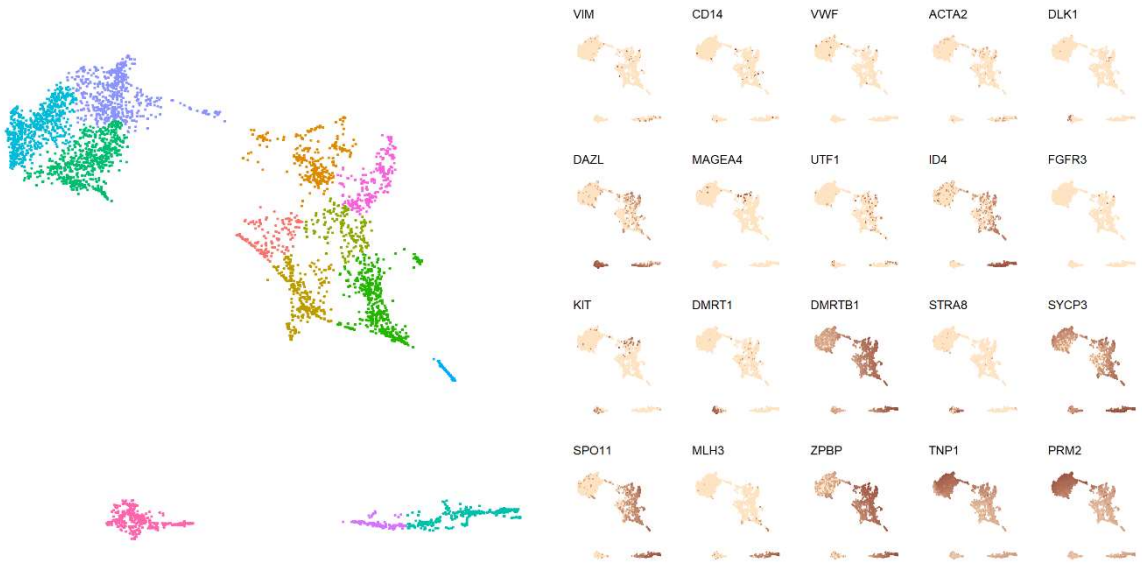
Figure 41



UMAPs produced from the exonic read counts produced by the refined novel pipeline, run through Seurat dimensionality reduction coloured by spermatid classification (a) determined in this study and (b) determined in the original study.

The spermatids were isolated and UMAP dimensionality reduction re-run in R. The resulting plot was coloured by the gene expression levels of *Zpbp, Tnp1* and *Prm2* to identify which were clusters of early, middle, and late spermatids.

Figure 42



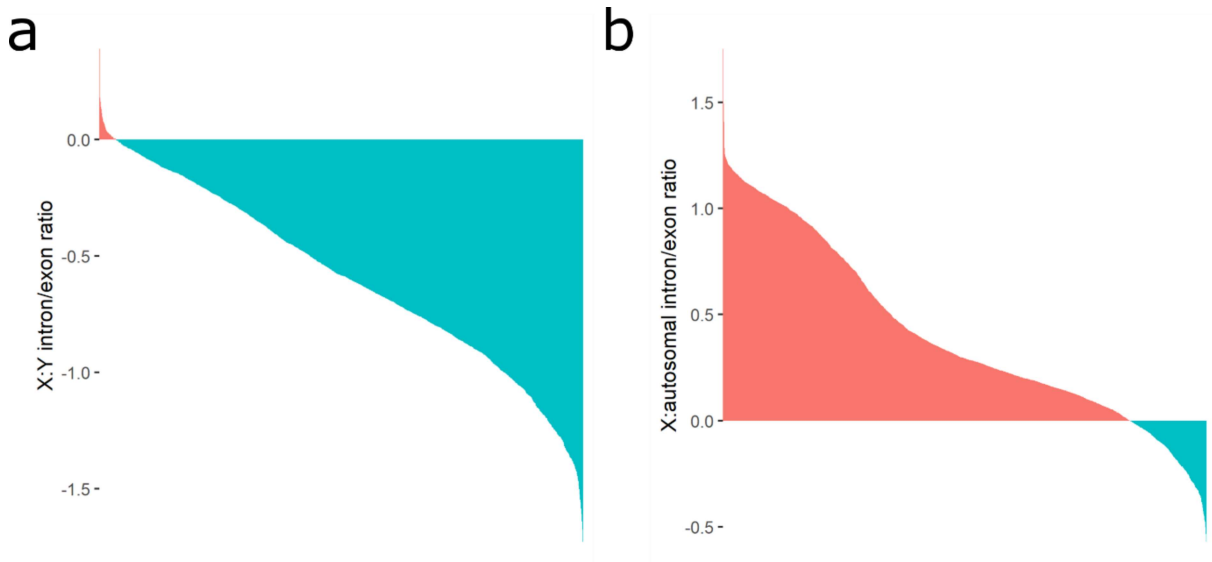
UMAP of just the spermatids, coloured by levels of ZPBP (early spermatids), TNP1 (mid spermatids) and PRM2 (late spermatids). It is likely that cluster 1,2,3,5 are early spermatids, with cluster 4 as mid spermatids, and cluster 6 as late spermatids.

With this data, instead of just comparing intronic X and Y ratios, as with the Chen *et al.* (2018) dataset, a slightly new calculation was run, which compared the intronic/exonic read ratio for the reads to the X chromosome with the intronic/exonic read ratio for the hits to the Y chromosome: [log2(Xintronic) + log2(Yexonic) – log2(Xexonic) – log2(Yintronic)]. This took into account the ratio of the intronic regions to the exonic regions, as a method of normalising the intronic reads, which was not accounted for in the analysis of the Chen *et al.* (2018) dataset.

Due to the nature of 10X sequencing, there was a bias towards Y-bearing cells when using the X:Y intron/exon ratio. This was simply because the Y chromosome is smaller, and the shorter length fragments in 10X sequencing means that some introns mapping to the larger X chromosome were not being included in the gene

expression counts. Because of this, an alternative calculation was applied, where the ratio of X reads to autosomal reads was used, rather than the ratio of X:Y.

Figure 43

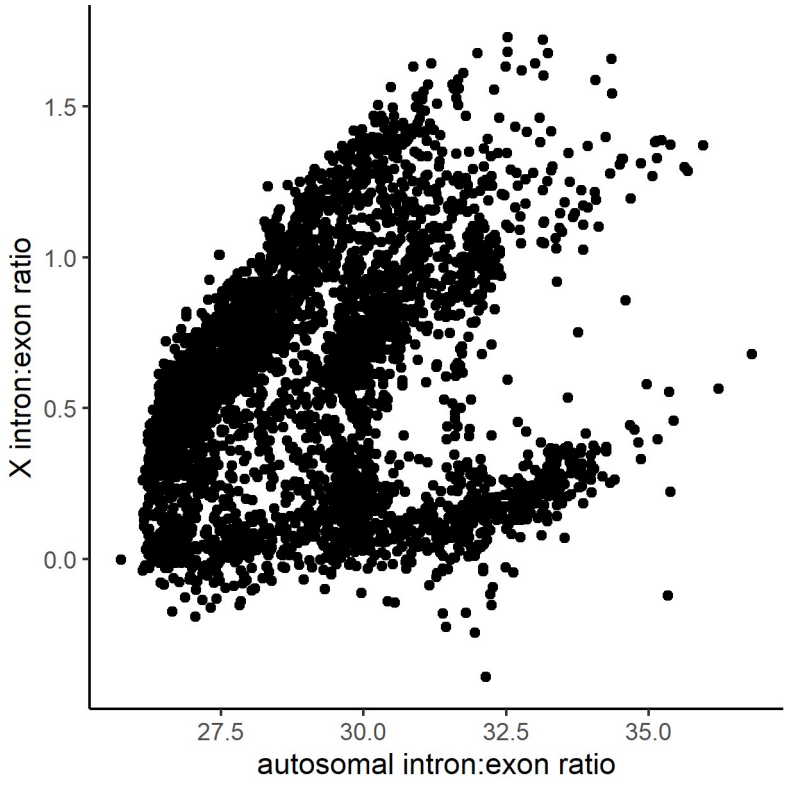


(a) X-Y intron:exon ratio calculation performed on the spermatids only from the Hermann dataset. (b) X-autosomal intron:exon ratio calculation performed on the spermatids only from the Hermann dataset. Positive values are a mixture of putative X and some Y bearing cells, and negative values are putative Y bearing cells.

Unfortunately, using just the X and autosomal read counts, from just the spermatids, there was still not a 50:50 ratio of putative X and putative Y bearing genes. In theory, those with a ratio of greater than 0 should be X bearing, and those less than 0 should be Y bearing. However, even though a bimodal separation was seen, it was not a 50:50 separation. Logically, there should be an almost equal split between X and Y bearing cells, especially with so many cells in a sample. Therefore, it appears that the pipeline is able to determine a putative Y bearing group and a mixed group of putatively X bearing and putatively Y bearing cells, when using 10X data. It was

not too surprising that there was not a 50:50 split, as 10X data has a very low sequencing depth.

Figure 44



Scatter plot of the intron:exon ratio for the X chromosome compared to the intron:exon ratio for the autosomes

At this stage, the pipeline was applied to a third dataset, which had fewer cells than the 10X dataset, but more cells than the Chen *et al.* (2018) dataset. It was also SmartSeq2 data, as was the Chen *et al.* (2018) dataset. It was hypothesised that using SmartSeq2 data would overcome the issue of sequencing depths. If genotyping could be achieved on SmartSeq2 data, it may then be possible to retrospectively identify X and Y bearing cells from 10X data by gene expression

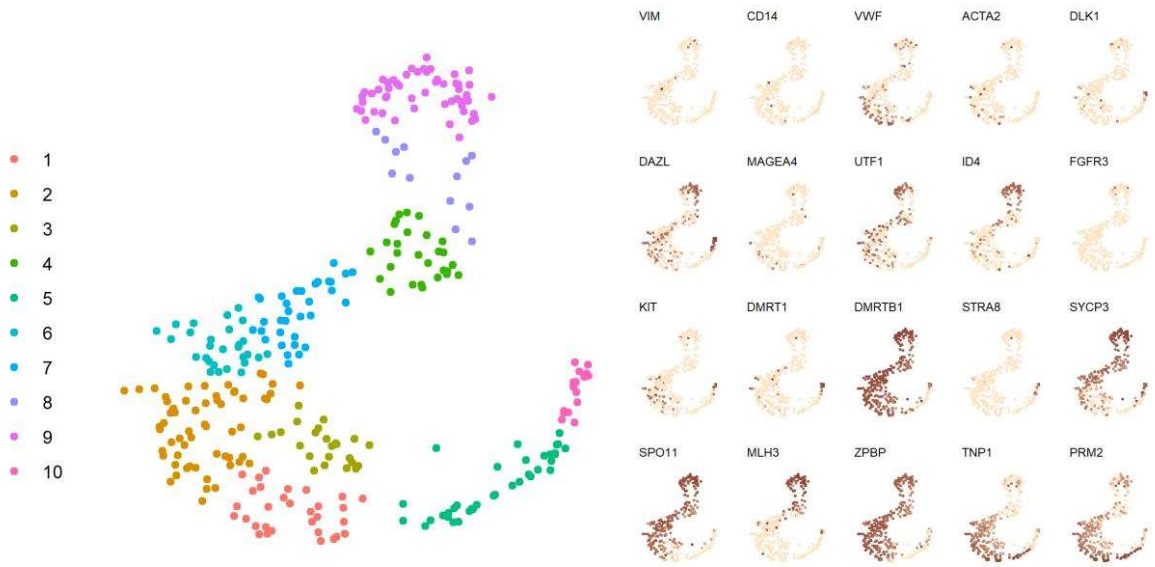
levels of key genes identified to be differentially expressed in X and Y bearing sperm from differential gene expression analysis on Smartseq2 data. However, first the pipeline needed to be applied to SmartSeq2 data with enough cells, to check the pipeline could work as anticipated.

4.3.5 Final SmartSeq Results with full Bhutani dataset

This dataset was Smart-Seq2 data (as was the Chen data) but contained more cells than the Chen study, with 398 cells. The data uploaded to the SRA project on NCBI initially only contained 96 individual cells with four technical replicate sequence runs per cell. The original authors of the study were contacted and asked to provide the full dataset.

On receipt of the full dataset, the novel pipeline was run, from FASTQ download through to featureCounts, and then UMAP dimensionality reduction on the exonic

Figure 45



UMAPs of full Bhutani dataset, coloured by normalised expression levels of 20 key genes.

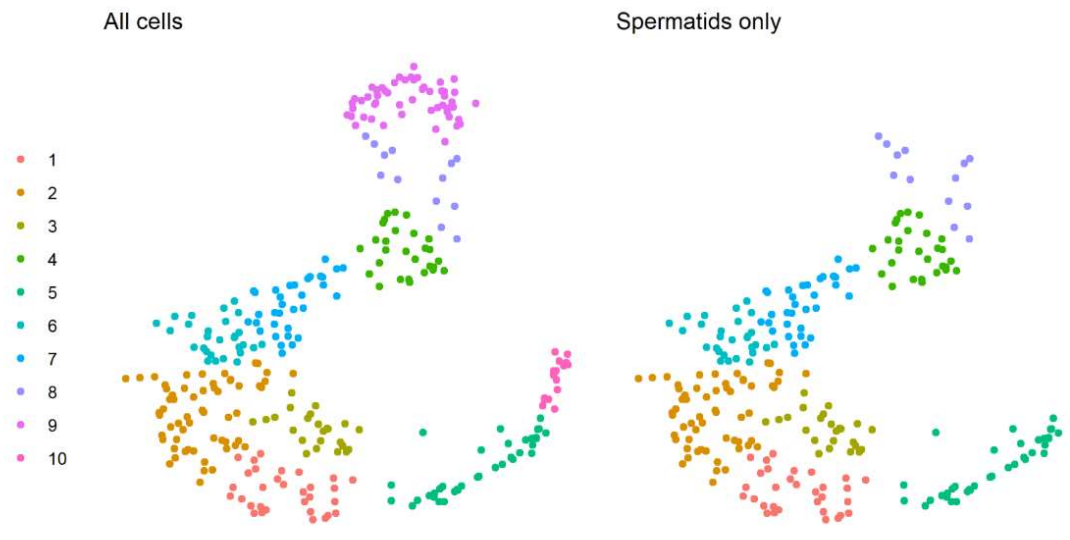
counts from each cell was performed. As with the 10X data analysis, UMAPs were coloured by gene expression level in each cell.

For spermatids, clusters needed to be either:

- *Zpbp1*-positive, but *Magea4, Dazl, Spo11, Mlh3*, and *Sycp3*-negative, which will be the early round spermatids
- *Tnp1*-positive (late round through to late elongating)
- *Prm2*-positive (elongating spermatid stage onwards)

Clusters 1, 2, 3, 4, 5, 6, 7 and 8 were determined to be spermatids from the UMAP.

Figure 46

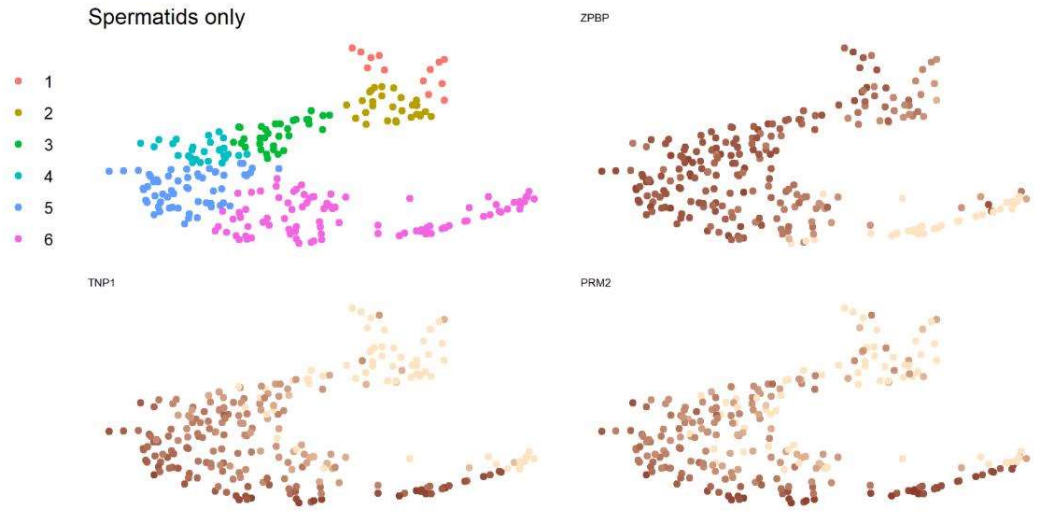


UMAP of the full dataset and the spermatids only

The UMAP of just the spermatids was coloured by key gene expression levels to show the progression of early-stage spermatids through to late-stage spermatids. For ease of interpretation, the clusters were renamed and reordered to have the

early-stage spermatids as cluster 1 progressing through to cluster 6 as the late spermatids.

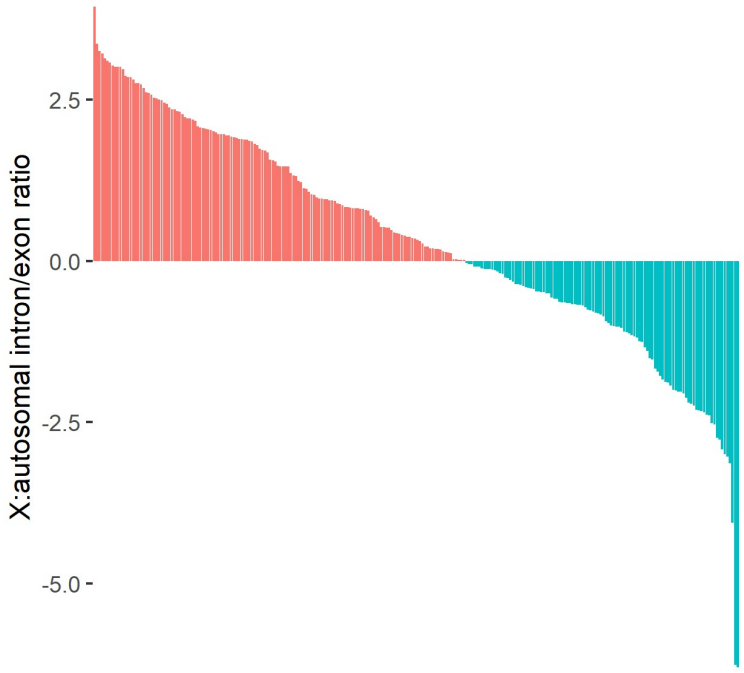
Figure 47



UMAP dimensionality reduction on just the spermatids, coloured by gene. Showing cluster 1 and 2 are the early spermatids, cluster 3 and 4 are the mid spermatids and clusters 5 and 6 are the late spermatids.

These spermatids were then isolated and the genotyping calculation of the ratio of X reads to autosomal reads performed, giving a bimodal distribution. The X intron/exon ratio was plotted against the autosomal intron/exon ratio, and cells formed two clusters. The cluster with higher X intron/exon ratios were determined as being X bearing as they had relatively more X transcripts compared to the autosomal transcripts.

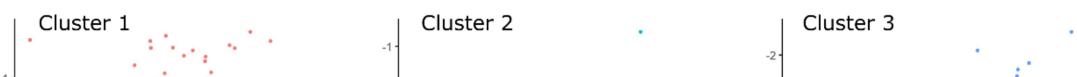
Figure 48



Genotyping calculation performed on the spermatids only population of the Bhutani dataset

The spermatids with the higher X:autosomal intronic ratios were called as putative X bearing cells and those with the lower X:autosomal intronic ratios were called as putative Y bearing cells, for differential gene expression analysis. This was using an adaptive threshold for each individual cluster, since each cluster varied slightly in their thresholds for the different clusters. The adaptive threshold was chosen where there was a clear separation between two clusters. Even a very small degree of separation on the plots represented a very large real-life difference as the plots used a logarithmic scale. Although this was done by eye, in future this thresholding could be automated, which would be more objective.

Figure 49



The cut-off of the adaptive threshold for each cluster is shown in Table 13. Cells above the threshold became putative X bearing spermatids, cells below the threshold became putative Y bearing spermatids. Because the intron:exon ratios are logarithmic, the cells in the higher up clusters have vastly greater reads mapping to the X chromosomes than those in the lower cluster, giving a high degree of certainty that they come from X bearing cells.

Table13: adaptive thresholds for each cluster of spermatids

Cluster 1	X intronic/exonic ratio = **-2.5**
Cluster 2	X intronic/exonic ratio = **-4**

Cluster 3	X intronic/exonic ratio = **-4**
Cluster 4	X intronic/exonic ratio = **-4**
Cluster 5	X intronic/exonic ratio = **-4**
Cluster 6	X intronic/exonic ratio = **-2.5**

4.4 Conclusion

The findings of this study show that there is potential to genotype sperm cells based upon their single cell RNA sequencing data, and this can be used in the next stage of research to investigate critical candidate cis-limited genes that may be directly involved in fertility.

So far it has been demonstrated that a simple RNA sequencing analysis pipeline as described in this study can be used to find a bimodal distribution of intronic hits to X and Y chromosomes in mouse sperm cells that have been sequenced using a Smartseq2 protocol. Provided the dataset has a large enough sample of cells, and the RNA sequencing method used was SmartSeq2, the pipeline can be used to determine whether the sperm cells are X bearing or Y bearing.

This has a number of implications – for starters it may be of use to the farming industry, as well as for IVF for selection of embryos to avoid sex-linked disease, but also for further research into candidate cis-limited genes that are acted upon by master regulator *Sly*, which has already been shown to cause sterility of mice when knocked out. This will help determine the genes and mechanisms involved in sterility, which are investigated using differential gene expression in the next chapter.

Putative X and Y bearing cells were unable to be identified from data sequenced using 10X protocol, however a group of putative Y and a group of mixed putative X and Y were identified. Genes escaping cis-limitation will be investigated in the next chapter. It may be possible in the future to look at the gene expression levels of

some key genes, and, by applying key gene signatures determined in the SmartSeq2 pathway to the 10X datasets, putative X and Y bearing cells may be able to be identified in 10X data as well.

5 Determining candidate cis-limited genes involved in sex ratio skew and sterility in mice, from differential gene expression analysis, gene ontology, enrichment analysis and genoinformativity

5.1 Introduction

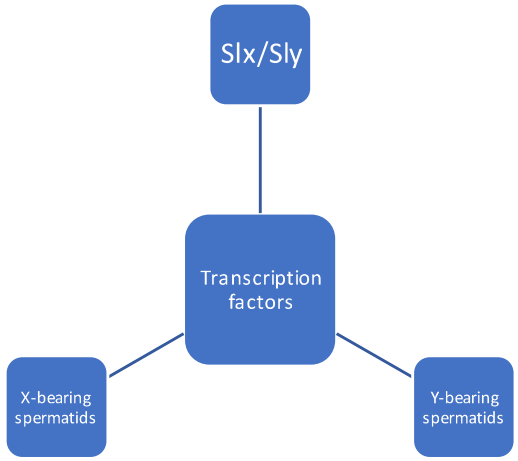
As previously mentioned, it is known that *Sly* and *Slx* are big drivers of infertility in mice, and they work differentially on downstream targets in X and Y bearing spermatids. As X and Y bearing spermatids are selectively impacted upon, it can be deduced that downstream targets of *Sly* and *Slx* must be cis-limited. Cis-limitation is the process by which transcripts from specific genes are not shared across the cytoplasmic brides that are formed during spermatogenesis.

5.1.1 Rationale

In order to determine the process by which sterility in mice occurs with *Sly* knockdown, it is important to elucidate what downstream genes *Sly* and *Slx* act upon; these are likely transcription factors that act differentially to express different genes in X-bearing and Y-bearing spermatids. It was determined from the morphology analysis in chapter 3 that a mismatch between the autosomal background and the X chromosome of a subspecies is a bigger driver of infertility than a mismatch between the X and Y backgrounds of subspecies. This has demonstrated the importance of looking at autosomal genes as targets of either *Slx* and *Sly,* or the autosomal targets of transcription factors that could be regulated by *Slx* and *Sly,* rather than just sex-linked genes. This chapter aims to identify

candidate cis-limited genes that *Sly/Slx* may act upon, that leads to functional differences in X bearing and Y bearing sperm cells.

Figure 50



Process by which Sly/Slx cause differences in X and Y bearing spermatids

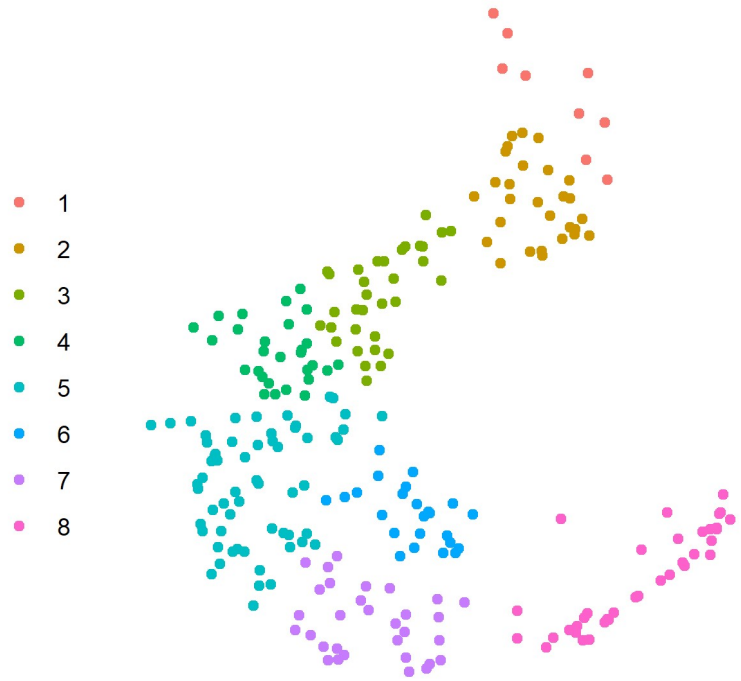
5.1.2 The novel genotyping pipeline allowed identification of 250 spermatids from SmartSeq2 data

Figure 51

UMAP of all cells present in Bhutani et al dataset (323 cells), where 250 were classified as spermatids

Chapter 4 identified a sample of 250 spermatids from the Bhutani *et al.* dataset (Bhutani *et al.*, 2021). Dimensionality reduction analysis was run on the sample of spermatids from the Bhutani *et al* data, to give a UMAP shape (Figure 51) which was then coloured by hierarchal cluster (Figure 52); clusters here represented the developmental stages of the spermatids. Previous analysis using gene markers (chapter 4) allowed identification of the early-stage spermatids at the bottom of the UMAP shape, through to the late-stage spermatids at the top of the UMAP shape.

Figure 52



UMAP dimensionality reduction on just the spermatids, showing cluster 1, 2 and 3 are the early spermatids at the top, cluster 4, 5 and 6 are the mid spermatids and moving up to cluster 7 and 8 are the late spermatids at the bottom of the plot.

5.1.3 The genotyping pipeline gave a roughly equal split of X bearing and Y bearing spermatids in the SmartSeq2 dataset

Of the 250 spermatids present, the genotyping pipeline gave a split of 132 spermatids determined to be X bearing, and 118 which were determined to be Y bearing. The UMAP of the spermatids was coloured by cells which were X bearing and cells which were Y bearing, and the X and Y bearing cells were evenly distributed across the shape.

Figure 53

Distribution of putative X bearing and putative Y bearing spermatids on the UMAP of all spermatids; putative X baring coloured in red and putative Y bearing coloured in blue

The cells were then split into six clusters based upon hierarchical clustering, to allow for easy identification of developmental stage, as detailed in Table 14.

Table14: number of cells in each cluster, which represents developmental stages. Cluster 1 is earliest stage spermatid, with cluster 6 being latest stage spermatid.

Cluster	1	2	3	4	5	6
X bearing	29	16	17	27	32	11
Y bearing	9	12	27	25	20	25

The cell IDs of each of each of the cells labelled as putative X and putative Y are shown in appendix 1.

5.1.4 Specific Aims

The specific aims of this final results chapter are to elucidate:

1) To get a differential gene expression profile for the X and Y bearing spermatids
2) To determine which of these differentially expressed genes could be candidate cis-limited genes
3) To investigate the function of these candidate cis-limited genes
4) To consider how these candidate cis-limited genes could be used in the future

5.2 Methods

The two groups of cells, putative X and putative Y, were compared against each other in a differential gene expression analysis. KEGG pathway maps were created for each of the lists of autosomal genes upregulated in X bearing cells and those

upregulated in Y bearing cells. STRING analysis produced a network of genes showing protein-protein interactions, to which k-means clustering can be applied. DAVID was used to show functional annotations and gene ontologies, as in STRING and KEGG pathways, amongst many other things, but a unique feature here is the protein domains.

Autosomal genes are equally as important as sex-linked genes to investigate in this analysis; chapter 3 demonstrated that X:autosomal mismatch is a bigger driver of poor morphology of sperm than X:Y mismatch, which suggests that there is something in the autosomes of significance. Of course, necessarily, the levels of autosomal genes should not be any different between X and Y bearing cells, other than due to temporal changes, or the influence of regulators such as transcription factors. Therefore, any differences seen in the autosomal gene counts is likely as a result of sex-linked transcription factors regulating these genes to be active or inactive at different stages of differentiation. It is important to consider these sex-linked transcription factors as potential cis-limited candidate gene targets of *Sly.* Therefore, this analysis begins by demonstrating that the differentially expressed autosomal genes in X and Y bearing spermatids have no functional significance, and then searches for potential transcription factors that may have the differentially expressed genes as their targets. These transcription factors, as well as the sex-linked differentially expressed genes may have functional relevance, or further regulatory relevance but all are candidate cis-limited genes.

Because of the multi-level nature of gene control, motif analysis was run twice, firstly on the set of autosomal DEGs, and then again on the list of potential transcription factors, to identify if the transcription factors may be regulated by second level transcription factors.

For this, the transcription factors identified as potentially binding to the genes upregulated in the X bearing spermatids, and the transcription factors identified as potentially binding to the genes upregulated in the Y bearing spermatids were given as the input to STREME.

TomTom (Gupta *et al.*, 2007) was used to find potential second level transcription factors that bind the motifs found in the candidate transcription factors that bind genes upregulated in X bearing spermatids, and in Y bearing spermatids.

Mean absolute read counts of each of these transcription factors identified in motif analysis were calculated for each stage of spermatid development, as genes are switched on and off transiently during spermatid development. This gave an overview of which transcription factors are selectively switched expressed at different times in different (X or Y bearing) spermatids.

During the genotyping stage, in chapter 4, it was shown that the SmartSeq2 data could be split into X-bearing and Y-bearing spermatids. However, it was also shown that 10X data did not split well into two equal sized groups. It was hypothesised that there could be a Y bearing group and an X-enriched group (i.e., a mixture X bearing and Y bearing).

In order to clarify whether or not the gene signatures from the SmartSeq2 data could be used to determine X and Y bearing spermatids, the gene signatures from the SmartSeq2 data can be used as the parameters upon which to perform dimensionality reduction in the 10X data. Dimensionality reduction was performed on just the expression levels of the autosomal genes identified to be upregulated in the Y bearing spermatids from the SmartSeq2 Bhutani *et al.* dataset. This meant that the clustering was entirely based upon the Y spermatid gene signature. It was

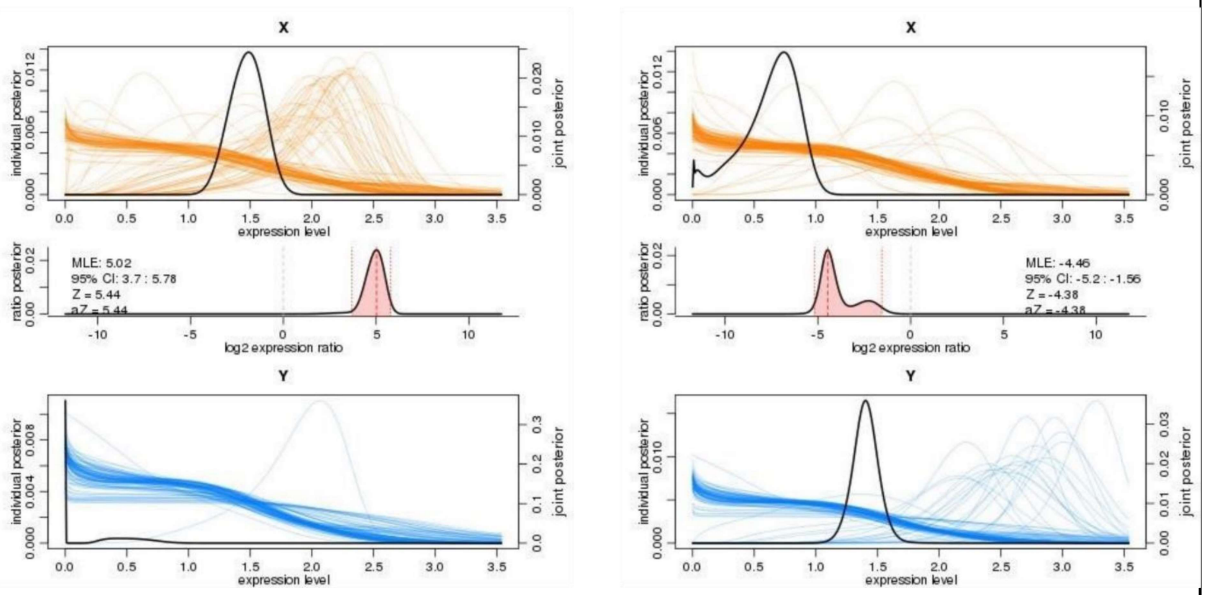
hypothesised that this gene signature could maybe distinguish two equal sized groups, which may be the X and Y bearing spermatids to be separated.

5.3 Results

5.3.1 SCDE analysis produced lists of genes upregulated in both the X bearing and the Y bearing cells

SCDE produces output charts per gene to show expression posteriors from individual cells, which are shown by the coloured lines, and from joint posteriors, for the two groups (X and Y bearing cells): these are shown on the top and bottom plots.

Figure 54

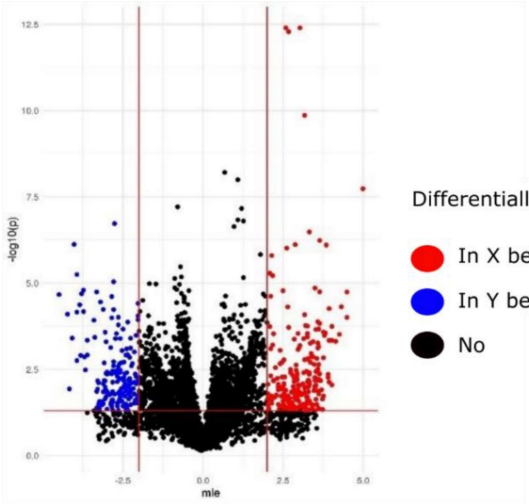


Output chart of SCDE, for genes (a) ENSMUSG00000086807, upregulated in X-bearing spermatids and (b) ENSMUSG00000022032, upregulated in Y-bearing spermatids. These output plots are produced for every DEG. Expression posteriors from individual cells, which are shown by the coloured lines, and from joint posteriors are shown for the two groups on the top and bottom plots. The middle plot shows the posterior of expression fold difference between the two defined (X and Y bearing) cell groups, with the 95% confidence intervals shown in red.

The middle plot shows the posterior of expression fold difference between the two defined X and Y cell groups, with the 95% confidence intervals shown in red.

Figure 54 shows an X-bearing spermatid DEG that has an average peak of expression level around 1.5, whereas a Y bearing spermatid DEG has a much smaller peak of around 0.5. As well as these charts which can be complex to interpret, SCDE also outputs a list of differentially expressed genes. Here, those genes positively differentially expressed are upregulated in the X bearing spermatids, and those negatively differentially expressed are upregulated in the Y bearing spermatids.

Figure 55



Volcano plot showing the genes significantly upregulated in X bearing spermatids (red) and the genes significantly upregulated in Y-bearing spermatids (blue). Significant upregulation is shown with logfc >2 and unadjusted p <0.05

Interpretation of differentially expressed gene data can take many forms, can be time consuming, and many data visualisation methods are highly critiqued. One of the most commonly used visualisations of differential gene expression analysis is a volcano plot, which shows for each gene, the negative base-10 log of the p-value against the base-2 log fold-change (McDermaid *et al.*, 2019). The higher up the y-axis a gene, the smaller the p-value. The closer the data points are to zero on the x-axis, the more similar the expression of the gene in the two groups: in this case the X and Y bearing groups. The genes upregulated in X-bearing cells (log2fc >2) are shown in red, and the genes upregulated in Y-bearing cells are shown in blue (log2fc > 2). Black points

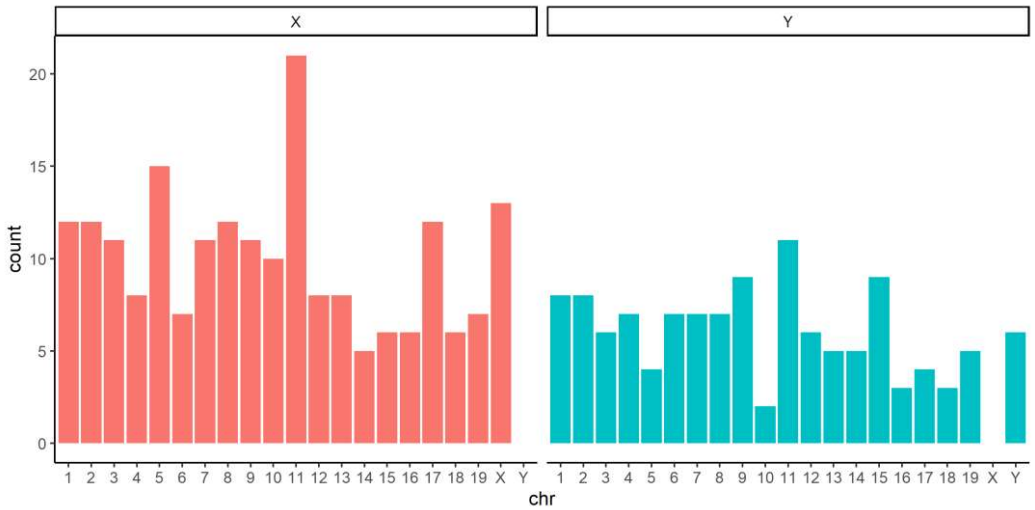
show genes which are not significantly upregulated in X or Y bearing – either because the log fold change is too small or because the p-value is too large.

It was found that 228 genes were significantly upregulated in the X bearing spermatids, with 188 of these being autosomal, and 162 were significantly upregulated in the Y bearing spermatids, with 116 of these being autosomal.

5.3.2 Analysis of the autosomal differentially expressed genes

A greater proportion of the differentially expressed genes found were autosomally encoded as opposed to sex-linked. The autosomal genes upregulated in the X-bearing spermatids came from all the autosomes, with chromosome 1, 2, 4, 11 and 17 having around as many genes upregulated as the X chromosome, or more. The autosomal genes upregulated in the Y-bearing spermatids came from all the autosomes, with chromosome 9, 11 and 15 having the highest number of upregulated genes (Figure 56).

Figure 56

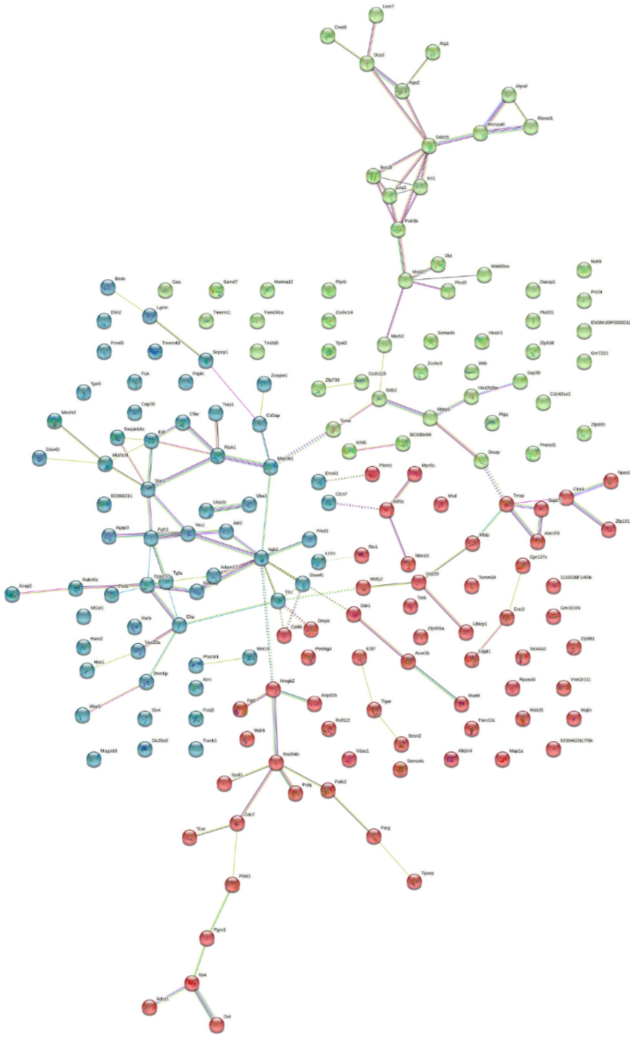


Frequency distribution of the number of significantly enriched genes (log2fc >2, p<0.05) mapping to each of the chromosomes in the X and Y bearing spermatids

5.3.2.1 Gene ontology analysis of the autosomal DEGs

The KEGG output for both the set of autosomal genes upregulated in the X-bearing spermatids and the autosomal genes upregulated in the Y-bearing spermatids was an empty data frame; meaning there were no biological pathways that were enriched in these gene lists by any more than would be expected by chance. This is what is to be expected of the autosomal genes, because any difference seen in autosomal gene levels between X and Y bearing spermatids should be due to regulation by sex-linked genes rather than due to functional enrichment.

Figure 57



STRING analysis network of the significantly enriched autosomal genes in X-bearing spermatids

The list of proteins encoded by the autosomal genes that were shown to be upregulated in X-bearing spermatids were uploaded into STRING (188 genes). This generated a map of protein-protein interactions, to which k-means clustering was applied.

This network showed 174 nodes (which represent the proteins encoded by the gene list), with 127 edges (protein-protein interactions). The average node degree was 1.46 and average local clustering coefficient was 0.382. The expected number of edges in the pathway was 119; the pathway did not have significantly more interactions than would be expected for a random set of proteins which showed that the proteins encoded by these genes are not involved in a common pathway – which is to be expected.

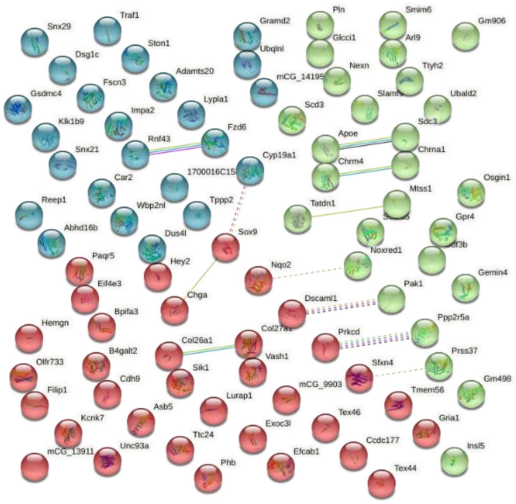
Gene ontology was performed in the STRING analysis, which showed the top biological processes, top molecular functions, and top cellular components functionally enriched in the network of proteins encoded by the autosomal genes upregulated in X-bearing spermatids. The significant processes are summarised in Table 15 and are determined to be significant if the strength of effect > 0.2 and FDR < 0.05.

Table 15: biological processes, molecular functions and cellular components of proteins encoded by the autosomal genes upregulated in X-bearing spermatids. Count in network has two numbers, the first referring to how many proteins in the input that are annotated with a particular term, and the second indicates how many proteins in total are associated. Strength is the log-10 ratio of the observed number of associated proteins to expected. False discovery rate is the p-values corrected for multiple testing using Benjamini-Hochberg procedure.

	GO-term	Description	Count in network	Strength of effect	False discovery rate
Biological process	GO:0006974	Cellular response to DNA damage stimulus	17 of 667	0.51	0.0192
	GO:0033554	Cellular response to stress	27 of 1493	0.36	0.0313
	GO:0044260	Cellular macromolecule metabolic process	64 of 4161	0.29	0.00018
	GO:0044267	Cellular protein metabolic process	47 of 3073	0.29	0.0066
	GO:0043170	Macromolecule metabolic process	79 of 5313	0.28	1.07e-5
	GO:0019538	Protein metabolic process	53 of 3634	0.27	0.0066
	GO:0010604	Positive regulation of macromolecule metabolic process	51 of 3467	0.27	0.0066
	GO:0051173	Positive regulation of nitrogen compound metabolic process	46 of 3119	0.27	0.0135
	GO:0006950	Response to stress	45 of 3045	0.27	0.0154
	GO:0009893	Positive regulation of metabolic process	52 of 3767	0.24	0.0173
	GO:0031325	Positive regulation of cellular metabolic process	46 of 3320	0.24	0.0441
	GO:0006807	Nitrogen compound metabolic process	77 of 5878	0.22	0.0021
	GO:0031323	Regulation of cellular metabolic process	71 of 5479	0.22	0.0066
	GO:0051171	Regulation of nitrogen compound metabolic process	66 of 5126	0.21	0.0102
	GO:0048523	Negative regulation of cellular process	60 of 4672	0.21	0.0238
	GO:1901564	Organonitrogen compound metabolic process	57 of 4475	0.21	0.0458
Molecular function	GO:1901363	Heterocyclic compound binding	65 of 4771	0.24	0.0031
	GO:0097159	Organic cyclic compound binding	65 of 4858	0.23	0.0039
	GO:0070013	Intracellular organelle lumen	56 of 4279	0.22	0.0115
Annotated keywords (UniProt)		WD repeat	10 of 266	0.68	0.0061
		DNA-damage	12 of 372	0.61	0.0060
		Magnesium	13 of 553	0.47	0.0293
		ATP-binding	26 of 1360	0.38	0.0055
		Nucleotide-binding	32 of 1749	0.37	0.0021
		Acetylation	41 of 3070	0.23	0.0283
		Repeat	55 of 4163	0.22	0.0060

The same analysis was performed for the 117 autosomal genes upregulated in the Y-bearing spermatids, where far fewer edges between nodes were found. This network showed 85 nodes (which represent the proteins encoded by the autosomal genes), with 11 edges (protein-protein interactions). The average node degree was 0.259 and average local clustering coefficient was 0.235. The expected number of edges in the pathway was 11.

Figure 58



STRING analysis network of the significantly enriched autosomal genes in Y-bearing spermatids

The pathway did not have significantly more interactions than would be expected for a random set of proteins. STRING did not find any significant functional enrichments in the network.

The top protein domains found by the Interpro database in the autosomal genes of both X-bearing and Y-bearing spermatids are summarised in Table 16.

Table 16: functional annotation chart for the autosomal genes upregulated in X-bearing and Y-bearing spermatids, identified by the DAVID program

	Category	Term	Gene count	p-value	Benjamini score
X-bearing spermatids	Uniprot sequence feature	REGION:Disordered	122	0.0005	0.1000
	Uniprot keyword post translational modification	Phosphoprotein	93	0.0120	0.2600
	GO term	cytoplasm	79	0.0001	0.0140
	GO term	nucleus	73	0.0001	0.0140
	Uniport sequence feature	COMPBIAS:Polar residues	67	0.0097	0.4900
	Uniprot keyword cellular component	Nucleus	59	0.0004	0.0130
	Uniport sequence feature	COMPBIAS:Basic and acidic residues	59	0.0020	0.2800
	GO term	protein binding	57	0.0510	1.0000
	Uniprot keyword domain	Repeat	56	0.0003	0.0034
	Uniprot keyword cellular component	Cytoplasm	54	0.0140	0.2200
	GO term	nucleoplasm	45	0.0002	0.0140
	GO term	cytosol	44	0.0140	0.5500
	GO term	metal ion binding	43	0.0170	0.8500
	Uniprot keyword post translational modification	Acetylation	41	0.0740	0.7800
	Uniprot keyword ligand	Nucleotide-binding	31	0.0049	0.0420
	GO term	nucleotide binding	29	0.0007	0.1200
	GO term	ATP binding	27	0.0006	0.1200
	Uniprot keyword ligand	ATP-binding	26	0.0047	0.0420
	Uniprot keyword molecular function	Transferase	25	0.0910	1.0000
	GO term	transferase activity	24	0.0270	1.0000
	Uniport sequence feature	COMPBIAS:Pro residues	23	0.0063	0.4000
	Uniport sequence feature	NP_BIND:ATP	22	0.0001	0.0370
Y-bearing spermatids	Uniport sequence feature	COMPBIAS:Polar residues	36	0.011	1
	GO term	membrane	35	0.012	0.63
	Uniprot domain	Transmembrane	30	0.019	0.13
	Uniprot domain	Transmembrane helix	29	0.011	0.13
	Uniprot	Glycoprotein	23	0.077	0.50
	Uniprot cellular component	Cell membrane	22	0.0094	0.24
	Uniprot sequence feature	CARBOHYD:N-linked (GlcNAc...) asparagine	20	0.078	1
	Uniprot sequence feature	TOPO_DOM:Cytoplasmic	18	0.081	1
	GO term	endoplasmic reticulum	15	0.0018	0.30
	Uniprot cellular component	Endoplasmic reticulum	10	0.03	0.39

5.3.2.2 The presence of motifs in the autosomal DEGs

All gene sequences upregulated in the X-bearing spermatids were given as the input to STREME, and 18 motifs with a p value < 0.05 were identified. Both the given and reverse complement strands were processed, and the test used was the Fisher exact test. Table 17 shows the output values from STREME.

Following this, the gene sequences upregulated in the Y-bearing spermatids were given as the input to STREME, and seven motifs with an unadjusted p value < 0.05 were identified. Both the given and reverse complement strands were processed, and the test used was the Fisher exact test. Table 18 shows the output.

Table 17 1: motif logo and percentage of genes containing the motif for motifs found in the genes upregulated in X-bearing spermatids

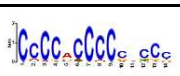
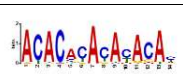
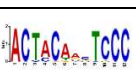
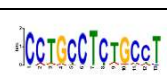
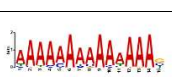
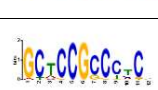
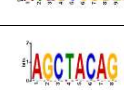
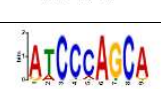
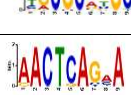
Motif	Motif logo	Percentage of genes containing motif site
CCCCACCCCCVCCC		70.9
AGCCAGGG		74.6
ACACACACAYACAY		26.2
AGGAAGGAGG		32.5
ACTACARCTCCC		16.5
CCTGCCTCTGCCT		21.4
AAAAAAAAAAAAAS		41.0
ACCAGGCTGGCCTCG		26.2
ATGGTGGCKCACAM		18.5
CCTGGAACTCACTS		18.2
GCTCCGCCCTCK		16.8
AGCTACAG		16.2
ATCCCAGCA		32.8
AAAAAGAAAGAA		39.9
CCCCGCCCC		61.0
TGCGCATGCGCA		21.7
AACTCAGAA		26.5
CGGMGCCGCGCG		17.1

Table 18: motif logo and percentage of genes containing the motif for motifs found in the genes upregulated in Y-bearing spermatids

Motif	Motif logo	Percentage of genes containing motif site
AAAAAAAAAAAAA		49.0
SCCCGCCCCT		52.4
CGCGGGGGCTGCCG		27.2
GTCCTGGAACTCA		29.3
CCCAGCACC		46.9
AGCCTGGTCYAC		27.2
CCAASTCGA		35.4

TomTom, part of the Meme Suite, was used to identify transcription factors that likely bind the motifs identified in the promoter regions of the differentially expressed genes. The output from TomTom is shown in Table 19, where transcription factors that may bind to the identified motifs are given.

Table 19: the novel motifs found in the genes upregulated in the X-bearing spermatids, and potential transcription factors that may bind these motifs

Motif	No. of matching transcription factors	Autosomal Transcription Factors	Sex linked transcription factors
CCCCACCCCCVCCC	35	*Zn281, Sp3, Klf5, Klf3, Sp2, Egr1, Sp1, Sall4, Zbt17, Maz, Klf4, Wt1, Srbp2, Egr2, Sp5, Klf15, Klf1, Sp4, Klf6, Zic1, E2f1, Srbp1, E2f6, E2f4, Ascl1, E2f7, E2f3, Mxi1, Myba, Htf4, Fli1, Vdr, Zbt7a*	*Gata1* (X) *Zfx* (X)
AGCCAGGG	11	*Sp2, Mxi1, Hen1, Zn335, Klf3, Klf6, Zbt7a, Bha15, Sp3, Ctcfl, Sp1*	
ACACACACAYACAY	7	*Egr2, Tbx21, Klf4, Brac, Egr1, Clock, Smad2*	
AGGAAGGAGG	28	*Elf5, Sp4, Etv2, Elf2, Ets2, Wt1, Ehf, Tead2, Sall4, Elf3, Erg, Nr1d2, Sp5, Etv6, Etv4, Klf15, Ptf1a, E2f3, Lyl1, Maz, Tyy1, Olig2, Hxa1, Spi1, Ets1, Htf4, Rorg*	*Taf1* (X)
ACTACARCTCCC	8	*Tha11, Zn143, Myba, Ascl1, Klf4, Snai2, Hsf1, Gabpa*	
CCTGCCTCTGCCT	39	*Klf6, Myod1, Snai2, Ascl1, Zn335, Klf15, Bha15, Sp4, Egr1, P53, Maz, Sp5, Prdm5, Spi1, Klf1, Po3f2, Myba, Maf, Smad2, Mafb, Wt1, P63, Lyl1, Nr1h3, Maff, Zbt17, Sp1, Etv6, Bach2, Zn281, Sp3, Cebpg, Klf3, Elf5, Atoh1, Myog, Jund, Egr2, Zfp57*	
AAAAAAAAAAAAAS	34	*Foxj3, Stat2, Irf3, Stat1, Hnf6, Foxq1, Foxl2, Lef1, Barx1, Cdx2, Evx2, Evi1, Foxj2, Irf2, Cux2, Tcf7, Foxd3, Irf1, Mef2c, Nanog, Sox4, Olig2, Irf8, Elf5, Mef2a, Foxa3, Irf4, Nr2e3, Spi1, Foxa2, Ets2, Aire*	*Sry* (Y) *Foxo4* (X)
ACCAGGCTGGCCTCG	17	*Gli1, Smad4, Tyy1, Esr2, Gcr, P63, Esr1, Zn335, P53, Prgr, Pax5, Err2, Vdr, Rarg, Nf1b, Smad2*	*Taf1* (X)
ATGGTGGCKCACAM	16	*Fosl2, Fos, Fosl1, Jun, Junb, Bach2, Atf3, Fosb, Jund, Mafb, Mafk, Maff, Zfp42, Tyy1, Smad4, Bach1*	

CCTGGAACTCACTS	26	*Pknx1, Nr1d2, Bcl6, Pbx1, Prdm5, Tead1, Vdr, Err2, Gfi1, Nr1h4, Hsf1, Err1, Zn322, Stat3, Myba, Ovol1, Relb, Batf, Elf2, Gfi1b, Sta5a, Nr1d1, Coe1, Stat4, Tead4*	*Mecp2* (X)
GCTCCGCCCTCK	31	*Sp4, Sp1, Klf3, E2f1, Sp3, Sp2, E2f4, Klf6, Egr1, Wt1, Klf4, Hen1, Egr2, Zbt17, Klf5, Ctcfl, Klf1, Klf15, Sp5, Cot2, Ctcf, Tyy1, Ptf1a, Pitx1, Mbd2, Maz, Nr1h3, Fli3, E2f2*	*Mecp2* (X) *Taf1* (X)
AGCTACAG	12	*Tfe2, Ascl1, Rfx6, Runx2, Ovol1, Myog, Runx3, Zn322, Dmrtb, Runx1, Bha15, Dmrt1*	
ATCCCAGCA	15	*Zn143, Tead4, Tha11, Zic2, Gfi1b, Gfi1, Stat3, Tead1, Sta5a, Stat4, Bach2, Mafb, Ari5b, Sta5b*	*Zic3* (X)
AAAAAGAAAGAA	39	*Foxj3, Stat1, Elf5, Lef1, Sox4, Tcf7, Irf3, Cux2, Foxq1, Foxl2, Nanog, Barx1, Otx2, Irf2, Batf3, Prdm1, Irf1, Stat2, Foxj2, Elf3, Spi1, Hnf6, Olig2, Sox10, Sox2, Spib, Irf8, Ets2, Etv6, Drmt1, Etc2, Irf4, Nr4a1, Smad2, Po2f1, Tf7l1*	*Sry* (Y) *Foxo4* (X) *Sox3* (X)
CCCCGCCCC	32	*Sp3, Sp2, Sp1, Sp4, E2f4, Egr1, E2f1, E2f7, Klf3, Klf6, Zbt17, E2f6, Egr2, Zn281, Klf15, Klf5, Klf1, E2f3, Maz, Sp5, Wt1, Srbp2, Sall4, Mxi1, Fli1, Klf4, Esr2, Mbd2, Rora, Zic1, Ctcfl*	*Gata1* (X)
TGCGCATGCGCA	4	*Nrf1, Max, Itf2, Arnt*	
AACTCAGAA	22	*Bach2, Nef2, Nf2l2, Bach1, Mafb, Mafk, Mafg, Nr2e3, Maff, Elf1, Maf, Nfya, Nfyc, Nkx21, Gfi1b, Gfi1, Foxi1, Sta5b, Zic2, Atf3, Nfyb, Batf3*	
CGGMGCCGCGCG	11	*Nfib, Ctcfl, Rfx1, Rfx3, Insm1, Rfx2, Hand1, Zeb1, Rfx6, Gli1, Nfic*	

TomTom was also used to find transcription factors that likely bind the motifs found in the genes differentially expressed in Y bearing spermatids. The output can be found in appendix 2.

These transcription factors identified in TomTom do not necessarily bind the genes differentially expressed in X-bearing and Y-bearing cells, but they are potential targets of binding sites that may be found in the differentially expressed genes

The potential transcription factors identified also showed common binding sites for upstream second level transcription factors. These motifs were identified by STREME and are detailed in appendix 2.

TomTom was used to find second level transcription factors that were likely to bind to the motifs identified in the transcription factors. These transcription factors are all summarised in appendix 2.

5.3.2.3 Motif analysis of just the long non-coding genes

Originally, motif analysis was run on the whole set of differentially expressed genes. However, both protein coding and long non-coding genes were found in the autosomal differentially upregulated gene lists.

11 genes were found to be long non-coding in the autosomal gene list upregulated in the X bearing spermatids: *4930463O16Rik, 9930104L06Rik, Gm7337, 4930417O22Rik, Gm11335, Platr14, 4933431E20Rik, 1700040D17Rik, C920006O11Rik, Gm6639* and *Zim3*.

30 genes were found to be long non-coding in the autosomal gene list upregulated in the Y bearing spermatids: *Gm5893, Gm8623, Gm15350, 4930579D09Rik, 1700125H03Rik, 4930538E20Rik, Gm13483, 1700112J05Rik, Platr20, Gm11650, 1700007P06Rik, 4921534H16Rik, 1700071M16Rik, Gm20389, A330023F24Rik, 4930448E22Rik, 4930528D03Rik, 4933417D19Rik, 1700024F13Rik,*

2810429I04Rik, 1700034G24Rik, 1700041M19Rik, Gm35549, Gm39336, 9430078K24Rik, 4930405A10Rik, Gm36245, Gm35554, 4930471G03Rik, 4930414N06Rik.

When the same motif analysis was run on just the long non-coding genes, the 11 input long non-coding genes differentially upregulated in the X bearing spermatids yielded 64 potential transcription factor binding sites. 269 potential transcription factors were identified that may bind these binding sites. Of these, eight were sex linked: *Sox3, Taf1, Elk1, Gata1, Tfe3, Mecp2, Klf8* (X-encoded) and *Sry* (Y-encoded). These sex-linked transcription factors may be implicated in functional differences between the X and Y bearing spermatids. The long-noncoding genes identified to be differentially expressed between X and Y bearing spermatids may bind these sex-linked transcription factors and, since the transcription factors are sex linked, they can cause differences in the X bearing and Y bearing spermatids.

Of the 30 input long-noncoding genes upregulated in the Y bearing spermatids, no significant motifs were found.

5.3.2.4 There were no differences in microRNA binding sites between the autosomal protein coding DEGs from the X bearing vs Y bearing spermatids

microRNA binding sites were identified using seedVicious and results are summarised below. No difference was found between the miRNA binding sites of the protein coding genes differentially expressed in the X bearing or Y bearing spermatids; to be expected but an important sanity check. The miRNA binding sites

found in the genes differentially upregulated in the X bearing and Y bearing spermatids are shown in Tables 20 and 21.

The percentage of the DEGs upregulated in X and Y bearing spermatids with at least one miRNA binding site for the five miRNAs identified is shown in Table 20.

Table 7: percentage of the DEGs upregulated in X and Y bearing spermatids with at least one miRNA binding site for the five miRNA identified using SeedVicious

miRNA	% DEGs Upregulated in X Bearing Spermatids with at least one binding site	% DEGs Upregulated in Y Bearing Spermatids with at least one binding site
miR-82-5p MIMAT0020323	59.3	51.1
miR-50-5p MIMAT0000021	23.1	26.7
miR-79-3p MIMAT0000051	39.5	34.9
miR-75-5p MIMAT0015108	14.1	15.1
miR-43-3p MIMAT0000014	55.9	46.5

Table21: miRNA binding sites found in the genes differentially upregulated in the X bearing spermatids

miRNA	No. of Genes (out of 177)	Which Genes
miR-82-5p MIMAT0020323	105 (59.3%)	*Ano4, Fmnl3, Tomm34, Clock, Cflar, Gm7337, Cep78, Madd, Usp33, Ripk1, Papln, Gpr137c, Rskr, Atxn7l3, 8030462N17Rik, Dazap1, Lgmn, Sema4c, Tmem43, Polq, Npas2, Ddr1, Cd2ap, Trank1, Zfp24, Hps1, Blmh, Wdfy2, Wdcp, Tiparp, Phactr4, Prkd3, Wdr25, Myo5c, Ticrr, Preb, Tonsl, Rbbp5, Fam13c, Plekhg4, Polr3b, Adam17, Ero1a, Zfp661, Abl2, 4933431E20Rik, Parg, Cgas, Eps15l1, Trrap, Acap2, Slc44a2, Zfp955a, 1110038F14Rik, Gm6639, E2f7, Vdac1, Ddx55, Uggt1, Mcoln2, Nbeal1, Ptprb, Vps33a, Ublcp1, Zfp738, Zcchc14, Dmxl1, Pkd2l1, Dbt, Zfp995, Palb2, Srcap, Zcchc9, Zcwpw1, Dnmbp, Rad54b, Pitrm1, Majin, Map3k1, Foxj3, Sesn2, Klhl5, C4a, Stat1, Pgm3, Wipf2, Wnt16, Noc3l, Slamf1, Heatr3, Atrn, Zfp976, Kif5b, Scpep1, Tfrc, Pbld2, Adcy1, Hormad2, Yes1, Mrpl27, Ago2, 4930463O16Rik, Fam241a, Dcp2, Dennd11*
miR-50-5p MIMAT0000021	41 (23.1%)	*Ano4, Ric1, Cflar, Uba3, Clock, Gm7337, Nek5, Usp33, Gpr137c, Ripk1, Rad54b, Preb, Mcoln2, Sema4b, Pgm3, Zfp24, Hps1, Noc3l, Sema4c, Zfp938, Ttc5, Map1a, Nbeal1, Pitrm1, Atrn, Fam13c, Zfp976, Kif5b, Trank1, Trrap, Tfrc, Vps33a, Il18, Ddx43, Gm10109, Elfn2, Ddr1, Yes1, Acap2, Ago2, Sesn2*
miR-79-3p MIMAT0000051	70 (39.5%)	*Zfp661, Ano4, Dazap1, Ptcd3, Fmnl3, Cflar, Uba3, Itgb1, Clock, Gm7337, Abl2, Foxj3, Gpr137c, Ripk1, Papln, Tyw5, Rskr, Lgmn, Klhl5, Rbbp5, Tmem43, Polq, Npas2, Sema4b, Pkd2l1, Cd2ap, E2f7, Trank1, Zfp24, Wnt16, Blmh, Ube2z, Sema4c, Wdfy2, Wdr6, Slamf1, Heatr3, Map3k1, Tiparp, Map1a, Wdr25, Nbeal1, Acap2, Itpa, C920006O11Rik, Rbmxl1, Kif5b, Tgfa, Parg, Atrn, Trrap, Dmxl1, Acvr1b, Tfrc, Palb2, Pbld2, Vdac1, Vmn2r111, Ptprb, Platr14, Six4, Adcy1, Gm10109, Hormad2, Dbt, H2bc21, Srcap, Ago2, Dcp2, Dennd11*
miR-75-5p MIMAT0015108	25 (14.1%)	*Srcap, Cflar, Ric1, Pkd2l1, Zfp24, Atrn, Tonsl, Papln, Cdc42se2, Samd7, Mcoln2, Ttc5, Ddr1, Gaa, Polq, E2f7, C920006O11Rik, Notch2, Zcchc14, Adam17, Npas2, Elfn2, Parg, Ago2, 1110038F14Rik*
miR-43-3p MIMAT0000014	99 (55.9%)	*Clock, Cflar, Gm7337, Madd, Pigq, 9930104L06Rik, Ripk1, Stat1, Nek5, Papln, Gpr137c, Agap3, Ptcd3, Tmem43, Polq, Npas2, Sema4b, Msrb2, Cd2ap, Zfp24, Fmnl3, Blmh, Wdfy2, Dmpk, Samd7, Tiparp, Prkd3, Ric1, Ticrr, Preb, Spdl1, Fcsk, Fam13c, Cdc42se2, Rbmxl1, Polr3b, Ero1a, Zfp661, Abl2, Ttc5, Acap2, Gm6639, Tnip1, Hmgb2, Zfp977, Nbeal1, Ptprb, Vps33a, Zfp738, Zcchc14, Dmxl1, Anp32b, Plekhg4, Pkd2l1, Eps15l1, Elfn2, Dbt, Fgfr3, Zfp995, Pgd, Srcap, Rad54b, Tgfa, Tmem11, Alg1, Mbip, Map3k1, Foxj3, Map1a, Ddb2, Pgm3, Klhl5, Itgb1, Mcoln2, Mvd, Alyref, Wnt16, Noc3l, Zfp938, Wdr6, Heatr3, Phactr4, Atrn, Zfp976, Trank1, Scpep1, Tfrc, Clcn7, Vmn2r111, Supt7l, Mageb3, Adcy1, Hormad2, 4930417O22Rik, Notch2, Ago2, 4930463O16Rik, Fam241a, Dennd11*

Table 8: miRNA binding sites found in the genes differentially upregulated in the Y bearing spermatids

miRNA	No. of Genes (out of 86)	Which Genes
miR-82-5p MIMAT0020323	44 (51.1%)	*1700012A03Rik, 1700034G24Rik, 1700071M16Rik, 1700125H03Rik, 4930448E22Rik, 4930471G03Rik, 9430078K24Rik, A330023F24Rik, Adamts20, Arl9, Asb5, Bpifa3, Car2, Col27a1, Dscaml1, Dus4l, Efcab1, Eif4e3, Filip1, Fzd6, Gemin4, Gm13483, Gm35554 Gm5893, Impa2, Lypla1, Mtss1, Nexn, Noxred1, Nqo2, Osgin1, Pak1, Platr20, Ppp2r5a, Prss37, Ptp4a1, Rnf43, Sfxn4, Smim6, Snx29, Ston1, Tlcd4, Tppp2, Unc93a*
miR-50-5p MIMAT0000021	23 (26.7%)	*Scara5, Gemin4, Ptp4a1, Col27a1, Nexn, Gria1, Hemgn, Asb5, Tlcd4, Snx29, Filip1, Adamts20, A330023F24Rik, Gm36245, Efcab1, Glcci1, 4921534H16Rik, Or4n4b, Ubqlnl, Ston1, Prkcd, Vash1, 1700071M16Rik*
miR-79-3p MIMAT0000051	30 (34.9%)	*9430078K, A330023F, Adamts20, Ccdc177, Col27a1, Dscaml1, Efcab1, Eif4e3, Exoc3l, Filip1, Gm35554, Gramd2, Hemgn, Impa2, Lypla1, Mtss1, Nexn, Prkcd, Prss37, Ptp4a1, Scd3, Sfxn4, Sik1, Snx29, Sox9, Ston1, Tatdn1, Tex46, Tlcd4, Ttc24*
miR-75-5p MIMAT0015108	13 (15.1%)	*Gria1, Filip1, Tatdn1, Chrm4, Apoe, 2810429I, Tlcd4, 1700016C, Glcci1, 1700012A, Scd3, Prkcd, Chrna1*
miR-43-3p MIMAT0000014	40 (46.5%)	*Filip1, Gpr4, Snx29, 4933417D19Rik, Col26a1, Gm906, Sfxn4, Fzd6, Sik1, Col27a1, 4930405A10Rik, Glcci1, Mtss1, Abhd16b, Adamts20, Or4n4b, Nqo2, Dscaml1, Eif4e3, Car2, Klk1b9, Scd3, Impa2, Fscn3, Tatdn1, Pln, Vash1, Pak1, Scara5, Reep1, B4galt2, Gm13483, Platr20, Tppp2, Tex46, Lypla1, Lurap1, Dsg1c, Ccdc177, Chrna1*

5.3.3 Key transcription factors were identified in common motif analysis of both the overall autosomal gene lists and the long non-coding genes

For the X-bearing spermatids, *Gata1, Zfx, Taf1, Foxo4, Mecp2, Zic3, Sox3, Klf8, Cdx4* and *Tfe* were found as potential transcription factors regulating the entire population of the differentially expressed genes. These are all located on the X chromosome. *Sry* was also found to be a potential transcription factor regulating the differentially expressed genes, which is located on the Y chromosome.

For the Y-bearing spermatids, *Sry* was found to be a potential transcription factor regulating the differentially expressed genes, which is located on the Y chromosome. *Foxo4, Gata1, Klf8, Zfx, Tfe3* and *Cdx4* were also found as potential transcription factors regulating the differentially expressed genes, but these are all located on the X chromosome.

Just because these transcription factors bind motifs found in the genes differentially expressed, does not mean that they will definitely be regulating these genes, as they have many targets. It simply means they may regulate. Both STREME and TomTom work to find potential binding sites and then transcription factors that may bind these potential motifs – so just because a transcription factor is identified does not mean it definitely binds the motif, and even if it does bind the motif, the motif may not even be present in the DEG. It is simply the best analysis tool available at this time. Working on the assumption that the differentially expressed genes in the Y bearing spermatids will not be regulated by transcription factors on the X chromosome, it is only *Sry* of interest in the Y-bearing spermatids.

The motif analysis of just the long non-coding genes provided a list of transcription factors of particular interest, *Sox3, Taf1, Elk1, Gata1, Tfe3, Mecp2, Klf8* and *Sry*. All of these were also identified in the original motif analysis of all the genes, except *Elk1*.

5.3.3.1 GO analysis of the putative transcription factors in X-bearing spermatids showed common functional pathways

KEGG analysis of the potential sex-linked transcription factors activated in X-bearing spermatids showed the presence of basal transcription factors, the Ras signalling

pathway, FoxO signalling pathway and signalling pathways regulating the pluripotency of stem cells.

String analysis showed that there were not significantly more interactions than expected in the proteins produced by these transcription factor genes, but showed strong interactions between *Zic3* and *Sox3*, and *Sox3* and *Zfx*. The associated annotated keywords were 'activator', 'DNA-binding', 'transcription regulation' and 'developmental protein'.

5.3.3.2 Refining the candidate gene list

Absolute read count differences between the X bearing and Y bearing spermatids removed some transcription factors from the candidate gene list.

Figure 59

	Early round spermatids		Mid round spermatids		Round spermatids		Late round spermatids		Elongating spermatids		Condensing spermatids	
Tfe3	0	0	0	0	0	0	14.8	7.9	17.5	0	0.4	22.0
Gata1	0	0	0	0	0	0	0	0.04	7.8	0	0	0
Taf1	5.1	0	0.6	0	5.1	0	1.6	0	0.9	0	3.7	0
Cdx4	0	0	0	0	0	0	0	0	0	0	0	0
Mecp2	0	0	0	0	0	12.9	0.3	0.04	1.7	0	0	0
Klf8	0	0	0	0	0	0	0.3	0	0	0	0	0
Foxo4	0.1	0	2.3	0	0.1	0	7.1	8.1	2.0	7.3	0	2.6
Sox3	0	0	0	0	0	0	0	0	0	0	0	0
Zic3	0	0	39.0	14.7	0	15.7	5.7	1.1	0.2	0.1	0	0
Zfx	0.6	0	7.1	3.6	0.6	0.1	0.2	8.4	0.6	3.5	0.1	0
Elk1	0.1	0	15.4	2.9	0.1	0.5	0	0	0.1	0	0	0
Sry	0	0	0	0	0	0.1	1.0	4.9	0	0	0	0

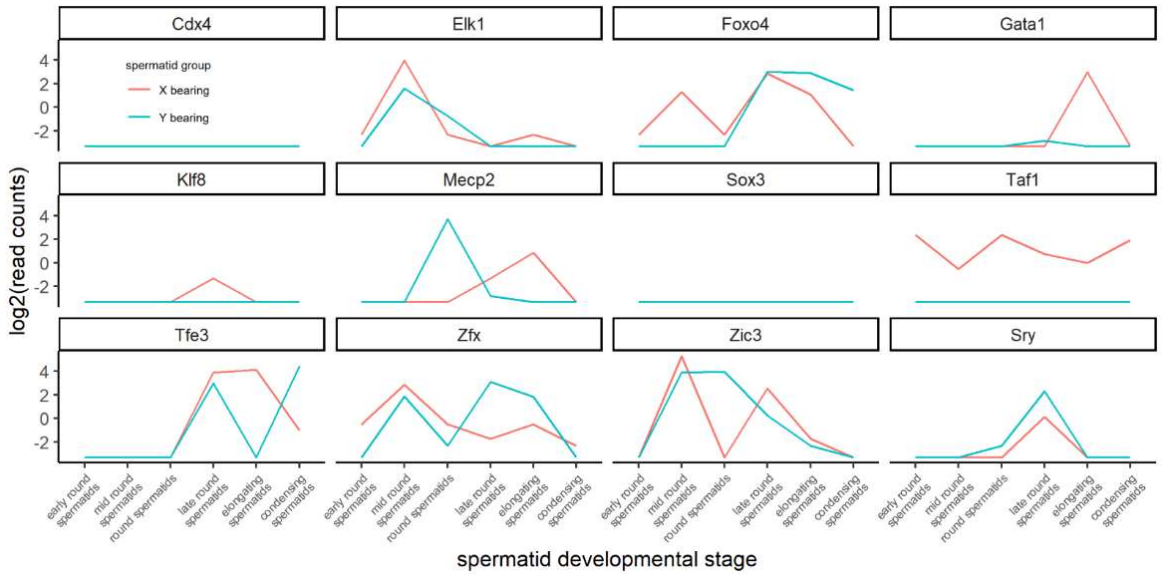
The absolute read counts of the X and Y bearing spermatids for each key gene for all the cells of each developmental stage

Figure 59 shows the read counts of the X and Y bearing spermatids for each key gene for all the cells of each developmental stage. It shows how the read counts change with developmental stage and between spermatid groupings.

Looking at the absolute read counts in Figure 59, as well as Figure 60, *Cdx4, Sox3* and *Klf8* are ruled out as candidate cis-limited genes as they are expressed to absolute minimal levels.

Tfe3 is also ruled out as a cis-limited gene as it appears to be being shared across the cytoplasmic bridges. It is present at the round spermatid stage onwards in both X and Y bearing spermatids. *Mecp2, Foxo4, Zfx, Zic3, Elk1* and *Sry* are excluded for the same reason. *Mecp2* is expressed in both X and Y spermatids at the round spermatid and elongating stage. *Foxo4* is expressed only in the X bearing spermatids in the early stages but from the round spermatid stage onwards it is expressed in both X and Y bearing spermatids; so, can be ruled out as a candidate cis-limited gene. *Zic3* is expressed in both X and Y bearing spermatids from the secondary spermatocyte stage onwards.

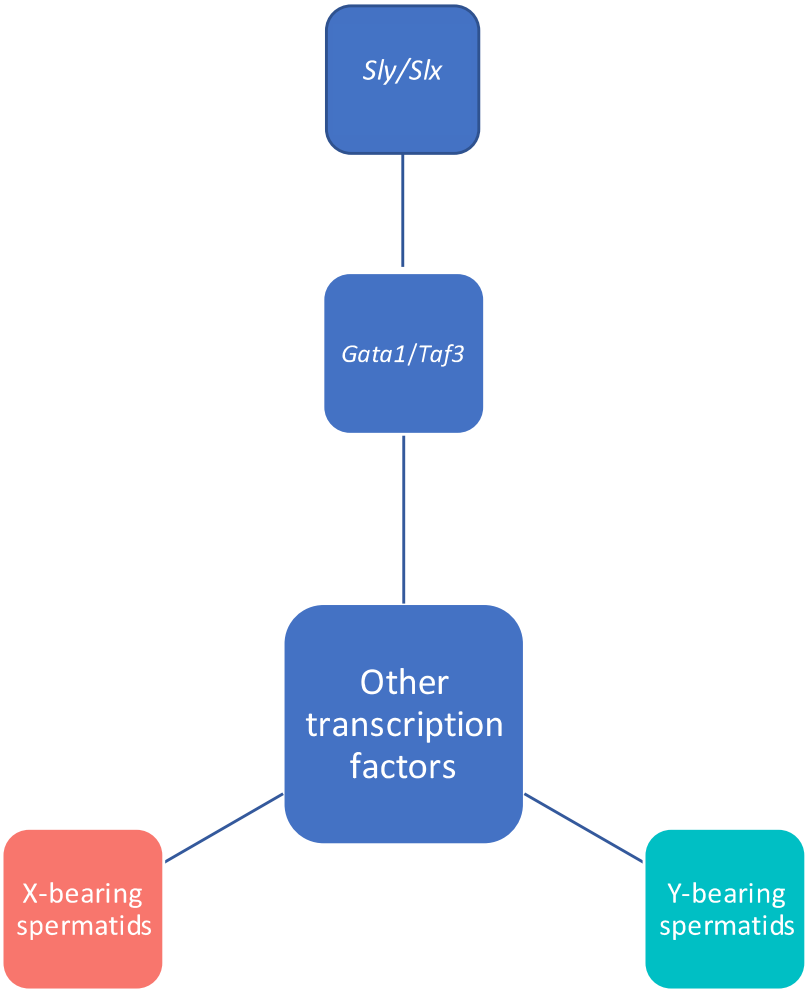
Figure 60



log2 of the read counts of each potential transcription factor in X bearing (red line) and Y bearing (blue line) spermatids.

Gata1 and *Taf1* are the only ones of the possible transcription factors identified that could be considered candidate cis-limited genes; so they will be examined more closely. Both had been identified as both a potential level transcription factor that binds common motifs found in the differentially expressed genes, and a potential transcription factor that binds the common motifs found in the potential transcription factors binding the DEGs, and a transcription factor that binds to common motifs in the long non-coding genes. This means if they are cis-limited, they are likely to regulate other transcription factors that may not be cis-limited.

Figure 61



Potential mechanism of regulatory pathways differentiating between X-bearing and Y-bearing spermatids

5.3.4 Analysis of the sex-linked differentially expressed genes

Whilst the autosomal genes were a critical first consideration, some sex-linked genes were also found by SCDE to be upregulated in X-bearing and Y-bearing cells; these genes could have either a functional or regulatory role. This is in contrast to the autosomal genes which could only have a regulatory role in differentiating between X and Y bearing spermatids.

Any of the genes that were identified in initial differential gene expression analysis that were sex linked, as well as any sex-linked transcription factors that may bind the differentially expressed genes became the candidate cis-limited gene list. When the list of potential candidate genes was identified, it was important to look deeper into their biological significance; could these candidate genes plausibly be causing morphological differences in X and Y bearing sperm? For example, might they be involved in the cytoskeleton remodelling or other processes involved in shaping the sperm?

5.3.4.1 GO analysis of the sex-linked DEGs

The sex-linked genes found to be upregulated in X and Y bearing cells are summarised in Table 23 below. Sex-linked genes could have either a functional or regulatory role, whereas any autosomal genes found could only have a regulatory role.

Table23: sex-linked genes upregulated in X-bearing and Y-bearing spermatids, and the roles of those genes

	Gene ID	Gene symbol	Role
Upregulated in X-bearing spermatids	ENSMUSG00000031066	*Usp11*	Involved in protein ubiquitination, which controls many intracellular processes such as the cell cycle and transcriptional activation
	ENSMUSG00000031134	*Rbmx*	RNA binding motif protein, enables chromatin binding activity, identical protein binding and nucleic binding
	ENSMUSG00000031271	*Serpina7*	Serine peptidase inhibitor, enables hormone binding activity, negative regulation of endopeptidase activity, response to vitamin A and thyroid hormone transport
	ENSMUSG00000031297	*Slc7a3*	Cationic amino acid transporter, allowed uptake of arginine from extracellular media into cells
	ENSMUSG00000031311	*Nono*	Enables chromatin binding, negative regulation of transcription, regulation of circadian rhythm
	ENSMUSG00000031357	*Syap1*	Expressed in nervous system, exact function unclear. Knockout mice are viable and fertile but have deficient motor behaviour (Von Collenberg *et al.*, 2019)
	ENSMUSG00000035725	*Prkx*	Predicted to enable cAMP-dependent protein kinase activity, cell adhesion, angiogenesis and endothelial cell proliferation
	ENSMUSG00000067215	*Usp51*	Regulates DNA damage response
	ENSMUSG00000085157	*4930515L19Rik*	Long non-coding RNA
	ENSMUSG00000085536	*Gm14950*	Predicted gene
	ENSMUSG00000086807	*Platr21*	Pluripotency associated transcript
	ENSMUSG00000095005	*Astx6*	Amplified spermatogenic transcripts X encoded
	ENSMUSG00000096374	*Gm6664*	Predicted gene
Upregulated in Y-bearing spermatids	ENSMUSG00000091749	*Tspy-ps*	Testis specific protein-Y encoded pseudogene
	ENSMUSG00000094575	*Gm20738*	Predicted gene
	ENSMUSG00000094575	*LOC108168660*	Gamete generation
	ENSMUSG00000095950	*Gm20737*	Predicted gene
	ENSMUSG00000095950	*Gm20793*	Predicted gene
	ENSMUSG00000101928	*Gm20824*	Predicted gene

Table24: functional annotation chart of the protein domains found in the sex-linked genes upregulated in X-bearing and Y-bearing spermatids, identified from the Interpro database as part of the DAVID program

		Term	Gene count	p-value	Benjamini score
X-bearing spermatids	GO term	Nucleus	6	0.064	1
	GO term	Chromosome	3	0.017	0.59
	GO term	Chromatin binding	3	0.017	0.62
	Uniprot keyword cellular component	Chromosome	3	0.018	0.22
	Uniprot sequence feature	ACT_SITE:Proton acceptor	3	0.027	0.21
Y-bearing spermatids	Go term	Methylated histone binding	2	0.0096	0.0096
	Interpro database	Spindlin/spermiogenesis-specific protein	2	0.011	0.022
	GO term	Gamete generation	2	0.014	0.056

5.3.5 Refining the candidate cis-limited gene list

The sex-linked differentially expressed genes in X and Y bearing spermatids and the sex-linked transcription factors identified are the initial candidate cis-limited genes. Their functions are summarised below in Table 25.

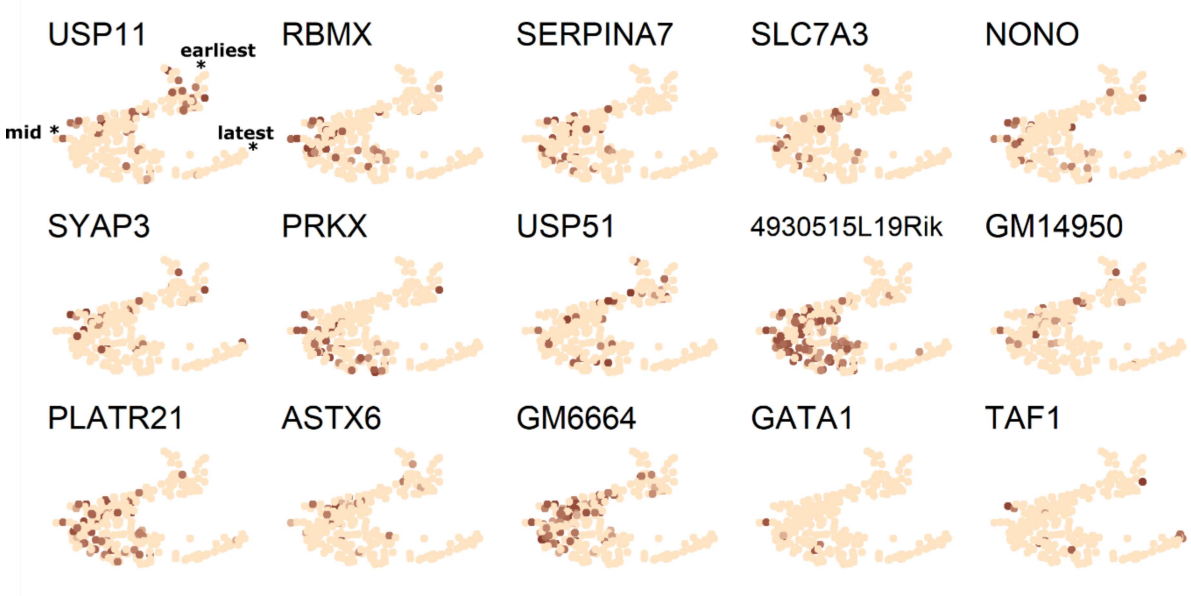
Table 25: functions of the sex-linked differentially expressed genes for X and Y bearing spermatids, and the functions of the sex-linked transcription factors that may bind the autosomal DEGs.

For X Bearing Spermatids		For Y Bearing Spermatids	
Usp11	Predicted to be involved in protein deubiquitination, be involved in proteolysis, be active in nucleus. Enables thiol-dependent deubiqutitinase and transcription corepressor binding activity	*Tspy-Ps*	Pseudogene
Rbmx	RNA binding motif protein, enables chromatin binding activity, nucleic acid binding activity, cellular response to interleukin-1 and regulation of RNA metabolic process. Predicted to act upstream of RNA splicing	*Gm20738*	Predicted gene 20738
Serpina7	Predicted to enable hormone binding activity, negative regulation of endopeptidase activity. Predicted to be located in extracellular region	*Loc108168660*	Predicted gene 21249
Slc7a3	Allows uptake of arginine from the extracellular media	*Gm20737*	Predicted gene 20737
Nono	Enables chromatin binding activity, located in the nucleus. Part of RNA polymerase II transcription regulator complex.	*Gm20793*	Predicted gene 20793
Syap1	Involved in TORC2 signalling, cellular response to epidermal grow factor, and positive regulation of fat cell differentiation	*Gm20824*	Predicted gene 20824
Prkx	Predicted to enable cAMP-dependent protein kinase activity, located in nucleus		
Usp51	Predicted to enable chromatin binding activity, regulates cell cycle and double-strand break repair		
4930515I19rik	Non-coding RNA		
Gm14950	Non-coding RNA		
Platr21	Non-coding RNA		
Astx6	Non-coding RNA		
Gm6664	Predicted gene 6664		
Gata1	Enables DNA binding activity, DNA-binding transcription activator activity, RNA-polymerase II specific and p53 binding activity. Located in the nucleus		
Taf1	Enables DNA binding activity, histone acetyltransferase activity and promoter specific chromatin binding activity. Located in pronucleus		

5.3.4.1 The candidate X-linked cis-limited genes are expressed to higher levels at the early-mid spermatogenesis stage

The gene expression levels of the candidate X-linked cis-limited genes are plotted onto the UMAP of the spermatids to identify when they are activated.

Figure 62



Read counts per cell of the candidate X-linked genes coloured onto the UMAP of the spermatids. The top of the UMAP plot is the earliest stage spermatids, with the latest stage at the bottom. The highest levels tend to be earlier/mid spermatogenesis stage

The highest levels of the X-linked candidate cis-limited genes tend to be earlier to mid spermatogenesis when the sperm cells are still developing in syncytia. The fact that they are evading being shared even at this stage suggests they are credible candidate cis-limited genes.

5.3.4.2 The candidate cis-limited genes share common gene ontologies including chromatin binding

These candidate cis-limited genes did share some common gene ontologies. The David database found three of these candidate genes to be involved in transcription from RNA polymerase II promoter (*Gata1, Rbmx, Taf1*) and three to be associated with ubiquitin-dependent protein catabolic processes (*Taf1, Usp11, Usp51*). It also found two genes were associated with myeloid cell differentiation (*Gata1, Prkx*), two with peptidyl-serine phosphorylation (*Taf1, Prkx*), two with cellular response to DNA damage stimulus (*Taf1, Nono*) and protein autophosphorylation (*Taf1, Prkx*), and two genes involved with protein deubiquitination (*Usp11, Usp51*).

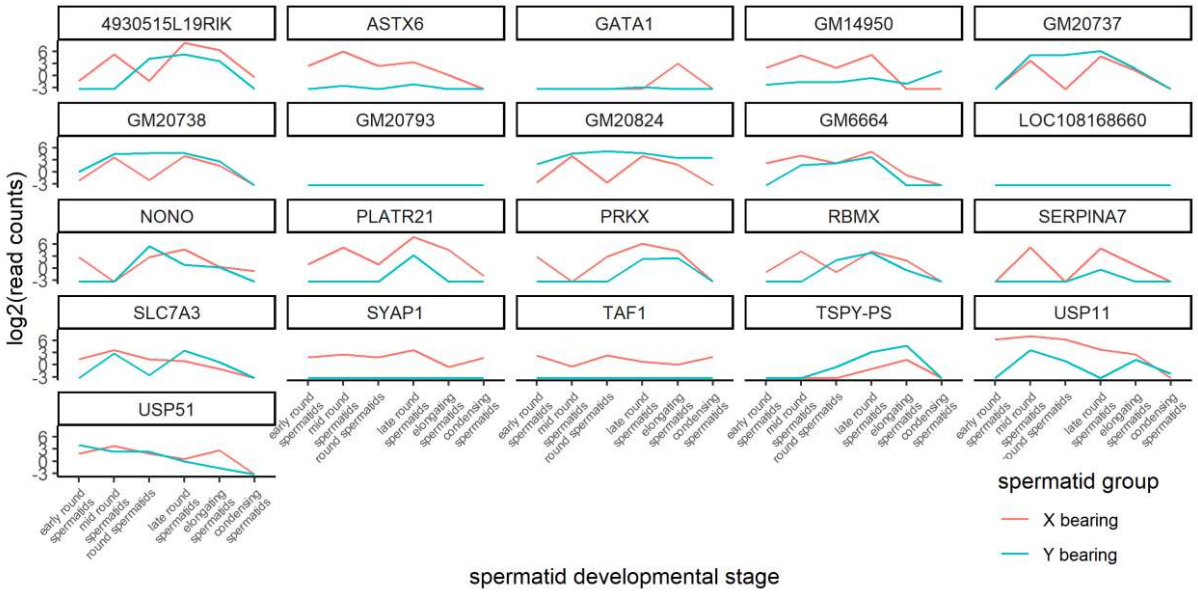
Three genes were associated with chromatin binding (*Gata1, Rbmx, Nono*), two with p53 binding (*Gata1, Taf1*), two with thiol-dependent ubiquitin-specific protease activity (*Usp11, Usp51*), one with cysteine-type peptidase activity (*Usp11, Usp51*), and two with RNA polymerase II sequence-specific DNA binding transcription factor binding (*Gata1, Taf1*). Three genes were also associated with the GO term ‘chromosome’ (*Nono, Usp11, Usp51*), seven associated with GO term ‘nucleus’ (*Gata1, Rbmx, Taf1, Nono, Platr21, Prkx, Usp11*), and five associated with GO term ‘nucleosome’ (*Gata1, Taf1, Nono, Prkx, Usp11*).

The KEGG pathway function identified *Rbmx* as being associated with the spliceosome, *Taf1* with basal transcription factors, and *Serpina7* with thyroid hormone synthesis.

5.3.4.3 The levels of each candidate gene were plotted to show the change in expression over time, in both the X and Y bearing spermatids

Plotting the levels of each candidate gene over time for both the X and Y bearing spermatids allowed more of the candidate cis-limited genes to be ruled out: *4930515l19rik, Gm20737, Gm20738, Gm20793, Gm20824, Gm6664, Loc108168660, Nono, Gm14950, Prkx, Rbmx, Slc7a3, Tspy-Ps, Usp11* and *Usp51*.

Figure 63



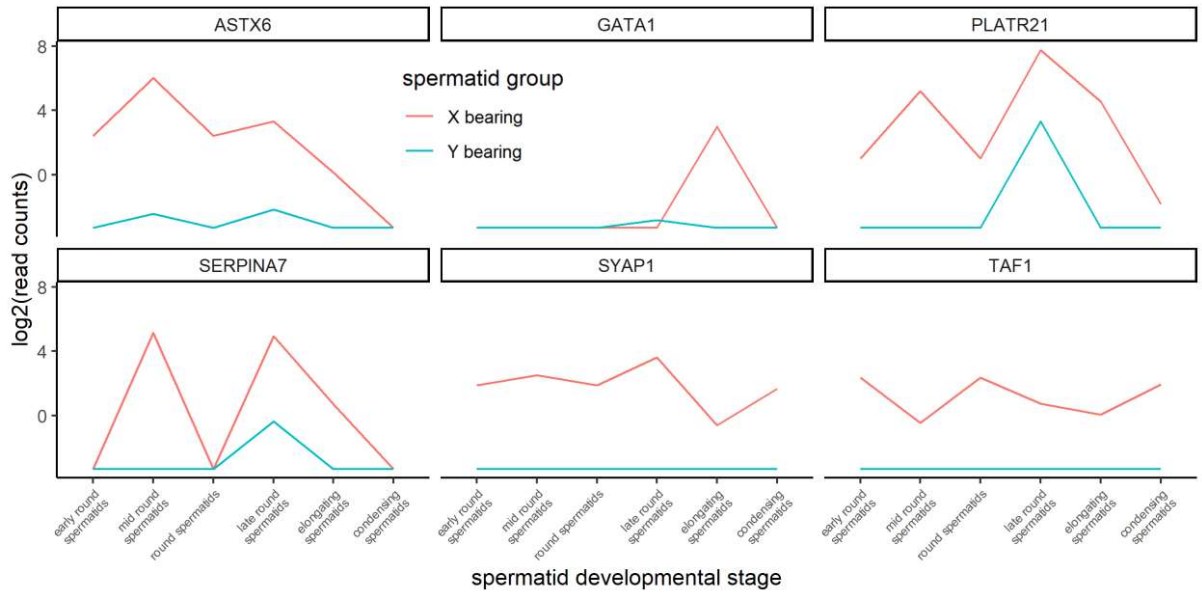
Expression level of each candidate cis-limited gene at for cells at different stages of development, by genotype. The red line represents X bearing spermatids and the blue line represents levels in the Y bearing spermatids. Some genes show expression in both the X vs Y bearing spermatids, so these are ruled out as candidate cis-limited genes

The reason for ruling these out is that, even though they were shown in SCDE to be differentially expressed, looking at the changes in expression levels over time, both spermatid groups are expressing the gene, in at least one developmental stage. Whilst at some time stages they may be differentially expressed, they cannot be truly

cis-limited if they are expressed in both X and Y bearing during at least one time stage.

The resultant candidate cis-limited gene list is: *Astx6, Gata1, Platr21, Serpina7, Syap1* and *Taf1.* Whilst *Platr21* has a peak for Y bearing spermatids, Figure 66 shows this is due to just one anomalous cell.

Figure 64

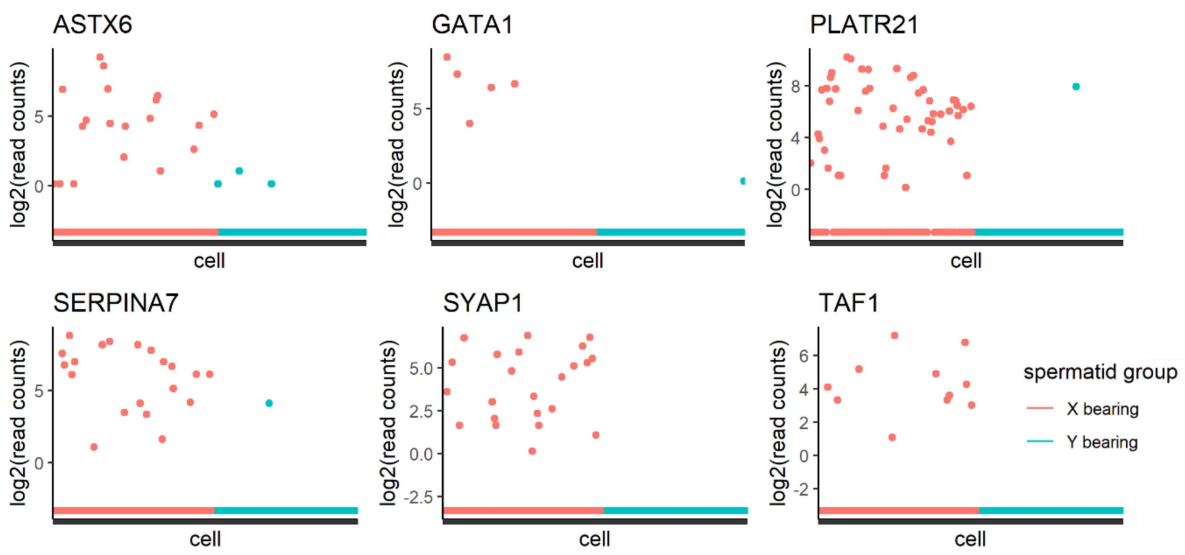


Expression level of each candidate cis limited gene for cells at different stages of development, by genotype. The red line represents X bearing spermatids and the blue line represents levels in the Y bearing spermatids.

5.3.4.4 Expression levels of the candidate genes in each individual cell shows the distribution between the X bearing and Y bearing spermatids

Scatter plots showing the log2-read counts of each X-linked candidate gene in every individual cell were plotted (Figure 66). No candidate Y-linked cis-limited genes were identified.

Figure 65



Expression levels per cell of the identified candidate cis-limited genes, coloured by X bearing (red) or Y bearing (blue) spermatid

5.3.4.5 comparing known cis-limited genes with the novel candidate cis-limited gene list

No common novel transcription factor binding sites were identified in *Smok2B* and *Spam1* when the parameters were adjusted in line with novel motif identification performed earlier in the analysis of the differentially expressed genes in STREME.

This meant that the motifs found in the novel candidate gene list could not be compared with motifs found in known cis-limited genes.

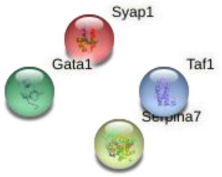
5.3.6 Biological significance of the candidate genes

The final candidate cis-limited gene list is:

- *Gata1*
- *Serpina7*
- *Syap1*
- *Taf1*
- *Astx6* (long non-coding RNA)
- *Platr21* (long non-coding RNA)

These are all X-linked genes, no Y-linked candidate cis-limited genes were found. Of the four protein coding candidate X-linked cis-limited genes, no significant pathways connecting the four were connected: they may all be involved in separate pathways.

Figure 66



STRING analysis pathway of proteins encoded by the candidate X-linked cis-limited genes and their interactions.

5.3.6.1 *Gata1*

Gata1 is a transcription factor found in this study to potentially bind many of the differentially expressed genes in X bearing spermatids. *Gata1* is known to be a hematopoietic transcription factor – it is essential for normal development of blood cells, both during embryogenesis and in adults. It has been shown to remain bound to chromatin during mitosis and this mitotic retention is critical for normal reactivation

of transcription after global transcription silencing. The loss of *Gata1* in the embryo has been shown to give a qualitative defect in erythroid cell differentiation, rather than a quantitative effect. *Gata1* regulates β1-tubulin; β1-tubulin knockout platelets have been shown to have abnormal cell shape (Italiano *et al.*, 2003). The marginal microtubule band of the blood platelets is what maintains the discoid shape; and the microtubules are polymers made from α-tubulin and β-tubulin. The intracellular force generated by microtubules drives many cell functions including morphogenesis and maintenance of cell shape. B1-tubulin has been shown to be critical for determining cell shape in platelets, and knockout of *Gata1* caused the same impacts as knockout of β1-tubulin (Italiano *et al.*, 2003). Knowing that the genes differentially expressed in X-bearing spermatids had *Gata1* binding sites in common, it could be hypothesised that *Gata1* also regulates shape maintenance in spermatids.

5.3.6.2 *Taf1*

Taf1 is a transcription factor found in this study to potentially bind many of the differentially expressed genes in X bearing spermatids. TAF proteins are encoded for by the *Taf* genes, and, along with TATA-binding protein are part of the TFIID multimeric protein complex and are involved in mediating the transcription preinitiation complex, which allows transcription of DNA to RNA. It has been shown that it is likely that mammalian germ cells require a highly specialised form of TFIID to programme their development (Gura *et al.*, 2020). *Taf4b*-deficiency in mouse is associated with reduction in the adult spermatogonial stem cell pool leading to progressive male infertility in early adulthood, and *Taf4b* has been shown to be just as important in humans for fertility. Whilst *Taf4b* is a separate gene from *Taf1* they are the same family and could plausibly operate in very similar mechanisms.

Taf1 has been shown to be exclusively nucleoplasmic in localisation, but the exact localisation within the nucleolus was dependent of spermatogenesis stage (Metcalf and Wassarman, 2007). *Taf1* functions in cell-cell adhesion in neural tube formation; it is possible that the role of *Taf1* in maintaining cell-cell interactions through modifying the actin cytoskeleton and regulating actin localisation (Rashid *et al.*, 2006) may extend to a role in sperm cell formation.

5.3.6.3 *Astx6*

Astx6 was found in a study to be one of the top 20 long-non coding RNAs to be differentially expressed between pachytene spermatocytes and round spermatids (Joshi and Rajender, 2020) and *Astx6* was found in this study to be differentially expressed between X bearing and Y bearing spermatids. Pachytene is the third stage of prophase in meiosis, where the chromosome pairs shorten and thicken, and the chromatids separate. Pachytene holds importance in terms of meiotic sex chromosome inactivation; during pachytene the sex body forms in the nucleus, and histone modifications occur. When this process goes wrong, cells cannot undergo prophase I. *Astx6* occurs to a greater extent in both pachytene spermatocytes and X-bearing spermatids.

5.3.6.4 *Platr21*

Platr21 was shown in this study to be upregulated in the X bearing spermatid population. *Platr21* is located in the nucleus. *Platr21* is pluripotency associated transcript 21: pluripotency is the ability of stem cells in the blastocyst to develop into the germ cells of the early embryo and all other cells of the adult body (Weinberger

et al., 2016). This means that *Platr21* likely plays a key role in cell differentiation. There are no human orthologues for *Platr21*, and research into *Platr21* function is limited. However, other pluripotency factors exist, including *Oct4, Sox2* and *Nanog*. Studies have shown where these have been knocked out, embryonic stem cells have failed to differentiate into different types of cells. The upregulation of *Platr21* in X bearing spermatids could therefore play a role in maintaining proper sperm cell shapes and functions in the X bearing spermatids of knockout models such as the Yq-del mice.

5.3.6.5 *Serpina7*

Serpina7 encodes TBG, the major thyroid hormone transport protein. Thyroid hormones play an important regulatory tole in growth and brain development. *Serpina7* has no inhibitory function, unlike other serpin genes. An excess or deficit of thyroid hormones in adult males has been shown to cause problems with testis function. Too many thyroid hormones are associated with reduced semen volume, reduced motility of sperm and poor morphology of sperm. Too little thyroid hormones are also associated with poor sperm morphology (La Vignera and Vita, 2018). Studies have shown that it would be prudent to include thyroid function tests as part of the diagnosis of infertile human males. The fact that *Serpina7* is differentially expressed in X bearing spermatids and is directly related to thyroid hormone function could provide an explanation as to why it is Y bearing sperm that are more significantly impacted morphologically in *Sly* knockout mice. *Sly* could upon *Serpina7* which would mean that when *Sly* is knocked out in Yq-del mice, *Sly* no longer acts upon *Serpina7* and the thyroid hormone levels become imbalanced.

5.3.6.6 *Syap1*

Syap is involved in a number of processes. These include *Torc2* signalling, response of cells to growth factor stimulus and cell response to peptide hormone stimulus. Both cell size and cell growth rate are impacted upon by *Torc2* signalling (Lucena *et al.*, 2018). Cell growth factor stimuli also impacts upon the growth of cells, and peptide hormone stimuli can lead to multiple activities of cells, including gene expression, growth and movement.

Whilst *Torc2* signalling is less understood than *Torc1* signalling, it has been shown that knockout of *Torc2* -specific components in C. elegans have shown that *Torc2* mediates lipid accumulation and growth. Mutant C. elegans without the gene *Rictor* (*Torc2* -specific component) are small, developmentally delayed, store more fat than wild-types and have reduced fecundity than wild-type C. elegans (Jones *et al.*, 2009). This echoes some of the effects of Yq-del Y-bearing sperm, where the Y-bearing sperm are short, blunt and rounder.

5.3.7 Gene signatures from the SmartSeq2 data cannot easily be used to genotype 10X data

It is likely that there are some genes that, when they are used to cluster the 10X cells upon, can give an X bearing and a Y bearing cluster. However, taking the full sample of DEGs identified in this study and using them to cluster the 10X data on, did not give the bimodal clustering needed to verify this would work. Therefore, this study has shown that applying the genotyping to 10X data is not as simple as just using those genes upregulated in the SmartSeq2 data as a gene signature. The upregulated genes from the SmartSeq2 data need to be filtered down to find the

ones that cause meaningful clusterings in the 10X data. which is likely to be a lengthy process but an important consideration for the future. The SmartSeq2 cells that the DEGs are identified from, and the 10X cells that the signature is applied to also need to be restricted to the same, very narrow differentiation period, to ensure that there are no 'early vs late stage' genes that influence the clustering.

Due to the dynamic nature of gene expression during differentiation through developmental stages, the gene expression profiles of X bearing vs Y bearing spermatids will vary over time. Batch effects were taken into account when running the SCDE program on all the cells, but to try and apply the analysis to 10X data and remove any early vs late-stage effects, the analysis was re-run on just the latest stage spermatids. Whilst this meant that some power would be lost, due to the smaller sample sizes, it provides a useful comparison to the global gene expression analysis data. For this analysis, any gene expressed in greater than ten cells was included.

Table 26: number of X-bearing and Y-bearing cells falling into each batch, described by developmental stage

	Early spermatids		**Mid spermatids**		**Late spermatids**	
Batch	**1**	**2**	**3**	**4**	**5**	**6**
X bearing	29	16	17	27	32	11
	46		44		43	
Y bearing	9	12	27	25	20	25
	21		52		45	

The absolute gene counts of the autosomal genes upregulated in late-stage X-bearing and late-stage Y-bearing spermatids were compared, and scatter plots with the average count per late-stage cell in X bearing spermatids against average count

per late-stage cell in Y bearing spermatids was plotted for each autosomal gene of interest (identified as log2fc > 2.0 and p<0.05).

The early spermatid stages are hardest to classify as X and Y bearing spermatids due to the rapidly changing nature of gene expression; at this stage 46 cells were classified as X-bearing spermatids and 21 as Y-bearing spermatids.

Clusters 1 and 2 were the earliest stage spermatids. 377 genes were found to be significantly upregulated (log2fc > 2, p <0.05) in the X-bearing spermatids and 104 genes were found to be upregulated in the Y bearing spermatids from cluster 1. 51 genes were found to be upregulated in the X-bearing spermatids and 23 genes were found to be upregulated in the Y bearing spermatids from cluster 2.

The mid spermatid stage is easier to classify X and Y bearing spermatids; at this stage 44 X-bearing spermatids and 52 Y-bearing spermatids were identified – from clusters 3 and 4.

42 genes were found to be upregulated in the X-bearing spermatids and 104 genes were found to be upregulated in the Y bearing spermatids from cluster 3. 69 genes were found to be upregulated in the X-bearing spermatids and 99 genes were found to be upregulated in the Y bearing spermatids from cluster 4.

The late spermatid stage is the optimum stage to classify X and Y bearing spermatids; at this stage 43 X-bearing spermatids and 45 Y-bearing spermatids were identified, from clusters 5 and 6. Both cluster 5 and cluster 6 were analysed together,

as analysing them separately did not give enough valid genes for cluster 6 to analyse properly.

123 genes were found to be upregulated in the late-stage X-bearing spermatids and 164 genes were found to be upregulated in the late-stage Y bearing spermatids, in clusters 5 and 6 together.

KEGG analysis and gene enrichment analysis was run upon these late stage differentially expressed genes in DAVID to identify any common functional pathways that may be present. No functional pathways were found in KEGG analysis for the genes upregulated in either late-stage X or late-stage Y-bearing spermatids. The enriched pathways identified in DAVID are summarised.

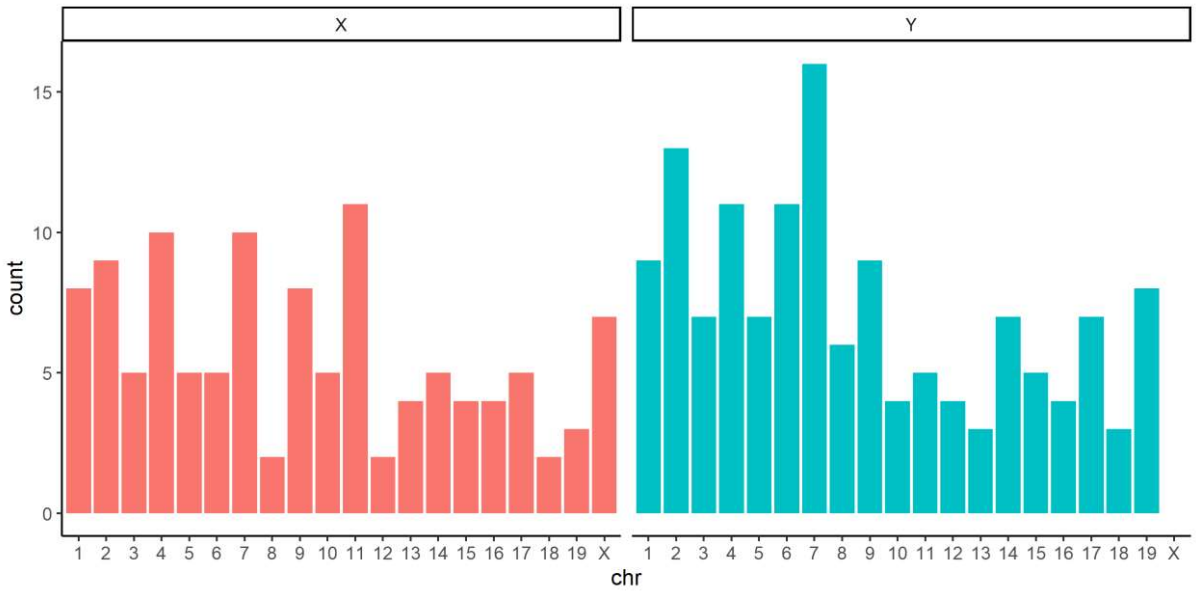
Table 27: functional annotation chart for the autosomal genes upregulated in late-stage X-bearing and late-stage Y-bearing spermatids, identified by the DAVID program

	Category	Term	Gene count	p-value	Benjamini score
X-bearing spermatids	Uniprot sequence feature	REGION:disordered	361	9.0e-12	1.5e-8
	Uniprot keyword post translational modification	Phosphoprotein	243	4.3e-7	4.7e-6
	GO term	Nucleus	227	9.0e-18	4.6e-15
	Uniprot sequence feature	COMPBIAS:Polar residues	196	3.4e-6	1.1e-3
	GO term	Cytoplasm	192	6.8e-4	3.5e-2
	Uniprot keyword cellular component	Nucleus	182	5.0e-14	2.3e-12
	Uniprot sequence feature	COMPBIAS:Basic and acidic residues	160	1.6e-5	4.4e-3
	GO term	Protein binding	158	8.7e-5	1.8e-2
	Uniprot keyword cellular component	Cytoplasm	154	2.6e-5	4.0e-4
	Uniprot keyword domain	Coiled coil	139	1.1e-8	2.6e-7
	Uniprot keyword domain	Repeat	138	2.6e-4	3.2e-3
	GO term	Nucleoplasm	133	2.6e-12	6.8e-10
	Uniprot keyword post translational modification	Acetylation	121	1.1e-6	8.1e-6
	GO term	Cytosol	118	8.5e-4	4.0e-2
Y-bearing spermatids	Go term	Methylated histone binding	2	0.0096	0.0096
	Interpro database	Spindlin/spermiogenesis-specific protein	2	0.011	0.022
	GO term	Gamete generation	2	0.014	0.056

It is likely that there is more transcriptional variability in the earlier stages of spermatid development, and with the smaller sample sizes there may be more false positive upregulated genes identified. The cytoplasmic bridges in developing spermatids break when spermatozoa are mature, so the late-stage spermatids are going to most informative in terms of the genes that may cause physiological differences.

The number of significantly upregulated genes mapping to each chromosome was plotted for both X-bearing and Y-bearing late-stage spermatids; distribution was similar to the distribution across the global sample.

Figure 62



Number of significantly differentially upregulated genes mapped to each chromosome in the late-stage X and Y-bearing spermatids. No significantly enriched genes mapped to the Y chromosome.

SCDE was performed on just these late-stage cells to determine a gene signature to apply back onto the 10X data, but a bimodal separation still was not seen. Further

work is needed to determine which of the genes are a key contributor to X or Y determination in 10X data.

Once a gene signature has been determined in the future, finding orthologues for other species and applying the clustering on single cell data from other species would be an important validation to demonstrate that this pipeline works on more than just mice.

5.4 Future Uses

Having developed a bioinformatics pipeline to be used on mice, and having shown differential gene expression analysis on populations of X vs Y bearing spermatids can be successfully performed, there are many future applications of this research pipeline. Firstly, the pipeline can be applied to any SmartSeq2 single-cell data set on any species. Key species for future investigation would be pigs and cows, for agricultural purposes and humans, for fertility improvements.

Secondly, the lab group has sent off Yq-del mice for RNA sequencing, with the intention of applying this analysis to the Yq-dels. This would allow the identification of genes *Sly* is acting upon. The results of the Yq-del sequencing can be compared with the results of this study – *Sly* works upon cis-limited downstream targets, and this study has identified some candidate cis-limited genes. If expression levels of the candidate cis-limited genes identified in this study are much lower in the Yq-del sequencing data then these candidate cis limited genes will have been verified as downstream targets of *Sly,* which is knocked out in Yq-del mice.

Unfortunately, due to time constraints, this is outside the scope of this project, since the data will not be produced in the desired timeframe for the purpose of this study.

However, this analysis has provided the key reference data for the Yq-del sequencing study when the time comes. When the Yq-del data is available, all the predictions made about candidate cis-limited genes identified in this study are entirely testable.

6 General Discussion

The overall aim of this project was to identify sex-linked and autosomal genes that are implicated in mouse sperm development. This project managed to identify a list of six sex-linked genes that are potentially cis-limited, which means they remain tethered in the developing spermatid, evading being shared between cytoplasmic brides between the cell and other developing spermatids. Whether or not these genes are actually cis-limited cannot be examined purely bioinformatically, which was beyond the scope of this project, but whether or not these genes are cis-limited and implicated in a broader pathway of sperm cell development is entirely testable and will hopefully be done in the future of this laboratory.

6.1 Differences between hybrids and knockouts

It was found that whilst two types of abnormal mouse sperm had the same morphological deformities, these deformities did not share a mechanism. Sperm from mouse hybridisations and sperm from mouse with a Y-chromosome partial knockout were investigated. Both 2/3 Yq-del sperm and the *musXdom* hybrids had sperm with a squashed shape and blunted hook. It is important to note that the *musXdom* hybrid mice are sterile, whereas 2/3 Yq-del mice showed a sex ratio skew but no major loss in fertility. In addition to this, the hybrid mice showed no loss of motility compared to the controls. The same morphological outcomes do not seem to share a mechanism. There are two different functional pathways that are leading to the same shape malformations.

A further consideration is that many ampliconic genes exist, not just *Sly* and *Slx,* which will be imbalanced in hybrid crosses. This study primarily considered *Sly* and

Slx but there are many more. The knockout of the Yq means not just knockout of *Sly* but other ampliconic genes. Yq-del mice have also total knockout of *Sly* and so have a much greater difference in *Sly* ampliconic copy number from X and Y chromosomal incompatibility than in hybrid mice which have a much more subtle mismatch in *Sly* copy number. Whilst *mus* strains have more *Sly* than *dom* strains, *dom* strains have more *Ssty1* than *mus*. In the process of Y introgression of a *dom* Y chromosome onto a *mus* background, there are opposing conflicts occurring: the Y chromosome copy number of *Sly* is slightly higher than the X and autosomal backgrounds, and the copy number of *Ssty1* slightly lower. Table 28 shows the ampliconic copy numbers of *Slx, Slxl1, Sstx, Sly, Ssty1* and *Ssty2* in *mus* and *dom* subspecies, highlighting that the hybrid sterility is a much more complex issue than just *Sly* knockout in Yq-del mice.

Table 28: ampliconic copy numbers of *Slx, Slxl1, Sstx, Sly, Ssty1* and *Ssty2* in *mus* and *dom* subspecies

Sub-species	Strain	*Slx*	*Slxl1*	*Sstx*	*Sly*	*Ssty1*	*Ssty2*
mus	PWK	50	38	33	213	93	135
	CZECHII	62	38	50	217	72	145
dom	LEWES	15	20	32	134	160	68
	WSB	13	25	32	127	167	71

The X-autosomal mismatch seen in the hybrid mice leads to reduced sperm count whilst the X-Y mismatch seen in the Yq-del mice leads to reduced swimming capability. The implication of X-autosomal mismatch in sterility was one of the main factors driving investigation into the autosomal differentially expressed genes rather

than just the sex-linked genes. It is likely that the differentially expressed autosomal genes are involved in a process leading to apoptosis of the sperm cells, reducing the sperm count, rather than any of the sex-linked cis-limited genes that likely lead to problems with sperm movement. From the candidate cis-limited gene list, some of the genes are sex-linked differentially expressed, whilst others are sex-linked transcription factors acting upon differentially expressed autosomal genes.

However, the fact that a common morphological defect occurs via (at least) two different mechanisms does pose an important consideration. In future, it would be great to use the automated Nuclear Morphology Analysis software in order to identify groups of malformed sperm cells. Ideally, specific morphological defects would be caused by specific genetic defects, so splitting out sperm cells that had a certain shape deformity would be identified as having a specified genetic defect. This study has shown this may not necessarily be plausible if multiple pathways lead to the same shape malformation, and so the software may be more for initial diagnosis of 'healthy' or 'unhealthy' sperm.

6.2 Scope for improvements and limitations of the pipeline

The pipeline developed provides a robust tool for genotyping cells: it yielded the same results that came from an already established tool and the results were conclusive: the two different clusters of logarithmic number of reads to the X-chromosome were distinctive enough to say that there was a significant and exponential difference in read counts to the X chromosome in the two different groups, that could not have been seen by chance. This must be a real effect due to the cells being X bearing or Y bearing as the magnitude of difference was so large.

Read counts were normalised for chromosome length, so this would not have confounded results. The pathway works well to yield the desired results, in data from cells sequenced using SmartSeq2 methodology. However, the first negative of the pathway is that it did not work using cells sequenced using 10X methodology. This greatly reduces the number of readily available datasets to sequence.

There are also two parts of the pipeline analysis whereby the analysis is of subjective nature. Firstly, the classification of spermatids, and secondly in downstream analysis where the pipeline relies upon adaptive thresholding. These two parts of the pipeline are subjective so could yield slightly different results dependent on who is running the pipeline. In this study, spermatid classification was able to be compared with the already published studies. Including some non-spermatid cells in the analysis also may not be the biggest problem, if the sample size is large enough. Where it could become a problem, however, is looking at genes with very low expression levels in either X bearing or Y bearing cells, whereby one or two misclassified cells could totally mask the differential expression. The adaptive thresholding should also not be a problem, since the magnitude of the difference in clusters should be large enough that the threshold is clear. However, in smaller datasets it may not be as clear. In earlier and later stages of spermatid development there also may not be a clear 50/50 split of X and Y bearing cells as expression levels vary to a higher degree at the very earliest and very latest stages, so users of the pipeline would have to be aware of this and avoid letting it affect their judgement on where to draw the threshold.

6.3 Candidate cis-limited genes and their functions

Gata1 is a hematopoietic transcription factor that regulates β1-tubulin which has implications for cell shape (Italiano *et al.*, 2003). *Taf1* functions in cell-cell adhesion in neural tube formation, affecting the actin cytoskeleton and regulating actin localisation (Rashid *et al.*, 2006). *Astx6* is differentially expressed between pachytene spermatocytes and round spermatids, and has a key role in development. *Platr21* is associated with pluripotency, so helps cells to differentiate and is therefore likely involved in cell growth and shape formation. *Serpina7* is directly related to thyroid hormone function, which is key for cell development. *Syap1* is involved in a number of processes, including *Torc2* signalling, response of cells to growth factor stimulus and cell response to peptide hormone stimulus. All of the candidate cis-limited genes identified in this study have the potential to be critical for cell differentiation and shape development, which is what is implicated in the knockout sperm cells or hybrid sperm cells, mediated by cis-limited genes. All six candidate cis-limited genes could plausibly be causing the phenotypes seen.

6.4 Applications of the candidate cis-limited gene list

The next step would be to experimentally identify if these genes are actually cis-limited. This will be achieved by running the pipeline on the Yq-del knockout mice. Knowing the six candidate cis-limited genes to look out for, the expression levels of these can be tested in the knockouts. *Sly* is a master regulator that would normally act to control the levels of these cis-limited genes. In the absence of *Sly*, as in the Yq-del mice, the cis-limited gene expression profiles should not match those of the wild-type sperm cells investigated in this thesis.

A candidate cis-limited gene list would give genetic targets for interventions when there is poor sperm formation. It could also help select for healthy sperm.

6.5 Applying this to 10X data

It would be of great benefit if the genotyping could be applied to 10X data. 10X data has some benefits over SmartSeq data and as such many published datasets use 10X sequencing methods. These benefits include, but are not limited to, reduced speed, reduced cost, and greater magnitude of data generation (Wang *et al.*, 2021). However, SmartSeq data detects more genes in cells, including low abundance transcripts, and has far less noise. The fact that 10X data can cover many more cells makes it a popular choice for many studies. As such, there are huge amounts of data available for potential analysis, provided there was a reliable method of genotyping cells from 10X data as well as SmartSeq.

The more datasets that can be analysed from various species, the greater the potential implications for medicinal and agricultural advances. A big hypothesis of this thesis is that the analysis pipeline can be applied to many different datasets of many different species in order to find some key genes for each species of interest, including, but not limited to: pigs, cows, humans. At this stage, there is no way of knowing what the results would be from other species such as cattle or pig, however, it is hypothesised that cis-limited genes affecting sperm formation may be able to be found. Having a robust pipeline that works well for 10X data, rather than just SmartSeq2 data means that, for very minimal cost, thousands of analyses could be run on datasets containing thousands of cells. For experimental advances such as the identification of a small number of key genes, data abundance is critical. The

more data, the more likely the genes identified are actually of interest. This could help hone in on the smallest, most accurate candidate lists, that could then be investigated experimentally in a wet lab. Running a bioinformatic pipeline would be far cheaper than running the wet lab experiments, and so having a better idea of a more refined gene list before performing the wet lab work would be of great benefit.

7 Conclusion and Evaluation

This thesis had three specific aims. (1) to assess the morphological differences of sperm and the sperm quality parameters between different hybrid mouse crosses. (2) to develop a pipeline to enable us to genotype sperm cells as X or Y bearing. (3) to perform differential expression analysis on RNA sequencing data from spermatids that are X bearing and Y bearing, to search for candidate cis-limited genes.

The morphological differences of the sperm nuclei of intraspecific control hybrids, intraspecific F1 hybrids and hybrids with Y-introgression were assessed, quantified and compared. The type of cross as well as the specific genotype were taken into account. Morphological parameters of the sperm nuclei were correlated with sperm quality parameters to show that sperm nuclei from crosses with more extreme morphological defects had lower relative testis weights and lower seminal vesicle weights. Sperm count and percentage motile sperm did not vary with morphological parameters.

A pipeline was developed and used to successfully genotype sperm cells as X or Y bearing. However, it could only be applied to SmartSeq RNA sequencing data, and not 10X data. There is no bioinformatic way of determining if the assignments of X bearing and Y bearing cells is correct: this would need to be done experimentally. Nevertheless, a number of sanity checks were performed along the way to give a high level of certainty that assignments were correct.

A shortlist of potential cis-limited genes was produced. With RNA-sequencing data incoming of the Yq-del mice, these candidate cis-limited genes will be able to be validated as being downstream targets of *Sly,* or potentially not. This means the candidate cis-limited gene list will soon be able to be a refined list of demonstrated

cis-limited genes. This will give insight into which genes are implicated in sperm development and will give potential targets to focus on when overcoming malformations and defects.

The results from this thesis will therefore aid with the improvement of sperm development in mice and will be able to be applied to other organisms to improve agriculture and human fertility.

8 References

Acharya, K. K. *et al.* (2006) 'cis-Requirement for the maintenance of round spermatid-specific transcription', *Developmental Biology*. Academic Press Inc., 295(2), pp. 781–790. doi: 10.1016/j.ydbio.2006.04.443.

Agarwal, A. (2003) 'Role of sperm chromatin abnormalities and DNA damage in male infertility', *Human Reproduction Update*. Oxford Academic, 9(4), pp. 331–345. doi: 10.1093/humupd/dmg027.

Alladin, N. *et al.* (2013) 'The three-dimensional image analysis of the chromocenter in motile and immotile human sperm', *Systems Biology in Reproductive Medicine*. Taylor & Francis, 59(3), pp. 146–152. doi: 10.3109/19396368.2013.772679.

Allsopp, R. C. *et al.* (1992) 'Telomere length predicts replicative capacity of human fibroblasts', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 89(21), pp. 10114–10118. doi: 10.1073/pnas.89.21.10114.

An, J. Y. *et al.* (2010) 'UBR2 mediates transcriptional silencing during spermatogenesis via histone ubiquitination', *Proceedings of the National Academy of Sciences*, 107(5), pp. 1912–1917. doi: 10.1073/pnas.0910267107.

Anderson, E. (1949) 'Introgressive hybridization.', *Introgressive hybridization.* John Wiley and Sons, Inc., New York, Chapman and Hall, Ltd., London.

Andolfi, L. *et al.* (2015) 'The application of scanning near field optical imaging to the study of human sperm morphology', *Journal of Nanobiotechnology*, 13(1), p. 2. doi: 10.1186/s12951-014-0061-5.

Aryee, S. N. D. *et al.* (2021) 'Sustainable genomic research for food security in sub-

Saharan Africa', *Agriculture and Food Security*. BioMed Central, 10(1), pp. 1–12. doi: 10.1186/s40066-021-00287-9.

Aston, K. I. *et al.* (2015) 'Aberrant sperm DNA methylation predicts male fertility status and embryo quality', *Fertility and Sterility*. Elsevier Inc., 104(6), pp. 1388-1397.e5. doi: 10.1016/j.fertnstert.2015.08.019.

Bachtrog, D. (2014) 'Signs of Genomic Battles in Mouse Sex Chromosomes', *Cell*. Cell Press, 159(4), pp. 716–718. doi: 10.1016/J.CELL.2014.10.036.

Bailey, T. L. (2021) 'STREME: accurate and versatile sequence motif discovery', *Bioinformatics*. Oxford Academic, 37(18), pp. 2834–2840. doi: 10.1093/BIOINFORMATICS/BTAB203.

Bailey, T. L. and Elkan, C. (1994) 'Fitting a mixture model by expectation maximization to discover motifs in biopolymers'.

Bailey, T. L. and Grant, C. E. (2021) 'SEA: Simple Enrichment Analysis of motifs', *bioRxiv*. doi: 10.1101/2021.08.23.457422.

Baird, D. M. *et al.* (2006) 'Telomere instability in the male germline', *Human Molecular Genetics*. Oxford Academic, 15(1), pp. 45–51. doi: 10.1093/hmg/ddi424.

Battulin, N. *et al.* (2015) 'Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach', *Genome Biology*. BioMed Central Ltd., 16(1). doi: 10.1186/s13059-015-0642-0.

Baur, J. A. *et al.* (2001) 'Telomere position effect in human cells', *Science*. American Association for the Advancement of Science, 292(5524), pp. 2075–2077. doi: 10.1126/science.1062329.

Bhattacharyya, T. *et al.* (2013) 'Mechanistic basis of infertility of mouse

intersubspecific hybrids', *Proceedings of the National Academy of Sciences of the United States of America*, 110(6). doi: 10.1073/PNAS.1219126110.

Bhutani, K. *et al.* (2021) 'Widespread haploid-biased gene expression enables sperm-level natural selection', *Science*. American Association for the Advancement of Science, 371(6533). doi: 10.1126/science.abb1723.

Birkhead, T. R. (1998) 'CRYPTIC FEMALE CHOICE: CRITERIA FOR ESTABLISHING FEMALE SPERM CHOICE', *Evolution*, 52(4), pp. 1212–1218. doi: 10.1111/j.1558-5646.1998.tb01848.x.

Biron-Shental, T. *et al.* (2018) 'Sub-fertile sperm cells exemplify telomere dysfunction', *Journal of Assisted Reproduction and Genetics*. Springer New York LLC, 35(1), pp. 143–148. doi: 10.1007/s10815-017-1029-9.

Bisht, S. and Dada, R. (2017) 'Oxidative stress: Major executioner in disease pathology, role in sperm DNA damage and preventive strategies', *Frontiers in Bioscience*, 9, pp. 420–447.

Blasco, M. A. (2007) 'The epigenetic regulation of mammalian telomeres', *Nature Reviews Genetics*. Nature Publishing Group, pp. 299–309. doi: 10.1038/nrg2047.

Bou, G. *et al.* (2014) 'A pre-breeding screening program for transgenic boars based on fluorescence in situ hybridization assay', *Transgenic Research*. Kluwer Academic Publishers, 23(4), pp. 679–689. doi: 10.1007/s11248-014-9801-4.

Bracke, A. *et al.* (2018) 'A search for molecular mechanisms underlying male idiopathic infertility', *Reproductive BioMedicine Online*. Elsevier Ltd, pp. 327–339. doi: 10.1016/j.rbmo.2017.12.005.

Braun, R. E. (1998) *Post-transcriptional control of gene expression during*

spermatogenesis When transcription stops and translation starts, *CELL & DEV ELOPMENTAL BI OLOGY*.

Canny, J. (1986) 'A Computational Approach to Edge Detection', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6), pp. 679–698. doi: 10.1109/TPAMI.1986.4767851.

Champroux, A. *et al.* (2018) 'Nuclear Integrity but Not Topology of Mouse Sperm Chromosome is Affected by Oxidative DNA Damage', *Genes*. MDPI AG, 9(10), p. 501. doi: 10.3390/genes9100501.

Chen, H. *et al.* (2006) 'Epididymal SPAM1 Is a Marker for Sperm Maturation in the Mouse1', *Biology of Reproduction*. Oxford University Press (OUP), 74(5), pp. 923–930. doi: 10.1095/biolreprod.105.048587.

Chen, Y. *et al.* (2018) 'Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis', *Cell Research*. Nature Publishing Group, 28(9), pp. 879–896. doi: 10.1038/s41422-018-0074-y.

Cho, C. L. and Agarwal, A. (2018) 'Role of sperm DNA fragmentation in male factor infertility: A systematic review', *Arab Journal of Urology*. Arab Association of Urology, 16(1), pp. 21–34. doi: 10.1016/j.aju.2017.11.002.

Cocquet, J. *et al.* (2012) 'A Genetic Basis for a Postmeiotic X Versus Y Chromosome Intragenomic Conflict in the Mouse', *PLOS Genetics*. Public Library of Science, 8(9), p. e1002900. doi: 10.1371/JOURNAL.PGEN.1002900.

Von Collenberg, C. R. *et al.* (2019) 'An essential role of the mouse synapse-associated protein Syap1 in circuits for spontaneous motor activity and rotarod balance', *Biology Open*. Company of Biologists Ltd, 8(6). doi:

10.1242/BIO.042366/VIDEO-2.

Cooke, H. J. and Smith, B. A. (1986) 'Variability at the telomeres of the human X/Y pseudoautosomal region', *Cold Spring Harbor Symposia on Quantitative Biology*. Cold Spring Harbor Laboratory Press, 51(1), pp. 213–219. doi: 10.1101/sqb.1986.051.01.026.

Cremer, T. and Cremer, C. (2001) 'Chromosome territories, nuclear architecture and gene regulation in mammalian cells', *Nature Reviews Genetics*. Nat Rev Genet, pp. 292–301. doi: 10.1038/35066075.

Dada, R. *et al.* (2012) 'Epigenetics and its role in male infertility', *Journal of Assisted Reproduction and Genetics*. Springer, pp. 213–223. doi: 10.1007/s10815-012-9715-0.

Dennis, G. *et al.* (2003) 'DAVID: Database for Annotation, Visualization, and Integrated Discovery.', *Genome biology*. BioMed Central, 4(5), pp. 1–11. doi: 10.1186/GB-2003-4-9-R60/TABLES/3.

Dobin, A. *et al.* (2013) 'STAR: Ultrafast universal RNA-seq aligner', *Bioinformatics*. Oxford University Press, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Dobzhansky, T. (2015) 'Genetic Nature of Species Differences', *https://doi.org/10.1086/280726*. Science Press , 71(735), pp. 404–420. doi: 10.1086/280726.

Eddy, E. M. (1998) 'Regulation of gene expression during spermatogenesis', *Seminars in Cell and Developmental Biology*. Elsevier Ltd, 9(4), pp. 451–457. doi: 10.1006/scdb.1998.0201.

Eggert-Kruse, W. *et al.* (1996) 'The Acridine Orange test: a clinically relevant

screening method for sperm quality during infertility investigation?', *Human Reproduction*, 11(4), pp. 784–789. doi: 10.1093/oxfordjournals.humrep.a019255.

Ellis, P. J. I. *et al.* (2005) 'Deletions on mouse Yq lead to upregulation of multiple X- and Y-linked transcripts in spermatids', *Human Molecular Genetics*. Oxford Academic, 14(18), pp. 2705–2715. doi: 10.1093/HMG/DDI304.

Ellis, P. J. I., Bacon, J. and Affara, N. A. (2011) 'Association of Sly with sex-linked gene amplification during mouse evolution: a side effect of genomic conflict in spermatids?', *Human Molecular Genetics*, 20(15), pp. 3010–3021. doi: 10.1093/hmg/ddr204.

Ellis, P. J.I., Bacon, J. and Affara, N. A. (2011) 'Association of Sly with sex-linked gene amplification during mouse evolution: a side effect of genomic conflict in spermatids?', *Human Molecular Genetics*. Oxford Academic, 20(15), pp. 3010–3021. doi: 10.1093/HMG/DDR204.

Ellis, P. J. I. and Griffin, D. K. (2020) 'Form from Function, Order from Chaos in Male Germline Chromatin', *Genes*. MDPI AG, 11(2), p. 210. doi: 10.3390/genes11020210.

Ellis, P. J. I. *et al.* (2011) 'REPRODUCTION REVIEW Transcriptional dynamics of the sex chromosomes and the search for offspring sex-specific antigens in sperm', *Reproduction*, 142, pp. 609–619. doi: 10.1530/REP-11-0228.

Elster, J. D. N. *et al.* (2000) 'Less is more: The risks of multiple births', *Fertility and Sterility*, pp. 617–623. doi: 10.1016/S0015-0282(00)00713-5.

Emrich, S. J. *et al.* (2007) 'Gene discovery and annotation using LCM-454 transcriptome sequencing', *Genome Research*. Cold Spring Harbor Laboratory Press, 17(1), pp. 69–73. doi: 10.1101/GR.5145806.

Evenson, D. P. *et al.* (1999) 'Utility of the sperm chromatin structure assay as a diagnostic and prognostic tool in the human fertility clinic', *Human Reproduction*. Oxford University Press, 14(4), pp. 1039–1049. doi: 10.1093/humrep/14.4.1039.

Evenson, D. P., Larson, K. L. and Jost, L. K. (2002) 'Sperm chromatin structure assay: Its clinical use for detecting sperm DNA fragmentation in male infertility and comparisons with other techniques', *Journal of Andrology*. American Society of Andrology Inc., 23(1), pp. 25–43. doi: 10.1002/j.1939-4640.2002.tb02599.x.

Fernández, J. L. *et al.* (2003) 'The sperm chromatin dispersion test: A simple method for the determination of sperm DNA fragmentation', *Journal of Andrology*. John Wiley & Sons, Ltd, 24(1), pp. 59–66. doi: 10.1002/j.1939-4640.2003.tb02641.x.

Fice, H. E. and Robaire, B. (2019) 'Telomere dynamics throughout spermatogenesis', *Genes*. MDPI AG, 10(7). doi: 10.3390/genes10070525.

Finch, K.A. *et al.* (2008) 'Nuclear organization in human sperm: preliminary evidence for altered sex chromosome centromere position in infertile males', *Human Reproduction*, 23(6), pp. 1263–1270. doi: 10.1093/humrep/den112.

Finch, K A *et al.* (2008) 'Nuclear organization in human sperm: preliminary evidence for altered sex chromosome centromere position in infertile males', *Human Reproduction*, 23(6), pp. 1263–1270. doi: 10.1093/humrep/den112.

Foster, H. A. *et al.* (2005) 'Non-random chromosome positioning in mammalian sperm nuclei, with migration of the sex chromosomes during late spermatogenesis', *Journal of Cell Science*, 118(9), pp. 1811–1820. doi: 10.1242/jcs.02301.

Franco, J. G. *et al.* (2008) 'Significance of large nuclear vacuoles in human spermatozoa: Implications for ICSI', *Reproductive BioMedicine Online*. Reproductive

Healthcare Ltd, 17(1), pp. 42–45. doi: 10.1016/S1472-6483(10)60291-X.

García-Vázquez, F. *et al.* (2016) 'Importance of sperm morphology during sperm transport and fertilization in mammals', *Asian Journal of Andrology*. Medknow Publications, pp. 844–850. doi: 10.4103/1008-682X.186880.

García-Vázquez, F. A. *et al.* (2015) 'Morphological study of boar sperm during their passage through the female genital tract', *Journal of Reproduction and Development*. The Japanese Society of Animal Reproduction (JSAR), 61(5), pp. 407–413. doi: 10.1262/jrd.2014-170.

Geraldes, A. *et al.* (2011) 'Higher differentiation among subspecies of the house mouse (Mus musculus) in genomic regions with low recombination', *Molecular Ecology*. John Wiley & Sons, Ltd, 20(22), pp. 4722–4736. doi: 10.1111/J.1365-294X.2011.05285.X.

Gerris, J. (2005) 'Single embryo transfer and IVF/ICSI outcome: a balanced appraisal', *Human Reproduction Update*, 11(2), pp. 105–121.

Gillies, E. A. *et al.* (2009) 'Hydrodynamic propulsion of human sperm', *Journal of Fluid Mechanics*. Cambridge University Press, 625, pp. 445–474. doi: 10.1017/S0022112008005685.

Godden, A. M. and Immler, S. (2022) 'The potential role of the mobile and non-coding genomes in adaptive response', *Trends in Genetics*. Elsevier Current Trends. doi: 10.1016/J.TIG.2022.08.006.

Good, J. M. *et al.* (2010) 'Widespread Over-Expression of the X Chromosome in Sterile F 1 Hybrid Mice', *PLoS Genet*, 6(9), p. 1001148. doi: 10.1371/journal.pgen.1001148.

Good, J. M., Handel, M. A. and Nachman, M. W. (2008) 'Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice', *Evolution; international journal of organic evolution*. Evolution, 62(1), pp. 50–65. doi: 10.1111/J.1558-5646.2007.00257.X.

Griffin, D. K. and Finch, K. A. (2005) 'The genetic and cytogenetic basis of male infertility', *Human Fertility*. Taylor & Francis, 8(1), pp. 19–26. doi: 10.1080/14647270400016407.

Gu, N. H. *et al.* (2019) 'Comparative analysis of mammalian sperm ultrastructure reveals relationships between sperm morphology, mitochondrial functions and motility', *Reproductive Biology and Endocrinology*. BioMed Central Ltd., 17(1), p. 66. doi: 10.1186/s12958-019-0510-y.

Gupta, S. *et al.* (2007) 'Quantifying similarity between motifs', *Genome Biology*. BioMed Central, 8(2), pp. 1–9. doi: 10.1186/GB-2007-8-2-R24/TABLES/3.

Gupta, S. and Ghosh Mazumdar, S. (2013) 'Sobel Edge Detection Algorithm', *International Journal of Computer Science and Management Research*, 2(2).

Gura, M. A. *et al.* (2020) 'Dynamic and regulated TAF gene expression during mouse embryonic germ cell development', *PLoS Genetics*. PLOS, 16(1). doi: 10.1371/JOURNAL.PGEN.1008515.

Haralick, R. M., Dinstein, I. and Shanmugam, K. (1973) 'Textural Features for Image Classification', *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(6), pp. 610–621. doi: 10.1109/TSMC.1973.4309314.

Harrison, R. G. and Larson, E. L. (2014) 'Hybridization, Introgression, and the Nature of Species Boundaries', *Journal of Heredity*. Oxford Academic, 105(S1), pp. 795–

809. doi: 10.1093/JHERED/ESU033.

Hazzouri, M. *et al.* (2000) 'Genome organization in the human sperm nucleus studied by FISH and confocal microscopy', *Molecular Reproduction and Development*. Wiley-Liss Inc., 55(3), pp. 307–315. doi: 10.1002/(SICI)1098-2795(200003)55:3<307::AID-MRD9>3.0.CO;2-P.

Hermann, B. P. *et al.* (2018) 'The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids In Brief', *Cell Reports*, 25, pp. 1650-1667.e8. doi: 10.1016/j.celrep.2018.10.026.

Homolka, D. *et al.* (2007) 'Chromosomal rearrangement interferes with meiotic X chromosome inactivation', *Genome Research*. Cold Spring Harbor Laboratory Press, 17(10), pp. 1431–1437. doi: 10.1101/gr.6520107.

Ichikawa, Y. *et al.* (2015) 'Nucleosome organization and chromatin dynamics in telomeres', *Biomolecular Concepts*. De Gruyter Mouton, pp. 67–75. doi: 10.1515/bmc-2014-0035.

Inagaki, A., Schoenmakers, S. and Baarends, W. M. (2010) 'DNA double strand break repair, chromosome synapsis and transcriptional silencing in meiosis'. doi: 10.4161/epi.5.4.11518.

Ioannou, D. and Griffin, D. K. (2010) 'Nanotechnology and molecular cytogenetics: the future has not yet arrived', *Nano Reviews*. Informa UK Limited, 1(1), p. 5117. doi: 10.3402/nano.v1i0.5117.

Ioannou, D. and Griffin, D. K. (2011) 'Male fertility, chromosome abnormalities, and nuclear organization', *Cytogenetic and Genome Research*, 133(2–4), pp. 269–279. doi: 10.1159/000322060.

Ioannou, Dimitris *et al.* (2011) 'Nuclear organisation of sperm remains remarkably unaffected in the presence of defective spermatogenesis', *Chromosome Res*, 19, pp. 741–753. doi: 10.1007/s10577-011-9238-z.

Italiano, J. E. *et al.* (2003) 'Mechanisms and implications of platelet discoid shape', *Blood*. American Society of Hematology, 101(12), pp. 4789–4796. doi: 10.1182/BLOOD-2002-11-3491.

Jameel, T. (2008) *Sperm swim-up: a simple and effective technique of semen processing for intrauterine insemination*.

Jan, S. Z. *et al.* (2017) 'Unraveling transcriptome dynamics in human spermatogenesis', *Development (Cambridge, England)*. Company of Biologists, 144(20), p. 3659. doi: 10.1242/DEV.152413.

JanouŠek, V. *et al.* (2012) 'Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between Mus musculus musculus and M. m. domesticus', *Molecular Ecology*. John Wiley & Sons, Ltd, 21(12), pp. 3032–3047. doi: 10.1111/J.1365-294X.2012.05583.X.

Jennings, C. and Powell, D. (1995) 'Genome organisation in the murine sperm nucleus', *Zygote*. Cambridge University Press, 3(2), pp. 123–131. doi: 10.1017/S0967199400002495.

Jones, K. T. *et al.* (2009) 'Rictor/TORC2 Regulates Caenorhabditis elegans Fat Storage, Body Size, and Development through sgk-1', *PLOS Biology*. Public Library of Science, 7(3), p. e1000060. doi: 10.1371/JOURNAL.PBIO.1000060.

Joshi, M. and Rajender, S. (2020) 'Long non-coding RNAs (lncRNAs) in spermatogenesis and male infertility', *Reproductive Biology and Endocrinology 2020*

18:1. BioMed Central, 18(1), pp. 1–18. doi: 10.1186/S12958-020-00660-6.

Kanehisa, M. and Goto, S. (2000) 'KEGG: Kyoto Encyclopedia of Genes and Genomes', *Nucleic Acids Research*. Oxford University Press, 28(1), p. 27. doi: 10.1093/NAR/28.1.27.

Kharchenko, P. V., Silberstein, L. and Scadden, D. T. (2014) 'Bayesian approach to single-cell differential expression analysis', *Nature methods*. Nat Methods, 11(7), pp. 740–742. doi: 10.1038/NMETH.2967.

Kimura, M. *et al.* (2008) 'Offspring's Leukocyte Telomere Length, Paternal Age, and Telomere Elongation in Sperm', *PLoS Genetics*. Edited by T. Gojobori. Public Library of Science, 4(2), p. e37. doi: 10.1371/journal.pgen.0040037.

Kitamura, A. *et al.* (2015) 'Epigenetic alterations in sperm associated with male infertility', *Congenital Anomalies*. Blackwell Publishing, 55(3), pp. 133–144. doi: 10.1111/cga.12113.

Kopania, E. E.K. *et al.* (2022) 'Molecular Evolution across Mouse Spermatogenesis', *Molecular Biology and Evolution*. Oxford Academic, 39(2). doi: 10.1093/MOLBEV/MSAC023.

Kopania, E.E K *et al.* (2022) 'The contribution of sex chromosome conflict to disrupted spermatogenesis in hybrid house mice', *Genetics*. Oxford Academic, 222(4). doi: 10.1093/GENETICS/IYAC151.

Krueger, F. (2015) 'A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files', 516, p. 517.

Kruger, A. N. *et al.* (2019) 'A Neofunctionalized X-Linked Ampliconic Gene Family Is Essential for Male Fertility and Equal Sex Ratio in Mice', *Current Biology*. Cell Press,

29(21), pp. 3699-3706.e5. doi: 10.1016/J.CUB.2019.08.057.

Kumar, N. and Singh, A. (2015) 'Trends of male factor infertility, an important cause of infertility: A review of literature', *Journal of Human Reproductive Sciences*. Medknow Publications, pp. 191–196. doi: 10.4103/0974-1208.170370.

Kumar, S. *et al.* (2005) 'Atomic force microscopy: A powerful tool for high-resolution imaging of spermatozoa', *Journal of Nanobiotechnology*. doi: 10.1186/1477-3155-3-9.

Kuroda, M. *et al.* (2004) 'Alteration of chromosome positioning during adipocyte differentiation', *Journal of Cell Science*. The Company of Biologists Ltd, 117(24), pp. 5897–5903. doi: 10.1242/jcs.01508.

de Lange, T. *et al.* (1990) 'Structure and variability of human chromosome ends.', *Molecular and Cellular Biology*. American Society for Microbiology, 10(2), pp. 518–527. doi: 10.1128/mcb.10.2.518.

Larson, E. L. *et al.* (2017) 'The Composite Regulatory Basis of the Large X-Effect in Mouse Speciation', *Molecular Biology and Evolution*. Oxford Academic, 34(2), pp. 282–295. doi: 10.1093/MOLBEV/MSW243.

Larson, E. L. *et al.* (2022) 'Stage-specific disruption of X chromosome expression during spermatogenesis in sterile house mouse hybrids', *G3 Genes|Genomes|Genetics*, 12(2). doi: 10.1093/g3journal/jkab407.

Lee, C. Y. *et al.* (2020) 'Telomere-led meiotic chromosome movements: recent update in structure and function', *Nucleus*. Taylor and Francis Inc., 11(1), pp. 111–116. doi: 10.1080/19491034.2020.1769456.

Liao, Y., Smyth, G. K. and Shi, W. (2014) 'FeatureCounts: An efficient general

purpose program for assigning sequence reads to genomic features', *Bioinformatics*. Oxford University Press, 30(7), pp. 923–930. doi: 10.1093/bioinformatics/btt656.

Lucena, R. *et al.* (2018) 'Cell size and growth rate are modulated by TORC2-dependent signals', *Current biology : CB*. NIH Public Access, 28(2), p. 196. doi: 10.1016/J.CUB.2017.11.069.

Luetjens, C. M., Payne, C. and Schatten, G. (1999) 'Non-random chromosome positioning in human sperm and sex chromosome anomalies following intracytoplasmic sperm injection', *Lancet*. Elsevier Limited, 353(9160), p. 1240. doi: 10.1016/S0140-6736(99)80059-2.

Van Der Maaten, L. and Hinton, G. (2008) *Visualizing Data using t-SNE*, *Journal of Machine Learning Research*.

Mahadevan, M. and Baker, G. (1984) 'Assessment and Preparation of Semen for In Vitro Fertilization', in *Clinical In Vitro Fertilization*. Springer London, pp. 83–97. doi: 10.1007/978-1-4471-3317-9_7.

La Manno, G. *et al.* (2018) 'RNA velocity of single cells', *Nature*, 560, pp. 494–498.

Marco, A. (2018) 'SeedVicious: Analysis of microRNA target and near-target sites', *PLOS ONE*. Public Library of Science, 13(4), p. e0195532. doi: 10.1371/JOURNAL.PONE.0195532.

Mardis, E. R. (2008) 'The impact of next-generation sequencing technology on genetics', *Trends in Genetics*, 24(3), pp. 133–141.

Maroto-Morales, A. *et al.* (2016) 'Current status and potential of morphometric sperm analysis', *Asian Journal of Andrology*. Medknow Publications, pp. 863–870. doi: 10.4103/1008-682X.187581.

Martin-DeLeon, P. A. (2006) ‘Epididymal SPAM1 and its impact on sperm function’, *Molecular and Cellular Endocrinology*. Elsevier, 250(1–2), pp. 114–121. doi: 10.1016/j.mce.2005.12.033.

Martinez, N. J. and Walhout, A. J. M. (2009) ‘The interplay between transcription factors and microRNAs in genome-scale regulatory networks’, *BioEssays : news and reviews in molecular, cellular and developmental biology*. NIH Public Access, 31(4), p. 435. doi: 10.1002/BIES.200800212.

McDermaid, A. *et al.* (2019) ‘Interpretation of differential gene expression results of RNA-seq data: review and integration’, *Briefings in Bioinformatics*. Oxford University Press, 20(6), p. 2044. doi: 10.1093/BIB/BBY067.

McInnes, L., Healy, J. and Melville, J. (2018) ‘UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction’.

Metcalf, C. E. and Wassarman, D. A. (2007) ‘Nucleolar colocalization of TAF1 and testis-specific TAFs during Drosophila spermatogenesis’, *Developmental Dynamics*. John Wiley & Sons, Ltd, 236(10), pp. 2836–2843. doi: 10.1002/DVDY.21294.

Mewborn, S. K. *et al.* (2010) ‘Altered Chromosomal Positioning, Compaction, and Gene Expression with a Lamin A/C Gene Mutation’, *PLoS ONE*. Edited by B. A. Sullivan. Public Library of Science, 5(12), p. e14342. doi: 10.1371/journal.pone.0014342.

Montjean, D. *et al.* (2015) ‘Sperm global DNA methylation level: association with semen parameters and genome integrity’, *Andrology*. Blackwell Publishing Ltd, 3(2), pp. 235–240. doi: 10.1111/andr.12001.

Moruzzi, J. F. *et al.* (1988) ‘Quantification and classification of human sperm

morphology by computer-assisted image analysis', *Fertility and Sterility*, 50(1), pp. 142–152. doi: 10.1016/S0015-0282(16)60022-5.

Moskovtsev, S. I. *et al.* (2009) 'Sperm DNA Damage: Correlation to Severity of Semen Abnormalities', *Urology*. Elsevier, 74(4), pp. 789–793. doi: 10.1016/j.urology.2009.05.043.

Mudrak, O., Tomilin, N. and Zalensky, A. (2005) 'Chromosome architecture in the decondensing human sperm nucleus', *Journal of Cell Science*. The Company of Biologists Ltd, 118(19), pp. 4541–4550. doi: 10.1242/jcs.02581.

MULLER and HJ (1942) 'Isolating mechanisms, evolution, and temperature', *Biol Symp*, 6, pp. 71–125.

Murugesan, S. and Prakash, B. (2016) 'Effect of dietary energy and crude protein on semen parameters and fertility in layer breeders males NAIP Project entitled "Value chain commercialization on maize and maize products in poultry nutrition". View project Feed resources of NEH region View project', *Article in Indian Journal of Poultry Science*. doi: 10.5958/0974-8180.2016.00064.7.

Olszewska, M., Wiland, E. and Kurpisz, M. (2008) 'Positioning of chromosome 15, 18, X and Y centromeres in sperm cells of fertile individuals and infertile patients with increased level of aneuploidy', *Chromosome Research*. Springer, 16(6), pp. 875–890. doi: 10.1007/s10577-008-1246-2.

Ostermeier, G. C. *et al.* (2001) 'Measurement of bovine sperm nuclear shape using Fourier harmonic amplitudes', *Journal of Andrology*, 22(4), pp. 584–594. doi: 10.1002/j.1939-4640.2001.tb02218.x.

Paigen, K. and Petkov, P. M. (2018) 'PRDM9 and Its Role in Genetic

Recombination', *Trends in Genetics*, 34(4). doi: 10.1016/j.tig.2017.12.017.

Palladino, F. *et al.* (1993) 'SIR3 and SIR4 proteins are required for the positioning and integrity of yeast telomeres', *Cell*. Cell Press, 75(3), pp. 543–555. doi: 10.1016/0092-8674(93)90388-7.

Panawala, L. (2017) 'Difference Between Spermatogenesis and Oogenesis', *PEDIAA*.

Perdrix, A. and Rives, N. (2013) 'Motile sperm organelle morphology examination (MSOME) and sperm head vacuoles: state of the art in 2013', *Human Reproduction Update*. Oxford Academic, 19(5), pp. 527–541. doi: 10.1093/humupd/dmt021.

Pereira, C. D. *et al.* (2019) 'Nuclear envelope dynamics during mammalian spermatogenesis: new insights on male fertility', *Biological Reviews*. Blackwell Publishing Ltd, 94(4), pp. 1195–1219. doi: 10.1111/brv.12498.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: A flexible suite of utilities for comparing genomic features', *Bioinformatics*. Oxford Academic, 26(6), pp. 841–842. doi: 10.1093/bioinformatics/btq033.

Rashid, D. *et al.* (2006) 'A requirement for NF-protocadherin and TAF1/Set in cell adhesion and neural tube formation', *Developmental Biology*. Academic Press Inc., 291(1), pp. 170–181. doi: 10.1016/J.YDBIO.2005.12.027.

Rathje, C. C. *et al.* (2019) 'Differential Sperm Motility Mediates the Sex Ratio Drive Shaping Mouse Sex Chromosome Evolution', *SSRN Electronic Journal*. Elsevier BV. doi: 10.2139/ssrn.3396496.

Ray, P. F. *et al.* (2017) 'Genetic abnormalities leading to qualitative defects of sperm morphology or function', *Clinical Genetics*, 91(2), pp. 217–232. doi:

10.1111/cge.12905.

Ribas-Maynou, J. and Benet, J. (2019) 'Single and Double Strand Sperm DNA Damage: Different Reproductive Effects on Male Fertility', *Genes*. MDPI AG, 10(2), p. 105. doi: 10.3390/genes10020105.

Rocca, M. S., Foresta, C. and Ferlin, A. (2018) 'Telomere length: lights and shadows on their role in human reproduction', *Biology of Reproduction*. Oxford University Press, 100(2), pp. 305–317. doi: 10.1093/biolre/ioy208.

Schneider, T. D. and Stephens, R. M. (1990) 'Sequence logos: a new way to display consensus sequences.', *Nucleic Acids Research*. Oxford University Press, 18(20), p. 6097. doi: 10.1093/NAR/18.20.6097.

Schulz, K. N. and Harrison, M. M. (2019) 'Mechanisms regulating zygotic genome activation', *Nature Reviews Genetics*. Nature Publishing Group, pp. 221–234. doi: 10.1038/s41576-018-0087-x.

Shah, K. *et al.* (2003) *The genetic basis of infertility*, *Reproduction*.

Shenk, M. K., Kaplan, H. S. and Hooper, P. L. (2016) 'Status competition, inequality, and fertility: Implications for the demographic transition', *Philosophical Transactions of the Royal Society B: Biological Sciences*. Royal Society of London, 371(1692). doi: 10.1098/rstb.2015.0150.

Shrivakshan, G. T. and Chandrasekar, C. (2012) *A Comparison of various Edge Detection Techniques used in Image Processing*.

Skinner, B. M., *et al.* (2019) 'A high-throughput method for unbiased quantitation and categorization of nuclear morphology†', *Biology of reproduction*. NLM (Medline), 100(5), pp. 1250–1260. doi: 10.1093/biolre/ioz013.

Skinner, B. M., *et al.* (2019a) 'Automated Nuclear Cartography Reveals Conserved Sperm Chromosome Territory Localization across 2 Million Years of Mouse Evolution', *Genes*. MDPI AG, 10(2), p. 109. doi: 10.3390/genes10020109.

Skinner, B. M., *et al.* (2019b) 'Automated Nuclear Cartography Reveals Conserved Sperm Chromosome Territory Localization across 2 Million Years of Mouse Evolution', *Genes*, 10(2), p. 109. doi: 10.3390/genes10020109.

Skinner, B. M. (2022) 'Nuclear Morphology Analysis 2.0.0: Improved image analysis software for measuring nuclear shape', *Journal of Open Source Software*, 7(79), p. 4767. doi: 10.21105/joss.04767.

Solov'eva, L. *et al.* (2004) 'Nature of telomere dimers and chromosome looping in human spermatozoa', *Chromosome Research*. Springer Netherlands, 12(8), pp. 817–823. doi: 10.1007/s10577-005-5513-1.

Sotolongo, B. *et al.* (2005) 'An Endogenous Nuclease in Hamster, Mouse, and Human Spermatozoa Cleaves DNA into Loop-Sized Fragments', *Journal of Andrology*. American Society of Andrology Inc., 26(2), pp. 272–280. doi: 10.1002/j.1939-4640.2005.tb01095.x.

Stark, R., Grzelak, M. and Hadfield, J. (2019) 'RNA sequencing: the teenage years', *Nature Reviews Genetics*, 20, pp. 631–656.

Suzuki, T. A. and Nachman, M. W. (2015) 'Speciation and reduced hybrid female fertility in house mice', *Evolution*. Society for the Study of Evolution, 69(9), pp. 2468–2481. doi: 10.1111/EVO.12747.

Svoboda, P. (2018) 'Mammalian zygotic genome activation', *Seminars in Cell and Developmental Biology*. Elsevier Ltd, pp. 118–126. doi:

10.1016/j.semcdb.2017.12.006.

Tanemura, K. *et al.* (2005) 'Dynamic rearrangement of telomeres during spermatogenesis in mice', *Developmental Biology*. Academic Press Inc., 281(2), pp. 196–207. doi: 10.1016/j.ydbio.2005.02.025.

Team, R. C. (2018) 'R: A Language and Environment for Statistical Computing'. Vienna, Austria.

Turner, K. J. *et al.* (2021) 'Telomere Distribution in Human Sperm Heads and Its Relation to Sperm Nuclear Morphology: A New Marker for Male Factor Infertility?', *International Journal of Molecular Sciences*, 22(14), p. 7599. doi: 10.3390/IJMS22147599.

Turner, L. M., Schwahn, D. J. and Harr, B. (2012) 'REDUCED MALE FERTILITY IS COMMON BUT HIGHLY VARIABLE IN FORM AND SEVERITY IN A NATURAL HOUSE MOUSE HYBRID ZONE', *Evolution*. John Wiley & Sons, Ltd, 66(2), pp. 443–458. doi: 10.1111/J.1558-5646.2011.01445.X.

Turner, S. *et al.* (2010) 'Telomere lengths in human oocytes, cleavage stage embryos and blastocysts', *Molecular Human Reproduction*. Oxford Academic, 16(9), pp. 685–694. doi: 10.1093/molehr/gaq048.

Turner, S. and Hartshorne, G. M. (2013) 'Telomere lengths in human pronuclei, oocytes and spermatozoa', *Molecular Human Reproduction*, 19(8), pp. 510–518.

Twyford, A. D. and Ennos, R. A. (2011) 'Next-generation hybridization and introgression', *Heredity 2012 108:3*. Nature Publishing Group, 108(3), pp. 179–189. doi: 10.1038/hdy.2011.68.

Urdinguio, R. G. *et al.* (2015) 'Aberrant DNA methylation patterns of spermatozoa in

men with unexplained infertility', *Human Reproduction*. Oxford University Press, 30(5), pp. 1014–1028. doi: 10.1093/humrep/dev053.

Vasilopoulos, E. *et al.* (2019) 'The association of female and male infertility with telomere length (Review)', *International Journal of Molecular Medicine*. Spandidos Publications, pp. 375–389. doi: 10.3892/ijmm.2019.4225.

La Vignera, S. and Vita, R. (2018) 'Thyroid dysfunction and semen quality', *International Journal of Immunopathology and Pharmacology*. SAGE Publications, 32. doi: 10.1177/2058738418775241.

Voldgorn, Y. I. *et al.* (2015) 'Cultivation and Differentiation Change Nuclear Localization of Chromosome Centromeres in Human Mesenchymal Stem Cells', *PLOS ONE*. Edited by T.-Y. Roh. Public Library of Science, 10(3), p. e0118350. doi: 10.1371/journal.pone.0118350.

Wang, L. (1990) *Texture Unit, Texture Spectrum, and Texture Analysis*, *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*.

Wang, X. (2007) 'Laplacian operator-based edge detectors', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5), pp. 886–890. doi: 10.1109/TPAMI.2007.1027.

Wang, X. *et al.* (2021) 'Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2', *Genomics, Proteomics & Bioinformatics*. Elsevier, 19(2), pp. 253–266. doi: 10.1016/J.GPB.2020.02.005.

Ward, W. S., Kimura, Y. and Yanagimachi, R. (1999) 'An Intact Sperm Nuclear Matrix May Be Necessary for the Mouse Paternal Genome to Participate in Embryonic Development1', *Biology of Reproduction*. Society for the Study of

Reproduction, 60(3), pp. 702–706. doi: 10.1095/biolreprod60.3.702.

Weinberger, L. *et al.* (2016) 'Dynamic stem cell states: naive to primed pluripotency in rodents and humans', *Nature Publishing Group*. doi: 10.1038/nrm.2015.28.

Wiblin, A. E. *et al.* (2005) 'Distinctive nuclear organisation of centromeres and regions involved in pluripotency in human embryonic stem cells', *Journal of Cell Science*. The Company of Biologists Ltd, 118(17), pp. 3861–3868. doi: 10.1242/jcs.02500.

Wienholds, E. and Plasterk, R. H. A. (2005) 'MicroRNA function in animal development'. doi: 10.1016/j.febslet.2005.07.070.

Wiland, E. *et al.* (2016) 'Topology of chromosome centromeres in human sperm nuclei with high levels of DNA damage', *Scientific Reports*. Nature Publishing Group, 6(1), pp. 1–12. doi: 10.1038/srep31614.

Wiland, E., Żegało, M. and Kurpisz, M. (2008) 'Interindividual differences and alterations in the topology of chromosomes in human sperm nuclei of fertile donors and carriers of reciprocal translocations', *Chromosome Research*. Springer Netherlands, 16(2), pp. 291–305. doi: 10.1007/s10577-007-1194-2.

Wright, D. L. (2001) 'Characterization of telomerase activity in the human oocyte and preimplantation embryo', *Molecular Human Reproduction*. Oxford University Press, 7(10), pp. 947–955. doi: 10.1093/molehr/7.10.947.

Wright, W. E. *et al.* (1996) 'Telomerase activity in human germline and embryonic tissues and cells', *Developmental Genetics*. Dev Genet, 18(2), pp. 173–179. doi: 10.1002/(SICI)1520-6408(1996)18:2<173::AID-DVG10>3.0.CO;2-3.

Wykes, S. M. and Krawetz, S. A. (2003) 'The structural organization of sperm

chromatin', *Journal of Biological Chemistry*. Elsevier, 278(32), pp. 29471–29477. doi: 10.1074/jbc.M304545200.

Xie, L. *et al.* (2010) 'Integration of sperm motility and chemotaxis screening with a microchannel-based device', *Clinical Chemistry*, 56(8), pp. 1270–1278. doi: 10.1373/clinchem.2010.146902.

Yates, A. D. *et al.* (2020) 'Ensembl 2020', *Nucleic Acids Research*. Oxford University Press, 48(D1), pp. D682–D688. doi: 10.1093/nar/gkz966.

Yazdani, F., Kazemi, A. and Ureizi-Samani, H. R. (2016) 'Studying the Relationship between the Attitude to Infertility and Coping Strategies in Couples Undergoing Assisted Reproductive Treatments.', *Journal of reproduction & infertility*, 17(1), pp. 56–60.

Yildirim, M. E. *et al.* (2015) 'What is harmful for male fertility: Cell phone or the wireless internet?', *Kaohsiung Journal of Medical Sciences*. Elsevier (Singapore) Pte Ltd, 31(9), pp. 480–484. doi: 10.1016/j.kjms.2015.06.006.

Younglai, E. V. *et al.* (2001) 'Sperm swim-up techniques and DNA fragmentation', *Human Reproduction*, 16(9), pp. 1950–1953. doi: 10.1093/humrep/16.9.1950.

Yu, Z. *et al.* (2003) 'Gene expression profiles in different stages of mouse spermatogenic cells during spermatogenesis', *Biology of Reproduction*. Biol Reprod, 69(1), pp. 37–47. doi: 10.1095/biolreprod.102.012609.

Zalenskaya, I. A., Bradbury, E. M. and Zalensky, A. O. (2000) 'Chromatin structure of telomere domain in human sperm', *Biochemical and Biophysical Research Communications*. Academic Press Inc., 279(1), pp. 213–218. doi: 10.1006/bbrc.2000.3917.

Zalenskaya, I. A. and Zalensky, A. O. (2004) ‘Non-random positioning of chromosomes in human sperm nuclei’, *Chromosome Research*, 12(2), pp. 163–173. doi: 10.1023/B:CHRO.0000013166.04629.97.

Zalensky, A. O. *et al.* (1993) ‘Organization of centromeres in the decondensed nuclei of mature human sperm’, *Chromosoma*. Springer-Verlag, 102(8), pp. 509–518. doi: 10.1007/BF00368344.

Zalensky, A. O. *et al.* (1995) ‘Well-defined genome architecture in the human sperm nucleus’, *Chromosoma*. Springer-Verlag, 103(9), pp. 577–590. doi: 10.1007/BF00357684.

9 Appendices

9.1 Appendix 1: the cell IDs of the spermatids labelled as X or Y bearing from the Bhutani *et al.* dataset (Bhutani *et al.*, 2021)

Table S1: the cell IDs of the spermatids labelled as X bearing and the spermatids labelled as Y bearing

Putative X bearing n=132			Putative Y bearing n=118		
SRR15905242	SRR15905788	SRR15905415	SRR15905238	SRR15905395	SRR15905682
SRR15905243	SRR15905789	SRR15905418	SRR15905240	SRR15905401	SRR15905683
SRR15905245	SRR15905790	SRR15905419	SRR15905253	SRR15905405	SRR15905684
SRR15905249	SRR15905793	SRR15905426	SRR15905255	SRR15905406	SRR15905689
SRR15905250	SRR15905794	SRR15905428	SRR15905261	SRR15905408	SRR15905691
SRR15905257	SRR15905816	SRR15905440	SRR15905264	SRR15905409	SRR15905695
SRR15905259	SRR15905904	SRR15905442	SRR15905267	SRR15905413	SRR15905702
SRR15905260	SRR15905937	SRR15905444	SRR15905271	SRR15905416	SRR15905705
SRR15905262	SRR15905275	SRR15905448	SRR15905277	SRR15905417	SRR15905708
SRR15905265	SRR15905276	SRR15905453	SRR15905283	SRR15905420	SRR15905709
SRR15905266	SRR15905278	SRR15905457	SRR15905284	SRR15905421	SRR15905713
SRR15905268	SRR15905279	SRR15905459	SRR15905286	SRR15905422	SRR15905714
SRR15905272	SRR15905280	SRR15905464	SRR15905288	SRR15905423	SRR15905716
SRR15905273	SRR15905282	SRR15905466	SRR15905302	SRR15905427	SRR15905724
SRR15905274	SRR15905285	SRR15905467	SRR15905305	SRR15905429	SRR15905732
SRR15905733	SRR15905300	SRR15905489	SRR15905308	SRR15905431	SRR15905737
SRR15905734	SRR15905301	SRR15905508	SRR15905313	SRR15905434	SRR15905739
SRR15905736	SRR15905309	SRR15905643	SRR15905316	SRR15905435	SRR15905744
SRR15905740	SRR15905310	SRR15905646	SRR15905318	SRR15905437	SRR15905746
SRR15905742	SRR15905322	SRR15905647	SRR15905323	SRR15905438	SRR15905749
SRR15905745	SRR15905324	SRR15905651	SRR15905326	SRR15905439	SRR15905750
SRR15905747	SRR15905327	SRR15905660	SRR15905330	SRR15905443	SRR15905751
SRR15905748	SRR15905329	SRR15905662	SRR15905332	SRR15905445	SRR15905759
SRR15905752	SRR15905331	SRR15905667	SRR15905333	SRR15905446	SRR15905770
SRR15905753	SRR15905334	SRR15905669	SRR15905337	SRR15905642	SRR15905772
SRR15905754	SRR15905335	SRR15905671	SRR15905340	SRR15905644	SRR15905774
SRR15905755	SRR15905338	SRR15905672	SRR15905341	SRR15905645	SRR15905780
SRR15905756	SRR15905339	SRR15905675	SRR15905342	SRR15905648	SRR15905783
SRR15905758	SRR15905343	SRR15905677	SRR15905346	SRR15905650	SRR15905792
SRR15905760	SRR15905344	SRR15905686	SRR15905348	SRR15905653	SRR15905926
SRR15905761	SRR15905345	SRR15905688	SRR15905353	SRR15905654	
SRR15905763	SRR15905349	SRR15905692	SRR15905367	SRR15905655	
SRR15905766	SRR15905350	SRR15905693	SRR15905368	SRR15905657	
SRR15905767	SRR15905351	SRR15905694	SRR15905371	SRR15905658	
SRR15905768	SRR15905352	SRR15905698	SRR15905376	SRR15905659	
SRR15905769	SRR15905366	SRR15905699	SRR15905377	SRR15905661	
SRR15905775	SRR15905373	SRR15905710	SRR15905378	SRR15905666	
SRR15905777	SRR15905375	SRR15905712	SRR15905379	SRR15905668	
SRR15905779	SRR15905387	SRR15905715	SRR15905382	SRR15905670	
SRR15905781	SRR15905399	SRR15905721	SRR15905383	SRR15905673	
SRR15905782	SRR15905407	SRR15905722	SRR15905384	SRR15905676	
SRR15905785	SRR15905410	SRR15905723	SRR15905389	SRR15905678	
SRR15905786	SRR15905411	SRR15905728	SRR15905390	SRR15905679	
SRR15905787	SRR15905412	SRR15905731	SRR15905393	SRR15905680	

9.2 Appendix 2: the motifs found in the differentially expressed genes and their candidate transcription factor targets

Table S2: the novel motifs found in the genes upregulated in the Y-bearing spermatids, and potential transcription factors that may bind these motifs

Motif	Number of matching transcription factors	Autosomal Transcription Factors	Sex linked transcription factors
AAAAAAAAAAAAA	31	*Foxj3, Foxq1, Stat1, Foxd3, Barx1, Cdx2, Evx2, Foxl2, Mef2c, Hnf6, Foxj2, Stat2, Cux2, Sox4, Lef1, Tcf7, Foxa2, Mef2a, Nanog, Olig2, Foxc1, Ets2, Irf3, Aire, Evi1, Foxa3, Elf5, Gata3, Sox5*	*Sry* (Y) *Foxo4* (X)
SCCCGCCCCT	27	*Klf6, Klf3, Sp1, Sp3, Sp4, Sp2, E2f4, Klf15, E2f1, Egr1, E2f7, Maz, Zbt17, E2f6, Wt1, Zn281, Klf5, Klf1, E2f3, Sp5, Egr2, Zic1, Sall4, Srbp2, Esr2, Mbd2*	*Gata1* (X)
CGCGGGGGCTGCCG	16	*Ap2b, Sp2, Mxi1, Zbt7a, Ascl1, Egr1, Myod1, Egr2, Srbp2, Hic1, Ctcfl, Klf6, Sp3, Bha15*	*Klf8* (X) *Zfx* (X)
GTCCTGGAACTCA	26	*Bcl6, Nkx28, Err2, Rara, Pknx1, Err1, Pbx1, Rxrg, Rxra, Stat4, St5a, Stat3, Nr5a2, Stf1, Tead4, Zn322, Tead1, Nfic, Prdm5, Sta5b, Hlf, Tf7l1, Nr1h4, Hsf1, Rfx6, Tead2*	
CCCAGCACC	19	*Klf5, Klf1, Klf3, Klf6, Tbx20, Sall4, Sp3, Sp2, Rest, Bha15, Egr1, Rfx6, Htf4, Sp1, Tbx3, Sp4, Tfe2, E2f1, Bach2*	
AGCCTGGTCYAC	4	*Zn322, Grhl2, Foxa2, Prdm5*	
CCAASTCGA	6	*Hnf4a, Mycn, Hnf4g, Stf1, Rfx3, Snai1*	

Table 9: motif logo and percentage of genes containing the motif for motifs found in the transcription factors that may bind the genes upregulated in X-bearing spermatids

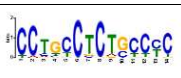
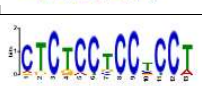
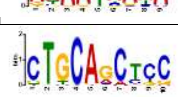
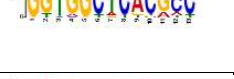
Motif	Motif logo	% of genes with motif site	Motif	Motif logo	% of genes with motif site
CCTGCCTCTGCCYC		42.9	GAAAGCAAA		39.7
ACACACACACA		22.3	AAACAAACAAAAA		52.2
TTAAAAAAAAAAAAA		37.5	AATTTTTWTAAAAAA		18.2
CCCCRCCCCCMC		65.5	CTGAGTTCRAGGC		22.6
TAATCCCAGCA		19.0	CTAGGCTT		41.0
CTCCCTCCCCC		20.9	TGACGTCAGTG		17.4
CGACGTTTCC		12.0	AGTGCTCTTA		7.9
CTGAGCAACA		17.7	AGCCAATCAG		18.2
AAGTGAGTTMCAG		23.4	CTCTCCTCCTCCT		14.9
GGGCTACWCAGAGAR		19.6	AAAATAAAA		39.4
GAAGTTAGGT		21.7	CTGCAGCTCC		34.0
ATGTAHDTACAT		16.0	CGCCCCCGCCGCSCC		31.5
ATACATCATKSAA		13.6	CCGCGCGCTCCCG		10.1
GGTGGCTCACGCC		13.3	AGAGAGAGAGAGAD		8.7
AGAGATGGCTCAGTG		10.3	TAGTTTAATRAT		8.7
GGAAGGAAGGA		29.3	CGCCTCCTGCCC		17.1
AAATGAAG		49.7	CAAGCAAGCC		18.2

Table 10: motif logo and percentage of genes containing the motif for motifs found in the transcription factors that may bind the genes upregulated in Y-bearing spermatids

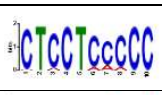
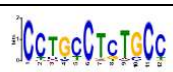
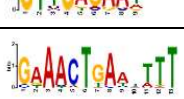
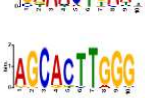
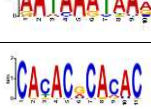
Motif	Motif logo	% of genes with motif site
CTCCTCCCCC		25.2
TAAAAAAAAAAAA		37.4
CCCCACCCCC		61.7
CCTGCCTCTGCC		37.4
AGTGAGTTCCAG		27.0
CCAATCAGA		22.5
ACGTCACCG		45.9
GTTGAGAAT		35.1
GAAACTGAATTTT		27.9
AGCACTTGGG		21.2
AATAAATAARTTY		22.1
CACACGCACAC		25.7

Table 11: the novel motifs found in the candidate transcription factors that bind genes upregulated in X bearing spermatids, and potential second-level transcription factors that may bind these motifs

Motif	No. of matching transcription factors	Autosomal Transcription Factors	Sex linked transcription factors
CCTGCCTCTGCCYC	51	*Maz, Sp5, Ascl1, Wt1, Zbt17, Klf15, Sp1, Sp4, Klf6, Sp2, Myog, Sp3, Myod1, Ptf1a, Snai2, Myba, Bha15, Klf5, Zn281, Egr1, Klf3, Elk4, Prdm5, E2f1, E2f7, Smad4, Klf1, Etv6, Spi1, Irf3, Ets1, Tal1, Smad3, Spib, Pknx1, Fli1, Elf5, Jund, Zn335, Erg, Ets2, Atoh1, Etv2, Nr2c2, Nr1h3, Pparg, Sox4, Tfe2, Lyl1*	*Gata1* (X) *Sox3* (X)
ACACACACACA	7	*Brac, Foxj3, Egr2, Klf4, Smad2, Clock, Tbx21*	
TTAAAAAAAAAAAAA	27	*Foxd3, Aire, Mef2d, Hnf1a, Otx2, Stat2, Sox4, Lef1, Prop1, Nanog, Barx1, Hnf6, Evx2, Cdx2, Stat1, Foxq1, Sox2, Foxc1, Nkx61, Cux2, Foxj3, Mef2c, Foxl2, Mef2a, Foxj2*	*Sry* (Y) *Foxo4* (X)
CCCCRCCCCCMC	34	*Zn281, Sp3, Zbt17, Egr1, Egr2, Sp2, Sall4, Klf3, Klf5, Sp1, Klf6, Klf4, Sp5, Srbp2, E2f1, Sp4, Maz, E2f4, Zic1, Mxi1, Wt1, Klf1, Klf15, Srbp1, Htf4, E2f7, E2f6, Myba, Pparg, Tbx21, E2f3, Usf2, Nr1d1*	*Klf8* (X)
TAATCCCAGCA	15	*Gfi1b, Gfi1, Tha1i, Zn143, Tead4, Runx3, Otx2, Pitx1, Alx1, Srbp1, Prop1, Crx, Ari5b, Zic2*	*Zic3* (X)
CTCCCTCCCCC	40	*Maz, Sp1, Wt1, Klf5, Zbt17, Zn281, Sp5, Klf15, Sp3, Klf3, Sall4, Klf6, E2f1, Ptf1a, E2f4, Egr2, E2f3, Fli1, E2f6, Ets1, Tal1, Sp2, Klf4, E2f7, Egr1, Tbx21, Zic1, Elf5, Srbp1, Sp4, Mbd2, Srbp2, Irf3, Esr2, Spib, Ets2, Gabpa, Nr1d1, Nr1d2*	*Gata1* (X)
CGACGTTTCC	9	*Xbp1, Zn143, Nr4a2, Foxc1, Pparg, Ppara, Usf1, Nfkb2, Cot1*	
CTGAGCAACA	13	*Rfx6, Hlf, Atf3, Nf2l2, Bach2, Nr1d1, Rora, Maff, Junb, Mtf1, Hic1, Sta5b, Nr1d2*	
AAGTGAGTTMCAG	24	*Nr1d2, Pknx1, Pbx1, Ovol1, Jun, Smad4, Hsf1, Ppara, Nr1d1, Twst1, Junb, Rorg, Elf2, Jund, Prop1, Rel, Pbx2, Gfi1, Atf3, Fosb, Pparg, Fosl1, Relb, Rora*	

GGGCTACWCAGAGAR	7	*Foxi1, Gli1, Hic1, Nfyb, Smad4, Tcf7, Smad2*	
GAAGTTAGGT	3	*Rorg, Vsx2, Elf2*	
ATGTAHDTACAT	10	*Po3f2, Foxa1, Pit1, Po3f1, P73, Foxl2, Poxj2, Cdx1, Foxk1, P63*	
ATACATCATKSAA	18	*Hnf6, Cdx1, Hxb8, Pbx2, Ddit3, Cebpg, Batf3, Evx1, Foxq1, Nfil3, Elf2, Hxa9, Hxb7, Hxa1, Meis1, Cux2, Atf4*	*Cdx4* (X)
GGTGGCTCACGCC	13	*Fos, Fosl1, Fosl2, Junb, Jund, Atf3, Jun, Fosb, E2f3, Bach1, Ap2a, E2f2*	*Zfx* (X)
AGAGATGGCTCAGTG	24	*Nr1d2, Mafk, Fosl1, Ndf1, Atf3, Fosb, Ndf2, Batf, Mafb, Bach2, Junb, Fosl2, Fos, Nr1d1, Cebpe, Jun, Maff, Ngn2, Nfe2, Rora, Atoh1, Zfp42, Bach1*	*Taf1* (X)
GGAAGGAAGGA	33	*Ets2, E2f6, Elf5, Sall4, Elf2, Sox4, Etv6, Ptf1a, E2f3, Olig2, Klf6, Nfac1, Gata6, Htf4, Wt1, E2f1, Maz, E2f7, Zbt17, Gata2, Etv4, Sp5, Klf15, Ets1, Sp1, Etv2, Elk4, E2f4, Klf5, Erg, Irf3, Elf3*	*Sox3* (X)
AAATGAAG	28	*Stat2, Creb1, Cux2, Zn143, Spi1, Spib, Hnf6, Irf8, Irf2, Stat1, Tyy1, Irf1, Atf4, Nf2l1, Ets2, Etv6, Poxj2, Batf3, Zfp42, Prdm1, Elf5, Maff, Tha11, Foxq1, Lhx3, Irf4, Foxl2*	*Taf1* (X)
CGCCTCCTGCCC	60	*Sp4, E2f3, Sp2, Sp1, Fli1, Myc,Elk4, Fev, Wt1, Erg, Ets1, E2f6, Insm1, Ctcfl, Mbd2, Klf15, Sp5, Htf4, Max, Zbt17, Elf2, Sp3, Zn281, Maz, Ctcf, Ets2, Snai1, Ptf1a, E2f7, Usf1, E2f1, Ngn2, Ascl1, Klf5, Nr1d1, Gabpa, Etv2, Etv4, Ap2a, Myf6, Ndf2, Usf2, Klf6, Bmal1, Snai2, Ascl2, Zic2, Pparg, Ap2b, Atoh1, Mitf, Ap2c, Rzra, Ndf1, Nr2c1, Myba*	*Klf8* (X) *Zic3* (X) *Zfx* (X) *Taf1* (X)
CAAGCAAGCC	6	*P53, Zn143, Tha11, P63, P73, Foxm1*	
GAAAGCAAA	33	*Irf9, Iff7, Foxm1, Po2f1, Lef1, Sox4, Bcl6, Irf2, Tcf7, Irf1, Ets2, Foxd3, Prdm1, Ahr, Olig2, Irf3, Foxj2, Rel, Foxa3, Runx1, Foxa1, Stat1, Barx1, Maf, Runx2, Foxa2, Irf8, Lyl1, Evx2, Six4, Cdx2*	*Sry* (Y) *Foxo4* (X)
AAACAAACAAAAA	38	*Foxj2, Foxq1, Foxd3, Foxj3, Barx1, Stat2, Hnf6, Cdx2, Stat1, Foxm1, Evx2, Foxp2, Sox5, Foxa2, Cux2, Foxa3, Foxc1, Aire, Foxi1, Nkx32, Pbx3, Hxb7, Foxa1, Nfya, Foxk1, Tcf7, Gata2, Foxl2, Elf3, Nfyc, Hxa13, Sp7, Irf3, Po2f2, Evi1, Sox4*	*Sry* (Y) *Foxo4* (X)

AATTTTTWTAAAAAA	15	*Barx1, Nfac4, Prop1, Nfac2, Mef2a, Otx2, Tbp, Foxj2, Pit1, Po5f1, Nanog, Aire, Sta5b, Sox2*	*Sry* (Y)
CTGAGTTCRAGGC	17	*Err2, Nr5a2, Rxra, Stf1, Rxrg, Err1, Rara, Cot2, Vdr, Zn322, Rxrb, Hnf4g, Ppara, Hnf4a, Batf, Tf7l1, Nkx28*	
CTAGGCTT	3	*P63, Kaiso, P73*	
TGACGTCAGTG	11	*Crem, Atf1, Creb1, Atf2, Vdr, Rxrb, Nr1h4, Mitf, Tgif1, Esr1*	*Tfe* (X)
AGTGCTCTTA	3	*Crx, Otx2, Prd14*	
AGCCAATCAG	16	*Nfyc, Nfyb, Nfya, Pbx3, Foxi1, Lhx6, Vsx2, Pbx1, Pknx1, Nobox, Pax6, Smad4, Nkx22, Pax5, Lhx3, Grhl2*	
CTCTCCTCCTCCT	30	*Zbt17, Wt1, Maz, Sp5, Klf15, Elf5, Sp1, Spib, Zn281, Ets2, Elf2, Spi1, Crx, Irf4, Sp4, Srbp2, Sp3, Olig2, E2f1, Irf8, Sp2, E2f7, Srbp1, Smad3, E2f4, Ehf, E2f6, Irf3, Htf4*	*Taf1* (X)
AAAATAAAA	46	*Mef2c, Mef2a, Foxj3, Stat2, Foxj2, Stat1, Mef2d, Batf3, Irf2, Hxa13, Irf1, Nr2e3, Foxq1, Foxl2, Irf7, Prdm1, Evi1, Cdx2, Evx2, Foxd3, Cdx1, Sox5, Twst1, Tcf7, Foxc1, Lef1, Sox2, Foxa3, Nanog, Tbp, Foxa1, Nfac1, Gata3, Foxa2, Tf7l2, Hnf6, Foxm1, Irf3, Cux2, Alx1, Barx1, Sox4, Gata4, Po5f1*	*Sry* (Y) *Foxo4* (X)
CTGCAGCTCC	17	*Ascl1, Tfe2, Myog, Bha15, Myf6, Snai2, Tha11, Ascl2, Myba, Prdm9, Hen1, Ptf1a, Itf2, Zn143, Myod1, Ngn2, Elk4*	
CGCCCCCGCCGCSCC	31	*Sp3, Sp1, Sp2, Klf3, Wt1, Klf6, Zbt17, Egr1, Zn281, Klf15, Usf2, Maz, Sp5, Sp4, Mxi1, Klf1, Egr2, Myog, Klf5, E2f6, Ctcf, Hen1, E2f3, Fli1, Myc, E2f7, Ctcfl, Mbd2, Klf4, Elk4*	*Zfx* (X)
CCGCGCGCTCCCG	20	*Sp1, Sp3, Klf3, Sp2, Zbt17, Sp4, Max, Egr1, Klf5, Klf6, Usf2, Prdm5, Ctcfl, Myc, Zn281, Sp5, Mycn, Wt1, E2f1, E2f3*	
AGAGAGAGAGAGAD	17	*Smad4, Olig2, Irf2, Irf4, Sp5, Myod1, Irf1, Irf8, Ets2, Smad3, Nr1d2, Tal1, Irf3, Prdm1, Maz, Spi1, Spib*	
TAGTTTAATRAT	17	*Alx1, Prop1, Batf3, Hnf1b, Hnf1a, Hxb8, Hnf6, Foxp2, Hxb7, Foxj3, Cux2, Pdx1, Stat2, Nfil3, Hxb4, Lhx2*	*Foxo4* (X)

Table 12: the novel motifs found in the candidate transcription factors that bind genes upregulated in Y bearing spermatids, and potential second-level transcription factors that may bind these motifs

Motif	No. of matching transcription factors	Autosomal Transcription factors	Sex linked trans-cription factors
CTCCTCCCCC	39	*Wt1, Sp5, Zn281, Zbt17, Maz, Sp1, Klf15, Egr1, Sp3, Sp4, Egr2, Hen1, Sp2, Ptf1a, Sall4, Rxrg, Rxra, Mbd2, E2f4, Ets2, E2f1, Srbp2, Rara, Zic1, Fli1, E2f3, Klf6, Klf3, Nr1d1, E2f6, E2f7, Fev, Ets1, Elf5, Ascl1, Ikzf1, Srbp1*	*Klf8* (X) *Gata1* (X)
TAAAAAAAAAAAAA	28	*Foxj3, Foxj2, Mef2c, Barx1, Foxq1, Foxl2, Foxd3, Foxc1, Nkx61, Nanog, Mef2a, Hnf6, Lef1, Evx2, Cux2, Aire, Cdx2, Stat1, Sox2, Mef2d, Stat2, Sox4, Foxa2, Sox5, Twst1, Hnf1a*	*Sry* (Y) *Foxo4* (X)
CCCCACCCCC	33	*Zbt17, Egr1, Zn281, Sp3, Sp2, Egr2, Sall4, Sp1, Maz, Sp4, E2f4, Klf4, E2f1, Klf3, Sp5, Zic1, Klf5, Wt1, Klf6, Klf1, Srbp2, Klf15, Mxi1, Htf4, Tbx21, E2f7, Zbt7a, E2f3, Srbp1, E2f6, Gli1*	*Klf8* (X) *Gata1* (X)
CCTGCCTCTGCC	45	*Myod1, Klf15, Myba, Ascl1, Bha15, Maz, Snai2, Wt1, Sp5, Sp4, Klf6, Zbt17, Egr1, Etv6, Tal1, Prdm5, Sp1, Sp2, Maf, Sp3, Zn335, Zn281, Klf3, Ets1, Gabpa, Mafb, Cebpg, Moyg, Nr1h3, Atoh1, Etv2, Klf1, Pknx1, Zfp57, Tfe2, Egr2, Klf4, Spi1, Ptf1a, Rfx1, Myb, Jund, Klf5, P53*	*Gata1* (X)
AGTGAGTTCCAG	15	*Hsf1, Pknx1, Nr1d2, Twst1, Vdr, Elf2, Pbx1, Bcl6, Jun, Relb, Ovol1, Gfi1, Smad4, Rel, Batf*	
CCAATCAGA	19	*Nfyc, Nfya, Nfyb, Foxi1, Pbx3, Lhx6, Smad2, Nobox, Nkx21, Hxa13, Pbx1, Lhx3, Lhx2, Pknx1, Prrx2, Gli1, Isl1, Barx1, Nkx25*	*Tfe3* (X)
ACGTCACCG	16	*Creb1, Crem, Atf2, Usf1, Tfeb, Atf1, Usf2, Esr2, Esr1, Mitf, Srbp1, Bhe40, Nr4a2, Err3, Rarg*	
GTTGAGAAT	8	*Hnf1a, Tbx21, Gata3, Nfac2, Hnf1b, Gfi1b, Ikzf1, Zn143*	
GAAACTGAATTTT	22	*Irf7, Prgr, Irf9, Prop1, Prdm1, Gcr, Maff, Irf1, Pbx2, Andr, Alx1, Irf2, Ovol1, Stat2, Nr4a2, Stat1, Err2, Six2, Rara, Grhl2, Cot1, Rorg*	
AGCACTTGGG	25	*Htf4, Max, Mycn, Myc, Bmal1, Atoh1, Itf2, Mxi1, Snai1, Bha15, Myb, Ascl2, Coe1, Tfe2, Arnt, Nkx31, Bhe40, Hnf4a, Myba, Ngn2, Ap2b, Nkx28, Nkx25, Nkx32, Ndf2*	
AATAAATAARTTY	28	*Pbx2, Hnf6, Cux2, Alx1, Foxj2, Prop1, Pit1, Foxq1, Foxj3, Foxp2, Cdx1, Hxb7, Hnf1b, Lhx2, Hxb8, Prrx2, Foxd3, Evx1, Po2f2, Foxl2, Hxa9, Foxo3, Nfac1, Foxa3, Foxa1, Barx1*	*Foxo4* (X) *Cdx4* (X)
CACACGCACAC	6	*Klf4, Egr2, Ahr, Clock, Egr1, Nrf1*	

9.3 Appendix 3: Published peer-reviewed paper

Telomere Distribution in Human Sperm Heads and Its Relation to Sperm Nuclear Morphology: A New Marker for Male Factor Infertility?

Kara J. Turner [1], Eleanor M. Watson [2], Benjamin M. Skinner [2] and Darren K. Griffin [1,*]

[1] School of Biosciences, University of Kent, Giles Lane, Canterbury CT2 7NH, UK; karacccu@gmail.com

[2] School of Life Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK; e.watson@essex.ac.uk (E.M.W.); b.skinner@essex.ac.uk (B.M.S.)

* Correspondence: d.k.griffin@kent.ac.uk

Abstract: Infertility is a problem affecting an increasing number of couples worldwide. Currently, marker tests for male factor infertility are complex, highly technical and relatively subjective. Up to 40% of cases of male factor infertility are currently diagnosed as idiopathic therefore, there is a clear need for further research into better ways of diagnosing it. Changes in sperm telomere length have been associated with infertility and closely linked to DNA damage and fragmentation, which are also known to be related to infertility (Fice and Robaire, 2019) However, telomere distribution is a parameter thus far underexplored as an infertility marker. Here, we assessed morphological parameters of sperm nuclei in fertile control and male factor infertile cohorts. In addition, we used 2D and 3D fluorescence in situ hybridization (FISH) to compare telomere distribution between these two groups. Our findings indicate that the infertile cohort sperm nuclei were, on average, 2.9% larger

in area and showed subtle differences in sperm head height and width. Telomeres were mainly distributed towards the periphery of the nuclei in the control cohort, with diminishing telomere signals towards the center of the nuclei. Sperm nuclei of infertile males, however, had more telomere signals towards the center of the nuclei, a finding supported by 3D imaging. We conclude that, with further development, both morphology and telomere distribution may prove useful investigative tools in the fertility clinic.

Keywords: telomere; sperm; nuclear organization; nuclear morphology; infertility

1. Introduction

Infertility under the current WHO definition is estimated to affect over 50 million (one in six) couples worldwide [1]. Approximately 50% is attributed to male factor issues, at least in part [2,3], up to 40% of which are currently diagnosed as idiopathic [4]. The causes of male factor infertility may include physiological or hormonal complications; however, in 15–30% of cases, a genetic cause can be attributed. These include chromosomal inversions, translocations, Y chromosome microdeletions, aneuploidy or DNA damage [5,6,7]. One largely under-explored genetic contribution to male factor infertility, however, is that of nuclear organization, the positioning and packaging of chromatin within the sperm nucleus. Although nuclear organization is relatively well-described in human sperm, the impact of aberrant nuclear organization and the exact mechanisms underlying how it might play a role in male factor infertility are currently unknown. A deeper understanding of this phenomenon in male factor infertility may shed light on the 30–40% of idiopathic

cases that demonstrate a ‘normal’ range in the standard semen parameters assessed (Bracke *et al.*, 2018; Cho and Agarwal, 2018).

The mature human sperm head conforms to a chromocentric model of nuclear organization. This describes the positioning of chromosomes such that the centromeres are located predominantly at the interior of the nucleus in chromocenters, while telomeres reside at the nuclear periphery in clusters (Zalensky *et al.*, 1993, 1995; Jennings and Powell, 1995). Organization of chromosomes in this way is enabled by the bending of chromosomes into hairpin structures and via close intra-chromosomal interactions at the telomere (Zalensky *et al.*, 1995; Solov’eva *et al.*, 2004).

The association of telomeres with the nuclear membrane in the sperm nucleus is thought to play an important role in anchoring chromosome territories in place to form a strict and highly organized functional nuclear landscape (Ward, Kimura and Yanagimachi, 1999; Zalenskaya, Bradbury and Zalensky, 2000; Sotolongo *et al.*, 2005; Rocca, Foresta and Ferlin, 2018; Lee *et al.*, 2020). For example, evidence has shown that chromosomes 1 and 6 occupy central-medial and medial-peripheral portions of the acrosomal half of the sperm nucleus, respectively (Zalenskaya and Zalensky, 2004; Mudrak, Tomilin and Zalensky, 2005), the X chromosome occupies the extreme anterior position of the nucleus in the acrosomal half (Luetjens, Payne and Schatten, 1999; Hazzouri *et al.*, 2000; Zalenskaya and Zalensky, 2004) and chromosomes 2, 5 and 18 occupy a region in the basal half of the nucleus nearer the tail (Luetjens, Payne and Schatten, 1999; Mudrak, Tomilin and Zalensky, 2005). This rigid arrangement of nuclear architecture in the sperm head evolves during the process of spermatogenesis (Foster *et al.*, 2005; Pereira *et al.*, 2019) and is thought to be fundamental in providing replication and transcriptional cues to the fertilized

oocyte (Zalenskaya, Bradbury and Zalensky, 2000; Sotolongo *et al.*, 2005). More specifically, it is postulated that the spatial location of chromatin determines the time of decondensation following fertilization that, in turn, dictates the epigenetic control of the paternal genome in the developing embryo (Wykes and Krawetz, 2003; Foster *et al.*, 2005).

The fact that a large proportion of infertile males exhibit evidence of impaired spermatogenesis lends credence to the hypothesis that altered nuclear organization might be a clinically relevant feature of male infertility. However, preliminary evidence from previous studies that have addressed this hypothesis is relatively scarce and inconclusive (K A Finch *et al.*, 2008; Olszewska, Wiland and Kurpisz, 2008; Wiland, Żegało and Kurpisz, 2008; Ioannou and Griffin, 2010; Alladin *et al.*, 2013; Perdrix and Rives, 2013). That being said, these studies have largely focused on centromere distribution, while telomere distribution within the human sperm nucleus remains under-investigated.

At present, the parameters set by the WHO that are used for male fertility assessment do not look at any information regarding sperm chromatin quality (Fice and Robaire, 2019). Despite this, there is ample evidence in the literature that indicates a clear relationship between sperm chromatin quality and male factor fertility. Such information provides considerable scope for elucidating hitherto undiscovered mechanisms. For example, there is significant evidence to suggest that there is a strong interplay between sperm chromatin quality and DNA damage (Ellis and Griffin, 2020). Using fluorescence in situ hybridization (FISH), confocal microscopy and 3D-reconstruction approaches on transgenic mouse models (*Gpx5*$^{-/-}$) of sperm nuclear oxidation, it was demonstrated that there is a significant impact on nuclear volume and surface area in the transgenic, oxidatively damaged

sperm, compared to wild-type sperm (Champroux *et al.*, 2018). This finding suggests that DNA fragmentation not only affects sperm nuclear organization (as demonstrated by Wiland et al. (Wiland *et al.*, 2016)) but also chromatin packaging (Wiland *et al.*, 2016; Champroux *et al.*, 2018). Human sperm have an increased vulnerability to oxidative stress due to a number of reasons; including the presence of high levels of polyunsaturated fatty acids (PUFAs) on their plasma membrane(which allow for many sites of free-radical-induced lipid peroxidation), low levels of antioxidants present due to the constraints on volume, a low level of cytoplasm and their inefficient DNA damage detection and repair system (Bisht and Dada, 2017). Indeed, DNA damage in sperm is seen in many instances among infertile men (Ribas-Maynou and Benet, 2019), and single-strand DNA breaks, in particular, tend to be related to oxidative stress. The latter usually leads to either a total lack of pregnancy or a vast increase in the time taken for conception to occur (Ribas-Maynou and Benet, 2019).

Taken together, these findings suggest that oxidation-induced DNA damage could lead to altered chromatin and telomere distribution in infertile men. Further research is, however, needed to explore this. In addition, further work is needed to determine the role that sperm nuclear size, density, texture and shape abnormalities play in male factor infertility (Ioannou and Griffin, 2011), especially as these abnormalities could, in part, be related to altered nuclear organization. Sperm heads consist almost solely of the sperm nucleus, and thus any change in distribution of genetic material in the nucleus could have a direct impact on the shape and structure of the sperm head. This may, in turn, impact the acrosome reaction, flagellum biogenesis and even the hydrodynamics of swimming (Gillies *et al.*, 2009). It is clear, therefore, that any abnormality in sperm nuclear size, density, texture or shape could have extreme

consequences on the sperm's function. In light of this, it seems a suitable time to revisit the organization of telomeres within the human sperm head and to investigate the overall shape of the sperm head in fertile control versus infertile males. Given that many current methods of assessing male fertility are clinically challenging due to the high levels of technology required, this study aims to assess whether telomere distribution or nuclear morphology could be a novel marker of male factor infertility (Fice and Robaire, 2019).

2. Results

All results were generated from a cohort of 13 males with normal semen parameters (henceforth termed "control cohort") and six males with compromised semen parameters (concentration, motility and/or morphology—henceforth termed "infertile cohort"). See materials and methods for details.

2.1. Nuclei from Infertile Men Have Altered Shape and Size

Sperm nucleus morphology from the infertile cohort differed significantly from the control cohort (Figures 1 and 2, Table 1). Measurements of the bounding heights, bounding widths and areas of each group of sperm nuclei (Figure 1) showed that in the infertile cohort, sperm were significantly taller and thinner with an overall larger area (Mann Whitney U tests; $p < 0.001$). The infertile cohort nuclei, on average, had a 3.34% smaller bounding width, 6.28% larger bounding height and 2.91% larger area. Bootstrapping confirmed that these results were not biased by individuals with more extreme values, with differences remaining significant following removal of any individual.

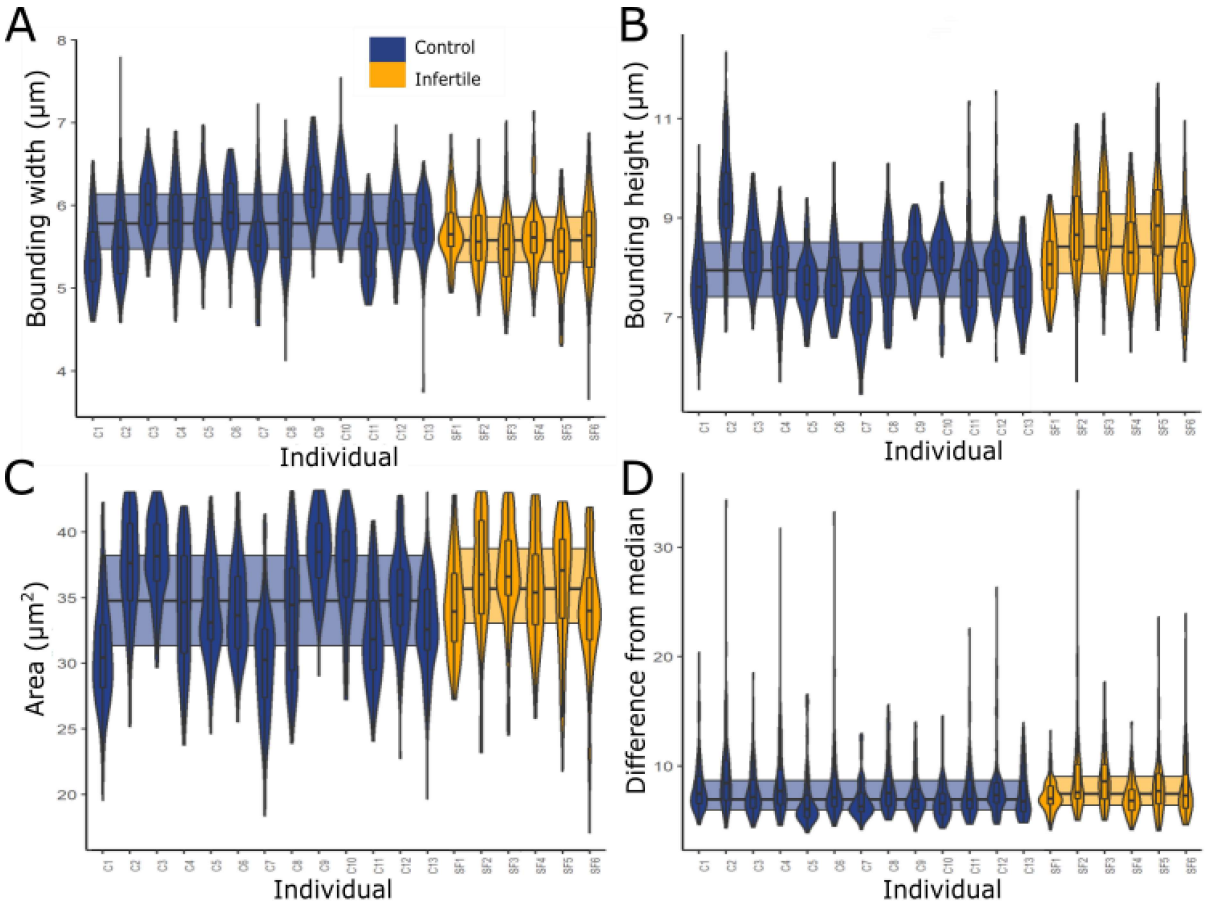


Figure 1. Comparison of nuclear morphologies between control and infertile individuals. (**A**) bounding widths; (**B**) bounding heights; (**C**) area shows infertile sperm nuclei are longer and thinner, and overall, larger. (**D**) Difference from median shows slightly increased variability of the nuclear shapes within each sample. The median and upper and lower quartiles are also shown for each cohort.

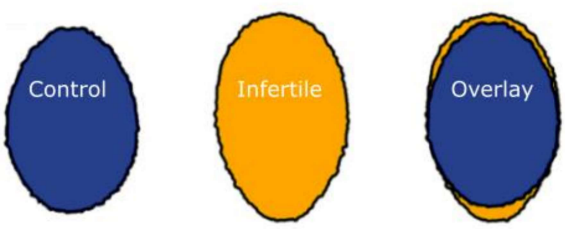


Figure 2. Consensus sperm nuclear shapes in control versus infertile males. The measurements that the infertile cohort nuclei are taller and thinner are further

demonstrated by overlaying the consensus outlines (the ‘average’ shape) of the control (blue) and infertile nuclei (yellow) nuclei.

Table 1. Sperm nuclear measurements in control versus infertile males. Mean values, standard deviations and standard error of the mean of the bounding widths, heights, areas and difference from median of the nuclei. * Denotes where the error bars overlap for the difference from median only.

	Controls				**Infertiles**			
Parameter	**Mean**	**Standard Deviation**	**Mean +/− SD**	**Mean +/− SEM**	**Mean**	**Standard Deviation**	**Mean +/− SD**	**Mean +/− SEM**
Bounding width (μm)	5.77	0.492	5.280–6.263	5.754–5.786	5.58	0.456	5.122–6.034	5.559–5.601
Bounding height (μm)	8.00	0.937	7.066–8.940	7.970–8.030	8.51	0.953	7.553-9.458	8.466–8.554
Area (μm^2)	34.45	4.616	29.831–39.063	34.300–34.600	35.45	4.415	31.033–39.863	35.246–35.654
Difference from median	7.96	3.085	4.874–11.045	7.860–8.060 *	8.19	3.081	5.108–11.270	8.048–8.332*

The variability of the nuclear shapes in each individual was also assessed via the difference from the median profile (for a technical explanation of this parameter, see (Skinner, Rathje, *et al.*, 2019)). This revealed that the infertile cohort had a slightly higher level (2.89%) of variability than the control cohort (Figure 1D), and therefore, less consistent morphology of the infertiles compared with the controls (Mann Whitney U test, $p = 0.01172$). Examples of the average and more varied-shaped nuclei are shown in Figure 3; the general morphological abnormalities seen are similar between the control and infertile cohorts, with apparent compression around the base of the nucleus. However, there were more of these abnormal sperm in the infertile cohort.

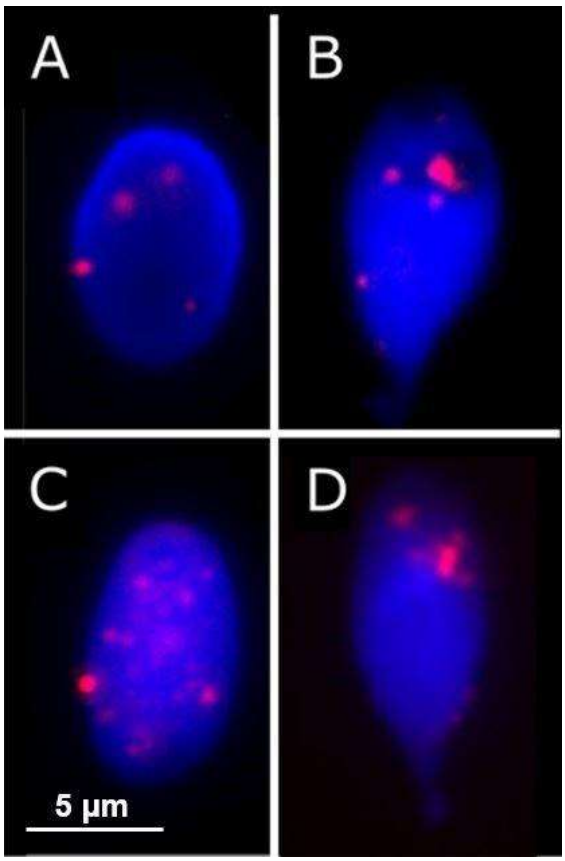


Figure 3. The nuclei with the smallest and largest differences from the average shape, for both control (**A**,**B**) and infertile cohorts (**C**,**D**), as measured by the difference in nucleus shape profiles from the median profile (see (Skinner, Rathje, *et al.*, 2019) for technical details). Shape abnormalities are similar in both but are more prevalent in the infertile cohort Red spots represent telomeres, whereas the nuclear volume is represented by a blue nucleic acid stain (DAPI).

2.2. Greater Number of Telomere Signals Observed in Infertile Cohort

The finding of shape and size differences between the nuclei in the cohorts indicated alterations in chromatin organization. Positions of telomeres were measured using 2D and 3D FISH. 2D FISH experiments showed that there was a greater number of telomere signals in the sperm nuclei of the infertile cohort compared to the controls, but there was also a higher variability in the infertile cohort: eight telomere signals on average per nucleus in the control cohort (SEM: 0.108) and 11 on average in the infertile cohort (SEM: 0.241). Telomeres were more internally located in sperm from the infertile cohort.

Results from 2D FISH experiments using a pan-telomere probe indicated a highly significant difference in the telomere distribution pattern of the sperm nuclei of control and infertile cohorts (chi-square, $p = < 0.001$). As shown in Figure 4, while telomeres could be seen across all areas of the nucleus in both groups, they were more likely to occupy more peripheral areas in the sperm nuclei of controls. In the infertile cohort, however, the opposite pattern of telomere distribution was observed. This finding was corroborated by signal warping the images of the telomere signals onto a common shape, generating a 'heat map' of where the signal was found in the nucleus [36]. Telomere signals in the control cohort appeared to be predominantly

evenly spread throughout the nucleus (Figure 5D), whereas those in the infertile cohort occupied more central zones of the nucleus (Figure 5E).

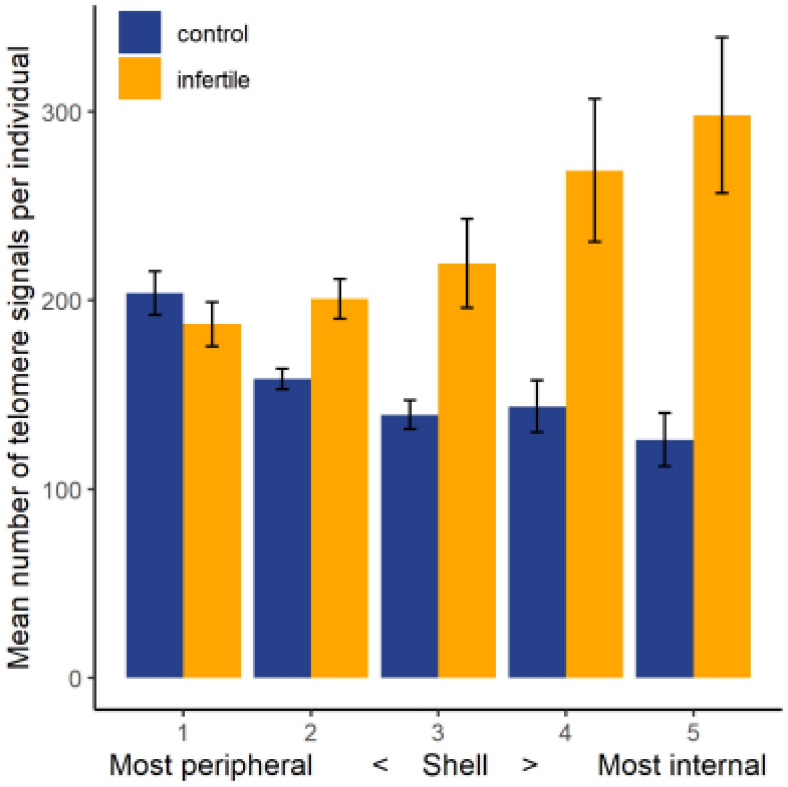


Figure 4. The total number of telomere signals per individual counted in each portion of the nuclear volume for control and infertile males. Shell 1 represents the outermost volume of the nuclei, while shell 5 represents the innermost volume of the nuclei. Error bars represent the standard error of the mean.

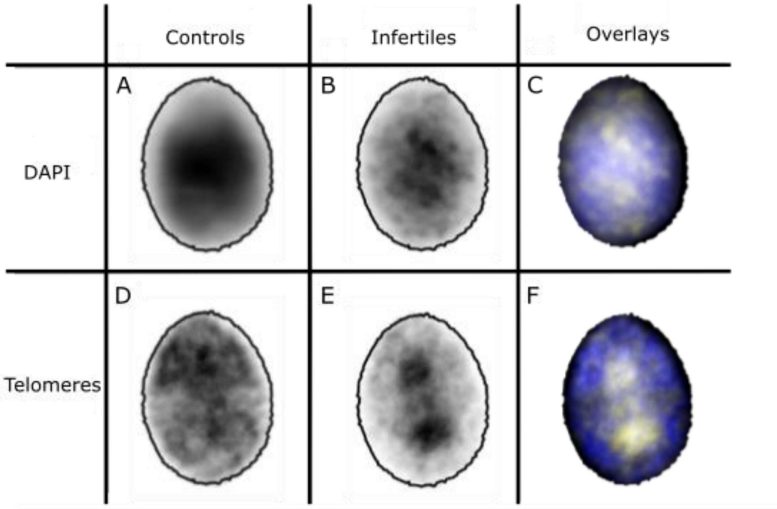


Figure 5. Telomere and DAPI (corresponding to chromatin) distributions in fertile and infertile cohorts. Panel **A**: DAPI heatmap for control cohort warped onto consensus nucleus for control cohort. Panel **B**: DAPI heatmap for infertile cohort warped onto consensus nucleus for control cohort. Panel **C**: Merged heatmaps of DAPI staining for both control (blue) and infertile (yellow) cohorts. Panel **D**: Telomere heatmap for control cohort warped onto consensus nucleus for control cohort. Panel **E**: Telomere heatmap for infertile cohort warped onto consensus nucleus for control cohort. Panel **F**: Merged heatmaps of telomere staining for both control (blue) and infertile (yellow) cohorts.

The difference also seemed to extend to the DAPI staining of chromatin in the nuclei (Figure 5A,B), which indicates that the chromatin distribution in the infertile cohort is more condensed toward the center of the nuclei compared to the controls. This suggests that the telomere distribution reflects a wider change in chromatin organization. This difference may be, in part, due to the larger size of the nuclei in the infertile cohort.

2.3. 3D Distribution of Telomeres

Two-dimensional imaging allowed for the rapid analysis of large numbers of nuclei but was susceptible to over-representing the proportion of internal signals. We performed 3D imaging on a subset of samples to validate our findings. 3D FISH experiments confirmed that telomeres are non-randomly distributed in the sperm heads of both the control and infertile cohorts. Moreover, 3D telomere distribution patterns remained significantly different between the two cohorts. The fractional distance between the center of the nucleus and the telomere signals was significantly shorter in the infertile cohort compared to the control cohort (Mann Whitney U test, $p < 0.001$). Overall, a greater proportion of telomeres occupied more internal regions of the nuclear volume in the sperm heads of the infertile cohort compared to controls (Figure 6). The longer tail on the density plot for the infertile cohort demonstrated a subtle shift in telomere positioning towards more internal regions of the nucleus in the infertile cohort.

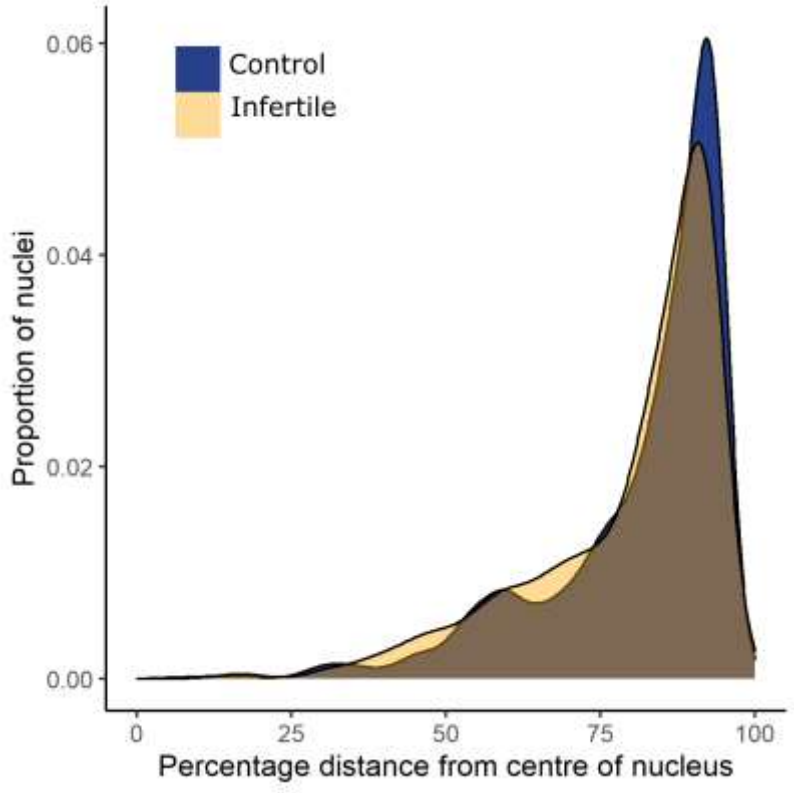


Figure 6. Frequency density of the number of nuclei for each cohort at each

fractional distance from the center of the nucleus. Telomeres in infertile males occupied more internal regions of the nucleus, a subtle but significant ($p < 0.001$) observation.

3. Discussion

This study provides improved understanding of chromatin and telomere organization in the sperm of men with compromised semen parameters. We demonstrated differences in the size and shape of sperm nuclei, in addition to differences in telomere distribution. Notably, the nuclei in the infertile cohort were larger and had a more internal distribution of telomeres.

In previous work, DNA damage in sperm has been linked to increased nuclear vacuole size [37,38], and therefore, abnormal nuclear morphology. This is unlikely, however, to be the sole cause for the greater observed nuclear area in the infertile cohort compared with the controls in this study due to the observed mean distribution of chromatin and telomeres in each cohort, and also, because observation of individual nuclei does not show the presence of oversized vacuoles in the infertile cohort. Nonetheless, this known link supports the idea of disaggregation of chromatin in sperm with a high level of damaged DNA, which fits with the findings of altered telomere distribution in our infertile cohort.

The strict order of chromosome territories within the somatic cell nucleus to form a functional nuclear landscape is strongly associated with normal gene expression patterns, which is, in turn, implicated in normal cellular function (Cremer and Cremer, 2001; Wiblin *et al.*, 2005; Voldgorn *et al.*, 2015). Thus, unsurprisingly, there is evidence to show that disruption in the organization of the genome within the nucleus of specific cell types leads to the pathogenesis of several disease

conditions. The most well-cited examples that demonstrate this phenomenon are illustrated in cases of laminopathies and certain types of cancers (Kuroda *et al.*, 2004; Mewborn *et al.*, 2010). However, whether nuclear organization forms a functional aspect in ensuring the reproductive potential of the gamete remains unclear. Two previous papers by our own group have investigated centromere localization in the sperm nuclei of infertile vs. fertile males; however, these present seemingly conflicting results. While the first shows a significant association between male infertility and sex chromosome positioning (K A Finch *et al.*, 2008), the second concludes that centromere distribution is largely stable despite impaired spermatogenesis (K A Finch *et al.*, 2008; Ioannou *et al.*, 2011). It was the aim of this study, therefore, to revisit this avenue of investigation by examining telomere distribution.

Consistent with the well-established centromeric model of nuclear organization in the human sperm nucleus, these data demonstrate that telomeres are preferentially distributed at the nuclear periphery. Moreover, the data highlight, for the first time, that telomere distribution is altered in the sperm nuclei of infertile males. While control males showed a highly significant non-random distribution of signals with a tendency towards the nuclear periphery, the sperm nuclei of the infertile cohort demonstrated the opposite telomere distribution pattern. Since telomere attachment to the nuclear periphery forms an anchorage point for chromosome territories, it is plausible that altered nuclear organization would, in turn, alter replication and transcriptional cues that are delivered to the oocyte upon fertilization (Svoboda, 2018; Schulz and Harrison, 2019).

In addition, evidence from experiments in yeast and mouse models show that by interacting with proteins at the nuclear membrane, telomeres are exposed to

epigenetic modifications that result in the silencing of genes nearby (known as the telomere position effect) (Palladino *et al.*, 1993; Baur *et al.*, 2001; Blasco, 2007). Thus, aberrant telomere distribution may alter the epigenetic control of the paternal genome and, in turn, hinder the developmental potential of the fertilized ovum. Indeed, several lines of evidence have demonstrated altered global DNA methylation patterns in the sperm of infertile males (Dada *et al.*, 2012; Aston *et al.*, 2015; Kitamura *et al.*, 2015; Montjean *et al.*, 2015; Urdinguio *et al.*, 2015).

Our results also highlight that telomere interactions appear to be perturbed in a greater proportion of sperm nuclei in the infertile cohort compared to controls. The average number of telomere signals in control males was eight, indicating that up to six telomeres may be clustered together in the native, highly condensed sperm nucleus. Interestingly, this association appears to be disrupted in the sperm nuclei of infertile males since the average number of telomere signals was 11. Although the variability of the infertile cohort was larger, with higher standard deviation of the mean values, it is plausible that this could, in part, be due to the fact that the infertile cohort was smaller in sample size than the control cohort. Moreover, although both groups possessed a subpopulation of sperm nuclei in which a highly dispersed pattern of signals was observed (as represented by more than twice the average number of signals), this subpopulation was significantly larger in the infertile cohort ($p = 0.002$).

While telomere dispersion in the sperm nuclei of male factor infertile patients has not been previously reported, others have shown that an increased number of telomere signals is associated with increased DNA fragmentation (Moskovtsev *et al.*, 2009), a factor known to be associated with male infertility (Evenson *et al.*, 1999; Evenson, Larson and Jost, 2002; Agarwal, 2003; Fernández *et al.*, 2003; Biron-Shental *et al.*,

2018). It is possible that this observation is linked to the fact that many of the sperm samples included in the infertile cohort possessed a high number of sperm with poor morphology. This may be indicative of impaired maturation during spermiogenesis, a process in which telomere distribution patterns are fluid. While telomeres are distributed at the nuclear periphery in spermatocytes, they are redistributed to a more central localization at the round spermatid stage, before finally returning to the nuclear periphery in the elongating spermatid, where they remain in the mature sperm head (Tanemura *et al.*, 2005). Since it is thought that telomere clustering at the nuclear periphery promotes homologous recombination-based mechanisms of telomere lengthening (Wright *et al.*, 1996), the dispersion of telomeres in the sperm nuclei of infertile males may result in the inheritance of shortened telomeres. This has the potential to result in chromosome instability and early developmental arrest in the resultant embryo. Indeed, evidence shows that following fertilization and throughout each cleavage event, telomere length shows a steady decline until the blastocyst stage when the embryonic genome becomes activated and telomerase expression is initiated (Wright, 2001; Turner *et al.*, 2010). It is thought, therefore, that the inheritance of apparently long telomeres from the paternal genome, as identified in several studies, acts as a buffer to this problem (Cooke and Smith, 1986; de Lange *et al.*, 1990; Allsopp *et al.*, 1992; Baird *et al.*, 2006; Kimura *et al.*, 2008). In support of this, several studies have demonstrated telomere shortening in the sperm nuclei of infertile males (for a review, see Vasilopoulos et al., 2019 (Vasilopoulos *et al.*, 2019)).

It has been hypothesized that the arrangement of histone-bound telomeres at the nuclear periphery of the cell provides a functional benefit following fertilization, as the periphery is more easily accessible by the oocyte than the middle of the nucleus

(Fice and Robaire, 2019). Instead of investigating a mechanism that could be causing the differential distribution of telomeres in the infertile cohort, suggesting that a pathway occurs due to infertility, an alternative explanation could be that poor telomere positioning will cause accessibility issues for the oocyte, and this could be the reason why infertility occurs. In this case, whilst telomere distribution would be a marker for infertility, it may not be a downstream effect in an infertile man, but rather a cause of infertility.

Given the 3D nature of the human sperm nucleus, it could be argued that it may be difficult to accurately extrapolate 3D information from a 2D image. Although some signals may appear in the center of the nucleus from a 2D image, it may be that those signals are in fact localized at the periphery at the top or bottom of the nucleus. Indeed, our own 3D data demonstrate a more pronounced preference for peripheral localization of telomeres in both cohorts. This suggests that results from 2D analysis may overestimate the true degree of the differences observed in telomere distribution patterns. However, it must be noted that an overestimation, whilst cause for caution, would have been an effect seen of equal size in both cohorts due to the strictly identical experimental procedures followed in both cohorts. While telomere signals were observed throughout all positions within the nuclear volume in 2D analysis, a preference for peripheral localization was observed.

Our 3D data are in support of other work in which a peripheral distribution of telomeres was demonstrated in the sperm heads of fertile males (Zalensky *et al.*, 1995); however, they are in contradiction of 3D observations made in more recent work by Ioannou et al. (Ioannou *et al.*, 2011), in which telomeres were not preferentially localized at the nuclear periphery. However, in the study by Ioannou et al. (Ioannou *et al.*, 2011), sperm nuclei were treated in order to swell the nuclear

volume in a bid to improve access for the pan-telomere probe. We did not employ any technique designed to swell the nucleus in the current study, so as to preserve the original nuclear architecture, and in order to carry out measurements of nuclear shape and size. We acknowledge, therefore, that it is possible that access for the pan-telomere probe was restricted to those telomeres in the most peripheral regions of the nuclear volume in our study. This may explain the discrepancy in the 3D distribution pattern of telomeres as well as in the total number of telomere signals observed between our own data and that of Ioannou et al. (Ioannou *et al.*, 2011). Similarly, it may explain why the total number of telomere signals in our 2D data is also lower than in other published work, in which nuclear swelling procedures were carried out (Zalensky *et al.*, 1995; Turner and Hartshorne, 2013).

In conclusion, these results suggest that nuclear organization is compromised in the sperm of infertile males, and although it may be difficult to spot subtle differences in the localization of a small number of specific individual sequences, when looking at the distribution of a larger number of telomere signals, a more obvious distinction emerges. It is possible that with the development of a more automated analysis approach, and further, more rigorous investigation, both telomere distribution and the morphological analysis of nuclei may prove a useful investigative tool in the fertility clinic.

4. Materials and Methods

4.1. Semen Samples

Semen samples were obtained from 13 control donors (Bridge Fertility Clinic, London, UK) and six males with compromised semen parameters (Embryogenesis, Athens, Greece) following approval from the University of Kent ethics committee.

Informed consent was obtained from all subjects involved in the study. Semen parameters of control and infertile donors were assessed by trained andrologists at the Bridge Fertility Clinic or Embryogenesis. Those samples with a sperm concentration of ≤20 × 10^6/mL and/or motility ≤20% and/or progressivity ≤5% and/or normal forms ≤10% were included in the infertile cohort since their semen parameters met WHO guidelines for infertility and they had presented at the fertility clinic having been unable to conceive within 12 months. Individual scores for each male included in the infertile cohort can be seen in Table 2. Samples were washed three times in sperm buffer (10 mM NaCl, 10 mM Tris pH 7), fixed in 3:1 methanol: acetic acid and stored at −20 °C until use.

Table 2. Infertile cohort semen parameters assessed by a trained andrologist.

Patient number	Concentration ($\times 10^{6}$/mL)	Motility (%)	Progressive Motility (%)	% Normal Forms
1	20	20	10	5
2	0.8	0	0	5
3	18	20	5	2
4	2	5	0	5
5	2	10	5	8
6	60	45	25	5

4.2. Telomeres and FISH

Telomere distribution was assessed by fluorescence in situ hybridization (FISH) using a biotin-labeled pan-telomere probe (Cambio) following the manufacturer's

protocol. Then, 2D fluorescence microscopy was carried out at magnification of 100× using a BX61 Olympus microscope equipped with a CCD camera (C10600-10B—ORCA-R2, Hamamatsu) and appropriate filters. Images of at least 100 sperm nuclei per patient were captured using SmartCapture3 (Digital Scientific). Next, 3D fluorescence microscopy was carried out using an Olympus IX71 microscope. After identification of the best plane of focus for the center of the nucleus, 61 images were captured in 0.1 μm increments throughout the nucleus (30 images below and 30 above the central plane) using Metamorph software.

Telomeres are the only parts of the sperm nucleus that remain histone bound and not protamine bound (Ichikawa *et al.*, 2015), so the cells were not swollen prior to FISH. This had the benefit of allowing for morphological analysis without extreme distortion of the nuclear shape, while retaining accessibility of the FISH probes to the majority of the telomeres.

4.3. 2D Telomere Distribution in Fertile vs. Infertile Males

In order to compare telomere distribution patterns in fertile and infertile males, the position of each telomere signal within each sperm nucleus was measured by splitting the nucleus into five rings of equal area using a custom-designed macro for the freely available software ImageJ (designed by Michael Ellis, Digital Scientific and previously described by Skinner et al. [36]. The total number of telomere signals in each ring was manually counted in all images for each sample. The average number of signals in each ring for each cohort was then calculated. A Chi^2 test was used to test the hypothesis that telomere distribution pattern is altered in the sperm nuclei of the infertile cohort compared to that of controls.

Telomere and chromatin distribution was also assessed using the morphometric software 'Nuclear Morphology Analysis' v1.20.0 developed by Skinner et al. (Skinner, Rathje, *et al.*, 2019), with nuclei detected via thresholding of the DAPI channel. Image warping was used to project the telomere and DAPI signal from each nucleus onto a consensus shape (Figure 2) [72]. The morphological parameters were also assessed using this software, focusing specifically on the areas, bounding heights and widths of the nuclei, as well as the 'difference from median' of each nucleus in each cohort, to assess the variability of shapes. Statistical analyses were performed in R [73]. Bounding height and width refer to the height and width of the rectangle enclosing each nucleus, after the nucleus has been rotated to place the apical-basal axis vertical.

4.4. 3D Telomere Distribution in Infertile vs. Control Cohorts

To ascertain the 3D localization of telomeres within the human sperm nucleus and to confirm the robustness of our 2D analysis, the 3D positions of telomere signals in the sperm nuclei of six fertile and five infertile males were measured. For each individual, 30–40 nuclei were captured and deconvolved using AutoQuantX3 advanced image deconvolution and visualization software (MediaCybernetics). Telomere position was calculated as the fractional distance of the telomere along a line starting at the center of the nucleus, through the center of the telomere signal, to the border of the nucleus. Fractional distances were grouped into frequency bins of width 0.2 for statistical comparisons. To compare the overall distribution of telomere signals within the sperm nuclei of fertile and infertile males, all data from each cohort were pooled and assessed using a Chi^2 test. In addition, a Mann-Whitney U test was performed to compare the distances between the telomere and the center of the nucleus, and the telomere and the nuclear border.

Author Contributions: Conceptualization, D.K.G. and B.M.S.; methodology, D.K.G. and B.M.S.; software and validation, B.M.S.; investigation, K.J.T. and E.M.W.; data curation and formal analysis, K.J.T. and E.M.W.; visualization, E.M.W.; supervision and project administration, D.K.G. and B.M.S.; writing—original draft, K.J.T. and D.K.G.; writing—review and editing, K.J.T., E.M.W., B.M.S. and D.K.G.; resources, D.K.G. and K.J.T.; funding acquisition, D.K.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a Medical Research Council Case studentship (grant number MRC G1000392) part supported by Digital Scientific UK.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Sciences Faculty Research Ethics Advisory Group (letter dated 9 April 2012, no number).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available since they are images from patient samples.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ray, P.F.; Toure, A.; Metzler-Guillemain, C.; Mitchell, M.J.; Arnoult, C.; Coutton, C. Genetic abnormalities leading to qualitative de-fects of sperm morphology or function. *Clin. Genet.* **2017**, *91*, 217–232.

2. Ribas-Maynou, J.; Benet, J. Single and Double Strand Sperm DNA Damage: Different Reproductive Effects on Male Fertility. *Genes* **2019**, *10*, 105, doi:10.3390/genes10020105.

3. Cho, C.-L.; Agarwal, A. Role of sperm DNA fragmentation in male factor infertility: A systematic review. *Arab. J. Urol.* **2018**, *16*, 21–34, doi:10.1016/j.aju.2017.11.002.

4. Bracke, A.; Peeters, K.; Punjabi, U.; Hoogewijs, D.; Dewilde, S. A search for molecular mechanisms underlying male idiopathic in-fertility. In *Reproductive BioMedicine Online*; Elsevier Ltd: 2018; Volume 36, pp. 327–339.

5. Fice, H.E.; Robaire, B. Telomere Dynamics throughout Spermatogenesis. *Genes* **2019**, *10*, 525, doi:10.3390/genes10070525.

6. Shah, K.; Sivapalan, G.; Gibbons, N.; Tempest, H.; Griffin, D.K. The genetic basis of infertility. *Reproduction* **2003**, *126*, 13–25, doi:10.1530/rep.0.1260013.

7. Griffin, D.K.; A Finch, K. The genetic and cytogenetic basis of male infertility. *Hum. Fertil.* **2005**, *8*, 19–26, doi:10.1080/14647270400016407.

8. Jennings, C.; Powell, D. Genome organisation in the murine sperm nucleus. *Zygote* **1995**, *3*, 123–131, doi:10.1017/s0967199400002495.

9. Zalensky, A.O.; Allen, M.J.; Kobayashi, A.; Zalenskaya, I.A.; Balhorn, R.; Bradbury, E.M. Well-defined genome architecture in the hu-man sperm nucleus. *Chromosoma* **1995**, *103*, 577–590.

10. Zalensky, A.O.; Breneman, J.W.; Zalenskaya, I.A.; Brinkley, B.R.; Bradbury, E.M. Organization of centromeres in the decondensed nuclei of mature human sperm. *Chromosoma* **1993**, *102*, 509–518, doi:10.1007/bf00368344.

11. Solov'eva, L.; Svetlova, M.; Bodinski, D.; Zalensky, A.O. Nature of telomere dimers and chromosome looping in human spermatozoa. *Chromosome Res.* **2004**, *12*, 817–823.

12. Sotolongo, B.; Huang, T.T.F.; Isenberger, E.; Ward, W.S. An Endogenous Nuclease in Hamster, Mouse, and Human Spermatozoa Cleaves DNA into Loop-Sized Fragments. *J. Androl.* **2005**, *26*, 272–280, doi:10.1002/j.1939-4640.2005.tb01095.x.

13. Ward, W.S.; Kimura, Y.; Yanagimachi, R. An Intact Sperm Nuclear Matrix May Be Necessary for the Mouse Paternal Genome to Participate in Embryonic Development1. *Biol. Reprod.* **1999**, *60*, 702–706, doi:10.1095/biolreprod60.3.702.

14. Zalenskaya, I.A.; Bradbury, E.; Zalensky, A.O. Chromatin Structure of Telomere Domain in Human Sperm. *Biochem. Biophys. Res. Commun.* **2000**, *279*, 213–218, doi:10.1006/bbrc.2000.3917.

15. Lee, C.Y.; Bisig, C.G.; Conrad, M.N.; Ditamo, Y.; de Almeida, L.P.; Dresser, M.E.; Pezza, R.J.; Lee, C.Y.; Bisig, C.G.; Conrad, M.N.; et al. Telomere-led meiotic chromosome movements: Recent update in structure and function. *Nucleus* **2020**, *11*, 111–116, doi:10.1080/19491034.2020.1769456.

16. Rocca, M.S.; Foresta, C.; Ferlin, A. Telomere length: Lights and shadows on their role in human reproduction. *Biol. Reprod.* **2018**, *100*, 305–317, doi:10.1093/biolre/ioy208.

17. Zalenskaya, I.A.; Zalensky, A.O. Non-random positioning of chromosomes in human sperm nuclei. *Chromosome Res.* **2004**, *12*, 163–173, doi:10.1023/b:chro.0000013166.04629.97.

18. Mudrak, O.; Tomilin, N.; Zalensky, A. Chromosome architecture in the decondensing human sperm nucleus. *J. Cell Sci.* **2005**, *118*, 4541–4550, doi:10.1242/jcs.02581.

19. Luetjens, C.M.; Payne, C.; Schatten, G. Non-random chromosome positioning in human sperm and sex chromosome anomalies following intracytoplasmic sperm injection. *Lancet* **1999**, *353*, 1240, doi:10.1016/s0140-6736(99)80059-2.

20. Hazzouri, M.; Rousseaux, S.; Mongelard, F.; Usson, Y.; Pelletier, R.; Faure, A.K.; Vourc'h, C.; Sele, B. Genome organization in the human sperm nu-cleus studied by FISH and confocal microscopy. *Mol. Reprod. Dev.* **2000**, *55*, 307–315.

21. Foster, H.; Abeydeera, L.R.; Griffin, D.K.; Bridger, J. Non-random chromosome positioning in mammalian sperm nuclei, with migration of the sex chromosomes during late spermatogenesis. *J. Cell Sci.* **2005**, *118*, 1811–1820, doi:10.1242/jcs.02301.

22. Pereira, C.D.; Serrano, J.B.; Martins, F.; da Cruz e Silva, O.A.B.; Rebelo, S. Nuclear envelope dynamics during mammalian spermato-genesis: New insights on male fertility. *Biol. Rev.* **2019**, *94*, 1195–1219.

23. Wykes, S.M.; Krawetz, S.A. The Structural Organization of Sperm Chromatin. *J. Biol. Chem.* **2003**, *278*, 29471–29477, doi:10.1074/jbc.m304545200.

24. Finch, K.; Fonseka, K.; Abogrein, A.; Ioannou, D.; Handyside, A.; Thornhill, A.; Hickson, N.; Griffin, D. Nuclear organization in human sperm: Preliminary evidence for altered sex chromosome centromere position in infertile males. *Hum. Reprod.* **2008**, *23*, 1263–1270, doi:10.1093/humrep/den112.

25. Olszewska, M.; Wiland, E.; Kurpisz, M. Positioning of chromosome 15, 18, X and Y centromeres in sperm cells of fertile individuals and infertile patients with increased level of aneuploidy. *Chromosome Res.* **2008**, *16*, 875–890.

26. Perdrix, A.; Rives, N. Motile sperm organelle morphology examination (MSOME) and sperm head vacuoles: State of the art in 2013. *Hum. Reprod. Updat.* **2013**, *19*, 527–541, doi:10.1093/humupd/dmt021.

27. Alladin, N.; Moskovtsev, S.I.; Russell, H.; Kenigsberg, S.; Lulat, A.G.-M.; Librach, C.L. The three-dimensional image analysis of the chromocenter in motile and immotile human sperm. *Syst. Biol. Reprod. Med.* **2013**, *59*, 146–152, doi:10.3109/19396368.2013.772679.

28. Wiland, E.; Olszewska, M.; Kurpisz, M. Interindividual differences and alterations in the topology of chromosomes in human sperm nuclei of fertile donors and carriers of reciprocal translocations. *Chromosome Res.* **2008**, *16*, 291–305, doi:10.1007/s10577-007-1194-2.

29. Ioannou, D.; Griffin, D.K. Nanotechnology and molecular cytogenetics: The future has not yet arrived. *Nano Rev.* **2010**, *1*, doi:10.3402/nano.v1i0.5117.

30. Ellis, P.J.I.; Griffin, D.K. Form from Function, Order from Chaos in Male Germline Chromatin. *Genes* **2020**, *11*, 210, doi:10.3390/genes11020210.

31. Champroux, A.; Damon-Soubeyrand, C.; Goubely, C.; Bravard, S.; Henry-Berger, J.; Guiton, R.; Saez, F.; Drevet, J.; Kocer, A. Nuclear Integrity but Not To-pology of Mouse Sperm Chromosome is Affected by Oxidative DNA Damage. *Genes* **2018**, *9*, 501.

32. Wiland, E.; Fraczek, M.; Olszewska, M.; Kurpisz, M. Topology of chromosome centromeres in human sperm nuclei with high levels of DNA damage. *Sci. Rep.* **2016**, *6*, 31614, doi:10.1038/srep31614.

33. Bisht, S.; Dada, R. Oxidative stress: Major executioner in disease pathology, role in sperm DNA damage and preventive strategies. *Front Biosci.* **2017**, *9*, 420–447.

34. Ioannou, D.; Griffin, D. Male Fertility, Chromosome Abnormalities, and Nuclear Organization. *Cytogenet. Genome Res.* **2010**, *133*, 269–279, doi:10.1159/000322060.

35. Gillies, E.A.; Cannon, R.M.; Green, R.B.; Pacey, A.A. Hydrodynamic propulsion of human sperm. *J. Fluid Mech.* **2009**, *625*, 445, doi:10.1017/s0022112008005685.

36. Skinner, B.M.; Rathje, C.C.; Bacon, J.; Johnson, E.E.P.; Larson, E.L.; Kopania, E.E.K; Good, J.M.; Yousafzai, G.; Affara, N.A.; Ellis, P.J.I. A high-throughput method for unbiased quantitation and categorization of nuclear morphology. *Biol. Reprod.* **2019**, *100*, 1250–1260.

37. Franco, J.G.; Baruffi, R.L.R.; Mauri, A.L.; Petersen, C.G.; A Oliveira, J.B.; Vagnini, L. Significance of large nuclear vacuoles in human spermatozoa: Implications for ICSI. *Reprod. Biomed. Online* **2008**, *17*, 42–45, doi:10.1016/s1472-6483(10)60291-x.

38. Cremer, T.; Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.* **2001**, *2*, 292–301, doi:10.1038/35066075.

39. Voldgorn, Y.I.; Adilgereeva, E.P.; Nekrasov, E.D.; Lavrov, A.V. Cultivation and Differentiation Change Nuclear Localization of Chromosome Centromeres in Human Mesenchymal Stem Cells. *PLoS ONE* **2015**, *10*, e0118350, doi:10.1371/journal.pone.0118350.

40. Wiblin, A.E.; Cui, W.; Clark, J.A.; Bickmore, W.A. Distinctive nuclear organisation of centromeres and regions involved in pluripo-tency in human embryonic stem cells. *J. Cell Sci.* **2005**, *118*, 3861–3868.

41. Kuroda, M.; Tanabe, H.; Yoshida, K.; Oikawa, K.; Saito, A.; Kiyuna, T.; Mizusawa, H.; Mukai, K. Alteration of chromosome positioning during adipocyte differentiation. *J. Cell Sci.* **2004**, *117*, 5897–5903, doi:10.1242/jcs.01508.

42. Mewborn, S.K.; Puckelwartz, M.J.; Abuisneineh, F.; Fahrenbach, J.P.; Zhang, Y.; MacLeod, H.; Dellefave, L.; Pytel, P.; Selig, S.; Labno, C.M.; et al. Altered Chromosomal Positioning, Compaction, and Gene Expression with a Lamin A/C Gene Mutation. *PLoS ONE* **2010**, *5*, e14342, doi:10.1371/journal.pone.0014342.

43. Ioannou, D.; Meershoek, E.J.; Christopikou, D.; Ellis, M.; Thornhill, A.R.; Griffin, D.K. Nuclear organisation of sperm remains remarkably unaffected in the presence of defective spermatogenesis. *Chromosome Res.* **2011**, *19*, 741–753, doi:10.1007/s10577-011-9238-z.

44. Schulz, K.N.; Harrison, M.M. Mechanisms regulating zygotic genome activation. *Nat. Rev. Genet.* **2019**, *20*, 221–234, doi:10.1038/s41576-018-0087-x.

45. Svoboda, P. Mammalian zygotic genome activation. *Semin. Cell Dev. Biol.* **2018**, *84*, 118–126, doi:10.1016/j.semcdb.2017.12.006.

46. Baur, J.; Zou, Y.; Shay, J.W.; Wright, W.E. Telomere Position Effect in Human Cells. *Science* **2001**, *292*, 2075–2077, doi:10.1126/science.1062329.

47. Blasco, M.A. The epigenetic regulation of mammalian telomeres. *Nat. Rev. Genet.* **2007**, *8*, 299–309, doi:10.1038/nrg2047.

48. Palladino, F.; Laroche, T.; Gilson, E.; Axelrod, A.; Pillus, L.; Gasser, S. SIR3 and SIR4 proteins are required for the positioning and integrity of yeast telomeres. *Cell* **1993**, *75*, 543–555, doi:10.1016/0092-8674(93)90388-7.

49. Dada, R.; Kumar, M.; Jesudasan, R.; Fernández, J.L.; Gosálvez, J.; Agarwal, A. Epigenetics and its role in male infertility. *J. Assist. Reprod. Genet.* **2012**, *29*, 213–223, doi:10.1007/s10815-012-9715-0.

50. Kitamura, A.; Miyauchi, N.; Hamada, H.; Hiura, H.; Chiba, H.; Okae, H.; Sato, A.; John, R.; Arima, T. Epigenetic alterations in sperm associated with male infertility. *Congenit. Anomalies* **2015**, *55*, 133–144, doi:10.1111/cga.12113.

51. Urdinguio, R.G.; Bayón, G.F.; Dmitrijeva, M.; Toraño, E.G.; Bravo, C.; Fraga, M.F.; Bassas, L.; Larriba, S.; Fernández, A.F. Aberrant DNA methylation patterns of sper-matozoa in men with unexplained infertility. *Hum. Reprod.* **2015**, *30*, 1014–1028.

52. Aston, K.I.; Uren, P.J.; Jenkins, T.G.; Horsager, A.; Cairns, B.R.; Smith, A.D.; Carrell, D.T. Aberrant sperm DNA methylation predicts male fertil-ity status and embryo quality. *Fertil. Steril.* **2015**, *104*, 1388–1397.e5.

53. Montjean, D.; Zini, A.; Ravel, C.; Belloc, S.; Dalleac, A.; Copin, H.; Boyer, P.; McElreavey, K.; Benkhalifa, M. Sperm global DNA methylation level:

Association with semen parameters and genome integrity. *Androl.* **2015**, *3*, 235–240, doi:10.1111/andr.12001.

54. Moskovtsev, S.I.; Willis, J.; White, J.; Mullen, J.B.M. Sperm DNA Damage: Correlation to Severity of Semen Abnormalities. *Urology* **2009**, *74*, 789–793.

55. Agarwal, A. Role of sperm chromatin abnormalities and DNA damage in male infertility. *Hum. Reprod. Updat.* **2003**, *9*, 331–345, doi:10.1093/humupd/dmg027.

56. Evenson, D.P.; Jost, L.K.; Marshall, D.; Zinaman, M.J.; Clegg, E.; Purvis, K.; De Angelis, P.; Claussen, O.P. Utility of the sperm chromatin structure assay as a diagnostic and prognostic tool in the human fertility clinic. *Hum. Reprod.* **1999**, *14*, 1039–1049, doi:10.1093/humrep/14.4.1039.

57. Evenson, D.P.; Larson, K.L.; Jost, L.K. Sperm Chromatin Structure Assay: Its Clinical Use for Detecting Sperm DNA Fragmentation in Male Infertility and Comparisons with Other Techniques. *J. Androl.* **2002**, *23*, 25–43, doi:10.1002/j.1939-4640.2002.tb02599.x.

58. Biron-Shental, T.; Wiser, A.; Hershko-Klement, A.; Markovitch, O.; Amiel, A.; Berkovitch, A. Sub-fertile sperm cells exemplify telo-mere dysfunction. *J. Assist. Reprod. Genet.* **2018**, *35*, 143–148.

59. Fernández, J.L.; Muriel, L.; Rivero, M.T.; Goyanes, V.; Vazquez, R.; Alvarez, J.G. The sperm chromatin dispersion test: A simple method for the determination of sperm DNA fragmentation. *J. Androl.* **2003**, *24*, 59–66.

60. Tanemura, K.; Ogura, A.; Cheong, C.; Gotoh, H.; Matsumoto, K.; Sato, E.; et al. Dynamic rearrangement of telomeres during sper-matogenesis in mice. *Dev. Biol.* **2005**, *281*, 196–207.

61. Wright, W.E.; Piatyszek, M.A.; Rainey, W.E.; Byrd, W.; Shay, J.W. Telomerase activity in human germline and embryonic tissues and cells. *Dev. Genet.* **1996**, *18*, 173–179.

62. Turner, S.; Wong, H.P.; Rai, J.; Hartshorne, G.M. Telomere lengths in human oocytes, cleavage stage embryos and blastocysts. *Mol. Hum. Reprod.* **2010**, *16*, 685–694, doi:10.1093/molehr/gaq048.

63. Wright, D.L. Characterization of telomerase activity in the human oocyte and preimplantation embryo. *Mol. Hum. Reprod.* **2001**, *7*, 947–955.

64. Allsopp, R.C.; Vaziri, H.; Patterson, C.; Goldstein, S.; Younglai, E.V.; Futcher, A.B.; Greider, C.W.; Harley, C.B. Telomere length predicts replicative capacity of human fibroblasts. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 10114–10118, doi:10.1073/pnas.89.21.10114.

65. Baird, D.M.; Britt-Compton, B.; Rowson, J.; Amso, N.; Gregory, L.; Kipling, D. Telomere instability in the male germline. *Hum. Mol. Genet.* **2005**, *15*, 45–51, doi:10.1093/hmg/ddi424.

66. Cooke, H.; Smith, B. Variability at the Telomeres of the Human X/Y Pseudoautosomal Region. *Cold Spring Harb. Symp. Quant. Biol.* **1986**, *51*, 213–219, doi:10.1101/sqb.1986.051.01.026.

67. de Lange, T.; Shiue, L.; Myers, R.M.; Cox, D.R.; Naylor, S.L.; Killery, A.M.; Varmus, H.E. Structure and variability of human chromosome ends. *Mol. Cell Biol.* **1990**, *10*, 518–527.

68. Kimura, M.; Cherkas, L.F.; Kato, B.S.; Demissie, S.; Hjelmborg, J.B.; Brimacombe, M.; Cupples, A.; Hunkin, J.L.; Gardner, J.P.; Lu, X.; et al. Offspring's Leukocyte Telomere Length, Paternal Age, and Telomere Elongation in Sperm. *PLoS Genet.* **2008**, 4, e37.

69. Vasilopoulos, E.; Fragkiadaki, P.; Kalliora, C.; Fragou, D.; Docea, A.O.; Vakonaki, E.; Tsoukalas, D.; Calina, D.; Buga, A.M.; Georgiadis, G.; et al. The association of female and male infer-tility with telomere length (Review). *Int. J. Mol. Med.* **2019**, *44*, 375–389.

70. Turner, S.; Hartshorne, G.M. Telomere lengths in human pronuclei, oocytes and spermatozoa. *Mol. Hum. Reprod.* **2013**, *19*, 510–518, doi:10.1093/molehr/gat021.

71. Ichikawa, Y.; Nishimura, Y.; Kurumizaka, H.; Shimizu, M. Nucleosome organization and chromatin dynamics in telomeres. *Biomol. Concepts* **2015**, *6*, 67–75, doi:10.1515/bmc-2014-0035.

72. Skinner, B.M.; Bacon, J.; Rathje, C.C.; Larson, E.L.; Kopania, E.E.K.; Good, J.M.; Affara, N.A.; Ellis, P.J.I. Automated Nuclear Cartography Reveals Conserved Sperm Chromosome Territory Localization across 2 Million Years of Mouse Evolution. *Genes* **2019**, *10*, 109.

73. R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

9.4 Appendix 4: Published peer-reviewed paper

The contribution of sex chromosome conflict to disrupted spermatogenesis in hybrid house mice

Emily E. K. Kopania*,[1], Eleanor M. Watson[‡], Claudia C. Rathje[§], Benjamin M. Skinner[‡], Peter J. I. Ellis[§], Erica L. Larson[†,1,2], Jeffrey M. Good*,[1,2]

*Division of Biological Sciences, University of Montana, Missoula, MT 59812

[‡]School of Life Sciences, University of Essex, Colchester, CO4 3SQ, U.K.

[§]School of Biosciences, University of Kent, Canterbury, CT2 7NJ, U.K

[†]Department of Biological Sciences, University of Denver, Denver, CO 80208

[1]Corresponding authors: emily.kopania@umconnect.umt.edu, erica.larson@du.edu, jeffrey.good@umontana.edu

[2]Co-senior authors

ORCID ID: 0000-0002-2710-2491 (EEKK)

ORCID ID: 0000-0002-1290-4360 (EMW)

ORCID ID: 0000-0002-7152-1167 (BMS)

ORCID ID: 0000-0001-9709-7934 (PJIE)

ORCID ID: 0000-0003-3006-645X (ELL)

ORCID ID: 0000-0003-0707-5374 (JMG)

Running Title: Sex Chromosomes and Hybrid Sterility

Keywords: speciation, intragenomic conflict, ampliconic genes, hybrid male sterility, testis expression, FACS, sex chromosomes

Corresponding Authors:

Emily E. K. Kopania

Division of Biological Sciences

University of Montana

Missoula, MT 59812

Email: emily.kopania@umconnect.umt.edu

Erica L. Larson

Department of Biological Sciences

University of Denver

Denver, CO 80208

Email: erica.larson@du.edu

Jeffrey M. Good

Division of Biological Sciences

University of Montana

Missoula, MT 59812

Phone: (406) 243 - 5122

Email: jeffrey.good@umontana.edu

Abstract

Incompatibilities on the sex chromosomes are important in the evolution of hybrid male sterility, but the evolutionary forces underlying this phenomenon are unclear. House mice (*Mus musculus*) lineages have provided powerful models for understanding the genetic basis of hybrid male sterility. X chromosome-autosome interactions cause strong incompatibilities in *Mus musculus* F1 hybrids, but variation in sterility phenotypes suggests a more complex genetic basis. Additionally, X-Y chromosome conflict has resulted in rapid expansions of ampliconic genes with dosage-dependent expression that is essential to spermatogenesis. Here we evaluated the contribution of X-Y lineage mismatch to male fertility and stage-specific gene expression in hybrid mice. We performed backcrosses between two house mouse subspecies to generate reciprocal Y-introgression strains and used these strains to test the effects of X-Y mismatch in hybrids. Our transcriptome analyses of sorted spermatid cells revealed widespread overexpression of the X chromosome in sterile F1 hybrids independent of Y chromosome subspecies origin. Thus, postmeiotic overexpression of the X chromosome in sterile F1 mouse hybrids is likely a downstream consequence of disrupted meiotic X-inactivation rather than X-Y gene copy number imbalance. Y-chromosome introgression did result in subfertility phenotypes and disrupted expression of several autosomal genes in mice with an otherwise non-hybrid genomic background, suggesting that Y-linked incompatibilities contribute to reproductive barriers, but likely not as a direct consequence of X-Y conflict. Collectively, these findings suggest that rapid sex chromosome gene family evolution driven by genomic conflict has not resulted in strong male reproductive barriers between these subspecies of house mice.

Introduction

Sex chromosomes are often involved in the evolution of reproductive isolation between animal species (Coyne and Orr 1989; Turelli and Orr 2000; Presgraves and Meiklejohn 2021), with hybrid sterility or inviability arising more often in the heterogametic sex (i.e., Haldane's Rule, Haldane 1922; Coyne and Orr 2004). Hybrid incompatibilities also tend to accumulate more rapidly on the X chromosome (Masly and Presgraves 2007), which is referred to as the large X-effect (Coyne and Orr 1989). Known as the two rules of speciation (Coyne and Orr 1989; Coyne and Orr 2004), these patterns have been supported across diverse taxa (Good *et al.* 2008a; Davis *et al.* 2015; Bi *et al.* 2019; Matute and Cooper 2021; Presgraves and

Meiklejohn 2021) and undoubtedly drive the early stages of intrinsic reproductive isolation in many systems. Both Haldane's rule and the large-X effect appear particularly strong when considering hybrid male sterility in XY systems, suggesting an important role for X chromosome evolution in both speciation and the evolution of spermatogenesis. However, it remains unclear to what extent these general patterns reflect common evolutionary processes, functional mechanisms unique to sex chromosomes, or a mixture of both (Meiklejohn and Tao 2010).

Intrinsic reproductive barriers between nascent species often arise as an indirect consequence of rapid evolution within populations (Dobzhansky 1937; Coyne and Orr 2004; Coughlan and Matute 2020), so the outsized contribution of sex chromosomes to male sterility may be an inevitable consequence of rapid molecular evolution on the X and Y chromosomes. For example, recurrent genomic conflict is thought to be rampant on the X and Y chromosomes because selfish genetic elements are more likely to arise on sex chromosomes (i.e., meiotic drive *sensu lato;* Frank 1991; Hurst and Pomiankowski 1991; Meiklejohn and Tao 2010; Lindholm *et al.* 2016). Hemizygosity of the X chromosome is also expected to promote more rapid adaptive molecular evolution relative to the autosomes across a broad range of conditions (i.e., the faster-X effect; Charlesworth *et al.* 1987; Vicoso and Charlesworth 2009). Note that hemizygosity on the X and Y chromosomes will also result in differential exposure of hybrid incompatibilities on the sex chromosomes in males if incompatibilities tend to be at least partially recessive (Turelli and Orr 1995; Turelli and Orr 2000). However, progress on understanding how often these diverse evolutionary processes contribute to the evolution of hybrid male sterility has been hampered by a lack of data on the genetic underpinnings of reproductive isolation.

From a mechanistic perspective, the X and Y chromosomes are also subject to unique regulatory processes during mammalian spermatogenesis that are critical for normal male fertility and shape patterns of molecular evolution (Larson *et al.* 2018a). Both the X and Y chromosomes are packaged into condensed chromatin early in meiosis, resulting in transcriptional silencing of most sex-linked genes known as meiotic sex chromosome inactivation (MSCI; McKee and Handel 1993). Repressive chromatin persists through the postmeiotic stages (Namekawa *et al.* 2006), although many essential X- and Y-linked genes are highly expressed in postmeiotic, haploid round spermatids (Mueller *et al.* 2008; Sin and

Namekawa 2013). Failure to broadly repress X-linked expression during these critical meiotic and postmeiotic stages can trigger spermatogenic disruption, reduced sperm production, and abnormal sperm morphology (Burgoyne *et al.* 2009; Turner 2015). Interestingly, sex chromosome repression during both stages appears prone to disruption in hybrid mammals (Mihola *et al.* 2009; Good *et al.* 2010; Campbell *et al.* 2013; Davis *et al.* 2015; Larson *et al.* 2017), which may reflect common regulatory pathways underlying the evolution of hybrid male sterility (Bhattacharyya *et al.* 2013; Larson *et al.* 2021). Understanding how these intermediate developmental sterility phenotypes relate to genomic conflict and the broader evolutionary dynamics of the sex chromosomes awaits more data.

House mice (*Mus musculus*) have emerged as predominant models for understanding both the basic molecular control of spermatogenesis and the evolution of hybrid male sterility in mammals (Phifer-Rixey and Nachman 2015). Closely related subspecies of mice, *Mus musculus* and *M. m. domesticus* (hereafter, "*musculus*" and "*domesticus*"), readily hybridize in both the lab and along a natural hybrid zone in Europe (Janoušek *et al.* 2012). Hybrid male sterility is the strongest and likely primary reproductive barrier isolating these incipient species in nature (Vyskočilová, et al. 2005; Turner, et al. 2012) and in the lab (Good *et al.* 2008b; Vyskočilová *et al.* 2009) following Haldane's rule (Haldane 1922; but see Suzuki and Nachman 2015). Male sterility is polymorphic with laboratory crosses yielding sterile, subfertile, or fertile male hybrids depending on genotype and cross direction (Good *et al.* 2008b; Balcova *et al.* 2016; Larson *et al.* 2018b; Widmayer *et al.* 2020); *musculus*♀ × *domesticus*♂ crosses usually result in sterile F1 males, while the reciprocal cross tends to be more fertile (Good *et al.* 2008b). This asymmetry is caused by epistatic incompatibilities that are exposed on the *musculus* X chromosome in hybrid males (Storchová *et al.* 2004; Good *et al.* 2008a; Turner and Harr 2014). House mice also remain the only mammalian system where the evolution of a specific gene, *Prdm9*, has been directly linked to the evolution of intrinsic reproductive barriers (Mihola *et al.* 2009; Bhattacharyya *et al.* 2013; Mukaj *et al.* 2020). *Prdm9* is an autosomal gene encoding a DNA-binding protein that directs double stranded breaks where meiotic recombination occurs (Grey *et al.* 2011). PRDM9 binding sites evolve rapidly (Oliver *et al.* 2009; Baker *et al.* 2015), leading to asymmetric binding in hybrid mice that triggers autosomal asynapsis and disruption of MSCI during early pachytene of Meiosis I (Mihola *et al.* 2009; Davies *et al.*

2016). *Prdm9*-related sterility depends on *Prdm9* heterozygosity and epistatic interactions with other unlinked factors, including a major incompatibility locus, *Hstx2*, located near the middle the *musculus* X chromosome (Forejt *et al.* 2021). This same X-linked region also influences hybrid male sterility in backcrossed consomic models (i.e., presumably independent of *Prdm9*; Storchová *et al.* 2004; Good *et al.* 2008a), and recombination rate variation between *M. m. musculus* and another subspecies, *M. m. castaneus* (Dumont and Payseur 2011).

This broad foundation on the genetics of hybrid male sterility provides an opportunity to further unravel the various evolutionary and mechanistic processes that contribute to the large X-effect in mice. *Prdm9*-related sterility plays a central role in the evolution of hybrid male sterility and the disruption of MSCI in F1 mouse hybrids (Forejt *et al.* 2021; Larson *et al.* 2021). However, X- and Y-linked hybrid sterility arises across a broader range of genetic architectures and phenotypes than can be easily ascribed to *Prdm9*-related interactions (Campbell *et al.* 2012; Campbell and Nachman 2014; Larson *et al.* 2018b; Larson *et al.* 2021). The mouse X and Y chromosomes also contain clusters of several high copy ampliconic genes (Mueller *et al.* 2008; Soh *et al.* 2014; Case *et al.* 2015; Morgan and Pardo-Manuel De Villena 2017; Larson *et al.* 2021) that appear to have evolved in response to intense intragenomic conflict (Cocquet *et al.* 2009; Ellis *et al.* 2011; Cocquet *et al.* 2012). These X- and Y-linked gene clusters are primarily expressed in postmeiotic cells with repressed sex chromatin (Namekawa *et al.* 2006; Sin *et al.* 2012) and thus increases in copy number may help counteract repressive chromatin (Ellis *et al.* 2011; Mueller *et al.* 2013; Sin and Namekawa 2013). Conflict arises because the maintenance of repressive postmeiotic sex chromatin appears to be controlled by dosage dependent interactions between X-linked (*Slx* and *Slxl1*) and Y-linked (*Sly*) gene families (Cocquet *et al.* 2012; Kruger *et al.* 2019). Experimental knockdowns of *Slx* and *Slxl1* showed increased sex chromosome repression, abnormal sperm head morphology, and an excess of male offspring. In contrast, knockdowns of *Sly* showed sex chromosome overexpression, abnormal sperm head morphology, and an excess of female offspring (Cocquet *et al.* 2009; Cocquet *et al.* 2012) due to reduced motility of Y-bearing sperm (Rathje *et al.* 2019). CRISPR-based deletions have further shown that sex-ratio distortion is primarily mediated by *Slxl1* versus *Sly* competition for the spindlin proteins (SPIN1, SSTY1/2; Kruger *et al.* 2019).

Copy numbers of *Slx*, *Slxl1,* and *Sly* genes have co-evolved in different mouse lineages (Ellis *et al.* 2011; Good 2012; Morgan and Pardo-Manuel De Villena 2017), such that hybrids could have copy number mismatch sufficient to generate dosage-based sterility phenotypes seen in genetic manipulation studies (Ellis *et al.* 2011). In support of this model, hybrid interactions between the *musculus* X and the *domesticus* Y have been shown to cause abnormal sperm head morphology (Campbell *et al.* 2012; Campbell and Nachman 2014), and male sterility is associated with extensive overexpression of the sex chromosomes in postmeiotic round spermatids in *musculus*♀ × *domesticus*♂ mice (Larson *et al.* 2017). These hybrids have proportionally higher numbers of *Slx* and *Slxl1* relative to *Sly* copies compared to non-hybrids and show patterns qualitatively consistent with the overexpression phenotypes observed in *Sly* knockdown and *Slx/Slxl1* duplication mice (Cocquet *et al.* 2012; Kruger *et al.* 2019). However, postmeiotic sex chromatin repression is thought to partially depend on repressive histone marks established during meiosis (Turner *et al.* 2006), and the same direction of the hybrid cross also shows disrupted MSCI in meiotic spermatocytes (Campbell *et al.* 2013; Larson *et al.* 2017). Thus, it remains unclear if the disruption of repressive postmeiotic chromatin is a consequence of X-Y mismatch or primarily a downstream epigenetic effect of deleterious interactions between the *musculus* X chromosome and *Prdm9* during meiosis (Larson *et al.* 2021).

Here, we advance understanding of the basis of hybrid male sterility in this system using a reciprocal backcrossing scheme to generate mice with the Y chromosome of one *Mus musculus* subspecies on the genomic background of another (Figure 1A). We used these Y-consomic genetic models to perform two reciprocal cross experiments while controlling for the effects of inbreeding. First, we tested for the potential rescue of sterility phenotypes in hybrid males with F1 autosomal genotypes but with matching X and Y chromosomes from the same subspecies (Hybrid F1 XY Match; Figure 1B). This experiment allowed us to tease apart X-Y interactions (i.e., *Slx* and *Slxl1* versus *Sly*) from X-autosomal interactions (i.e., *Prdm9*-related sterility). Second, we tested the effects of X-Y mismatch on different subspecific backgrounds (Non-hybrid XY Mismatch; Figure 1B). This experiment allowed us to test for incompatibilities exposed on introgressed Y chromosomes that occur independently of other hybrid interactions. We used genome sequencing to quantify X- and Y-linked gene copy numbers, quantified male reproductive phenotypes (testis weight and

high-resolution sperm head morphology), and used Fluorescence-Activated Cell Sorting (FACS) to isolate cell populations enriched for either early meiotic leptotene-zygotene spermatocytes or postmeiotic round spermatids. We used these experiments to address three main questions: (i) Does X-Y mismatch cause abnormal male reproductive traits? (ii) Do differences in copy number predict differences in ampliconic gene family expression levels during late spermatogenesis? (iii) Is X-Y mismatch associated with disrupted gene expression during late spermatogenesis, particularly on the sex chromosomes?

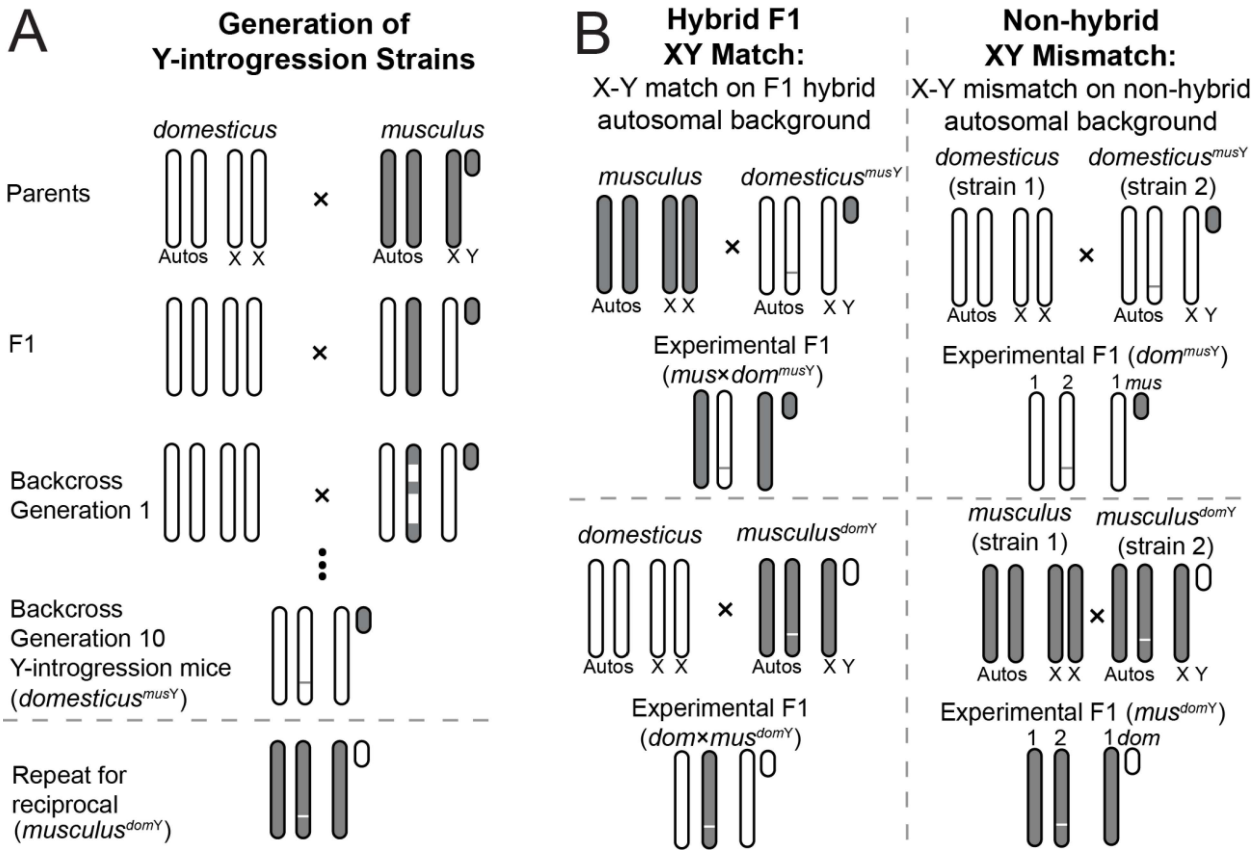


Figure 1: Experimental design. (A) Backcrosses used to generate Y-introgression mouse strains. We performed 10 generations of backcrosses in reciprocal directions to generate mice with a *Mus musculus domesticus* (*domesticus*) genetic background and *Mus musculus (musculus)* Y chromosome (*domesticus*^{musY}) and mice with a *musculus* genetic background and *domesticus* Y chromosome (*musculus*^{domY}). The thin horizontal line on the autosomes represents residual autosomal introgression, which is theoretically expected to represent

about 0.1% of the autosomes. (B) Crosses were performed with Y-introgression mice to produce two types of experimental F1 mice. For Hybrid F1 XY Match, we crossed Y-introgression males to females from the other subspecies to generate F1 mice with hybrid autosomes but matched sex chromosomes. For Non-hybrid XY Mismatch, we crossed Y-introgression males to females from a different strain but the same subspecies to generate F1 mice with X-Y mismatch and non-hybrid autosomes. Autos = autosomes, X = X chromosome, Y = Y chromosome.

Materials and Methods

Mouse resources and experimental design

We sought to test the effects of X-Y mismatch independent of the effects of X-autosomal incompatibilities and inbreeding. To do so, we conducted two experiments: (1) a "Hybrid F1 XY Match" experiment to test if matching the subspecies origin of the X and Y rescued expression and reproductive phenotypes on an otherwise F1 hybrid autosomal background, and (2) a "Non-hybrid XY Mismatch" experiment to test if introgressed X-Y subspecies origin mismatch disrupted expression and reproductive phenotypes on a non-hybrid autosomal background. To breed mice for these experiments, we first generated reciprocal consomic introgression strains with the Y chromosome from one subspecies on the genetic background of the other by backcrossing *musculus* (PWK) and *domesticus* (LEWES) for 10 generations, which we refer to as *musculus*^{domY} and *domesticus*^{musY} (Figure 1A). We tested to ensure our Y-introgression strains had copy number mismatch representative of that expected in natural hybrids. We used publicly available whole genome sequence data to estimate copy number in wild house mice (PRJEB9450 for *domesticus,* 14 males and 1 female, Pezer *et al.* 2015; PRJEB11742 for *musculus,* 5 males and 11 females, Harr *et al.* 2016) and wild-derived inbred laboratory mouse strains representing *musculus* (PWK/PhJ and CZECHII/EiJ) and *domesticus* (LEWES/EiJ and WSB/EiJ; PRJNA732719; one male individual per strain; Larson *et al.* 2021). We then used these Y-introgression strains to perform two experiments and test the effects of X-Y mismatch on hybrid sterility independent of X-autosomal incompatibilities (Figure 1B).

Experiment 1, Hybrid F1 XY Match: To test the effects of X-autosomal F1 incompatibilities without the effect of sex chromosome mismatch, we crossed Y-introgression males to females with the same autosomal and X chromosome type as the male Y chromosome (LEWES or PWK). This generated mice with an F1 hybrid autosomal background and X-autosomal mismatch but X and Y chromosomes from the same subspecies. Throughout the text, we refer to these mice as *mus*×*dom*$^{mus\text{Y}}$ and *dom*×*mus*$^{dom\text{Y}}$. We compared these mice to standard F1 hybrid mice with the same X chromosome and autosomal background but no Y chromosome introgression (PWK♀ × LEWES♂, hereafter "*mus*×*dom*" and LEWES♀ × PWK♂, hereafter "*dom*×*mus*").

Experiment 2, Non-hybrid XY Mismatch: To test the effects of X-Y mismatch while controlling for inbreeding effects, we crossed Y-introgression males to females from the same subspecies but a different strain from the genomic background of the Y-introgression strain (CZECHII or WSB). This generated mice with a non-hybrid (intrasubspecific) F1 autosomal background and mismatched sex chromosomes (i.e., no X-autosomal mismatch), which we will refer to as *mus*$^{dom\text{Y}}$ and *dom*$^{mus\text{Y}}$. We compared these to intrasubspecific F1 mice with the same autosomal background as these F1 Y-introgression mice, but without sex chromosome mismatch (CZECHII♀ × PWK♂, hereafter "*mus*" and WSB♀ × LEWES♂, hereafter "*dom*"). Note that these Non-hybrid XY Mismatch mice had X chromosomes from different laboratory strains than the Hybrid F1 XY Match mice of the same subspecies as a necessary consequence of breeding mice with a heterozygous F1 background.

All mice from wild-derived inbred strains, Y-introgression strains, and experimental crosses were maintained in breeding colonies at the University of Montana (UM) Department of Laboratory Animal Resources (IACUC protocols 002-13, 050-15, and 062-18), which were initially purchased from The Jackson Laboratory, Bar Harbor, ME in 2010. Replacement stock of LEWES/EiJ mice were ordered in 2013, and these mice were used for the backcrosses to generate the *dom*$^{mus\text{Y}}$ Y-introgression strains, as dames in the *dom* intrasubspecific F1s, and as sires in the *dom*×*mus* and *dom*×*mus*$^{dom\text{Y}}$ crosses.

Whole genome sequencing and copy number estimation

We sequenced whole genomes from one male mouse of each Y-introgression strain to estimate ampliconic gene family copy numbers. We extracted DNA from mouse liver using a Qiagen DNeasy kit and sent samples to Novogene (Novogene Corporation Inc., Sacramento, California) for library preparation and sequencing using Illumina HiSeq paired-end 150bp. Libraries were prepared and sequenced twice to increase unique coverage. We trimmed raw reads with Trimmomatic version 0.39 (Bolger *et al.* 2014). We mapped reads to the mouse reference genome build GRCm38 using bwa mem version 0.7.17 (Li and Durbin 2009) and used picard version 2.18.29 to fix mates and mark duplicates (Picard Toolkit). Data from the two sequencing runs were then merged for each sample.

To identify paralogs of ampliconic gene families, we extracted known X (*Slx*, *Slxl1*, *Sstx*), Y (*Sly, Ssty1, Ssty2*), and autosomal (*Speer,* and α*-takusan*) ampliconic gene sequences from the mouse reference GRCm38 using Ensembl annotation version 102 (Yates *et al.* 2019). We used the predicted gene *Gm5926* for *Sstx* because *Sstx* was not annotated in this version of Ensembl. For the autosomal gene families, we used the longest annotated genes in the gene family (α7*-takusan* and *Speer4f2*). We performed Ensembl BLAT searches with these sequences against the GRCm38 mouse reference, allowing up to 1000 hits. We then extracted all BLAT hits with greater than or equal to 97% sequence identity and an e-value of 0.0 and considered these filtered BLAT hits to be gene family paralogs for downstream copy number estimation.

We estimated copy numbers using a relative coverage approach similar to (Morgan and Pardo-Manuel De Villena 2017) and AmpliCoNE (Vegesna *et al.* 2020). For the relative coverage approach, we used Mosdepth v0.3.2 (Pedersen and Quinlan 2017) to estimate coverage across paralogous regions and divided this sum by half the genome-wide average coverage to account for hemizygosity of the sex chromosomes in males.

AmpliCoNE also estimates copy number based on relative coverage, while also controlling for GC content and only using informative regions based on repeat masking and mappability. AmpliCoNE was developed for estimating copy number on the

assembly and annotation of the human Y, so we made some modifications to allow AmpliCoNE to work with the mouse sex chromosomes (Larson *et al.* 2021); https://github.com/ekopania/modified-AmpliCoNE). Specifically, we replaced AmpliCoNE's method for identifying informative sites with an approach more suitable for the mouse assembly. For each ampliconic gene family, we extracted all k-mers of length 101bp from the sequence of one gene representing the ampliconic family and mapped these back to the mouse reference genome using Bowtie2 and allowing up to 500 multiple mapping hits. For each gene, we identified the most frequent number of times (*m*) k-mers mapped to the mouse genome and kept only k-mers that mapped *m* times. We identified all locations where these k-mers mapped with 2 or fewer mismatches. We considered the start locations of these k-mer mapping hits to be "informative sites."

A small amount of autosomal material (~0.1%) is expected to have introgressed along with the Y chromosome in our backcross experiments. To test this theoretical expectation and identify regions of introgression, we mapped whole genome sequence data from Y-introgression strains to both parental genomes using bwa mem v0.7.17-r1188 (Li and Durbin 2009) and called variants with GATK HaplotypeCaller v4.2.2.0. We then counted the number of variants in 100kb windows across the autosomes and identified regions where the number of variants when mapped to the maternal parent (autosomal background) genome exceeded the number of variants when mapped to the paternal parent (Y-introgression) genome. We repeated this analysis using whole genome sequence data from PWK and LEWES samples in our mouse colony. We excluded regions that had more variants when mapped to the opposite strain than when mapped to the same strain, as these are likely regions where genotype calls are unreliable due to assembly issues. After excluding these regions, 100kb windows with at least two more variants when mapped to the maternal parent compared to the paternal parent were considered introgressed in Y-introgression strains, reflecting the 95th percentile of differences in the number of variants within a window.

Reproductive phenotypes

We phenotyped unmated male mice that were weaned at 21 days post-partum (dpp) into same-sex sibling groups and housed individually starting at 45 dpp to minimize effects of social dominance. Phenotypes were collected from at least six individuals for each cross type; sample sizes for each phenotype and cross type are in Table 1. We weighed paired testes and paired seminal vesicles and calculated their mass relative to body weight. We compared offspring sex ratios from Y-introgression mice by recording the number of offspring of each sex at weaning. We then tested for a significant difference from an even sex ratio using a Pearson's chi-squared test in R, and did a power analysis for this chi-squared test using the *pwr.chisq.test* function in the pwr package in R.

To quantify sperm morphology, we extracted sperm from each cross type from cauda epididymides diced in 1mL Dulbecco's PBS (Sigma) and incubated at 37℃ for 10 minutes. Sperm were fixed in 2% PFA, then dropped onto a slide with DAPI solution to stain the sperm nuclei. We imaged greater than 400 nuclei per genotype and analyzed the images using the Nuclear Morphology Analysis software (Skinner *et al.* 2019). We used two microscopes but performed clustering analysis on combined nuclei imaged from both microscopes to ensure that nuclei imaged on one scope were not clustering separately from those taken on the other microscope (Supplemental Material, Figure S1). The Nuclear Morphology Analysis software uses a Canny edge detection algorithm to detect objects (nuclei) within images, orients and aligns the nuclei, and uses a modification of the Zahn-Roskies transformation of the nucleus outlines to automatically detect landmarks. The software estimates area, perimeter, bounding height, bounding width, regularity, difference from median, and a consensus shape of the nuclei for each genotype. We tested for significant differences among cross types for each of these parameters using a Wilcoxon rank sum test in R. Using this automated morphology analysis software, we were able to analyze 5652 nuclei and detect subtle but significant differences that may not be measurable by eye or qualitative analysis.

Testis sorting and RNA sequencing

We collected testes from mice immediately following euthanization and isolated cells at different stages of spermatogenesis using Fluorescence-Activated Cell Sorting (FACS; Getun *et al.* 2011). The full FACS protocol is available on GitHub (https://github.com/goodest-goodlab/good-protocols/tree/main/protocols/FACS). Briefly, we decapsulated testes and washed them twice with 1mg/mL collagenase (Worthington Biochemical), 0.004mg/mL DNase I (Qiagen), and GBSS (Sigma), followed by disassociation with 1mg/mL trypsin (Worthington Biochemical) and 0.004mg/mL DNase I. We then inactivated trypsin with 0.16mg/mL fetal calf serum (Sigma). For each wash and disassociation step, we incubated and agitated samples at 33°C for 15 minutes on a SciGene Model 700 Microarray Oven at approximately 10rpm. We stained cells with 0.36mg/mL Hoechst 33324 (Invitrogen) and 0.002mg/mL propidium iodide and filtered with a 40µm cell filter. For Hybrid F1 XY Match, we sorted using a FACSAria Fusion flow cytometer, and for Non-hybrid XY Mismatch we sorted cells using a FACSAria IIu cell sorter (BD Biosciences), both at the UM Center for Environmental Health Sciences Fluorescence Cytometry Core. We periodically added 0.004mg/mL DNase I as needed during sorting to prevent DNA clumps from clogging the sorter. We sorted cells into 15µL beta-mercaptoethanol (Sigma) per 1mL of RLT lysis buffer (Qiagen) and kept samples on ice whenever they were not in the incubator or the cell sorter. We performed cell sorting on four individuals of each cross type and focused on two cell populations: early meiotic spermatocytes (leptotene/zygotene) and postmeiotic round spermatids. We extracted RNA using the Qiagen RNeasy Blood and Tissue Kit and checked RNA integrity with a TapeStation 2200 (Agilent). Only two samples had RNA integrity numbers (RIN) less than 8 (RIN = 7 and 7.1; Supplemental Material, Table S1). We prepared RNAseq libraries using the KAPA mRNA hyperprep kit and sequenced samples with Novogene (Illumina NovaSeq6000 PE 150). Samples were prepared and sequenced together, but Hybrid F1 XY Match mice and Non-hybrid XY Mismatch mice were sorted on different FACS machines, so to minimize experimental batch effects we analyzed these two experiments separately unless otherwise noted.

Gene expression analyses

We performed gene expression analyses on FACS expression data representing two cell populations: early meiosis (leptotene-zygotene, hereafter “early”) and postmeiosis (round spermatids, hereafter “late”). For the early cell type, a few samples did not group with others of the same cross type in multidimensional scaling (MDS) plots (Supplemental Material, Figure S2). These samples were likely contaminated with other cell types based on their relative expression levels of cell-type marker genes from *Mus musculus* testes single-cell RNAseq experiments (Supplemental Material, Figure S3; Green *et al.* 2018; Hunnicutt *et al.* 2021), and were therefore removed from expression analyses. Because sex chromosome ampliconic genes are primarily expressed in late spermatogenesis (Mueller *et al.* 2013; Larson *et al.* 2018a), and disrupted sex chromosome expression in hybrid males primarily occurs after the early cell type stage (Larson *et al.* 2017), we focus on data from the late cell type in the main text and report results from the early cell type in the Supplemental Material.

We performed gene expression analyses using mice from both our Hybrid F1 XY Match and Non-hybrid XY Mismatch experiments, and reanalyzed expression data from (Larson *et al.* 2017), which generated spermatogenesis cell-type enriched gene expression data from the same F1 hybrid crosses (PWK$^{♀}$ × LEWES$^{♂}$ and LEWES$^{♀}$ × PWK$^{♂}$) and intrasubspecific F1 crosses (CZECHII$^{♀}$ × PWK$^{♂}$ and WSB$^{♀}$ × LEWES$^{♂}$) used in this study.

We trimmed RNAseq reads using trimmomatic v0.39 (Bolger *et al.* 2014). One sample (PP.LL30.7MLZ) had about an order of magnitude more reads than any other sample (> 900 million raw reads), so we downsampled to the mean number of reads after trimming using fastq-sample version 0.8.3 and verified that reads were properly paired after downsampling using fastq_pair (Edwards and Edwards 2019). We quantified reads using a kmer-based quasi-mapping approach implemented in salmon v1.4.0 (Patro *et al.* 2017) and a salmon index based on the mouse reference transcriptome version GRCm38. We then converted from transcript-level counts to gene-level counts using the R packages tximport 1.14.2 and EnsDb.Mmusculus.v79. We used EdgeR version 3.32.1 to normalize expression data. First, we filtered out genes with low expression by only including genes that had an FPKM > 1 in at least four samples. Then, we normalized expression data following the recommendations in the tximport documentation.

We quantified expression levels of ampliconic gene families by calculating transcripts per million (TPM) for each gene separately then summing TPM values for all paralogs of a gene family (≥97% sequence identity). We used linear mixed-effect models to test if gene family expression level was significantly associated with copy number for *Slx, Slxl1, Sly, Ssty1, Ssty2,* and *α-takusan.* We compared disrupted expression levels on the autosomes, X chromosome, and Y chromosome by subtracting normalized FPKM values in control mice from normalized FPKM values in X-Y mismatch mice and control mice for every gene (Good *et al.* 2010). We then used a Mann-Whitney U test to compare the distribution of normalized FPKM differences among the chromosome types. To identify Differentially Expressed (DE) genes between cross types, we used the likelihood ratio test approach with false-discovery rate (FDR) correction in EdgeR and visualized overlaps in DE genes among cross types using the R package UpSetR (Conway *et al.* 2017). We removed DE genes in autosomal regions we identified as putatively introgressed, because these genes may be DE due to introgressed autosomal variants rather than incompatibilities resulting from mismatching sex chromosomes. For ampliconic genes with high sequence similarity, some reads are expected to map multiply but will only be assigned to one member of the ampliconic gene family. Therefore, individual genes within gene families may sometimes be identified as DE, even though their paralogs are not, due to differences in read assignment across paralogs.

We further investigated genome-wide expression differences among cross types using weighted correlation network analyses (WGCNA; Langfelder and Horvath 2008). We identified correlated expression modules significantly associated with different cross types using a linear model and Tukey's honest significant difference (HSD) test. We used R version 4.0.3 for all statistical tests and to implement all R packages (R Core Team).

Data availability

Whole genome sequence data from Y-introgression strains and RNAseq data from testes cell sort populations are publicly available on the Sequence Read Archive, accession numbers PRJNA816542 (whole genome) and PRJNA816886 (RNAseq). Raw phenotype data are available in the Supplemental Material, Table S2.

Scripts used to modify the AmpliCoNE program for copy number estimation are publicly available at: https://github.com/ekopania/modified-AmpliCoNE. Scripts used for gene expression analyses are available at: https://github.com/ekopania/xy_mismatch_expression_analyses.

Results

Copy Number Imbalance in Y-introgression Mice

We first estimated ampliconic gene family copy numbers in wild mice, wild-derived inbred strains, and Y-introgression mice using whole genome sequencing. The samples that we sequenced had genome-wide average coverages of 10-15×, and samples with publicly available data all had coverage >5×. We found that *musculus* tended to have higher *Slx* and *Sly* copy numbers than *domesticus* (median *Slx* copy number in *musculus*: 62, in *domesticus*: 17, FDR-corrected Wilcoxon rank sum $P < 0.01$; median *Sly* copy number in *musculus*: 226, in *domesticus*: 109, FDR-corrected Wilcoxon rank sum $P < 0.01$), qualitatively consistent with previous studies (Ellis *et al.* 2011; Case *et al.* 2015; Morgan and Pardo-Manuel De Villena 2017; Figure 2A). *Slxl1* copy numbers also tended to be higher in *musculus,* but there was high copy number variation for this gene family in *domesticus* with some samples reaching copy numbers as high as those found in *musculus* (median *Slxl1* copy number in *musculus*: 37, in *domesticus*: 31, FDR-corrected Wilcoxon rank sum $P < 0.01$; Figure 2B). *Slx, Slxl1,* and *Sly* copy numbers for wild-derived inbred strains were representative of those found in wild mice (Figures 2A and 2B; Supplemental Material, Table S3), consistent with previous results (Larson *et al.* 2021). Our Y-introgression mice retained copy numbers similar to those of pure strains with the same X and Y chromosome genotypes, so they had *Slx-Sly* and *Slxl1-Sly* dosage imbalance similar to that expected in natural hybrids (Figures 2A and 2B; Supplemental Material, Table S3).

Additional ampliconic gene families showed copy number differences between *musculus* and *domesticus* that were also represented in our Y-introgression mice. *Sstx* had similar copy numbers in *musculus* and *domesticus,* but its two Y-linked homologs showed differences between subspecies, with *Ssty1* having more copies in *domesticus* and *Ssty2* having more copies in *musculus* (median *Sstx* copy number in *musculus*: 48, in *domesticus*:

39, FDR-corrected Wilcoxon rank sum P = 0.57; median *Ssty1* copy number in *musculus*: 74, in *domesticus*: 139, FDR-corrected Wilcoxon rank sum P < 0.01; median *Ssty2* copy number in *musculus*: 145, in *domesticus*: 92, FDR-corrected Wilcoxon rank sum P < 0.01; Figure 2C, 2D).

We also estimated copy number for *α-takusan* and *Speer*, two autosomal ampliconic gene families thought to be regulated by sex chromosome ampliconic genes (Moretti *et al.* 2020). In both males and females, *α-takusan* showed a high correlation in copy number with *Slx* (r = 0.95; Pearson's correlation P < 0.001), suggesting that it was co-amplified with the *Slx* gene family (Figure 2E). Note that correlation tests were performed without phylogenetic correction, because we wanted to test if gene families were co-amplified regardless of whether this was a result of shared evolutionary history. *Speer* copy number was more difficult to estimate using our approaches due to lower sequence similarity among *Speer* paralogs compared to other ampliconic gene families, but our estimates suggested that *Speer* may also have higher copy number in *musculus* relative to *domesticus* (Supplemental Material, Table S3). To verify our computational copy number estimates, we also performed digital droplet PCR (ddPCR) on a subset of *dom* samples using the *Slxl1* primers from (Kruger *et al.* 2019). We found 15 *Slxl1* copies with ddPCR, consistent with findings in (Kruger *et al.* 2019). While our computational estimates are higher than this, we found similar results if we imposed a stricter cutoff for considering genes paralogs (98-99% sequence identity), likely reflecting a high specificity of the primers we used. We also found similar results using a different computational approach based on relative coverage (Supplemental Material, Table S3; Larson *et al.* 2021).

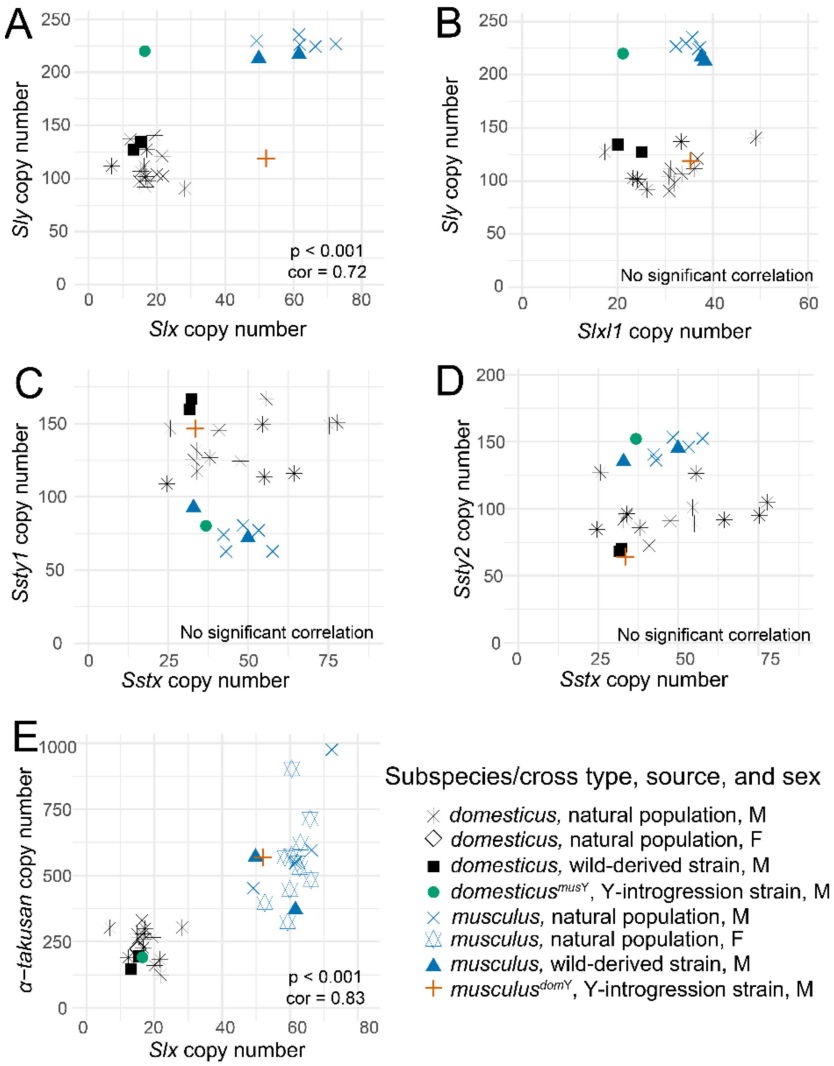


Figure 2: Copy number estimates for ampliconic gene families in wild mice, wild-derived inbred strains, and Y-introgression strains. Copy number was estimated using a 97% identity cutoff for paralogs. (A-D) show copy numbers in male mice, with Y chromosome genes on the y-axis and their X chromosome homologs on the x-axis. (E) includes both males and females and shows haploid copy number for the autosomal gene family *α-takusan* on the y-axis and haploid copy number for the X-linked family *Slx* on the x-axis. Note that (A) and (B) show the same information on the y-axis and (C) and (D) show the same information on the x-axis to compare copy numbers for ampliconic gene families that have two different homologous gene families on the opposite sex chromosome. Correlations and p-values are based on a Pearson's correlation test. P-values were FDR-corrected for multiple tests.

Residual Autosomal Introgression in Y-introgression Strains

We identified putative introgressed regions by mapping samples to both subspecies reference genomes, dividing the reference genome autosomal regions into 24,639 100kb windows, and identifying SNPs in these windows. We found evidence for introgression in 105 windows in *domesticus*^{musY}, and 33 windows in *musculus*^{domY}, representing 0.43% and 0.13% of the autosomal windows that passed filtering, respectively (Supplemental Material, Table S4). Thus, the *domesticus*^{musY} strain had approximately four times more introgression than the theoretical expectation of 0.1% based on the number of backcross generations. The relatively large difference in percentages of introgression between the strains was primarily due to an ~7.6 Mbp introgressed region on chromosome 2 in *domesticus*^{musY} (Supplemental Material, Figure S4). This large introgressed region had an average difference of 958 SNPs, in contrast to the median difference of eight SNPs across all other putatively introgressed autosomal regions. Thus, the introgressed region on chromosome 2 in the *domesticus*^{musY} strain likely represents the only large track of autosomal introgression, with some evidence for additional, smaller amounts of introgression throughout the autosomes in both reciprocal Y-introgression strains.

Some of the putatively introgressed regions we identified may be prone to introgression more generally. The large area on chromosome 2 overlapped with a region with evidence for introgression from *musculus* into the *domesticus* wild-derived inbred strains STRA and STRB (Mukaj *et al.* 2020). We used the Mouse Phylogeny Viewer (Yang *et al.* 2011) to identify an additional nine mouse inbred strains with introgression from *musculus* into a *domesticus* background in this region (Supplemental Material, Figure S4C). In one area of the mouse hybrid zone, a SNP contained within this introgressed region showed evidence for excess of the *musculus* allele in mice with primarily *domesticus* backgrounds, suggesting that introgression of this region from *musculus* into *domesticus* may have occurred in wild populations (Teeter *et al.* 2010). This region is also adjacent to *R2d2*, a copy number variant in mice that shows transmission ratio distortion in females heterozygous for the high copy number *R2d2* drive allele (Didion *et al.* 2016). We also identified 5 different 100kb windows near each other on chromosome 14 with evidence for introgression in *musculus*^{domY} mice that overlap with a region in the *musculus* wild-derived

strain PWD with evidence for introgression from *domesticus* (41.3-41.4Mb, 41.8-41.9Mb, 42.2-42.3Mb, 42.3-43.4Mb, and 44.2-44.3Mb; Mukaj *et al.* 2020).

X-Y Mismatch Contributed to Male Sterility Phenotypes

We next asked if X-Y mismatch was associated with male sterility phenotypes (Table 1). For Hybrid F1 XY Match, where we compared hybrid mice both with and without sex chromosome mismatch, hybrids with a *musculus*♀ × *domesticus*♂ background had lower relative testes mass than hybrids with the reciprocal *domesticus*♀ × *musculus*♂ background regardless of whether they had X-Y mismatch or not (Figure 3A). These results were consistent with previous studies showing more severe hybrid sterility in the *musculus*♀ × *domesticus*♂ direction of this cross (Good *et al.* 2008b; Good *et al.* 2010; Campbell *et al.* 2012; Larson *et al.* 2017). Although *domesticus*♀ × *musculus*♂ showed much less severe sterility phenotypes than the reciprocal F1 hybrid, we still considered these mice to be potentially subfertile because their relative testes mass and sperm morphology parameters were significantly different from those of either pure *dom* or pure *mus* (Figure 3, Table 1), and even subtle reductions in fertility may be important in nature, where sperm competition is high for house mice (Dean *et al.* 2006). For Hybrid F1 XY Match mice, *dom×mus*^{domY} mice had higher relative testis mass than *dom×mus* mice, suggesting that X-Y match partially rescued relative testes mass in some mice with a hybrid autosomal background (Figure 3A). In the reciprocal direction, however, X-Y match had no significant effect on relative testes mass (Figure 3A). For Non-hybrid XY Mismatch, we found that mice with X-Y mismatch had reduced relative testis mass compared to control mice with the same non-hybrid X and autosomal background (Figure 3A). In summary, we found little effect of X-Y mismatch on testis mass in the most sterile F1 cross (*musculus*♀ × *domesticus*♂), where sterility is therefore likely due to X-autosomal or autosomal-autosomal incompatibilities (Campbell and Nachman 2014). However, in the reciprocal and more fertile F1 direction X-Y mismatch seemed to have an important effect on testis mass. Furthermore, in the absence of any autosomal or X-autosomal incompatibilities, X-Y mismatch resulted in slightly but significantly decreased relative testis mass.

We saw severe sperm head abnormalities in our Hybrid F1 XY Match crosses with a *musculus*$^{♀}$ × *domesticus*$^{♂}$ background (*mus×dom* and *mus×dom*^{musY}). Sperm from both these cross types had significantly lower bounding height and bounding width compared to all other cross types (FDR-corrected Wilcoxon rank sum P << 0.0001; Table 1), largely due to their shortened hook and consistent with hybrid sterility in this direction of the cross (Figure 3B, 3C). This was also consistent with previous manual (categorical) observations of abnormal sperm head morphology in this cross type in other studies (Good *et al.* 2008a; Campbell and Nachman 2014; Larson *et al.* 2017; Larson *et al.* 2018b). The reciprocal *dom×mus* F1 hybrids had sperm with higher bounding height and bounding width compared to sperm from all other cross types, including the reference subspecies (FDR-corrected Wilcoxon rank sum P < 0.01; Table 1; Figure 3B, 3C). This direction of the cross is generally considered more fertile but sometimes shows reduced fertility compared to non-hybrid mice (Larson *et al.* 2018b). It is possible that the larger overall size of these sperm may reflect abnormal nuclear packaging and could contribute to reduced fertility in *domesticus*$^{♀}$ × *musculus*$^{♂}$ F1 mice. When comparing X-Y match mice to F1 hybrids with abnormally small sperm heads, *mus×dom*^{musY} mice had significantly higher bounding width and bounding height than *mus×dom* mice (FDR-corrected Wilcoxon rank sum P < 0.01; Table 1; Figure 3B, 3C). These results suggest that X-Y match rescued some of the aberrant sperm head morphology associated with hybrid sterility in *musculus*$^{♀}$ × *domesticus*$^{♂}$ F1s, but the effects of X-Y match rescue were subtle, consistent with previous observations (Campbell and Nachman 2014). In the reciprocal cross direction, *dom×mus*^{domY} had lower bounding width and bounding height than the abnormally large *dom×mus* sperm heads (FDR-corrected Wilcoxon rank sum P << 0.0001; Table 1; Figure 3B, 3), so X-Y match rescued some of the oversized sperm head morphology we observed in *dom×mus*.

In Non-hybrid XY Mismatch, we observed subtle effects of X-Y mismatch consistent with our Hybrid F1 XY Match observations. Sperm from *mus*^{domY} mice had slightly lower bounding height and bounding width compared to sperm from *mus* (FDR-corrected Wilcoxon rank sum P < 0.01; Table 1; Figure 3B, 3C), consistent with lower bounding height and bounding width in sperm from *mus×dom* mice that also had a *mus* X chromosome and *dom* Y chromosome. However, *mus*^{domY} sperm were more similar in size to *mus* sperm than *mus×dom* sperm and qualitatively had a hook morphology more similar to that of fertile

mus than sterile *mus×dom* mice, so the contribution of X-Y mismatch to sperm head morphology is small compared to the effect of X-autosomal interactions. In the reciprocal direction, dom^{musY} mice had sperm with higher bounding height and bounding width compared to sperm from *dom* mice (FDR-corrected Wilcoxon rank sum P << 0.0001; Table 1; Figure 3B, 3C), consistent with the higher bounding height and bounding width in *dom×mus* hybrids. Sperm from dom^{musY} mice also had smaller areas (FDR-corrected Wilcoxon rank sum P << 0.0001; Table 1; Supplemental Material, Figure S5), so the larger bounding height and bounding width are primarily the result of a slightly elongated hook rather than an overall increase in the sperm head size. Other sperm head morphology parameters, including area, perimeter, and differences from median, showed similar subtle differences or no differences among cross types (Table 1; Supplemental Material, Figures S1 and S5).

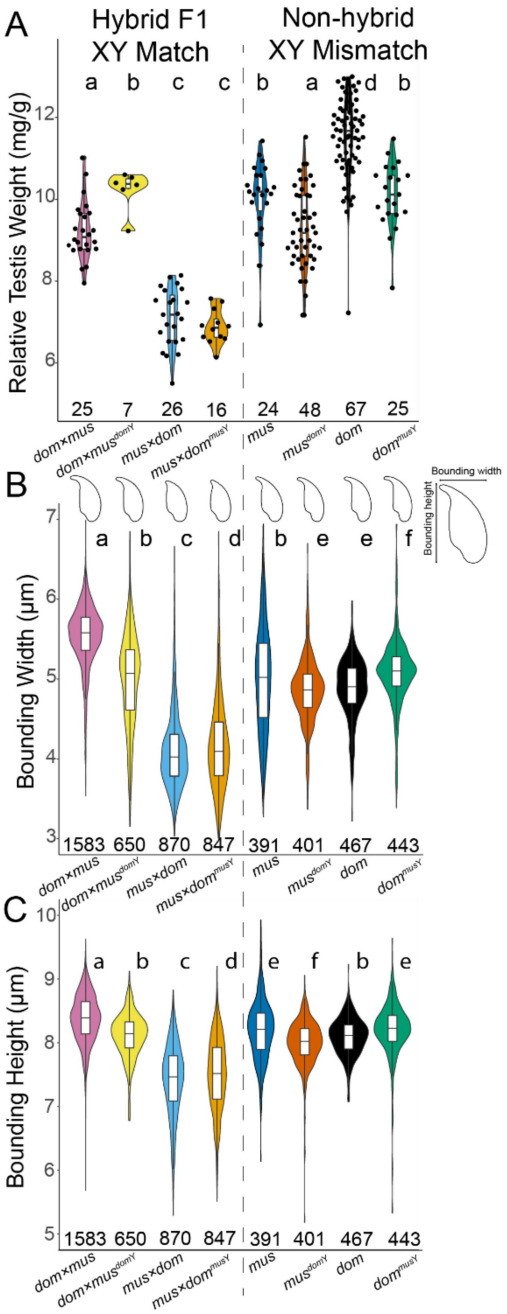


Figure 3: (A) Relative testes mass (mg/g), (B) sperm nucleus bounding width (µm), and (C) sperm nucleus bounding height (µm) by cross type. Letters above each violin plot indicate significant differences (FDR-corrected $P < 0.05$) based on a Welch's t-test (relative testes mass) or Wilcoxon rank-sum test (bounding width and height). Sample size for each cross type is indicated below each violin plot. Bounding width and height sample sizes indicate the number of sperm nuclei observed. Representative sperm nuclei morphologies for each cross type are depicted above each violin plot in (B).

Genetic manipulation studies have shown offspring sex ratio skews under *Slxl1-Sly* dosage imbalance, contributing to evidence for *Slxl1-Sly* intragenomic conflict. Male mice with an excess of *Sly* relative to *Slxl1* produce more male offspring, while mice with an excess of *Slxl1* produce more female offspring (Cocquet *et al.* 2012; Kruger *et al.* 2019) due to reduced motility of Y-bearing sperm (Rathje *et al.* 2019). We asked if more subtle imbalances in relative copy numbers expected in natural hybrid mice also result in sex ratio skews and did not see a significant difference from a 50:50 sex ratio for offspring of X-Y mismatch mice (Supplemental Material, Table S5). A more extreme dosage imbalance than that seen in our X-Y mismatch experimental mice (and in natural hybrids) is probably required to produce a large sex ratio skew. However, it is important to note that we had very little power to detect differences in sex ratio, with type II error probabilities over 0.8 (Supplemental Material, Table S5).

Table 1: Reproductive phenotypes for experimental X-Y mismatch mice and controls. Median values are presented +/- 1 standard error. Sample sizes are in parentheses. For sperm morphology parameters (bounding height, bounding width, area, perimeter, difference from median [a measure of the variability of nuclear shapes within the sample]), sample sizes indicate the number of sperm heads observed, and variance is depicted in violin plots (Figure 3; Supplemental Material, Figure S5). Gray boxes indicate significant differences (FDR-corrected Wilcoxon rank sum test $P < 0.05$) between X-Y mismatch cross types and control cross types with the same autosomal background. (‡) Indicates phenotypes with significant differences (FDR-corrected pairwise Wilcoxon rank sum test $P < 0.05$) between *mus×dom* F1 hybrids and both parental subspecies (*mus* and *dom*). (*) Indicates phenotypes with significant differences (FDR-correct pairwise Wilcoxon rank sum test $P < 0.05$) between *dom×mus* F1 hybrids and both parental subspecies (*mus* and *dom*). Testes and seminal vesicle weights are both paired. SV = seminal vesicle

Phenotype	Hybrid F1 XY Match				Non-hybrid XY Mismatch			
	dom*×*mus	***dom*× *mus*^{domY}**	***mus*×*dom***	***mus*× *dom*^{musY}**	***mus***	***mus*^{domY}**	***dom***	***dom*^{musY}**
Body mass (g)	20 +/- 0.3 (24)	19.6 +/- 0.3 (7)	17.9 +/- 0.4 (24)	18 +/- 0.4 (12)	19 +/- 0.4 (23)	18.4 +/- 0.3 (47)	18 +/- 0.2 (67)	19 +/- 0.5 (21)
Testes mass (mg)‡	186.4 +/- 3 (24)	200.7 +/- 2 (6)	123.9 +/- 2 (23)	125.6 +/- 3 (12)	193.2 +/- 5 (23)	172.7 +/- 2 (47)	209.1 +/- 3 (67)	189.3 +/- 6 (21)
Relative testes mass (mg/g)*‡	9.1 +/- 0.1 (24)	10.4 +/- 0.2 (6)	7.2 +/- 0.1 (23)	6.9 +/- 0.1 (12)	10.2 +/- 0.2 (23)	9.2 +/- 0.1 (47)	11.7 +/- 0.1 (67)	10.1 +/- 0.2 (21)
Relative SV mass (mg/g)	6.6 +/- 0.2 (23)	7.3 +/- 0.6 (6)	5.2 +/- 0.3 (24)	5.3 +/- 0.3 (12)	6 +/- 0.3 (23)	6.7 +/- 0.2 (47)	5.2 +/- 0.2 (65)	5.9 +/- 0.3 (21)
Bounding height‡*	8.39 (1583)	8.14 (650)	7.46 (870)	7.52 (847)	8.21 (391)	8.02 (401)	8.11 (467)	8.23 (443)
Bounding width‡*	5.58 (1583)	5.07 (650)	4.02 (870)	4.09 (847)	5.02 (391)	4.87 (401)	4.9 (467)	5.11 (443)
Area‡*	24.5 (1583)	21.6 (650)	20.1 (870)	20.1 (847)	22.1 (391)	20 (401)	21.3 (467)	20.4 (443)
Perimeter‡*	23.8 (1583)	22.7 (650)	19.8 (870)	20.2 (847)	22.7 (391)	21.9 (401)	22.3 (467)	23 (443)
Difference from median‡*	6.22 (1583)	8.67 (650)	8.22 (870)	10.8 (847)	8.88 (391)	5.77 (401)	5.86 (467)	6.72 (443)

Slx- and *Slxl1-Sly* Dosage Imbalance Did Not Lead to Ampliconic Gene Family Overexpression

Copy number imbalance of *Slx* and *Slxl1* relative to *Sly* is thought to disrupt expression of these gene families in late spermatogenesis, with particularly strong evidence for *Slx* and *Slxl1* overexpression when *Sly* is knocked down (Cocquet *et al.* 2009; Cocquet *et al.* 2012) and *Slxl1* overexpression when *Slx* and *Slxl1* are duplicated (Kruger *et al.* 2019). *Slx*, *Slxl1,* and *Sly* appear to be involved in the regulation of sex chromatin which impacts the regulation of many genes during late spermatogenesis (Kruger *et al.* 2019). Therefore, we predicted that their misregulation may disrupt the expression of additional genes, including

additional Y-linked ampliconic gene families *Ssty1/2* and the autosomal ampliconic gene family *α-takusan* (Larson *et al.* 2017; Moretti *et al.* 2020). To test if *Slx, Slxl1, Sly, Ssty1, Ssty2,* and *α-takusan* expression was disrupted under less extreme copy number differences in hybrid mice, we compared ampliconic gene family expression levels in round spermatids among cross types. We did not directly quantify copy number for the mice that were FACS sorted, so we used our previous copy number estimates from pure strains sharing the same sex chromosomes as our experimental mice (Larson *et al.* 2021). For all six gene families, expression level was significantly associated with copy number based on a linear mixed-effects model with experiment as a random effect to control for batch effects (FDR-corrected $P < 0.05$; Figure 4). However, for *Slxl1,* this association was negative, suggesting that copy number was not the primary determinant of *Slxl1* expression. This is interesting given that we found high overlap in the range of *Slxl1* copy numbers in naturally occurring *musculus* and *domesticus* (Figure 2B), and the previous demonstration that *Slxl1* plays a more direct role in sex ratio bias than *Slx* (Kruger *et al.* 2019). We then tested if X-Y mismatch had a significant effect on expression level using a linear mixed-effects model with both copy number and presence of X-Y mismatch as fixed effects and experiment as a random effect. We used an ANOVA to compare this model to a null model with copy number as the only fixed effect and experiment as a random effect. For all six genes, X-Y mismatch was not significantly associated with ampliconic gene expression levels (FDR-corrected ANOVA $P > 0.05$). When we specified the direction of X-Y mismatch (i.e., *musculus* X and *domesticus* Y, the direction with an excess of *Slx* relative to *Sly*), only *Ssty2* expression was significantly associated with X-Y mismatch in this direction (FDR-corrected ANOVA $P > 0.05$).

We also tested if X-autosomal background was significantly associated with expression levels using the same mixed-effects model approach. For *Slx, Slxl1, Sly*, *Ssty1,* and *Ssty2,* the sterile hybrid background (*musculus*♀ × *domesticus*♂) was significantly associated with expression levels after FDR-correction (*Slx* ANOVA $P << 0.0001$; *Slxl1* $P < 0.001$; *Sly* $P = 0.01$; *Ssty1* $P < 0.001$; *Ssty2* $P = 0.001$). We observed overexpression of *Slx, Slxl1, Sly, Ssty1,* and *Ssty2* relative to their copy numbers for mice with *musculus*♀ × *domesticus*♂ backgrounds (*mus×dom* and *mus×dom*^{musY}; Figure 4A-E), consistent with previous studies showing that these hybrid mice exhibit widespread overexpression on the sex chromosomes

(Good *et al.* 2010; Campbell *et al.* 2013; Larson *et al.* 2017). Both *mus×dom* and *mus×dom*^{musY} mice in our study overexpressed *Slx, Slxl1,* and *Sly* (Figure 4A, 4B, and 4C), suggesting that matching X and Y chromosomes from *musculus* did not rescue *Slx, Slxl1,* or *Sly* upregulation, and that the overexpression we observed likely results from X-autosomal incompatibilities that disrupt MSCI rather than *Slx*- or *Slxl1-Sly* dosage imbalance. Additionally, *mus*^{domY} mice from our Non-hybrid XY Mismatch also had a *musculus* X and *domesticus* Y, the same X and Y chromosome combination found in sterile hybrids that results in an excess of *Slx* and *Slxl1* copies relative to *Sly* copies. If *Slx*- or *Slxl1-Sly* dosage imbalance contributed to *Slx, Slxl1,* and *Sly* overexpression, we would expect *mus*^{domY} mice to have higher expression than *mus* controls. We observed the opposite effect, with *mus*^{domY} mice showing slightly lower *Slx, Slxl1,* and *Sly* expression levels (Figure 4A, 4B, and 4C). This result provides further evidence that postmeiotic *Slx, Slxl1*, and *Sly* overexpression in sterile F1 hybrids is unlikely to be primarily due to *Slx*- or *Slxl1-Sly* dosage imbalance, and that X-Y mismatch in the absence of autosomal mismatch is not sufficient to cause overexpression of *Slx, Slxl1,* and *Sly*.

Given that *Slx, Slxl1,* and *Sly* are thought to regulate the *α-takusan* ampliconic family, we predicted that *α-takusan* expression levels would also be associated with a *musculus*♀ × *domesticus*♂ background. Surprisingly, this association was not significant (ANOVA P = 0.40). Instead, we observed that *α-takusan* was overexpressed in all cross types with an F1 autosomal background regardless of cross direction (Figure 4F), and that expression was significantly associated with an F1 autosomal background (ANOVA P < 0.01). This suggests that *α-takusan* regulation likely involves autosomal loci in addition to SLX, SLXL1, SLY, SSTY1, and SSTY2 (Moretti *et al.* 2020).

Sex-linked ampliconic genes are primarily expressed during postmeiotic spermatogenesis, in mice and more generally across mammals (Cocquet *et al.* 2012; Mueller *et al.* 2013; Sin and Namekawa 2013). Our non-hybrid expression data supported this, with little to no expression of *Slx, Slxl1, Sly,* or *Ssty1/2* in early meiotic cells in our *mus* and *dom* samples. However, we did detect some meiotic expression of *Slx, Slxl1, Sly,* and *Ssty2* in mice with hybrid autosomal backgrounds, and expression levels of these gene families in early meiosis was significantly associated with F1 autosomal background (ANOVA P < 0.05, Supplemental Material, Figure S6). X chromosome expression has been shown to be

disrupted throughout spermatogenesis in F1 hybrids, although the effect was smaller during earlier spermatogenic stages (Larson *et al.* 2017). Our results suggest that disruption of early spermatogenesis regulatory networks may result in spurious expression of sex-linked ampliconic genes during early meiotic stages when they are normally silenced.

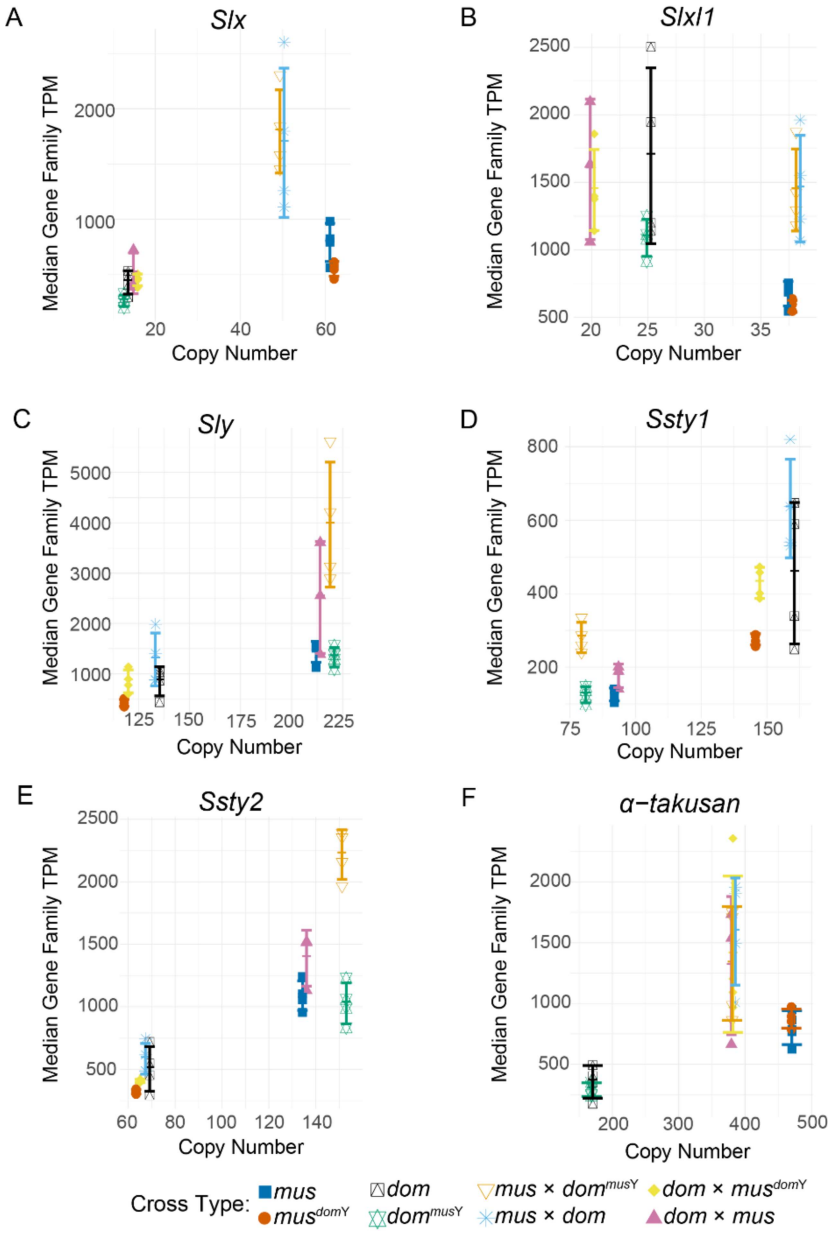


Figure 4: Normalized expression levels of *Slx* (A), *Slxl1* (B), *Sly* (C), *Ssty1* (D), *Ssty2* (E), and *α-takusan* (F) ampliconic gene families in different cross types plotted against their copy numbers. Copy number estimates are based on estimates from wild-derived strains used in experimental and control crosses (see Figure 2). Cross types with the same sex chromosome and therefore same copy number estimate are jittered slightly along the x-axis for clarity. Expression level was calculated by summing transcripts-per million (TPM) for each paralog

of the gene family with at least 97% sequence identity to the ampliconic gene. Points represent values for individual samples, and lines indicate median and standard deviation for each cross type.

X-Y Mismatch Was Not Associated with Sex Chromosome Overexpression in Sterile F1 Hybrids

Next we sought to differentiate if widespread postmeiotic overexpression in sterile hybrids was a direct result of sex chromosome mismatch, a continuation of disrupted meiotic sex chromosome inactivation (MSCI), or a combination of both (Larson *et al.* 2017; Larson *et al.* 2021). We first reanalyzed data from (Larson *et al.* 2017) and repeated their result showing sex chromosome upregulation in late spermatogenesis in sterile F1 hybrids (*mus×dom*, Figure 5A and 5D). We then tested if upregulation was due to X-Y mismatch by comparing relative expression levels in F1 hybrids to those in our Hybrid F1 XY Match mice, which had sex chromosomes from the same subspecies. If X-Y mismatch contributed to sex chromosome upregulation in sterile hybrids, we would expect to see some rescue from disrupted postmeiotic expression in these Hybrid F1 XY Match mice, with *mus×dom*^{musY} mice having lower expression on the X chromosome relative to *mus×dom* F1s. Contrary to this prediction, the X chromosome showed similar expression levels when comparing expression in these two cross types. Therefore, restoring matching sex chromosomes did not rescue expression levels on the *musculus* X chromosome from overexpression in hybrids (Figure 5B). We further tested the effects of sex chromosome mismatch using our Non-hybrid XY Mismatch mice, which had introgressed Y chromosomes on a non-hybrid autosomal background. If mismatch between a *musculus* X chromosome and *domesticus* Y chromosome was sufficient to induce postmeiotic sex chromosome overexpression, then we would expect to see higher X chromosome expression in *mus*^{domY} mice. Instead, we observed slight under expression on the X chromosome compared to the autosomes in *mus*^{domY} mice, confirming that sex chromosome mismatch does not cause X chromosome overexpression in late spermatogenesis (Figure 5C).

We also found evidence that sex chromosome mismatch does not contribute to Y chromosome overexpression in late spermatogenesis in sterile *musculus*♀ × *domesticus*♂

hybrids. The Y chromosome was upregulated in *mus×dom* sterile hybrids relative to *dom×mus*^{domY} mice. This could be due to rescue of *domesticus* Y chromosome expression when paired with the *domesticus* X, but it could also be due to overall lower sex chromosome expression in mice with a *domesticus*♀ × *musculus*♂ background (Figure 5E). In Non-hybrid XY Mismatch, we saw that *mus*^{domY} mice had lower expression on the Y chromosome compared to *dom* controls, in contrast to the Y chromosome overexpression observed in *mus×dom* hybrids (Figure 5F). Thus, X-Y mismatch does appear to influence Y chromosome expression, but in the opposite direction of that observed in sterile hybrids.

In the reciprocal cross (*domesticus*♀ × *musculus*♂ F1 hybrids), we found some evidence that X-Y mismatch may contribute to disrupted expression of X-linked genes. Here Y chromosome expression was not different from that on the autosomes (Figure 5G), but the X chromosome tended to be downregulated (Figure 5J; Larson *et al.* 2017). There was no evidence that X-Y match restored normal X chromosome expression levels in *dom×mus*^{domY} (Hybrid F1 XY Match), with this cross type showing similar or even slightly lower expression levels on the X chromosome relative to *dom×mus* hybrids (Figure 5K). However, in Non-hybrid XY Mismatch we observed lower expression on the X chromosome in *dom*^{musY} mice relative to *dom* controls (Figure 5L). Therefore, a *domesticus* X paired with a *musculus* Y can result in suppression of X-linked gene expression even in the absence of autosomal incompatibilities.

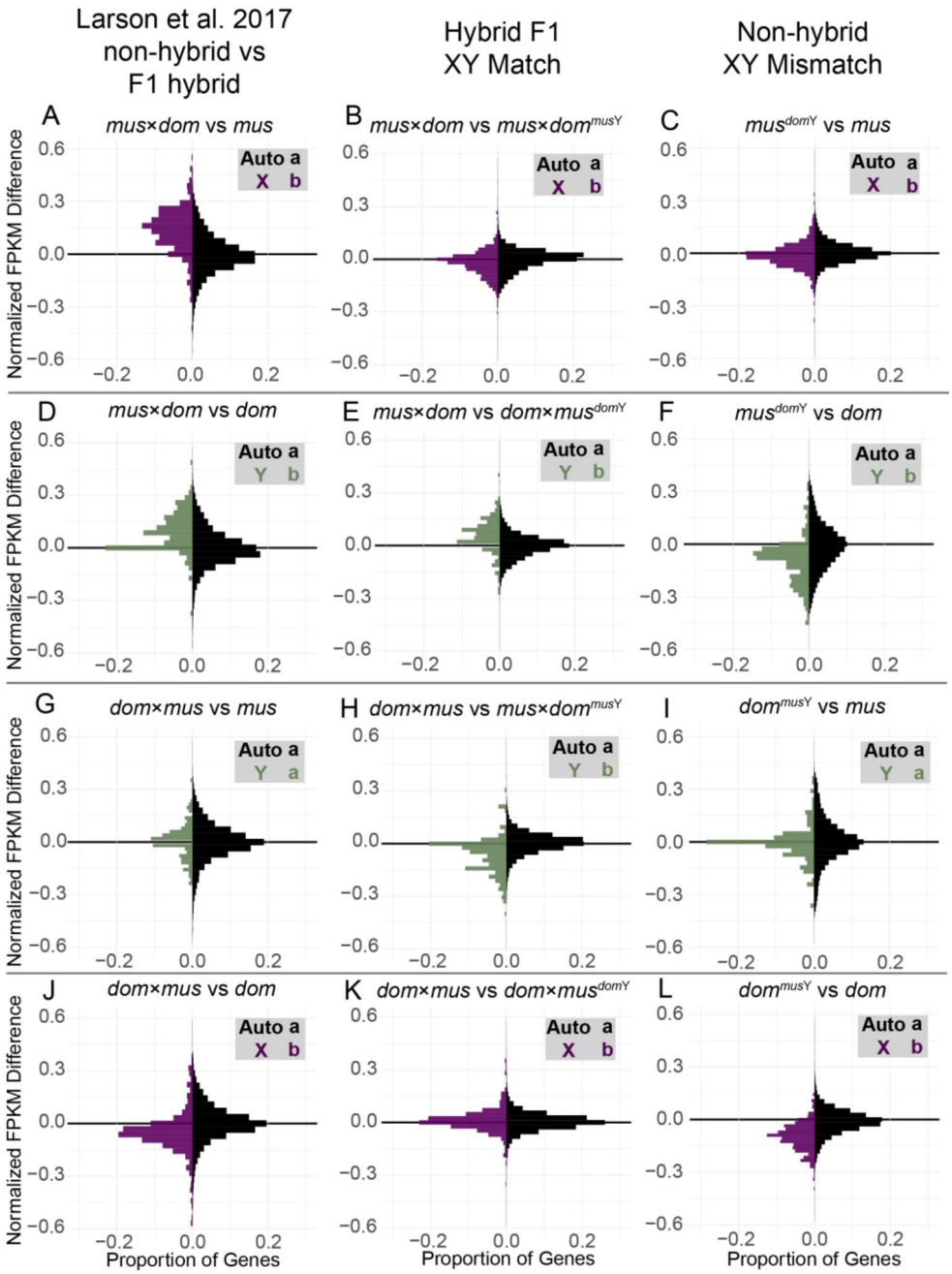


Figure 5: Histograms of relative expression levels between experimental cross types and control mice. (A-C) Contrasts that all have a *musculus* X chromosome, (D-F) contrasts with a *domesticus* Y chromosome (G-I) contrasts with a *musculus* Y chromosome, and (J-L) contrasts with a *domesticus* X chromosome. (A-F) represent sex chromosome mismatch present in sterile hybrids (*musculus* X and *domesticus* Y), while (G-L) represent sex chromosome mismatch present in more fertile hybrids (*domesticus* X and *musculus* Y). The first column (A, D, G, and J) shows data reanalyzed from (Larson *et al.* 2017). The second column (B, E, H, K) tests if gene expression levels are rescued when the sex chromosomes are matched but on a hybrid autosomal background (Hybrid F1 XY Match). The third column

(C, F, I, L) tests for disrupted expression due to sex chromosome mismatch alone, on a non-hybrid autosomal background (Non-hybrid XY Mismatch). The y-axis shows the difference in normalized expression levels between the two cross types being compared. The x-axis shows the proportion of genes in each expression difference bin. Black bars represent the autosomes, purple bars represent the X chromosome, and green bars represent the Y chromosome. Letters indicate significant differences in median expression differences among the chromosome types based on a Mann-Whitney U test (FDR-corrected $P < 0.05$).

X-Y Mismatch Disrupted the Expression of Several Genes during Late Spermatogenesis

We also tested for effects of X-Y mismatch on individual genes by identifying differentially expressed (DE) genes in X-Y mismatch mice compared to controls. In our reanalysis, we identified many more overexpressed genes in sterile *mus×dom* hybrids compared to *mus* and many more underexpressed genes in the reciprocal *dom×mus* hybrids compared to *dom* on the X chromosome (Table 2), consistent with previous results (Larson *et al.* 2017) and with our observations of overall expression differences (Figure 5). We then asked if any of these X-linked DE genes were associated with X-Y mismatch. If so, then we would expect our Hybrid F1 XY Match *mus×dom*^{musY} to rescue some of the disrupted X-linked expression, and thus manifest as DE genes in comparisons between *mus×dom* and *mus×dom*^{musY}. These genes should also overlap with genes DE between *mus×dom* and *mus*. However, there were only two X-linked DE genes in the *mus×dom* versus *mus×dom*^{musY} comparison (Table 2), and only one was also DE in the *mus×dom* versus *mus* comparison (Figure 6). This gene is a predicted protein coding gene, *Gm10058*, that shares 97% sequence identity with *Slx* and is therefore likely a paralog of this gene family. The other DE gene was *Btbd35f17*, another ampliconic gene with a protein-protein binding domain that is specifically expressed in male reproductive tissues (Smith *et al.* 2019). In Non-hybrid XY Mismatch, we only observed one X-linked DE gene in *mus*^{domY} compared to *mus*, and this gene was not DE in any other comparisons. Taken together, both Hybrid F1 XY Match and Non-hybrid XY Mismatch results suggest that almost all DE genes on the X chromosome in sterile *musculus*♀ × *domesticus*♂ hybrids are disrupted due to X-autosomal or autosomal-autosomal incompatibilities, rather than Y-linked incompatibilities.

On the X chromosome, very few DE genes were shared across multiple comparisons. However, 57 DE genes were shared between the *mus×dom* versus *mus* and *dom×mus* versus *dom* comparisons. When we looked at DE genes separated by direction of expression difference, only eight were shared between these two comparisons (Supplemental Material, Figure S7), so most of the overlap represented genes overexpressed in *mus×dom* but underexpressed in *dom×mus*. This could indicate that similar regulatory networks are disrupted in reciprocal F1 hybrids, but in ways that disrupt gene expression levels in opposite directions.

In contrast to the X chromosome, more Y-linked DE genes were shared across comparisons (Figure 6). Sterile *mus×dom* hybrids had 17 Y-linked DE genes that showed a clear bias towards overexpression (Table 2). Of these 17 DE genes, 5 were shared with the Hybrid F1 XY Match comparison *mus×dom* versus *dom×mus*^{domY}, so having *domesticus* X and Y chromosomes partially rescued expression levels on the Y chromosome in *dom×mus*^{domY} mice. However, none of the 17 Y-linked genes DE in sterile hybrids were also DE in the Hybrid F1 XY Match comparison (*mus*^{domY} versus *dom*), so it is unlikely that X-Y mismatch alone disrupts expression of these genes. Instead, there may be a complex interaction between X-Y mismatch and a hybrid autosomal background that disrupts Y chromosome expression. Consistent with this, we found the most Y-linked DE genes in comparisons between cross types with reciprocal hybrid autosomal backgrounds but the same Y chromosome (Table 2). Of these, 78 Y-linked DE genes were shared between these two comparisons (Figure 6), suggesting that reciprocal hybrid autosomal backgrounds may have resulted in disrupted expression for many of the same Y-linked genes, regardless of the subspecies origin of the Y chromosome.

We also found several autosomal genes that were DE between cross types with the same autosomal background but different sex chromosome combinations (Table 2). We excluded autosomal genes that overlapped with putatively introgressed regions, so the DE that we detected was unlikely to result from *cis*-regulatory effects of variants from the opposite subspecies that introgressed along with the Y chromosome. In Hybrid F1 XY Match, 104 autosomal genes were DE when comparing *mus×dom* to *dom×mus*^{domY} and 494 autosomal genes were DE when comparing *dom×mus* to *mus×dom*^{musY} (Table 2). These comparisons involved reciprocal crosses with the same autosomal and Y chromosome

genotypes, and so DE presumably resulted from X-autosomal incompatibilities. Although overexpression on the X chromosome tends to be the most notable expression pattern associated with X-autosomal incompatibilities, previous studies have shown disrupted postmeiotic autosomal expression in sterile hybrids as well (Larson *et al.* 2017). We detected only six (non-overlapping) DE genes in each comparison with different Y chromosomes but the same autosomal and X chromosome genotypes (*mus×dom* versus *mus×dom*^{musY} and *dom×mus* versus *dom×mus*^{domY}; Table 2).

In Non-hybrid XY Mismatch, we identified some autosomal DE genes in comparisons that had different Y chromosomes but the same autosomal and X backgrounds, suggesting that interactions involving the Y chromosome disrupted some autosomal expression, but the number of autosomal DE genes was not enriched relative to the number of X-linked DE genes (Fisher's Exact Test P > 0.05; Table 2).These autosomal DE genes tended to be underexpressed in the cross type with X-Y mismatch regardless of the direction of the cross (Table 2) and must result from direct interactions with the Y chromosome or indirect interactions with X-Y mediated expression changes. Only one autosomal gene, *Babam2*, was DE in both reciprocal comparisons. It is a member of the BRCA1-A complex, which is involved in DNA double-strand break repair (The Uniprot Consortium 2020).

Table 2: Number of differentially expressed genes in round spermatids for different cross type comparisons. "Higher" indicates higher expression (i.e., overexpressed) in the cross type with X-Y mismatch (F1 hybrids in Larson et al. 2017 and Hybrid F1 XY Match, Y-introgression F1 crosses in Non-hybrid XY Mismatch). "Lower" indicates lower expression (i.e., underexpressed) in the cross type with X-Y mismatch. For comparisons in the "Other Contrasts" category, "higher" indicates higher expression in the first cross type listed (*mus* or *mus×dom*). Gray boxes indicate chromosomes that are from the same subspecies in the two cross types being compared. Reciprocal F1s were considered as having the same autosomal backgrounds. Autosomal DE genes overlapping with putatively introgressed regions were excluded from comparisons involving Y-introgression mice.

		Autosomes		X Chromosome		Y Chromosome	
		Higher	**Lower**	**Higher**	**Lower**	**Higher**	**Lower**
Larson et al. 2017	*mus*×*dom* vs *mus*	1518	1476	252	13	109	66
	mus×*dom* vs *dom*	1357	1241	190	55	15	2
	dom×*mus* vs *mus*	1360	1009	62	73	6	8
	dom×*mus* vs *dom*	1237	878	27	73	69	86
Hybrid F1 XY Match	*mus*×*dom* vs *mus*×*dom*^{musY}	3	3	2	0	74	70
	mus×*dom* vs *dom*×*mus*^{domY}	21	83	38	96	68	84
	dom×*mus* vs *mus*×*dom*^{musY}	372	122	44	101	76	85
	dom×*mus* vs *dom*×*mus*^{domY}	2	4	1	0	71	66
Non-hybrid XY Mismatch	*mus*^{domY} vs *mus*	13	34	1	0	52	63
	mus^{domY} vs *dom*	1820	2269	28	179	3	69
	dom^{musY} vs *mus*	1634	1679	70	55	10	7
	dom^{musY} vs *dom*	13	63	0	10	14	70
Other Contrasts	*mus* vs *dom*	1536	1774	38	139	46	96
	mus×*dom* vs *dom*×*mus*	42	19	85	36	5	1

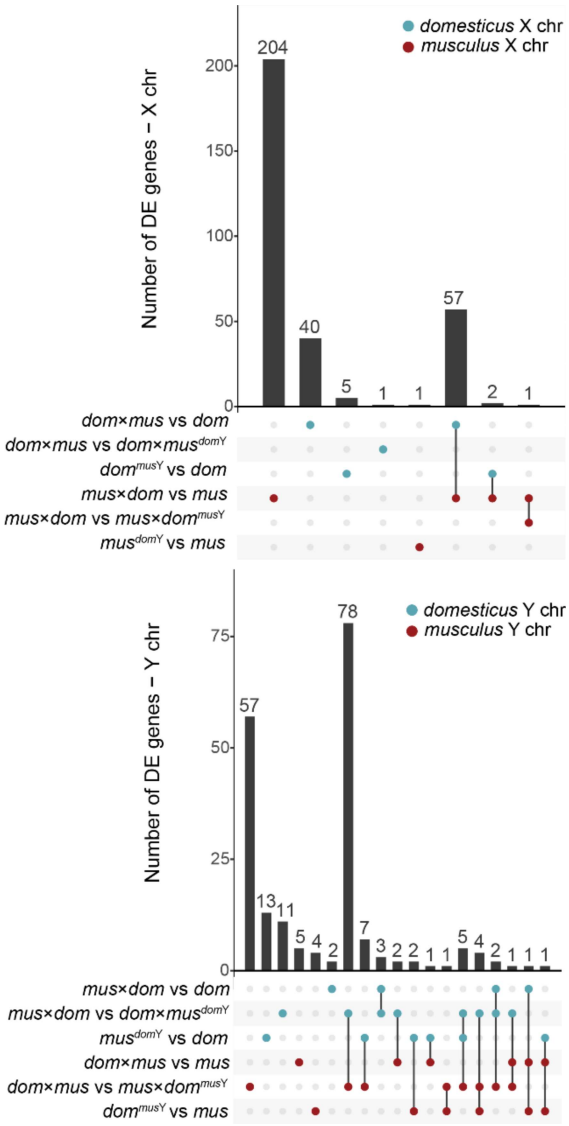


Figure 6: Upset plots showing the number of differentially expressed (DE) genes in each cross type comparison, and genes that are DE across multiple comparisons. (A) DE genes on the X chromosome. (B) DE genes on the Y chromosome. Bars corresponding to multiple dots connected by lines indicate genes that are DE across multiple comparisons. Bars corresponding to single dots indicate genes that are DE in only one comparison. Blue dots indicate comparisons on the *domesticus* X chromosome (A) or *domesticus* Y chromosome (B), and red dots indicate comparisons on the *musculus* X chromosome (A) or *musculus* Y chromosome (B). Genes that were DE in opposite directions across multiple comparisons of the same sex chromosome were excluded.

Finally, we tested if DE genes tended to be in the same co-expression networks using weighted correlation network analysis (WGCNA). We found one module in Hybrid F1 XY Match associated with the *mus×dom* autosomal background, one module in Non-hybrid XY Mismatch associated with the *musculus* background, and one module in Non-hybrid XY Mismatch associated with the *domesticus* background (Figure 7A, B, D). These modules were significantly enriched for genes DE between cross types with different autosomal backgrounds (Table 3). There were also multiple modules enriched for DE genes despite not having a significant association with cross type (Table 3). For example, Module 5 was significantly enriched for DE genes in all pairwise comparisons in Hybrid F1 XY Match. Although we did not detect a significant cross type association for this module, there was a trend towards an autosomal background by sex chromosome effect for this module, with *mus×dom* background cross types tending to have lower module membership in general, but with *mus×dom*^{musY} mice tending to have higher module membership than *mus×dom* mice (Figure 7E). Another Hybrid F1 XY Match module showed a similar pattern (Module 3, Figure 7C) and was enriched for genes DE between *dom×mus* and *mus×dom*^{musY} (Table 3). In Non-hybrid XY Mismatch, Module 5 was enriched for genes DE between *mus*^{domY} and either subspecies (*mus* or *dom*; Table 3), and X-Y mismatch mice tended to have lower associations with this module (Figure 7). We likely did not have enough power to detect significant module associations with complex autosome by sex chromosome interactions given our sample size, especially because these effects on gene expression tended to be subtle and affect relatively few genes (Figure 5, Table 2). Despite low power, the fact that certain modules were enriched for DE genes suggests that groups of genes were disrupted in similar ways in X-Y mismatch mice, and that particular gene networks may be disrupted under X-Y mismatch. Additionally, we found a significant positive correlation in module eigengene values between Hybrid F1 XY Match and Non-hybrid XY Mismatch (Module 5 in both experiments, $r = 0.64$; FDR-corrected Pearson's correlation $P < 0.001$; Supplemental Figure S8) and a significant overlap in genes (279 genes, FDR-corrected Fisher's Exact Test $P < 0.001$), suggesting that these two modules represent genes with similar expression patterns between the two experiments. Interestingly, these modules trended towards a negative association with cross types that had a *musculus* X chromosome and *domesticus* Y chromosome (Figure 7E, 7F), and may represent genes with similar expression patterns

under X-Y mismatch regardless of autosomal background. All DE genes and their module memberships are listed in Supplemental Material, Tables S6 and S7.

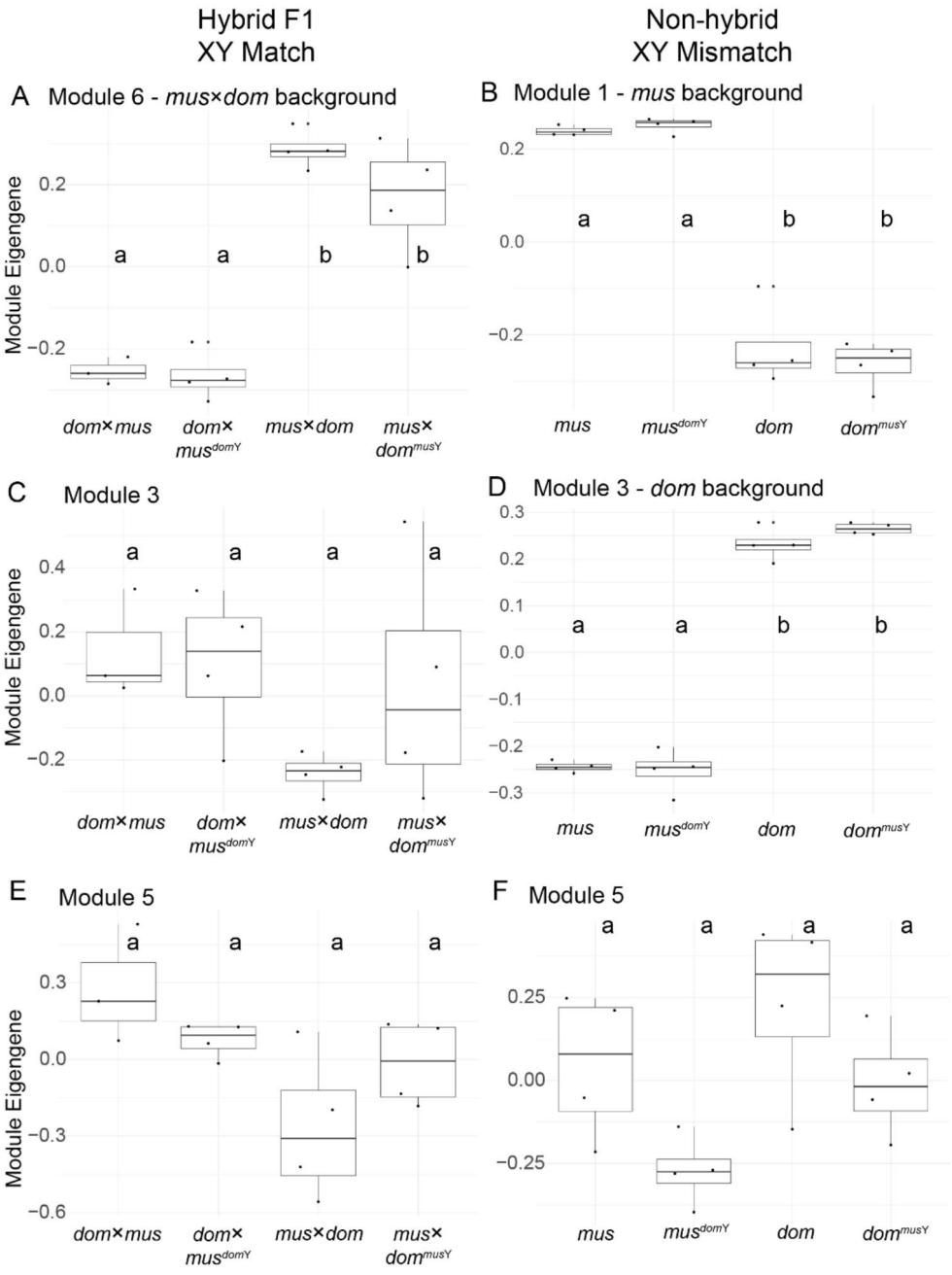


Figure 7: Example WGCNA module eigengene values plotted by cross type. Note that WGCNA was performed separately for each experiment, so there is not necessarily a relationship between Hybrid F1 XY Match and Non-hybrid XY Mismatch modules with the same number. Modules that were significantly associated with cross types are also labeled based on these associations (A, B, and D). Other modules shown were not significantly associated with a cross type but trended towards an association with X-autosomal

background by Y chromosome type interaction and were enriched for DE genes in at least one comparison (C, E, and F; Table 3). Letters indicate significant differences in module association based on linear models with post-hoc Tukey tests ($P < 0.05$).

Table 3: Number of differentially expressed genes in each WGCNA module. Rows indicate WGCNA modules and columns indicate comparisons between cross types used to identify differentially expressed (DE) genes. Module associations with cross types are based on linear models with post-hoc Tukey tests. Shaded boxes indicate a significant enrichment for DE genes based on a hypergeometric test with FDR-correction ($P < 0.05$). Note that there is not necessarily a relationship between Hybrid F1 XY Match and Non-hybrid XY Mismatch modules with the same module number.

			Number of DE genes in module			
Hybrid F1 XY Match	**Module**	**Significant cross type associations**	***mus*×*dom* vs *mus*×*dom*^{musY}**	***mus*×*dom* vs *dom*×*mus*^{domY}**	**dom×*mus* vs *mus*×*dom*^{musY}**	***dom*×*mus* vs *dom*×*mus*^{domY}**
	1	none	0	13	9	0
	2	none	1	19	35	2
	3	none	1	10	170	1
	4	none	4	3	11	2
	5	none	7	21	155	5
	6	*mus*×*dom* background	2	87	102	1
Non-hybrid XY Mis-match	**Module**	**Significant cross type associations**	***mus*^{domY} vs *mus***	***mus*^{domY} vs *dom***	***dom*^{musY} vs *mus***	***dom*^{musY} vs *dom***
	1	*mus* background	0	1039	972	40
	2	none	0	168	133	4
	3	*dom* background	9	913	970	4
	4	none	4	91	358	2
	5	none	23	532	28	9
	6	none	3	329	17	5
	7	none	1	220	55	1
	8	none	1	22	77	8

	9	none	0	106	3	0
	10	none	1	1	104	1
	11	none	0	111	24	0
	12	none	1	3	6	0
	13	none	0	1	0	0
	14	none	1	1	12	1
	15	none	0	0	0	0
	16	none	0	0	5	0
	17	none	0	18	0	1
	18	none	0	0	0	0

Discussion

The large X-effect and Haldane's rule are prevalent patterns observed in intrinsic hybrid incompatibilities across diverse taxa and suggest that sex chromosomes play a predominant role in speciation, but the evolutionary forces underlying rapid sex chromosome divergence that leads to hybrid incompatibilities remain unclear (Presgraves and Meiklejohn 2021). One compelling hypothesis is that hybrid incompatibilities are a consequence of intragenomic conflict between sex chromosomes (Frank 1991; Hurst and Pomiankowski 1991; Lindholm *et al.* 2016). In this study, we showed that intragenomic conflict between the sex chromosomes may contribute to some hybrid incompatibilities in house mice, but not in a simple dosage-dependent manner, and with subtle effects relative to other components of F1 hybrid incompatibilities. Notably, we find that X-Y conflict does not appear to contribute to postmeiotic disruption of sex chromosome repression, a major regulatory phenotype associated with hybrid sterility in house mice (Larson *et al.* 2017). Below, we discuss the implications of our findings for the genetic basis of house mouse male hybrid sterility and the potential role of intragenomic conflict in speciation.

Insights into the Genetic Basis of Mouse Male Hybrid Sterility

Our results did not support the model of *Slx*- and *Slxl1-Sly* dosage imbalance leading to X chromosome overexpression in mouse F1 hybrids. In Hybrid F1 XY Match, we showed that X-Y match on an F1 background did not restore postmeiotic X chromosome repression (Figure 5). In Non-hybrid XY Mismatch, we directly tested the effects of X-Y mismatch in the absence of X-autosomal mismatch on postmeiotic spermatogenesis gene expression. We found some evidence for disrupted expression in X-Y mismatch mice (Figure 5, Table 2), but the effects were relatively subtle and often in the opposite direction than expected based on genetic manipulation studies (Cocquet *et al.* 2012; Kruger *et al.* 2019) or disrupted expression in sterile F1 mice (Larson *et al.* 2017; Figures 4, 5, and 6).

Our results indicate that genetic manipulation studies, which performed nearly complete knockdowns or duplications, are not representative of the more subtle copy number differences expected to occur in natural hybrids (Cocquet *et al.* 2009; Cocquet *et al.* 2012; Kruger *et al.* 2019). Another important difference from genetic manipulation studies is that we used wild-derived inbred strains instead of the C57BL/6J classic laboratory mouse, which has a mostly *domesticus* background but some *musculus* introgression throughout, including the Y chromosome (Nagamine *et al.* 1992). Because C57BL/6J is mostly *domesticus* with a *musculus* Y chromosome, it has a similar genetic composition as our wild-derived dom^{musY} mice and therefore may show some of the same subtle disruptions to gene expression and sperm morphology that we observed compared to pure *domesticus* mice. We also introgressed the entire Y chromosome, so there should not have been dosage imbalances among ampliconic genes on the same sex chromosome. However, our Y-introgression mice also had imbalance between all Y-linked ampliconic genes and interacting genes on the X chromosome and autosomes, so it is unclear if introgressing the entire Y chromosome should cause larger or smaller effects on postmeiotic spermatogenesis expression.

SLX, SLXL1, and SLY proteins interact with other sex-linked and autosomal ampliconic genes, including *Ssty1/2, α-takusan,* and *Speer,* so additional gene families may be involved in intragenomic conflict with *Slx, Slxl1,* and *Sly* (Kruger *et al.* 2019; Moretti *et al.* 2020). Our autosomal gene family expression results seem to further complicate understanding of the consequences of ampliconic gene conflict as we found that the *α-takusan* gene family is overexpressed in F1 hybrids regardless of cross direction or sex chromosome type (Figure

4F). Sex chromosome mismatch, however, did not disrupt *α-takusan* expression when the autosomal background was non-hybrid. This was somewhat puzzling because protein products of sex-linked ampliconic genes are thought to regulate *α-takusan* expression in late spermatogenesis, perhaps again indicating that copy number differences between subspecies are too subtle to generate strong regulatory phenotypes. Another surprising expression result was that *Slxl1* expression levels were not correlated with *Slxl1* copy numbers (Figure 4B). Other genes are likely involved in the regulation of *Slxl1* (Moretti *et al.* 2020), and it is possible that the evolution of these *trans*-acting factors may play a more important role in determining overall *Slxl1* expression levels than *Slxl1* copy number *per se*.

On balance, our results suggest that differences in *Slx*- or *Slxl1-Sly* dosage do not result in strong hybrid incompatibilities. We did not observe sex chromosome overexpression with an excess of *Slx* and *Slxl1* copies or underexpression with an excess of *Sly* copies as predicted under the conflict model (Larson *et al.* 2017). Therefore, the primary mechanisms underlying postmeiotic X chromosome overexpression in sterile F1 hybrids likely do not involve X-Y interactions. Instead, disrupted postmeiotic repression is likely a continuation of *Prdm9*-mediated MSCI disruption (Bhattacharyya *et al.* 2013; Bhattacharyya *et al.* 2014; Mukaj *et al.* 2020).

Although X-Y copy number imbalance is unlikely to explain disrupted postmeiotic repression in F1 hybrids, sex chromosome interactions may play a role in house mouse hybrid sterility. We showed that X-Y mismatch can lead to disrupted expression of ampliconic genes and other genes throughout the genome (Figure 4, Figure 6, Table 2), and some of these genes are essential for spermatogenesis. For example, *Taf7l* knockouts have abnormal sperm morphology (Cheng *et al.* 2007), *Prdx4* knockouts have reduced sperm counts (Iuchi *et al.* 2009), and both these genes were differentially expressed in *dom*^{musY} mice. We also showed that hybrid interactions involving the Y-chromosome are associated with subfertility phenotypes (Table 1), consistent with previous studies (Campbell *et al.* 2012; Campbell and Nachman 2014). Here we have focused on interactions between the sex chromosomes because the ampliconic gene conflict model established a clear prediction for X-Y incompatibilities, but we could not distinguish X-Y incompatibilities from Y-autosomal incompatibilities in our experimental crosses. We note that several of our observations

could result from Y-autosomal interactions. Indeed, introgression of the Y chromosome (Non-hybrid XY Mismatch) induced autosomal regulatory phenotypes.

We observed some autosomal regions that co-introgressed with the Y chromosome, and some of these regions have been shown to introgress in other mouse hybrids (Supplemental Material, Figure S4). These may be regions that are incompatible with the Y chromosome from the opposite subspecies, and therefore must co-introgress for mice to be viable or fertile. The large introgressed region we identified on chromosome 2 is adjacent to a multicopy gene, *R2d2*, involved in meiotic drive during female meiosis (Didion *et al.* 2016). *R2d2* has only been shown to act in females (Didion *et al.* 2016), but our crossing scheme only involved backcrossing hybrid males. We also generated Y-intogression mice using the LEWES/EiJ strain, which is fixed for the low copy number allele of *R2d2*, and PWK/PhJ, which also appears to have low *R2d2* copy number (Didion *et al.* 2016), so it is unlikely that this introgression is a direct result of *R2d2* drive as previously described. Nevertheless, the exact functions of *R2d2* are unresolved, so this large region of introgression may be related to *R2d2,* but probably not through a direct meiotic drive mechanism.

Our results are likely important in the context of mouse speciation in nature. Mice sampled from the European hybrid zone are often advanced generation hybrids with complex patterns of ancestry from both *musculus* and *domesticus*, and true F1 genotypes are exceptionally rare (Teeter *et al.* 2010; Turner *et al.* 2012). Therefore, understanding mechanisms of hybrid incompatibility in addition to F1 X-autosomal incompatibilities is essential for understanding the complex genetic basis of mouse speciation occurring in nature. The Non-hybrid XY Mismatch experiment demonstrated that disrupted gene expression phenotypes can occur in the absence of an F1 autosomal background. Previous studies have shown that advanced intercrosses of hybrid mice show different sterility phenotypes than F1s (Campbell *et al.* 2012), and *Prdm9*-mediated hybrid sterility requires an F1 autosomal background, leading others to speculate that genetic incompatibilities underlying hybrid sterility may be different in later hybrid generations (Campbell and Nachman 2014; Mukaj *et al.* 2020). Our results show that Y chromosome introgression can contribute to reduced fertility (consistent with (Campbell *et al.* 2012) and some disrupted spermatogenesis gene expression in later generation hybrids with non-F1 autosomal backgrounds.

What Is the Contribution of Sex Chromosome Conflict to Speciation?

Several studies have proposed a link between intragenomic conflict and hybrid incompatibilities (Tao *et al.* 2001; Phadnis and Orr 2009; Wilkinson *et al.* 2014; Zanders *et al.* 2014; Case *et al.* 2015; Zhang *et al.* 2015; Larson *et al.* 2017), but it remains unknown how prevalent these systems are in natural populations or if intragenomic conflict is the primary selective force behind the evolution of these incompatibilities. While X-autosomal incompatibilities are known to play a central role in house mouse hybrid sterility, previous work has shown that house mouse speciation likely has a more complex genetic basis (Vyskočilová *et al.* 2005; Good *et al.* 2008b; Turner *et al.* 2012; Turner and Harr 2014; Larson *et al.* 2018b) and may involve sex chromosome intragenomic conflict (Ellis *et al.* 2011; Campbell *et al.* 2012; Larson *et al.* 2017). The exact mechanisms underlying reduced fertility associated with Y chromosome mismatch is unknown, and it is still unclear what role, if any, sex chromosome intragenomic conflict may play (Ellis *et al.* 2011; Campbell *et al.* 2012; Larson *et al.* 2017).

Ampliconic genes are a common feature of mammalian sex chromosomes, and they tend to be expressed specifically during spermatogenesis (Li *et al.* 2013; Soh *et al.* 2014; Skinner *et al.* 2016; Lucotte *et al.* 2018; Bellott *et al.* 2017; Hughes *et al.* 2020; reviewed in Larson *et al.* 2018a). Although difficult to quantify, evolution of ampliconic gene families involved in spermatogenesis is arguably one of the most rapidly evolving components of mammalian genomes (Mueller *et al.* 2013; Soh *et al.* 2014; Lucotte *et al.* 2018; Cechova *et al.* 2020; Vegesna *et al.* 2020). Intragenomic conflict among sex chromosome ampliconic genes has been proposed as a mechanism through which hybrid incompatibilities have evolved in at least three mammalian groups (Davis *et al.* 2015; Dutheil *et al.* 2015; Larson *et al.* 2018a; Kruger *et al.* 2019). In cats, loci associated with hybrid sterility tend to be in or near high copy number genes (Davis *et al.* 2015). In great apes, sex chromosome amplicon copy number can evolve rapidly (Lucotte *et al.* 2018; Cechova *et al.* 2020), and ampliconic regions on the X chromosome are thought to have experienced selective sweeps as a result of strong selection pressures imposed by intragenomic conflict with the Y chromosome (Nam *et al.* 2015). These regions also overlap sections of the modern human X chromosome that lack Neandertal introgression, and therefore may represent regions involved in genetic

incompatibilities between modern humans and Neandertals (Dutheil *et al.* 2015). However, most of these connections remain speculative and the X chromosome is clearly a hotspot of the evolution of hybrid incompatibilities (Masly and Presgraves 2007; Good *et al.* 2008a).

Theoretical work introducing the idea that sex chromosome intragenomic conflict could contribute to hybrid incompatibilities focused on this phenomenon as an explanation for Haldane's rule and the large X-effect (Frank 1991; Hurst and Pomiankowski 1991). However, genetic conflict between the sex chromosomes during reproduction cannot explain some observations, such as the applicability of Haldane's rule and the large X-effect to hybrid inviability or the important role of the X chromosome in many incompatibilities that occur in homogametic hybrids (Coyne 1992). In this study, we showed that X-Y conflict may have a small effect on male hybrid sterility, but *Prdm9*-mediated incompatibilities probably play the most important role in the observations consistent with Haldane's rule and the large X-effect in house mice. Interactions among *Prdm9, Hstx2,* and other autosomal and X-linked loci in hybrids result in failed or delayed double strand break repair, which eventually leads to meiotic arrest and male sterility (Forejt *et al.* 2021). The rapid divergence of *Prdm9* and its binding sites is likely the result of PRDM9 haplotype selection, leading to biased gene conversion and hotspot erosion (Baker *et al.* 2015). Thus, intragenomic conflict is unlikely to be the primary underlying cause of house mouse hybrid male sterility.

It remains unknown if the recurrent evolution of ampliconic genes is a consequence of intragenomic conflict across mammals, but this is generally assumed to be the case. If so, intragenomic conflict may be much more important in the evolution of hybrid incompatibility loci than once thought (Johnson and Wu 1992; Coyne and Orr 2004). Some recent empirical studies support this hypothesis in both flies and mammals (Presgraves and Meiklejohn 2021), however, our study did not provide direct support for this hypothesis. X-Y mismatch likely contributes to hybrid male sterility and disrupted expression, but in more complex ways than the *Slx*, *Slxl1,* and *Sly* dosage-based conflict model, and with relatively small effects on hybrid sterility. In particular, we note that *Ssty1* and *Sly* show opposing copy number patterns between subspecies, such that replacing a *musculus* Y with a *domesticus* Y simultaneously increases *Ssty1* while decreasing *Sly*, and vice versa. It is possible that higher copy number of one gene can compensate for reduced copy number of the other in

regulating postmeiotic sex chromatin. This work is thus distinct from previous work focused on deletions (that reduce copy number of both genes) or RNA interference (that selectively targets one gene).

Further work is required to identify loci involved in X-Y or Y-autosomal incompatibilities, but it is plausible that intragenomic conflict among ampliconic genes still plays a role given that these genes are the primary sex chromosome genes expressed in the postmeiotic stages during which spermatogenesis expression is highly disrupted (Sin and Namekawa 2013; Larson *et al.* 2017). Copy number mismatch between these gene families may play important roles in reproductive outcomes in nature, as has been implied from slight sex ratio skews in regions of the hybrid zone (Macholán *et al.* 2008). Even subtle differences in fertility could have important effects on fitness, especially given that sperm competition appears to be common in mice (Dean *et al.* 2006). However, our work suggests that such effects do not manifest as a major reproductive barrier between populations.

Data Availability

The sequence data reported in this paper are available through the National Center for Biotechnology Information Sequence Read Archive under accession numbers PRJNA816542 and PRJNA816886**.** The phenotype data are available in the Supplemental Material, Table S2.

Acknowledgements

We would like to thank Sara Keeble for assistance with animal husbandry, Pamela K. Shaw and the UM Fluorescence Cytometry Core supported by an Institutional Development Award from the NIGMS (P30GM103338 and S10-OD025019) and the UM Lab Animal Resources staff. This study included research conducted in the University of Montana Genomics Core, supported by a grant from the M. J. Murdock Charitable Trust (to J.M.G.). Computational resources and support from the University of Montana's Griz Shared Computing Cluster (GSCC), supported by grants from the National Science Foundation (CC-2018112 and OAC-1925267, J.M.G. co-PI), contributed to this research. We thank members

of the Good Lab, Polly Campbell, Lila Fishman, Doug Emlen, and Travis Wheeler for helpful discussions and comments during the development of this study. We thank three anonymous reviewers for helpful comments on an earlier version of this manuscript. This work was supported by grants from the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health (R01-HD073439, R01-HD094787 to J.M.G.). E.E.K.K. was supported by the National Science Foundation Graduate Research Fellowship Program (DGE-1313190), a grant from the National Science Foundation (DEB-1754096 to J.M.G.), and a Rosemary Grant Award for Graduate Student Research from the Society for the Study of Evolution. E.L.L. was supported by the National Science Foundation (DEB-2012041). C.C.R. was supported by the BBSRC (BB/N000463/1 to P.E.). P.E. also acknowledges funding from the Leverhulme Trust (RPG-2019-414 194). E.M.W. was supported by a faculty studentship from the University of Essex. B.M.S was supported by UKRI (University of Essex). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the National Institutes of Health, or the Society for the Study of Evolution.

Author Contributions

J.M.G. and E.L.L. conceived and funded the project. E.E.K.K., E.L.L., and J.M.G. designed the experiments. E.L.L. and E.E.K.K. did the mouse husbandry and breeding. E.E.K.K. performed the mouse dissections, cell sorts, and sequencing library preparation. E.M.W., C.C.R., B.M.S., and P.J.I.E. performed and analyzed the sperm morphology assays. E.E.K.K. analyzed the data. E.E.K.K., E.L.L., and J.M.G. wrote the manuscript with input from all authors.

Conflicts of Interest

The authors declare no conflicts of interest.

References

Baker, C. L., S. Kajita, M. Walker, R. L. Saxl, N. Raghupathy *et al.*, 2015 PRDM9 Drives Evolutionary Erosion of Hotspots in *Mus musculus* through Haplotype-Specific Initiation of Meiotic Recombination. PLOS Genetics 11: e1004916.

Balcova, M., B. Faltusova, V. Gergelits, T. Bhattacharyya, O. Mihola *et al.*, 2016 Hybrid Sterility Locus on Chromosome X Controls Meiotic Recombination Rate in Mouse. PLOS Genetics 12: e1005906.

Bellott, D. W., H. Skaletsky, T.-J. Cho, L. Brown, D. Locke *et al.*, 2017 Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. Nat Genet 49: 387-394.

Bhattacharyya, T., S. Gregorova, O. Mihola, M. Anger, J. Sebestova *et al.*, 2013 Mechanistic basis of infertility of mouse intersubspecific hybrids. Proceedings of the National Academy of Sciences of the United States of America 110: E468-E477.

Bhattacharyya, T., R. Reifova, S. Gregorova, P. Simecek, V. Gergelits *et al.*, 2014 X Chromosome Control of Meiotic Chromosome Synapsis in Mouse Inter-Subspecific Hybrids. PLOS Genetics 10: e1004088.

Bi, Y., X. Ren, R. Li, Q. Ding, D. Xie *et al.*, 2019 Specific Interactions Between Autosome and *X* Chromosomes Cause Hybrid Male Sterility in *Caenorhabditis* Species. Genetics 212: 801-813.

Bolger, A. M., M. Lohse and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

Burgoyne, P. S., S. K. Mahadevaiah and J. M. A. Turner, 2009 The consequences of asynapsis for mammalian meiosis. Nat Rev Genet 10: 207-216.

Campbell, P., J. M. Good, M. D. Dean, P. K. Tucker and M. W. Nachman, 2012 The Contribution of the Y Chromosome to Hybrid Male Sterility in House Mice. Genetics 191: 1271-1281.

Campbell, P., J. M. Good and M. W. Nachman, 2013 Meiotic Sex Chromosome Inactivation Is Disrupted in Sterile Hybrid Male House Mice. Genetics 193: 819-828.

Campbell, P., and M. W. Nachman, 2014 X–Y Interactions Underlie Sperm Head Abnormality in Hybrid Male House Mice. Genetics 196: 1231-1240.

Case, L. K., E. H. Wall, E. E. Osmanski, J. A. Dragon, N. Saligrama *et al.*, 2015 Copy number variation in Y chromosome multicopy genes is linked to a paternal parent-of-origin effect on CNS autoimmune disease in female offspring. Genome Biology 16: 28.

Cechova, M., R. Vegesna, M. Tomaszkiewicz, R. S. Harris, D. Chen *et al.*, 2020 Dynamic evolution of great ape Y chromosomes. Proceedings of the National Academy of Sciences: 202001749.

Charlesworth, B., J. A. Coyne and N. H. Barton, 1987 The Relative Rates of Evolution of Sex Chromosomes and Autosomes. American Naturalist 130: 113-146.

Cheng, Y., M. G. Buffone, M. Kouadio, M. Goodheart, D. C. Page *et al.*, 2007 Abnormal Sperm in Mice Lacking the *Taf7l* Gene. Molecular and Cellular Biology 27: 2582-2589.

Cocquet, J., P. J. Ellis, S. K. Mahadevaiah, N. A. Affara, D. Vaiman *et al.*, 2012 A genetic basis for a postmeiotic X versus Y chromosome intragenomic conflict in the mouse. PLoS Genet 8: e1002900.

Cocquet, J., P. J. Ellis, Y. Yamauchi, S. K. Mahadevaiah, N. A. Affara *et al.*, 2009 The multicopy gene Sly represses the sex chromosomes in the male mouse germline after meiosis. PLoS Biol 7: e1000244.

Conway, J. R., A. Lex and N. Gehlenborg, 2017 UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics 33: 2938-2940.

Coughlan, J. M., and D. R. Matute, 2020 The importance of intrinsic postzygotic barriers throughout the speciation process. Philosophical Transactions of the Royal Society B: Biological Sciences 375: 20190533.

Coyne, J. A., 1992 Genetics and Speciation. Nature 355: 511-515.

Coyne, J. A., and H. A. Orr, 1989 Two rules of speciation, pp. 180–207 in *Speciation and its consequences*, edited by D. Otte and J. Endler. Sinauer Associates, Sunderland (Massachusetts).

Coyne, J. A., and H. A. Orr, 2004 *Speciation*. W.H. Freeman.

Davies, B., E. Hatton, N. Altemose, J. G. Hussin, F. Pratto *et al.*, 2016 Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. Nature 530: 171-176.
Davis, B. W., C. M. Seabury, W. A. Brashear, G. Li, M. Roelke-Parker *et al.*, 2015 Mechanisms Underlying Mammalian Hybrid Sterility in Two Feline Interspecies Models. Molecular Biology and Evolution 32: 2534-2546.
Dean, M. D., K. G. Ardlie and M. W. Nachman, 2006 The frequency of multiple paternity suggests that sperm competition is common in house mice (*Mus domesticus*). Molecular Ecology 15: 4141-4151.
Didion, J. P., A. P. Morgan, L. Yadgary, T. A. Bell, R. C. McMullan *et al.*, 2016 *R2d2* Drives Selfish Sweeps in the House Mouse. Molecular Biology and Evolution 33: 1381-1395.
Dobzhansky, T., 1937 Genetic Nature of Species Differences. The American Naturalist 71: 404-420.
Dumont, B. L., and B. A. Payseur, 2011 Genetic Analysis of Genome-Scale Recombination Rate Evolution in House Mice. PLOS Genetics 7: e1002116.
Dutheil, J. Y., K. Munch, K. Nam, T. Mailund and M. H. Schierup, 2015 Strong Selective Sweeps on the X Chromosome in the Human-Chimpanzee Ancestor Explain Its Low Divergence. PLOS Genetics 11: e1005451.
Edwards, J. A., and R. A. Edwards, 2019 Fastq-pair: efficient synchronization of paired-end fastq files. bioRxiv: 552885.
Ellis, P. J., J. Bacon and N. A. Affara, 2011 Association of Sly with sex-linked gene amplification during mouse evolution: a side effect of genomic conflict in spermatids? Hum Mol Genet 20: 3010-3021.
Forejt, J., P. Jansa and E. Parvanov, 2021 Hybrid sterility genes in mice (*Mus musculus*): a peculiar case of PRDM9 incompatibility. Trends in Genetics.
Frank, S. A., 1991 Divergence of Meiotic Drive-Suppression Systems as an Explanation for Sex- Biased Hybrid Sterility and Inviability. Evolution 45: 262-267.
Getun, I. V., B. Torres and P. R. J. Bois, 2011 Flow Cytometry Purification of Mouse Meiotic Cells. Journal of Visualized Experiments: JoVE: 2602.
Good, J. M., 2012 The Conflict within and the Escalating War between the Sex Chromosomes. PLoS Genet 8: e1002955.
Good, J. M., M. D. Dean and M. W. Nachman, 2008a A Complex Genetic Basis to X-Linked Hybrid Male Sterility Between Two Species of House Mice. Genetics 179: 2213-2228.
Good, J. M., T. Giger, M. D. Dean and M. W. Nachman, 2010 Widespread Over-Expression of the X Chromosome in Sterile F_1 Hybrid Mice. PLoS Genet 6: e1001148.
Good, J. M., M. A. Handel and M. W. Nachman, 2008b Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. Evolution 62: 50-65.
Green, C. D., Q. Ma, G. L. Manske, A. N. Shami, X. Zheng *et al.*, 2018 A Comprehensive Roadmap of Murine Spermatogenesis Defined by Single-Cell RNA-Seq. Developmental Cell 46: 651-667.e610.
Grey, C., P. Barthès, G. Chauveau-Le Friec, F. Langa, F. Baudat *et al.*, 2011 Mouse PRDM9 DNA-Binding Specificity Determines Sites of Histone H3 Lysine 4 Trimethylation for Initiation of Meiotic Recombination. PLOS Biology 9: e1001176.
Haldane, J. B. S., 1922 Sex ratio and unisexual sterility in hybrid animals. Journal of Genetics 12: 101-109.
Harr, B., E. Karakoc, R. Neme, M. Teschke, C. Pfeifle *et al.*, 2016 Genomic resources for wild populations of the house mouse, *Mus musculus* and its close relative *Mus spretus*. Sci Data 3: 160075.
Hughes, J. F., H. Skaletsky, T. Pyntikova, N. Koutseva, T. Raudsepp *et al.*, 2020 Sequence analysis in Bos taurus reveals pervasiveness of X–Y arms races in mammalian lineages. Genome Research 30: 1716-1726.
Hunnicutt, K. E., J. M. Good and E. L. Larson, 2021 Unraveling patterns of disrupted gene expression across a complex tissue. Evolution 76: 275-291.

Hurst, L. D., and A. Pomiankowski, 1991 Causes of sex ratio bias may account for unisexual sterility in hybrids: a new explanation of Haldane's rule and related phenomena. Genetics 128: 841-858.

Iuchi, Y., F. Okada, S. Tsunoda, N. Kibe, N. Shirasawa *et al.*, 2009 Peroxiredoxin 4 knockout results in elevated spermatogenic cell death via oxidative stress. Biochemical Journal 419: 149-158.

Janoušek, V., L. Wang, K. E. N. Luzynski, P. Dufková, M. VyskočIlová Martina *et al.*, 2012 Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between *Mus musculus musculus* and *M. m. domesticus*. Molecular Ecology 21: 3032-3047.

Johnson, N. A., and C. I. Wu, 1992 An empirical test of the meiotic drive models of hybrid sterility: sex-ratio data from hybrids between *Drosophila simulans* and *Drosophila sechellia*. Genetics 130: 507-511.

Kruger, A. N., M. A. Brogley, J. L. Huizinga, J. M. Kidd, D. G. de Rooij *et al.*, 2019 A Neofunctionalized X-Linked Ampliconic Gene Family Is Essential for Male Fertility and Equal Sex Ratio in Mice. Current Biology 29: 3699-3706.e3695.

Langfelder, P., and S. Horvath, 2008 WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9: 559.

Larson, E. L., S. Keeble, D. Vanderpool, M. D. Dean and J. M. Good, 2017 The Composite Regulatory Basis of the Large X-Effect in Mouse Speciation. Molecular Biology and Evolution 34: 282-295.

Larson, E. L., E. E. K. Kopania and J. M. Good, 2018a Spermatogenesis and the Evolution of Mammalian Sex Chromosomes. Trends Genet 34: 722-732.

Larson, E. L., E. E. K. Kopania, K. E. Hunnicutt, D. Vanderpool, S. Keeble *et al.*, 2021 Stage-specific disruption of X chromosome expression during spermatogenesis in sterile house mouse hybrids. G3 Genes|Genomes|Genetics 12: jkab407.

Larson, E. L., D. Vanderpool, B. A. J. Sarver, C. Callahan, S. Keeble *et al.*, 2018b The Evolution of Polymorphic Hybrid Incompatibilities in House Mice. Genetics 209: 845–859.

Li, G., B. W. Davis, T. Raudsepp, A. J. Pearks Wilkerson, V. C. Mason *et al.*, 2013 Comparative analysis of mammalian Y chromosomes illuminates ancestral structure and lineage-specific evolution. Genome Research 23: 1486-1495.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754-1760.

Lindholm, A. K., K. A. Dyer, R. C. Firman, L. Fishman, W. Forstmeier *et al.*, 2016 The Ecology and Evolutionary Dynamics of Meiotic Drive. Trends in Ecology & Evolution 31: 315-326.

Lucotte, E. A., L. Skov, J. M. Jensen, M. C. Macià, K. Munch *et al.*, 2018 Dynamic Copy Number Evolution of X- and Y-Linked Ampliconic Genes in Human Populations. Genetics 209: 907-920.

Macholán, M., S. J. Baird, P. Munclinger, P. Dufková, B. Bímová *et al.*, 2008 Genetic conflict outweighs heterogametic incompatibility in the mouse hybrid zone? BMC Evolutionary Biology 8: 271.

Masly, J. P., and D. C. Presgraves, 2007 High-Resolution Genome-Wide Dissection of the Two Rules of Speciation in *Drosophila*. PLOS Biology 5: e243.

Matute, D. R., and B. S. Cooper, 2021 Comparative studies on speciation: 30 years since Coyne and Orr. Evolution 75: 764-778.

Meiklejohn, C. D., and Y. Tao, 2010 Genetic conflict and sex chromosome evolution. Trends in Ecology & Evolution 25: 215-223.

Mihola, O., Z. Trachtulec, C. Vlcek, J. C. Schimenti and J. Forejt, 2009 A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase. Science 323: 373-375.

Moretti, C., M. Blanco, C. Ialy-Radio, M.-E. Serrentino, C. Gobé *et al.*, 2020 Battle of the Sex Chromosomes: Competition between X and Y Chromosome-Encoded Proteins for Partner Interaction and Chromatin Occupancy Drives Multicopy Gene Expression and Evolution in Muroid Rodents. Molecular Biology and Evolution 37: 3453–3468.

Morgan, A. P., and F. Pardo-Manuel de Villena, 2017 Sequence and Structural Diversity of Mouse Y Chromosomes. Molecular Biology and Evolution 34: 3186-3204.
Mueller, J. L., S. K. Mahadevaiah, P. J. Park, P. E. Warburton, D. C. Page *et al.*, 2008 The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. Nat Genet 40: 794-799.
Mueller, J. L., H. Skaletsky, L. G. Brown, S. Zaghlul, S. Rock *et al.*, 2013 Independent specialization of the human and mouse X chromosomes for the male germ line. Nat Genet 45: 1083-1087.
Mukaj, A., J. Piálek, V. Fotopulosova, A. P. Morgan, L. Odenthal-Hesse *et al.*, 2020 *Prdm9* Intersubspecific Interactions in Hybrid Male Sterility of House Mouse. Molecular Biology and Evolution 37: 3423-3438.
Nagamine, C. M., Y. Nishioka, K. Moriwaki, P. Boursot, F. Bonhomme *et al.*, 1992 The musculus-type Y Chromosome of the laboratory mouse is of Asian origin. Mammalian Genome 3: 84-91.
Nam, K., K. Munch, A. Hobolth, J. Y. Dutheil, K. R. Veeramah *et al.*, 2015 Extreme selective sweeps independently targeted the X chromosomes of the great apes. Proc Natl Acad Sci U S A 112: 6413-6418.
Namekawa, S. H., P. J. Park, L.-F. Zhang, J. E. Shima, J. R. McCarrey *et al.*, 2006 Postmeiotic Sex Chromatin in the Male Germline of Mice. Current Biology 16: 660-667.
Oliver, P. L., L. Goodstadt, J. J. Bayes, Z. Birtle, K. C. Roach *et al.*, 2009 Accelerated Evolution of the *Prdm9* Speciation Gene across Diverse Metazoan Taxa. PLOS Genetics 5: e1000753.
Patro, R., G. Duggal, M. I. Love, R. A. Irizarry and C. Kingsford, 2017 Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. Nature methods 14: 417-419.
Pedersen, B. S., and A. R. Quinlan, 2017 Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics 34: 867-868.
Pezer, Ž., B. Harr, M. Teschke, H. Babiker and D. Tautz, 2015 Divergence patterns of genic copy number variation in natural populations of the house mouse (*Mus musculus domesticus*) reveal three conserved genes with major population-specific expansions. Genome Research 25: 1114-1124.
Phadnis, N., and H. A. Orr, 2009 A Single Gene Causes Both Male Sterility and Segregation Distortion in *Drosophila* Hybrids. Science 323: 376-379.
Phifer-Rixey, M., and M. W. Nachman, 2015 Insights into mammalian biology from the wild house mouse *Mus musculus*. Elife 4: e05959.
Presgraves, D. C., and C. D. Meiklejohn, 2021 Hybrid Sterility, Genetic Conflict and Complex Speciation: Lessons From the *Drosophila simulans* Clade Species. Frontiers in Genetics 12.
Rathje, C. C., E. E. P. Johnson, D. Drage, C. Patinioti, G. Silvestri *et al.*, 2019 Differential Sperm Motility Mediates the Sex Ratio Drive Shaping Mouse Sex Chromosome Evolution. Current Biology 29: 3692-3698.e3694.
Sin, H. S., Y. Ichijima, E. Koh, M. Namiki and S. H. Namekawa, 2012 Human postmeiotic sex chromatin and its impact on sex chromosome evolution. Genome Res 22: 827-836.
Sin, H. S., and S. H. Namekawa, 2013 The great escape: Active genes on inactive sex chromosomes and their evolutionary implications. Epigenetics 8: 887-892.
Skinner, B. M., C. C. Rathje, J. Bacon, E. E. P. Johnson, E. L. Larson *et al.*, 2019 A high-throughput method for unbiased quantitation and categorization of nuclear morphology. Biology of Reproduction 100: 1250-1260.
Skinner, B. M., C. A. Sargent, C. Churcher, T. Hunt, J. Herrero *et al.*, 2016 The pig X and Y Chromosomes: structure, sequence, and evolution. Genome Research 26: 130-139.
Smith, C. M., T. F. Hayamizu, J. H. Finger, S. M. Bello, I. J. McCright *et al.*, 2019 The mouse Gene Expression Database (GXD): 2019 update. Nucleic acids research 47: D774-D779.

Soh, Y. Q., J. Alfoldi, T. Pyntikova, L. G. Brown, T. Graves *et al.*, 2014 Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. Cell 159: 800-813.

Storchová, R., S. Gregorová, D. Buckiová, V. Kyselová, P. Divina *et al.*, 2004 Genetic analysis of X-linked hybrid sterility in the house mouse. Mammalian Genome 15: 515-524.

Suzuki, T. A., and M. W. Nachman, 2015 Speciation and reduced hybrid female fertility in house mice. Evolution 69: 2468-2481.

Tao, Y., D. L. Hartl and C. C. Laurie, 2001 Sex-ratio segregation distortion associated with reproductive isolation in *Drosophila*. Proceedings of the National Academy of Sciences 98: 13183-13188.

Teeter, K. C., L. M. Thibodeau, Z. Gompert, C. A. Buerkle, M. W. Nachman *et al.*, 2010 The variable genomic architecture of isolation between hybridizing species of house mice. Evolution 64: 472-485.

The UniProt Consortium, 2020 UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49: D480-D489.

Turelli, M., and H. A. Orr, 1995 The dominance theory of Haldane's rule. Genetics 140: 389-402.

Turelli, M., and H. A. Orr, 2000 Dominance, Epistasis and the Genetics of Postzygotic Isolation. Genetics 154: 1663-1679.

Turner, J. M. A., 2015 Meiotic Silencing in Mammals. Annual Review of Genetics 49: 395-412.

Turner, J. M. A., S. K. Mahadevaiah, P. J. I. Ellis, M. J. Mitchell and P. S. Burgoyne, 2006 Pachytene Asynapsis Drives Meiotic Sex Chromosome Inactivation and Leads to Substantial Postmeiotic Repression in Spermatids. Developmental Cell 10: 521-529.

Turner, L. M., and B. Harr, 2014 Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions. eLife 3: e02504.

Turner, L. M., D. J. Schwahn and B. Harr, 2012 Reduced Male Fertility Is Common but Highly Variable in Form and Severity in a Natural House Mouse Hybrid Zone. Evolution 66: 443-458.

Vegesna, R., M. Tomaszkiewicz, O. A. Ryder, R. Campos-Sánchez, P. Medvedev *et al.*, 2020 Ampliconic genes on the great ape Y chromosomes: Rapid evolution of copy number but conservation of expression levels. Genome Biology and Evolution 12: 842–859.

Vicoso, B., and B. Charlesworth, 2009 Effective population size and the faster-X effect: An extended model. Evolution 63: 2413-2426.

Vyskočilová, M., G. Pražanová and J. Piálek, 2009 Polymorphism in hybrid male sterility in wild-derived *Mus musculus musculus* strains on proximal chromosome 17. Mammalian Genome 20: 83.

Vyskočilová, M., Z. Trachtulec, J. Forejt and J. Piálek, 2005 Does geography matter in hybrid sterility in house mice? Biological Journal of the Linnean Society 84: 663-674.

Widmayer, S. J., M. A. Handel and D. L. Aylor, 2020 Age and Genetic Background Modify Hybrid Male Sterility in House Mice. Genetics 216: 585-597.

Wilkinson, G. S., S. J. Christianson, C. L. Brand, G. Ru and W. Shell, 2014 Haldane's Rule Is Linked to Extraordinary Sex Ratios and Sperm Length in Stalk-Eyed Flies. Genetics 198: 1167-1181.

Yang, H., J. R. Wang, J. P. Didion, R. J. Buus, T. A. Bell *et al.*, 2011 Subspecific origin and haplotype diversity in the laboratory mouse. Nat Genet 43: 648-655.

Yates, A. D., P. Achuthan, W. Akanni, J. Allen, J. Allen *et al.*, 2019 Ensembl 2020. Nucleic Acids Research 48: D682-D688.

Zanders, S. E., M. T. Eickbush, J. S. Yu, J.-W. Kang, K. R. Fowler *et al.*, 2014 Genome rearrangements and pervasive meiotic drive cause hybrid infertility in fission yeast. eLife 3: e02630.

Zhang, L., T. Sun, F. Woldesellassie, H. Xiao and Y. Tao, 2015 Sex Ratio Meiotic Drive as a Plausible Evolutionary Mechanism for Hybrid Male Sterility. PLOS Genetics 11: e1005073.

9.5 Appendix 5: Working title and co-authors for paper in preparation

Genetic background modulates the effect of X-Y and Y-autosomal mismatch on male reproductive phenotypes in hybrid mice

Eleanor M. Watson*[,1], Emily E. K. Kopania[§], Claudia C. Rathje[ǂ], Erica L. Larson[†], Jeffrey M. Good[§], Peter J. I. Ellis[ǂ], Benjamin M. Skinner*[,1]

*School of Life Sciences, University of Essex, Colchester, CO4 3SQ, U.K.

[ǂ]School of Biosciences, University of Kent, Canterbury, CT2 7NJ, U.K

[§]Division of Biological Sciences, University of Montana, Missoula, MT 59812

[†]Department of Biological Sciences, University of Denver, Denver, CO 80208

[1]Corresponding authors: e.watson@essex.ac.uk, b.skinner@essex.ac.uk

ORCID ID: 0000-0002-1290-4360 (EMW)

ORCID ID: 0000-0002-2710-2491 (EEKK)

ORCID ID: 0000-0003-3006-645X (ELL)

ORCID ID: 0000-0001-9709-7934 (PJIE)

ORCID ID: 0000-0003-0707-5374 (JMG)

ORCID ID: 0000-0002-7152-1167 (BMS)

Possible journals: BMC Genomics, Proc B, PLOS One, BMC Genomic Data

Data availability:

Keywords:

hybrid male sterility, nuclear morphologies, sex chromosomes, sperm nuclei, chromosomal mismatch, introgression