

Genomic approaches in medical and population genetics



Hilary Martin
Balliol College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity Term, 2015

Acknowledgements

Recognition of those who have contributed to specific projects is given in each chapter. Here, I thank the people who have helped me more generally throughout my PhD. Firstly, I thank my supervisor, Peter Donnelly, for the invaluable guidance, advice and opportunities he has given me over the last four years, and for all that he has taught me. His thorough approach has been immensely instructive, and has set a high standard I hope to follow throughout my scientific career.

Many thanks also to Liz Batty, with whom I've worked closely on the platypus project. Liz, thank you for being so patient and uncomplaining through every round of tedious mapping and variant calling, even on the umpteenth version of the reference genome. I've been extremely grateful for your help, particularly in the last few months, when you stepped in just at the right moment during frenzied preparation for conferences. Hopefully it will all be worth it, eventually! Julie Hussin has also been a valued colleague and mentor, and has helped with the recombination and platypus projects. Julie, your support, curiosity and enthusiasm for the science kept me going when I needed it. Ryan Christ also gave invaluable assistance with the Bayesian models on the recombination project, and taught me a huge amount, for which I'll be forever grateful. Ryan, I love your energy and dedication! I also thank Jared O'Connell for his patience in helping me with duoHMM (Chapter 5). He went far beyond the call of duty.

To Davis McCarthy, Edouard Hatton, Alex Young, Clare Bycroft, Winni Kretschmar, Robbie Davies, Holly Trochet, Na Cai, Carme Mont and Kiran Garimella, thank you for the encouragement you've given me, for the interesting conversations (scientific and otherwise), and for your good humour. It's been fun going through this with you! I am likewise grateful to Oliver Venn, Anjali Hinch and Ellen Leffler for all their useful advice on the platypus project. Thanks also to Colin Freeman, for his tolerance with my often stupid computational questions, and to Kate Distin-Harvey for her endless patience in often trying circumstances. I thank Alistair Pagnamenta and Jude Craft for their energy and willingness to help in the WGS500 project, as well as Stefano Lise, John Broxholme, Andy Rimmer and Jean-Baptiste Cazier for their help early in my DPhil. Thanks also to Simon Engeldow, Paolo Piazza and Rory Bowden for putting up with all the chopping and changing over the platypus project. To Dorret Boomsma, Hamdi Mbarek, Gonnecke Willemsen, Jouke-Jan Hottenga, Kerrie

McAloney and Scott Gordon, thank you for your help and useful comments on the recombination project.

I am very grateful to Gil McVean for his wise guidance throughout my DPhil, and to Ed Blair and Jenny Taylor for their efforts on the WGS500 case studies described in Chapter 2. Chris Spencer, Jonathan Marchini, Simon Myers, Jonathan Flint and Julian Knight have also been sources of useful advice and encouragement at various times. I thank Radu Aricescu for pointing out the significance of the *GRIA3* finding (Chapter 2), and for his enthusiasm throughout the project. Thanks also to Ben Davies his advice and support, and for not giving up during those many failed attempts at making the *GRIA3* mouse. I also sincerely thank my funders for giving me this great opportunity to come to Oxford: the Nuffield Department of Medicine, the Clarendon Fund, the Oxford-Australia Fund and Balliol College.

To my housemates, many thanks for your friendship and support during these busy last couple of months, particularly to Freya Shearer for help with proofreading. To Nikki Trott, Drummond Bone, Carol Clark, Doug Dupree and Bruce Kinsey, thank you for your good humour and wise counsel, and for the many unique opportunities you've given me at Balliol. As for my college friends, I can't tell you how much you've meant to me. Will Clegg, Gabi De Wit, Phoebe Downing, Sara Dutta, David Favara, Tom Harty, Martha Imprialou, Charlotte Krass, Emily Magnuson, Rahul Nath, Aoife O'Gorman, Mireia Crispin Ortuzar, Andrew Paverd, Indrek Pernik, Seb Pike, Dan Rosmarin, Fiona Skerman, Helen Taylor, Elisabetta Tollardo, John Walmesley, and Robin Winkler - these have been golden years. I won't forget the fun we've, the inspiration you've given me, nor your sound advice, your propensity for mischief, and your generosity.

To Stein Belderok, thank you for your support, unfailing encouragement to better myself, and for your dedication to learning what you know is important to me. I couldn't have asked for more. To my sister, Felicity, who is wise beyond her years, thank you for your tolerance during those many dinner table conversations about genetics, your sense of humour, your strength and for your sensible advice. Finally, I must thank my parents, Georgia Chenevix-Trench and Nick Martin, who have prepared me for this my whole life. I can only hope to prove worthy of all the advantages they have given me.

Genomic approaches in medical and population genetics

Hilary Martin
Balliol College

Doctor of Philosophy
Trinity Term, 2015

Abstract

Over the last fifteen years, rapid advances in genotyping and DNA sequencing technologies have revolutionised genetic and biomedical research. In this thesis, we present some applications of these technologies in studying rare disease, population genetics and meiotic recombination. We begin by reviewing previous research in these areas in Chapter 1. Then in Chapter 2, we present some case studies of Mendelian neurological disorders that were carried out as part of a large clinical whole-genome sequencing project, WGS500. These led to the discovery of several new genes for a type of severe early-onset epilepsy called Ohtahara Syndrome, and of a particularly interesting mutation that tentatively suggests a role for a glutamate receptor gene, *GRIA3*, in circadian rhythm control. In Chapter 3, we examine some general lessons learnt from the WGS500 project, including the utility of sequencing family members to reduce the number of candidate pathogenic variants, and the perils of focusing on candidate genes. Chapter 4 describes a population sequencing project on the platypus, in which we sequenced 58 samples from across the whole species range. Our results provide insights into the population structure and history of this fascinating mammal, and also into the ongoing evolution of its remarkable chain of ten sex chromosomes. Finally, in Chapter 5, we describe a study of the effect of maternal age on meiotic recombination, the largest of its kind to date. Our results from multiple cohorts suggest a small but significant positive effect of maternal age on the number of crossovers, but with substantial heterogeneity between cohorts that is likely due to sampling noise, though confounders may also play a role. These studies illustrate the power of genomic approaches for investigating fundamental biological processes at the population, individual and cellular levels.

Contents

1	Introduction	1
1.1	The genomics revolution	1
1.2	Finding genes for rare disease	3
1.3	Population genetics	11
1.4	Meiotic recombination	15
2	Case studies from the WGS500 project	19
2.1	Introduction	19
2.2	Methods	20
2.2.1	Case recruitment	20
2.2.2	Sequencing and variant calling pipeline	20
2.2.3	Variant filtering strategies	20
2.2.4	Linkage analysis	23
2.2.5	Variant validation and functional follow-up	23
2.3	Results	24
2.3.1	Early-onset epilepsy	24
2.3.1.1	Introduction	24
2.3.1.2	Patients 1 and 3: <i>KCNQ2</i> and <i>SCN2A</i>	25
2.3.1.3	Patient 2: <i>KCNT1</i>	26
2.3.1.4	Patient 4: <i>PIGQ</i>	29
2.3.1.5	Patient 5: <i>CSNK1G1</i>	34
2.3.1.6	Patient 6: <i>CBL</i>	34
2.3.1.7	Mutation screening in other epileptic encephalopathy cases	36
2.3.1.8	Family 7: uncertain cause	36
2.3.2	Putatively novel syndromes	37
2.3.2.1	Mental retardation and disturbed sleep-wake cycle	38

2.3.2.2	Spastic paraplegia, learning disability and pseudopseudohypoparathyroidism	44
2.4	Discussion	48
2.5	Acknowledgements	52
3	General lessons from the WGS500 project	53
3.1	Introduction	53
3.2	Methods	58
3.2.1	Samples	58
3.2.2	Variant annotation for overarching analysis	58
3.2.3	Searching for variants in candidate genes	59
3.3	Results	61
3.3.1	Use of parental information	61
3.3.2	Screening against WGS500	66
3.3.3	Utility of candidate gene lists	69
3.3.4	Regulatory variants	71
3.4	Discussion	80
3.5	Acknowledgements	84
4	Whole-genome sequencing of the platypus	86
4.1	Introduction	86
4.1.1	Natural history of the platypus	86
4.1.2	The platypus genome project	89
4.1.3	Previous work on platypus population genetics	91
4.1.4	Platypus sex chromosomes	94
4.1.5	Aims of the platypus population sequencing project	98
4.2	Methods	104
4.2.1	Samples and sequencing	104
4.2.2	Improving the reference genome	104
4.2.2.1	Mate pair data	104
4.2.2.2	BACs and fosmids	107
4.2.2.3	Molecule data	107
4.2.2.4	<i>De novo</i> assembly	107
4.2.2.5	Rescaffolding and gap filling	107
4.2.3	Mapping	109
4.2.4	Variant calling	109
4.2.5	Identifying relatives	111

4.2.6	Population genetic analyses	111
4.2.7	Identifying X-linked scaffolds	112
4.2.8	Identifying regions homologous to the Y chromosomes	114
4.2.9	Gene identification and synteny	115
4.3	Results	115
4.3.1	Improving the platypus reference genome	115
4.3.2	Variant calling in population samples	120
4.3.3	Population structure, diversity and history	126
4.3.4	Sex chromosomes	138
4.4	Discussion	147
Appendices		157
4.A	Appendices	157
4.A.1	Supplementary Figures	157
4.A.2	Supplementary Tables	178
5	Multi-cohort analysis of the maternal age effect on recombination	191
5.1	Introduction	191
5.2	Methods	197
5.2.1	Cohorts	197
5.2.2	Auxiliary data on alternative reproductive technologies and oral contraception	197
5.2.3	Preparation of genotype data	199
5.2.4	Calling crossovers	199
5.2.5	Linear mixed models and meta-analysis	201
5.2.6	Bayesian hierarchical models	201
5.2.7	Posterior predictive checks	206
5.2.8	Assessing the frequentist properties of the Bayesian method	206
5.3	Results	208
5.3.1	Comparison of duoHMM to NFTOOLS	208
5.3.2	Linear mixed models and meta-analysis	208
5.3.3	Bayesian analysis of the maternal age effect	211
5.3.4	Posterior predictive checks	218
5.4	Discussion	218

Appendices	224
5.A Appendices	224
5.A.1 Supplementary Figures	225
5.A.2 Supplementary Tables	248
6 Conclusion	260
Bibliography	263

Chapter 1

Introduction

1.1 The genomics revolution

Over the last fifteen years, rapid advances in genotyping and DNA sequencing technologies have revolutionised genetics and medicine. Single nucleotide polymorphism (SNP) genotyping was initially conducted one variant at a time, and cost $\sim \$1$ per SNP. The invention of multiplex SNP genotyping using DNA microarrays (Wang *et al.*, 1998) was followed by a huge increase in throughput: capacity per array increased from hundreds of SNPs to hundreds of thousands by 2005, and then to a million SNPs by 2007 (Altshuler *et al.*, 2008). Improvements in DNA sequencing have been even more staggering (Figure 1.1), driven by the transition from Sanger-based (dideoxy chain termination) sequencing to “next-generation” technologies in 2008. Whereas the original human genome project cost \$3 billion and took eleven years from its launch in 1990 to the publication of the initial paper (Lander *et al.*, 2001), it is now standard practice to sequence a human genome in a day for about \$1500 (Wetterstrand, 2015).

Initially, multiplex SNP genotyping was primarily used for linkage analysis (Middleton *et al.*, 2004; Sawcer *et al.*, 2004), where it facilitated the process of gene discovery for rare diseases by providing much denser marker sets than microsatellite panels. Since 2005, SNP chips have been widely used in genome-wide association studies (GWAS) of complex traits (reviewed in Visscher *et al.*, 2012)¹. Meanwhile, since 2008, the reduced cost of DNA sequencing has led to a massive acceleration of Mendelian disease gene discovery (Section 1.2). In Chapter 2 of this thesis, we

¹The GWAS revolution is not described in this chapter since it is not directly relevant to this thesis.

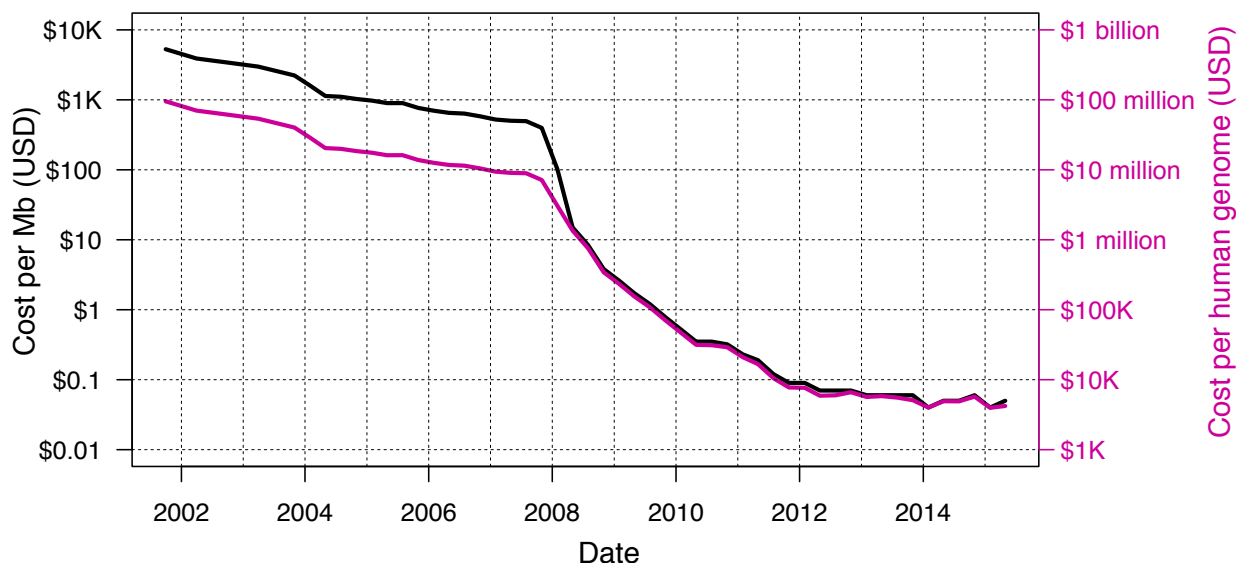


Figure 1.1: DNA sequencing costs from September 2001 to April 2015.

These plots are based on data from the National Human Genome Research Institute (Wetterstrand, 2015). The black line represents the cost per megabase (Mb) of sequence (lefthand Y-axis), and the pink line shows the cost for a whole human genome (righthand Y-axis). The data from 2001 till October 2007 represent the costs of “first-generation” technologies (Sanger-based chemistries and capillary-based instruments), assuming that 6X coverage with average read length 500-600bp was required for re-sequencing a human genome of size 3Gb. From January 2008 onwards, they represent the costs of “next-generation” sequencing platforms, assuming 30X coverage of Illumina or SOLiD sequencing (average read length 75-150bp) or 10X coverage with 454 sequencing (average read length 300-400bp). The costs include reagents, library preparation, sequencing instruments (spread over three years), and informatics related to sequence production. Note that the current costs according to Wetterstrand are rather than higher than those quoted by sequencing centres, which are closer to \$1500 on the Illumina X10.

describe some examples of this, conducted as part of WGS500, a large clinical whole-genome sequencing (WGS) study in Oxford. In Chapter 3, we examine some more general lessons learned in WGS500 about the merits of different study designs and analysis strategies.

In addition to these applications in the study of human disease, whole-genome SNP genotyping and sequencing have also been applied in population genetics, for example, to infer the population structure and demographic history of humans and other organisms. A review of this is given in Section 1.3, and in Chapter 4, we describe a WGS study of the platypus. Finally, these genomic approaches have facilitated research into the fundamental molecular process of meiotic recombination. We introduce this in Section 1.4, and in Chapter 5, we describe a study of patterns of recombination in human pedigrees.

1.2 Finding genes for rare disease

The first successful attempts to find genes causing Mendelian disorders required extensive prior knowledge of the disease biology, for example, the role of blood type antigens in peptic ulceration (Aird *et al.*, 1954), and of the haemoglobin β protein in sickle-cell anaemia (Ingram, 1956). Botstein *et al.* (1980) proposed that restriction fragment length polymorphisms (RFLPs), discovered simultaneously by Lawn *et al.* (1978) and Kan and Dozy (1978), could be used to construct a linkage map for testing the co-segregation of genetic markers with a disease within families. The first disease gene mapped was the one for Huntington's disease (HD), identified by Gusella *et al.* (1983). They found an RFLP marker linked to the disorder in two large families and mapped this to chromosome 4 using human-mouse somatic cell hybrids. It was another ten years before the causal mutation in this gene was actually identified as an unstable triplet repeat (MacDonald *et al.*, 1993).

Linkage followed by positional cloning was the standard method for finding Mendelian disease genes in the 1980s, 1990s and early 2000s (reviewed by Duncan *et al.*, 2014). The first step was to genotype markers (initially RFLPs, later microsatellites, then SNPs) in family members to identify a region that showed statistically significant linkage with the disease, indicating that it was close to the disease-causing gene. Then,

since there was no human genome sequence available, partially overlapping DNA fragments were isolated from genomic libraries to progress along the chromosome towards the gene (chromosome walking), which was eventually sequenced in cases and in controls to identify the causal mutation(s) (Puliti *et al.*, 2007). Some early successes were chronic granulomatous disease (Royer-Pokora *et al.*, 1986), Duchenne muscular dystrophy (Monaco *et al.*, 1986) and retinoblastoma (Lee *et al.*, 1987). These discoveries were facilitated by the presence of deletions of part of the gene. In contrast, the cystic fibrosis (CF) gene was the first to be discovered by positional cloning without the availability of such structural variants. The CF locus was mapped to 7q31 by three groups simultaneously (Beaudet *et al.*, 1986; Knowlton *et al.*, 1985; White *et al.*, 1985), and Riordan *et al.* (1989) then isolated the gene from cDNA libraries and identified the first disease-causing mutation.

The earliest human genetic map, which was based on ~ 400 RFLP markers (Donis-Keller *et al.*, 1987), facilitated the identification of 42 Mendelian disease genes by 1995 (see review by Collins (1995)). Linkage studies using this map usually resulted in regions 10-20cM (10-20Mb) long, which often contained hundreds of genes. In the 1990s, RFLPs began to be replaced by microsatellites, discovered by two groups simultaneously in 1989 (Litt and Luty, 1989; Weber and May, 1989). These markers had many more alleles and were hence more informative for linkage (Weissenbach, 1993). By 2001, at least one disease-related mutation had been recorded in 1100 genes in OMIM (Online Mendelian Inheritance in Man) (Peltonen and McKusick, 2001). The publication of the human genome sequence (Lander *et al.*, 2001) subsequently greatly accelerated the gene discovery process by alleviating the need for painstaking chromosome walking. High density multiplex SNP genotyping began to be widely employed for linkage studies of both rare and common disease (Craig and Stephan, 2005), since it had a higher and more evenly distributed information content than the panels of ~ 400 microsatellites typically used previously (Evans and Cardon, 2004; Schaid *et al.*, 2004).

Another microarray technology developed in the late 1990s was comparative genome hybridisation (arrayCGH) (Pinkel *et al.*, 1998), which is used to detect large copy number variants (CNVs). Multiple structural variants linked to Mendelian disease had already been detected using traditional cytogenetics, and under linkage peaks using fluorescence *in situ* hybridization (FISH), pulsed-field gel electrophoresis (PFGE), or quantitative Southern analysis (reviewed in Hurles *et al.*, 2008; Stankiewicz and

Lupski, 2010). Examples include duplications of *PMP22* in Charcot-Marie-Tooth syndrome type 1A (Lupski *et al.*, 1991) and deletions of *ELN* in Williams-Beuren syndrome (Ewart *et al.*, 1993). ArrayCGH allows more rapid detection of structural variants in large cohorts of patients. High-throughput screening in clinical genetics laboratories led to the discovery of several new syndromes (“genomic disorders”) caused by particular recurrent rearrangements, such as the deletion of 16p11.2 in autism (Marshall *et al.*, 2008; Kumar *et al.*, 2008; Weiss *et al.*, 2008), and reciprocal deletions and duplications of 1q21.1, which cause microcephaly or macrocephaly and developmental and behavioural abnormalities (Brunetti-Pierri *et al.*, 2008). These rearrangements are often caused by nonallelic homologous recombination during meiosis (Turner *et al.*, 2008).

Despite the many discoveries made through positional cloning and arrayCGH, by 2010, only 43% (2893 out of 6641) of all known or suspected Mendelian disorders had been attributed to mutations in a specific gene (Kaiser, 2010). The traditional linkage approach was not tractable for disorders in which only a few small affected families were available, or which were due to *de novo* mutations or had reduced reproductive fitness, lower penetrance, or locus heterogeneity. Thanks to the development of high-throughput sequencing technologies and targeted exon capture methods, whole-exome sequencing (WES), and, to an extent, WGS, greatly accelerated the pace of gene discovery from 2009 onwards. The mean number of new genes discovered per year increased from ~ 166 between 2005 and 2009, to ~ 236 between 2010 and 2014 (Chong *et al.*, 2015).

The first use of WES was to sequence the exomes of eight HapMap individuals and four unrelated patients with the dominant disorder Freeman-Sheldon syndrome (FSS) (Ng *et al.*, 2009). The study demonstrated that *MYH3*, the gene known to cause FSS, could be easily identified, as it was the only gene containing a coding mutation in all patients that was absent from dbSNP (Sherry *et al.*, 2001) and the eight HapMap exomes. The following year, the same authors reported the first discovery of new Mendelian disease gene using WES. They identified *DHODH* as the gene causing Miller syndrome by sequencing the exomes of four affected individuals in three families (Ng *et al.*, 2010a). Since then, hundreds of studies have applied WES or, in some cases, WGS, to various Mendelian disorders, shedding light on their pathogenesis and also revealing more about the normal function of many genes (reviewed in Ng *et al.*, 2010b; Bamshad *et al.*, 2011; Ku *et al.*, 2012; Biesecker and Green, 2014; Chong *et al.*,

2015).

These technologies have been particularly useful for studying disorders caused by *de novo* mutations (reviewed by Veltman and Brunner, 2012; Ronemus *et al.*, 2014), such as Schinzel-Giedion syndrome (Hoischen *et al.*, 2010b) and Baraitser-Winter syndrome (Riviere *et al.*, 2012), which did not recur in families and thus would have been intractable to linkage. A few studies have also reported mosaic disease-causing mutations that occurred postzygotically but early in development, for example in brain malformation syndromes (Pagnamenta *et al.*, 2012b; Riviere *et al.*, 2012; Jammuar *et al.*, 2014), overgrowth syndromes (Lindhurst *et al.*, 2011, 2012) and skin disorders (Couto *et al.*, 2015; Levinsohn *et al.*, 2015). These examples show the efficacy of next-generation sequencing for detecting mutations present at a low level or only in a single tissue, with the proportion of reads carrying the alternate allele typically ranging from 10% to 20% (Yang *et al.*, 2014).

Several groups have managed to identify disease-causing mutations by sequencing only a single proband from a consanguineous family, combined with homozygosity mapping using SNP chips. For example, Bilgüvar *et al.* (2010) identified mutations in *WDR62* as the cause of various cerebral cortical malformations by sequencing one of two affected individuals from a small Turkish family. They subsequently found other *WDR62* mutations in unrelated patients with a similar phenotype. Cabral *et al.* (2012) found a putatively pathogenic mutation in *CHST8* in a consanguineous family with peeling skin syndrome, and then provided further evidence for its causality through experiments in keratinocytes. In a large study of 136 consanguineous families with autosomal-recessive intellectual disability (ID), Najmabadi *et al.* (2011) performed targeted capture of exons within homozygous linkage intervals. They identified likely causal mutations in known ID genes in 26 families, and putatively pathogenic mutations in 50 new genes in 52 families. Only two of the new genes had mutations in more than one family, and Najmabadi *et al.* regarded the rest as tentative, given the absence of supporting functional data.

Other large studies of disorders with high locus heterogeneity have been similarly cautious. Three papers published in parallel in 2012 conducted WES on several hundred trios in which the child had sporadic autism (O’Roak *et al.*, 2012b; Sanders *et al.*, 2012; Neale *et al.*, 2012). They each reported only a handful of genes with recurrent *de novo* mutations that they thought were likely to be causal (or at least contribute to

the phenotype), but they demonstrated an enrichment for *de novo* mutations in candidate autism genes, genes in particular pathways, and genes expressed in the brain. In a similar study, Gilissen *et al.* (2014) sequenced the whole genomes of 50 trios in which the child had intellectual disability (ID). They found significant enrichment for *de novo* coding SNVs and CNVs in genes previously linked to ID-related disorders, compared to healthy controls from published studies. They reached conclusive diagnoses for twenty patients who had *de novo* mutations affecting known or candidate ID genes.

For Mendelian diseases, the rate of diagnosis based on phenotypic features and conventional testing is only ~50% (Shashi *et al.*, 2013). A particularly fruitful application of WES/WGS has been for improving diagnosis (reviewed in Ku *et al.*, 2012; Boycott *et al.*, 2013). In an attempt to evaluate this, Need *et al.* (2012) sequenced twelve patients with undiagnosed and suspected genetic conditions, and were able to make a likely genetic diagnosis in six. Three of these findings revealed that the patients actually showed an atypical presentation of a particular syndrome known to be caused by mutations in that gene (*TCF4* - Pitt-Hopkins syndrome; *SMAD4* - Myhre syndrome; *EFTUD2* - mandibulofacial dysostosis). Another broadened the set of phenotypes already associated with that gene (*SCN2A* - seizure disorders, sometimes accompanied by severe intellectual disability). WES had been particularly useful in this respect, since previously, clinical geneticists ascertained patients based on similarity of presentation, which tended to bias away from discovering pleiotropic effects of genes (Beaulieu *et al.*, 2014; Chong *et al.*, 2015).

Conversely, WES also greatly helped in disorders that can be caused by mutations in any one of many genes, which are laborious to screen clinically with standard Sanger sequencing (Ku *et al.*, 2012). For examples, more than 80 genes are known to cause Charcot-Marie Tooth (CMT) disease (Timmerman *et al.*, 2014), and 70% of patients in CMT type 2, the less severe form, still have no known genetic cause. The traditional molecular diagnostic strategy was a one-by-one approach. In contrast, WES and WGS have been used to rapidly identify causal mutations in known genes, *GJB1* (Montenegro *et al.*, 2011) and *SH3TC2* (Lupski *et al.*, 2010). To increase the number of patients that can be screened given limited resources, targeted gene panels have been developed for several diseases, such as congenital disorders of glycosylation (Jones *et al.*, 2011), primary ciliary dyskinesia (Berg *et al.*, 2011), epilepsy (Carvill

et al., 2013; Della Mina *et al.*, 2015) and autism (O’Roak *et al.*, 2012a).

A small number of discoveries from WES/WGS studies have led to better treatments (reviewed in Boycott *et al.*, 2013). In some cases, the genetic findings prompted the use of an existing pharmacological therapy, resulting in clinical improvements (Bainbridge *et al.*, 2011; St. Hilaire *et al.*, 2011; Bosch *et al.*, 2012; Rilstone *et al.*, 2013; Venkateswaran *et al.*, 2014). Another oft-quoted example is that of Worthey *et al.* (2011), who sequenced the exome of an infant with a Crohn’s disease-like illness, and found a novel, hemizygous missense mutation in the *XIAP* (X-linked inhibitor of apoptosis) gene, which plays a role in the NOD signalling pathway, involved in inflammatory bowel disease. Following this discovery, the child received an allogeneic haematopoietic progenitor cell transplant, the recommended treatment for XIAP deficiency, and this led to remission of his gastrointestinal disease.

The standard approach for finding causal mutations in these studies has been to filter for coding variants (nonsynonymous, nonsense, frameshifting, or splice site-altering) that are rare or absent in healthy normal people, then restrict to those that fit the right mode of inheritance, using family data when available. Early on, variants were often ranked by conservation scores and by the severity of their predicted effect on protein structure (Ng *et al.*, 2010b) using methods such as PolyPhen (Adzhubei *et al.*, 2010) and MutationTaster (Schwarz *et al.*, 2014). However, these have become less popular due to their conflicting results (Li *et al.*, 2010; Thusberg *et al.*, 2011), lack of discriminative power (Dong *et al.*, 2015), and because we can now exclude more variants on the basis of their frequency in public databases. Initially, dbSNP (Sherry *et al.*, 2001) and the 1000 Genomes pilot data (1000 Genomes Project Consortium *et al.*, 2010) were used to check the frequency of variants. The number of publicly available healthy controls greatly increased in subsequent phases of 1000 Genomes, and thanks to the release of data from the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP)² (Tennessen *et al.*, 2012; Fu *et al.*, 2013), and, recently, the Exome Aggregation Consortium (ExAC; <http://exac.broadinstitute.org/>). After bioinformatic filtering to find candidates, researchers have typically either sought genetic validation by identifying other patients

²The NHLBI ESP was intended to study complex heart, lung and blood disorders in European and African Americans, and contains a mixture of controls, individuals with the extremes of some traits (lipids and blood pressure), and patients with specific diseases (early onset myocardial infarction and early onset stroke). It is generally regarded as a reasonable control set for severe Mendelian disease unrelated to these phenotypes.

with plausibly causal mutations in the same gene who showed a similar phenotype, or by functionally testing candidates in cell lines or model organisms. Several databases have been established for sharing phenotypic and genotypic data, such as the Leiden Open Variation Database (Fokkema *et al.*, 2011), DECIPHER (Swaminathan *et al.*, 2012), GeneMatcher (Sobreira *et al.*, 2015) and PhenomeCentral (Buske *et al.*, 2015). These have facilitated candidate identification (e.g. Utami *et al.*, 2014) and the validation process by connecting researchers who have patients with similar features or with mutations in the same gene (e.g. Barge-Schaapveld *et al.*, 2011; Thevenon *et al.*, 2013; Au *et al.*, 2015).

In 2011 and 2012, several large WES studies were launched to evaluate the clinical and diagnostic utility of this technology in a variety of rare diseases. These include the Baylor College of Medicine WES study (Yang *et al.*, 2014), the FORGE (Finding of Rare Disease Genes) Canada consortium (Beaulieu *et al.*, 2014), and the Deciphering Developmental Disorders (DDD) study (Wright *et al.*, 2015; Deciphering Developmental Disorders Study, 2015). The Baylor study sequenced 2,000 patients, and reached a molecular diagnosis in 504 (25.2%)³. Strikingly, 30% of these mutations were in disease genes reported since 2011, emphasising both the pace of new discoveries and the importance of consulting recent literature. The FORGE Consortium investigated 264 disorders (783 samples), and identified disease-causing variants in 146 (55%) of these, including 67 novel genes of which 41 have been either genetically or functionally validated. They found that the most effective strategies were either to sequence multiple unrelated individuals with the same very rare but highly recognisable clinical condition, and look for pathogenic variants in a common gene (30/32 disorders solved), or, for consanguineous families with mapping data, to sequence one or two affected individuals and focus on mapped regions (42/60 disorders solved). The DDD study conducted WES on 1,133 trios in which the child (and in ~15% of trios, one or both parents) had an undiagnosed developmental disorder. Focusing on known developmental disorder genes, Wright *et al.* (2015) reported a diagnostic rate of 27%, which increased to 31% when 12 novel genes were included (Deciphering Developmental Disorders Study, 2015). These novel genes were identified through a combination of high-throughput gene knock-down experiments in zebrafish, and statistical tests that used estimates of gene-specific mutation rates (Samocha *et al.*, 2014) to construct the null distribution for the expected number of *de novo* variants

³A molecular diagnosis was deemed to be made if a pathogenic or likely pathogenic variant was detected in a known gene for a Mendelian disease that overlapped the phenotype of the patient.

in each gene.

In 2010, researchers in Oxford launched the WGS500 project in collaboration with Illumina. Conducted in parallel with the large WES studies described above, this project aimed to sequence the genomes of 500 people with a variety of clinical phenotypes, including Mendelian syndromes, immunological disorders and cancers. It was decided to sequence whole genomes rather than just exomes for a number of reasons. Firstly, there are problems associated with the target capture technology used in WES (Asan *et al.*, 2011): it produces poor coverage of some splice sites and exons, often those with high GC content (common at the start of genes) (Hoischen *et al.*, 2010a; Manase *et al.*, 2014), and in some cases, only the main transcript of a gene is targeted so some exons are missing entirely. Secondly, WGS is better for the detection of structural variants, particularly those not associated with gain or loss, such as inversions or translocations. Lastly, it was anticipated that some of the cases, particularly in the more complex immunological disorders, would be due at least in part to noncoding mutations, which are obviously inaccessible to WES. The WGS500 project included 94 independent cases or families in 31 case studies of Mendelian disorders, several of which are presented in Chapter 2 of this thesis. It also provided an opportunity to evaluate the efficacy of different analysis strategies, as described in Chapter 3.

There is a lack of consensus in the field about the difference between a “variant” and a “mutation”, and the correct use of these terms to describe a change in DNA that causes or contributes to a disease. In this thesis, for a *de novo* variant, we shall use “mutation” when it is thought likely to contribute to a patient’s disease, and “variant” when there is no specific implication of involvement in a disease. For inherited homozygous variants, in some cases, we still refer to likely pathogenic ones as “variants”. Sometimes, we will begin by talking about a particular “variant” and then, having presented evidence for its pathogenicity, we will call it a “mutation”. If a mutation is described as “causal”, we mean to imply that we think there is reasonable evidence that it is responsible for a patient’s disease. By “pathogenic”, we imply that the mutation is either fully or partially responsible for the disease.

1.3 Population genetics

Genomic methods have also had a significant impact in the study of the structure and history of populations. Until the early 2000s, population genetic inference was mainly based on the distribution of markers such as HLA types or blood groups (Cavalli-Sforza and Bodmer, 1999), matrilineal mitochondrial DNA (mtDNA) or the patrilineal Y chromosome (reviewed in Castro *et al.*, 1998; Stumpf and Goldstein, 2001). For example, mitochondrial RFLPs were used to infer an African root in the human mtDNA tree (Cann *et al.*, 1987) and to study genetic structure in different *Drosophila* species (Hale and Singh, 1987; Afonso *et al.*, 1990). Many groups performed genealogical inference using microsatellites or a few hundred SNPs in the non-recombining part of the human Y chromosome (NRY) (e.g. Underhill *et al.*, 2000; de Knijff, 2000), and these data were used to confirm the African origin of anatomically modern humans (Seielstad *et al.*, 1999). The advantages of data from mtDNA and the Y chromosome were that the absence of recombination allowed reconstruction of a gene tree⁴, that the small effective population size (N_e) allowed better discrimination between populations than did autosomal DNA, and that it was low cost and could thus be collected from a large number of individuals (Veeramah and Hammer, 2014). However, the fact that these sequences represent a single genealogy means that they provide limited information about the underlying population history, and because genealogies are stochastic given a particular demographic model, a genealogy built from a single locus may not reflect the historical relationships between populations under study (Novembre and Ramachandran, 2011). Indeed, in a recent paper analysing 940 samples from the Human Genome Diversity Panel (HGDP), Emery *et al.* (2015) demonstrated that mtDNA-haplogroup membership was insufficient information even to assign individuals to continental groups. Thus, approaches which considered multiple autosomal loci have become more widely used to infer demographic parameters and population structure (Rosenberg and Nordborg, 2002).

One of the most popular methods to examine population structure was developed by Pritchard *et al.* (2000). This method, STRUCTURE, is a model-based clustering algorithm that identifies a specified number of subgroups (clusters) with distinct allele frequencies, and assigns individuals to those clusters. In the admixture version of the model, individuals can have membership in multiple clusters, with admixture proportions summing to 1. Rosenberg *et al.* (2002) applied this method to infer

⁴A phylogenetic tree showing the evolutionary relationships between DNA sequences.

population structure in over 1000 individuals from 52 human populations, using 377 autosomal microsatellites. They showed that these clusters generally corresponded to major geographic regions, but that some individuals had partial membership in multiple clusters, reflecting either admixture or continuous gradations in allele frequencies across regions. Since then, STRUCTURE and extensions to it (Falush *et al.*, 2003a; Alexander *et al.*, 2009; Hubisz *et al.*, 2009) have been used in thousands of studies of multiple different species, ranging from chickens (Rosenberg *et al.*, 2001) to *Helicobacter pylori* (Falush *et al.*, 2003b) and sunflowers (Harter *et al.*, 2004).

The development of affordable SNP arrays was catalysed by the International HapMap Project, which aimed to discover SNPs and describe patterns of linkage disequilibrium (LD) in different human populations in order to choose tag SNPs for association studies (International HapMap Consortium, 2005). Genome-wide SNP analyses began to be widely used for investigating genetic diversity, adaptation and population structure (reviewed in Pool *et al.*, 2010; Novembre and Ramachandran, 2011; Crisci *et al.*, 2012). In a landmark paper, Novembre *et al.* (2008) proposed an alternative to inference with the admixture model of STRUCTURE, demonstrating that Principal Component Analysis (PCA) could be used to describe continuous forms of population structure more flexibly. They showed that a plot of the first two principal components (PCs) based on $\sim 500,000$ SNPs in $\sim 3,200$ European individuals from the Population Reference Samples (POPRES) bore a striking resemblance to the map of Europe. Novembre and Stephens (2008) argued that PCs correlated with geography cannot necessarily be interpreted as migrations, as originally argued by Cavalli-Sforza (1994), who had applied PCA to protein polymorphisms, but rather may simply reflect isolation-by-distance (i.e. there is only gene exchange among neighbouring populations) (see News and Views by Reich *et al.*, 2008). They also pointed out the dependence of PCA results on the sample sizes and distribution of sampling locations. Despite these drawbacks, PCA became widely used to provide a multi-dimensional summary of population genetic data, and particularly to account for population structure in GWAS (Tian *et al.*, 2008).

Several haplotype-based methods have also been developed to infer population structure and perform demographic inference. These are often based on the model of Li and Stephens (2003), in which each individual’s phased haplotypes are modelled as a mosaic of chunks “copied” from donor individuals, separated by recombination events (now referred to as “chromosome painting”). Price *et al.* (2009) used this in

their method, HAPMIX, which infers local ancestry along chromosomes using unphased data from an admixed individual and phased haplotypes from one or two reference populations. Lawson *et al.* (2012) built a coancestry matrix summarising the chromosome painting, and developed a method to partition the dataset into clusters and assign individuals to these. This method was used by Leslie *et al.* (2015) to infer detailed population structure in the British Isles, and to explore patterns of shared ancestry between the different British groups and populations from continental Europe, which revealed signals of historical migrations.

Inferring historical demography from genomic data is of considerable interest, both for its own sake, and in order to construct a realistic null model that can be used to reveal outlier loci that may be targets of selection⁵. Many methods exist for doing demographic inference (for a recent review, see Gattepaille *et al.*, 2013). One class of methods is based on the site frequency spectrum (SFS). An example is *δaδi* (Gutenkunst *et al.*, 2009), which uses a diffusion approach to approximate the joint multi-population SFS, and then infers parameters of the demographic model (which may include expansions, contractions, migrations and admixture) using a maximum composite likelihood. A drawback of SFS-based methods is that their application to SNP chip data is rendered difficult by the ascertainment bias inherent in how the SNPs were chosen for the arrays, which is difficult to correct for (Veeramah and Hammer, 2014). There is an alternative set of methods based on LD and haplotypic summary statistics. For example, Atzmon *et al.* (2010) exploited the relationship between the frequency and length of identical-by-descent (IBD) tracts to provide evidence for a bottleneck in Ashkenazi Jews twenty-three generations ago that was followed by a large expansion. McEvoy *et al.* (2011) analysed patterns of LD at different genetic distances to infer changes in human effective population size over time.

Sequence data avoid the problems of ascertainment bias associated with SNP chips, and several methods exist for using them to infer demographic parameters. These include nonparametric “skyline plot” methods that estimate historical effective population size (N_e) trajectories by reconstructing the underlying genealogy from nonrecombining loci, such as mtDNA (reviewed in Ho and Shapiro, 2011). These methods are too computationally intensive when the number of loci grows, and so

⁵Many studies have used genomic data to search for signatures of various types of natural selection (e.g. reviewed in Scheinfeldt and Tishkoff, 2013). These are not used in this thesis and therefore not described in this chapter.

new approaches were needed in the next-generation sequencing era. A landmark paper by Li and Durbin (2011) demonstrated that it was possible to estimate historical N_e using a single diploid genome sequence. Their method, the Pairwise Sequentially Markovian Coalescent (PSMC), examines how the local density of heterozygous sites changes along the genome, reflecting segments of constant time to most recent common ancestor ($T_{MRC A}$) between the two chromosomes separated by historical recombination events. Knowing the coalescence rate in a particular epoch allows estimation of N_e at that time. This method has been extensively used to study population size changes in different human groups and in other species, but, as is often emphasised (Veeramah *et al.*, 2013), care should be taken when interpreting the results, since N_e estimates can be affected by other demographic processes other than true growth or shrinkage, such as admixture. Additionally, false positive or false negative variant calls in the sequence data can lead to misestimation of the times, since they alter the effective mutation rate. Also, PSMC is not informative about the very recent or very ancient past, since there are insufficient coalescences between two sequences in those epochs. Several other new methods have been developed that extend this approach to multiple individuals in order to look at more recent time points (Sheehan *et al.*, 2013; Schiffels and Durbin, 2014).

The reduced cost of high-throughput sequencing has facilitated many studies into the genetic diversity and population structure of humans and other organisms. The 1000 Genomes Project (1000 Genomes Project Consortium *et al.*, 2010, 2012) and the NHLBI ESP (Fu *et al.*, 2013) described the distribution of variants of different functional classes and frequencies across various human populations. In one of the earliest studies to use Illumina short-read sequencing on non-humans, Xia *et al.* (2009) conducted low-coverage sequencing of 40 domesticated and wild silkworm genomes, and used the patterns of genetic diversity to infer the size of the bottleneck that occurred during domestication, and to identify genomic regions that may have been under selection. Ai *et al.* (2015) undertook a similar project in pigs, and identified highly differentiated loci that they hypothesised may have played a role in regional adaptations to different latitudes. Some of these loci showed evidence of having been introgressed from an extinct pig species. There have also been several population sequencing studies in primates. Auton *et al.* (2012) studied patterns of fine-scale recombination in chimpanzees, and Leffler *et al.* (2013) used the same WGS data to identify signatures of long-term balancing selection. Prado-Martinez *et al.* (2013) sequenced the genomes of 79 individuals from six great ape species, demonstrated

genetically distinct populations within each species, and noted how fluctuating population size over time had affected genetic diversity. Xue *et al.* (2015) extended this work in gorillas, and discovered high homozygosity and an increased load of deleterious variation in mountain gorillas, resulting from prolonged population decline.

In Chapter 4 of this thesis, we describe our work on population sequencing of the platypus, a member of the monotremes, a group of egg-laying mammals. This is the first study to investigate genetic diversity in this species using whole-genome data.

1.4 Meiotic recombination

Finally, genomic approaches have allowed investigation of the process of meiotic recombination, both at a molecular and population scale. Recombination, the exchange of DNA between homologous chromosomes during meiosis, is important for several key reasons. Firstly, it produces different combinations of alleles at different loci, creating more diversity in the population for positive selection to act upon. It also increases the ability of selection to eliminate deleterious mutations (Muller, 1964; Hill and Robertson, 1966) and incorporate beneficial ones (Fisher, 1930; Muller, 1932).⁶ The other role of recombination, to ensure the correct segregation of chromosomes into the gametes (Smith and Nicolas, 1998), is most relevant to the work described in Chapter 5.

The early recombination maps in humans were based on pedigrees. Polymorphic markers were genotyped in families and recombination events were inferred based on the patterns of allele transmission to offspring. The first of these maps was made from ~400 RFLPs genotyped in twenty-one three-generation families. Donis-Keller *et al.* (1987) combined mathematical linkage analysis with the physical localisation of clones to arrange these markers into 23 linkage groups, corresponding to the 23 human chromosomes. This RFLP map was superseded in 1996 by a 5,264-marker microsatellite map (Dib *et al.*, 1996). The Marshfield map (Broman *et al.*, 1998), based on >8000 microsatellites and 188 meioses, became the gold standard until Kong *et al.*

⁶Much has been written about the importance of recombination for adaptation, and its effect on genetic diversity and base composition (see Felsenstein, 1974; Otto and Lenormand, 2002; Keightley and Otto, 2006; Spencer *et al.*, 2006; Coop and Przeworski, 2007; Sella *et al.*, 2009; McGaugh *et al.*, 2012; Gossmann *et al.*, 2014), but since this is not directly relevant to the work in this thesis, it will not be reviewed here.

(2002) published a much higher-resolution map constructed from 5,136 microsatellites and 1,257 meiotic events in the deCODE dataset. The most recent deCODE map, based on genome-wide SNP data from 15,257 meioses (Kong *et al.*, 2010), has a resolution of $\sim 10\text{kb}$, 300 times better than the 2002 map.

Methods based on LD are an alternative to pedigree approaches, and allow estimation of historical recombination rates on a scale of kilobases from population genetic SNP data. One such approach, LDHat (McVean *et al.*, 2002, 2004), relies on the composite likelihood method of Hudson (2001) to allow different recombination rates between each pair of SNPs, and then estimates a variable recombination rate using a Markov Chain Monte Carlo scheme. This was first applied to the whole human genome by Myers *et al.* (2005), and subsequently to the HapMap data (International HapMap Consortium, 2005). Most recently, it has been used to investigate recombination in chimpanzees (Auton *et al.*, 2012), other primates (Stevenson *et al.*, 2015) and dogs (Auton *et al.*, 2013).

Both pedigree- and LD-based methods have revealed much about the recombination process. Whereas pedigree studies are useful for characterising variation among individuals and for comparing patterns between males and females, LD-based methods estimate rates that are averaged across both sexes and reflect historical recombination events, and are typically finer-scale than estimates from families (Auton and McVean, 2012). We will review some of the key findings of both methods here.

In mammals, the genetic map length tends to be proportional to the number of chromosome arms (Pardo-Manuel de Villena and Sapienza, 2001), although a more recent study in humans (Fledel-Alon *et al.*, 2009) suggested that only one crossover per chromosome rather than per arm is needed to ensure proper segregation, and that additional crossovers occur in proportion to the chromosome length. Crossovers do not occur independently along chromosomes, but are rather subject to interference, such that a crossover at one location discourages others from occurring nearby in the same meiosis (Broman and Weber, 2000).

The early pedigree-based linkage maps showed that human recombination rates vary by an order of magnitude over megabase scales along chromosomes, and are higher nearer the telomeres than the centromere (Kong *et al.*, 2002). They are also highly variable on finer scales. Sperm-typing experiments, in which individual or

pooled sperm are analysed using a PCR-based method, found that most crossovers are concentrated in short segments of 1-2kb known as “hotspots” (Arnheim *et al.*, 2003; Jeffreys *et al.*, 2001). Studies based on LD patterns found that there is an historical hotspot about every 50kb in humans, and 90% of crossovers occur in 10-20% of the sequence (Myers *et al.*, 2005). These hotspots were found to be determined, at least in humans and mice, by a zinc-finger protein, PRDM9, which binds to particular motifs in the DNA and trimethylates histone H3K4, promoting double-stranded break formation in those regions (Baudat *et al.*, 2010; Myers *et al.*, 2010; Parvanov *et al.*, 2010). It is now thought that all, or almost all, autosomal recombination in humans occurs in hotspots (Simon Myers, personal communication).

There is extensive variation in recombination patterns between sexes and individuals, and between gametes from the same individual (Broman *et al.*, 1998; Kong *et al.*, 2002; Coop and Przeworski, 2007; Kong *et al.*, 2010). These differences were revealed in pedigree studies and through experiments in human oocytes and sperm, in which antibodies were used to visualise crossover-associated proteins such as MLH1 (e.g. Tease and Hulten, 2004; Lenzi *et al.*, 2005). It should be noted, however, that many of the gametes observed in MLH1 analyses may not lead to viable offspring. Pedigree studies give a more accurate sense of variability in live births.

The genetic map is 1.6 times longer in human females than in males (Broman *et al.*, 1998; Kong *et al.*, 2002; Cheung *et al.*, 2007), with average crossover counts typically estimated at about 39-42 per meiosis in females and 25-27 in males, depending on the resolution of markers used. The number of crossovers is also more variable in females than males (26-58 per meiosis *versus* 18-35) (Fledel-Alon *et al.*, 2011; Hussin *et al.*, 2011), and has a higher heritability (0.25 in females *versus* 0.14 in males) (Fledel-Alon *et al.*, 2011). Variation in several genes, including *PRDM9* (Hinch *et al.*, 2011) and *RNF212* (Kong *et al.*, 2008; Fledel-Alon *et al.*, 2011), has been associated with mean recombination rate. In the largest study to date, Kong *et al.* (2014) reported thirteen variants associated with recombination rate, of which ten showed an effect only in one sex.

Another factor that has been reported to affect female but not male recombination rates is age (Kong *et al.*, 2004; Coop *et al.*, 2008; Hussin *et al.*, 2011; Bleazard *et al.*, 2013). It has been proposed that this may be related to the increase in aneuploidy rates in the offspring of older mothers (Henderson and Edwards, 1968). However, the

evidence about the direction and magnitude of the maternal age effect on the number of crossovers is conflicting. We revisit this controversial question in Chapter 5 of this thesis.

Chapter 2

Case studies from the WGS500 project

2.1 Introduction

This chapter and the next describe work conducted as part of WGS500, a large project that began in Oxford in 2010 with the aim of sequencing the genomes of 500 people with different diseases in order to evaluate the clinical utility of whole-genome sequencing (WGS). Here, we focus on case studies of some Mendelian neurological disorders, and in the next chapter we describe an evaluation of different strategies for pinpointing the mutations causing putatively monogenic Mendelian diseases.

Clinical geneticists in Oxford recruited nine families of which one or more members had a severe, early-onset neurological disease (Table 2.1). Patients 1-6 had no affected relatives, suggesting a *de novo* mutation or recessive inheritance pattern. The remaining families had more than one affected child, so a recessive mutation was suspected. The whole genomes of the healthy parents and affected children from Families 1-8 and of a single affected individual from Family 9 were sequenced to at least 25X coverage. Each family was screened for candidate *de novo*, simple recessive, compound heterozygous and X-linked variants as appropriate. Since previous applications of whole-exome sequencing (WES) and WGS for finding disease genes were reviewed in Chapter 1, in this chapter we will simply introduce and report results from each case study in turn.

2.2 Methods

2.2.1 Case recruitment

The families were recruited to participate in WGS500 through the Oxford Clinical Genetics department. A summary of the main clinical features is given in Table 2.1, along with a list of the genetic tests the patients had undergone before entering this study (all of which were negative).

2.2.2 Sequencing and variant calling pipeline

These methods are described in detail in the WGS500 paper (Taylor *et al.*, 2015) and thus are described only briefly here. WGS was conducted on the Illumina HiSeq platform to a coverage of at least 25X. The reads were mapped to the human reference genome (build 37d5) with Stampy (Lunter and Goodson, 2011) and single nucleotide variants (SNVs) and small indels were called with Platypus (Rimmer *et al.*, 2014). Variants were annotated relative to RefSeq transcripts using ANNOVAR (Wang *et al.*, 2010) and relative to Ensembl transcripts using an in-house tool called VariantAnno, similar to the Ensembl Variant Effect Predictor, written by Peter Humburg.

Several methods were used to search for copy number variants (CNVs). These included a) generating count profiles in 10kb bins along the genome and correcting for noise due to biases in sequence composition, b) OncoSNP-SEQ (Yau, 2013) and c) ExomeDepth (Plagnol *et al.*, 2012). These methods are described in more detail in Taylor *et al.* (2015). In addition to these, GenomeSTRiP (Handsaker *et al.*, 2011) was used to search for intermediate-sized deletions (50-50,000bp) that would have been missed by arrayCGH and Platypus. We filtered the calls based on their frequency in other WGS500 samples, using the same criteria as for the SNV/small indel calls, described in the next section. We considered the closest gene to each deletion as the affected gene, and, for the compound recessive model, treated large deletions and SNVs/small indels together.

2.2.3 Variant filtering strategies

To identify *de novo* mutations in the trios, we first screened for variants that were called as homozygous reference in the parents but heterozygous in the child. We then

Table 2.1: **Summary of case studies from WGS500.** In the trios, the child was the only affected family member. In the quartets, both children were affected. OS: Ohtahara syndrome, DD: developmental delay; NSEO: nonsyndromic early-onset epilepsy; MR: mental retardation; LD: learning disability; PDA: patent ductus arteriosis; ASD: anterior septal defect.

Family	Family members sequenced	Phenotype	Family history	Ancestry	Previous genetic tests
1	trio	OS; severe DD	no	Caucasian	arrayCGH, <i>FRAXA</i> , <i>STXBP1</i> , <i>MECP2</i> , <i>CDKL5</i> , <i>POLG</i> , <i>ARX</i>
2	trio	OS; metopic synostosis; severe DD	paternal great-grandmother and her sister had epilepsy	Caucasian	arrayCGH, <i>FRAXA</i> , <i>MECP2</i> , <i>CDKL5</i> , <i>STXBP1</i>
3	trio	OS; severe DD	no	Caucasian	arrayCGH, <i>CDKL5</i> , <i>ARX</i> , <i>STXBP1</i>
4	trio	OS; severe DD	Mothers cousin died of seizures at age 2	West African	arrayCGH, <i>MECP2</i> , <i>ARX</i> , <i>STXBP1</i>
5	trio	severe NSEO; microcephaly; severe DD	no	Indian	arrayCGH, <i>MECP2</i> , <i>UBE3A</i> , <i>TCF4</i>
6	trio	severe NSEO; severe DD; PDA and ASD as neonate; cutaneous hypopigmentation	no	Indian	arrayCGH, <i>MECP2</i> , <i>CDKL5</i> , <i>STXBP1</i>
7	quartet	severe MR, NSEO, stereotypic hand movements; mother has milder LD	no	Caucasian	<i>RET</i>
8	quartet	severe MR, DD, disturbed sleep-wake cycle	no	Caucasian	<i>FRAXA</i>
9	one of four affecteds from a consanguineous family	LD, spastic paraplegia, hypoparathyroidism	proband's sister, uncle and aunt also affected	Pakistani	none

filtered these based on the genotype log likelihood ratio (GLLR)¹, requiring this to be below -5 in all three individuals. To remove those likely to be due to technical artifacts or incorrect calling of parental genotypes, we removed variants that had been seen before in unrelated individuals in WGS500, the 1000 Genomes Project (1000G) or the NHLBI Exome Sequencing Project (ESP). The GLLR threshold of -5 was chosen because, although it was quite lenient and gave far too many putative *de novos* genome-wide even after the application of the latter filters (range 138-1388, average 546 across the six EOE trios), there was a manageable number of *de novo* coding candidates to examine (1-5 per trio), and a more stringent threshold of -10 seemed to be too strict for some trios (since the possible GLLs depend on coverage levels), giving 5-200 putative *de novos* genome-wide.

We then prioritised variants that were predicted to alter the protein sequence (nonsynonymous SNVs, stop loss or gain variants, indels, or splice site mutations), particularly those that were highly conserved across the 46 vertebrate species in the UCSC conservation track. The value of these different filters is discussed in Chapter 3 and in the WGS500 paper (Taylor *et al.*, 2015).

For the quartets, we considered the possibility that both affected children had inherited the same *de novo* mutation from one parent. This could happen if the mutation were mosaic in the parent’s germline (Campbell *et al.*, 2014). To check for this, we searched for variants that were heterozygous in both offspring but homozygous reference in both parents, with the genotype log likelihood ratio less than -5 in all individuals. In all cases, when we examined these candidate *de novo* variants in the Integrated Genomics Viewer (Robinson *et al.*, 2011), none of them looked real; all showed at least one read carrying the variant in a parent, or showed evidence of being an artefact in a difficult region e.g. with short tandem repeats.

We considered a simple recessive or compound heterozygous model in all families, and an X-linked model where appropriate, again focusing on protein-altering mutations at conserved positions. For the compound heterozygous model, we required two such variants in the same gene, one inherited from each heterozygous parent. Since the diseases under study are extremely rare (frequency $< 1/100,000$), we excluded variants with a frequency greater than 0.005 in 1000G or ESP, or for which there

¹The GLLR is the difference between the log likelihoods for the most likely and the second most likely genotypes.

were any homozygotes or hemizygotes or more than five heterozygotes amongst the other WGS500 samples ($n = 294$ at the time of analysis). We refer to variants not previously seen in dbSNP, 1000G, ESP or WGS500 as “novel”.

For families 7 and 9, we also considered noncoding variants, in the absence of any good coding candidates. We applied the frequency filters specified above and annotated rare variants using the Ensembl Variant Effect Predictor. To narrow the search, we then focused on variants that were annotated as being in regulatory regions, which were identified by the Ensembl curators using information about various chromatin marks and transcription factor binding sites, and that were within conserved blocks of DNA identified using phastCons (Siepel *et al.*, 2005).

2.2.4 Linkage analysis

In Family 9, the female proband (LDS_10) was sequenced with WGS, and her unaffected mother (LDS_30) and three affected relatives (LDS_27, LDS_28 and LDS_29) were genotyped on Illumina chips (Figure 2.8). Since the parents of both pairs of affected siblings were first cousins, we first considered the simple recessive model, focusing on regions that were identical-by-descent for both alleles (IBD2) in all the affecteds, and homozygous. However, a compound recessive model would also fit the pedigree, and so we also considered regions that were IBD2 and heterozygous respectively. To identify these regions in the genotyped individuals, we selected SNV markers for linkage with linkdatagen (Bahlo and Bromhead, 2009) and then ran MERLIN (Abecasis *et al.*, 2002) using the Kong and Cox exponential model (Kong and Cox, 1997), specifying a recessive model. We then identified regions that were also IBD2 in the sequenced individual. Candidate variants were genotyped in the other unaffected siblings by Sanger sequencing to check segregation.

2.2.5 Variant validation and functional follow-up

All laboratory work, including validation of putatively causal variants by Sanger sequencing, was completed by others, as credited in the Acknowledgements at the end of this chapter. The methods are not covered here, but those relating to the functional work on the epilepsy trios are described in Martin *et al.* (2014).

2.3 Results

2.3.1 Early-onset epilepsy

This section is largely taken from Martin *et al.* (2014).

2.3.1.1 Introduction

Severe early-onset epilepsy (EOE) encompasses a broad spectrum of phenotypes which are highly heterogeneous at the clinical and molecular levels (Nordli, 2012). Whilst some causative genes have been identified for many of these sub-phenotypes, the limitations of current technologies mean that genetic testing is largely confined to the genes associated with the specific presenting phenotype. However, it is increasingly being recognised that a given gene can cause multiple phenotypes (Nordli, 2012), and that more comprehensive genetic testing may improve molecular diagnostic yield (Kodera *et al.*, 2013). This is useful clinically not only because it can help make or confirm a diagnosis, but also because it may allow counseling on recurrence risk and prenatal testing.

In its most severe form, EOE involves frequent seizures beginning in the first three months of life, with abundant epileptic activity that contributes to significant cognitive and motor delay (Nordli, 2012; Berg *et al.*, 2010). It is frequently associated with gross structural brain abnormalities and occasionally with metabolic disorders, which are often genetic in origin (Ohtahara and Yamatogi, 2003). Ohtahara Syndrome (OS) is a severe form of EOE characterised by a distinctive electroencephalogram (EEG) pattern known as burst-suppression, which consists of periodic high voltage bursts of slow waves mixed with spikes, followed by marked attenuation (Pavone *et al.*, 2012). The frequency of OS is about 1 in 100,000 live births (Hino-Fukuyo *et al.*, 2009). Children with OS typically have multiple seizure types including tonic spasms and focal seizures, which are often refractory to anti-epileptic drugs (Tavyev Asher and Scaglia, 2012). Affected children may progress onto other epilepsy syndromes such as West Syndrome (Nordli, 2012), or they may die in infancy.

Several genes have been implicated in severe EOE. The first reported for OS was the X-linked *ARX* gene, which encodes a developmental transcription factor (Kato *et al.*, 2007). *De novo* mutations in *STXBP1* (Saitsu *et al.*, 2008; Deprez *et al.*, 2010), encoding a protein involved in synaptic vesicle release, in *CDKL5* (Melani *et al.*, 2011), encoding a serine/threonine kinase, and in ion channel genes *KCNQ2*

(Saitsu *et al.*, 2012), *SCN2A* (Nakamura *et al.*, 2013; Touma *et al.*, 2013) and *SCN8A* (Veeramah *et al.*, 2013) have also been implicated, as have recessive mutations in the glutamate transporter *SLC25A22* (Molinari *et al.*, 2005). However, many OS patients test negative for mutations in these genes, indicating that other genes have yet to be identified. Multiple additional genes have been associated with the broader range of early-onset epilepsies, including genes encoding cytoskeletal components (*SPTAN1* (Saitsu *et al.*, 2010)) and proteins involved in signaling (*PLCB1* (Kurian *et al.*, 2010)), DNA repair (*PNKP* (Shen *et al.*, 2010)), and neurotransmitter synthesis (*PNPO* (Mills *et al.*, 2005)). However, clinical testing is limited by the availability and costs of conventional single-gene tests, and thus tends to be restricted to genes that have been associated with the specific type of epilepsy. There is therefore scope to apply a more comprehensive approach to diagnosis using whole-genome methods.

WGS500 included six trios and one quartet in which the children had severe EOE and the parents were healthy. The patients were selected because traditional clinical molecular genetic approaches had failed to uncover the causal mutation. Four of the trio children had been diagnosed with OS, and the other two had severe non-syndromic early-onset epilepsy (NSEOE). The two daughters in the quartet had mental retardation and severe NSEOE.

2.3.1.2 Patients 1 and 3: *KCNQ2* and *SCN2A*

Two of the four OS cases had *de novo* nonsynonymous mutations in genes encoding ion channel subunits that had previously been implicated in EOE: *KCNQ2* and *SCN2A*. Mutations in *KCNQ2* and *SCN2A* were originally reported to cause benign neonatal seizures (Heron *et al.*, 2002; Berkovic *et al.*, 1994), but more recently they have been found in children with more severe epilepsy (Weckhuysen *et al.*, 2012; Ogiwara *et al.*, 2009). The *KCNQ2* mutation, NM_004518:c.C827T:p.T276I, falls in the highly conserved fifth transmembrane segment of the channel that forms part of the pore. It is just two amino acids away from the T274M mutation reported in another EOE patient with a burst-suppression EEG (Weckhuysen *et al.*, 2012), and other mutations in this pore region are known to cause benign neonatal epilepsy and infantile seizures (Singh *et al.*, 1998, 2003; Zhou *et al.*, 2006). The *SCN2A* mutation, NM_001040143:c.A5558G:p.H1853R, is in the cytosolic C-terminal region of the protein, after the final transmembrane domain. Different *de novo* mutations in the

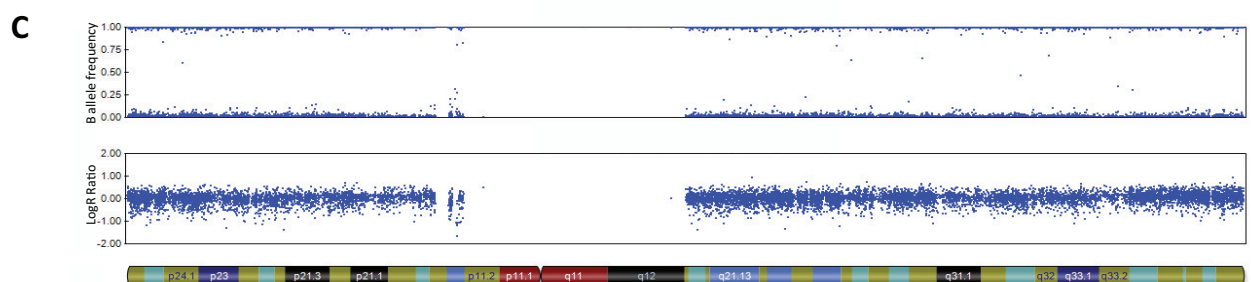
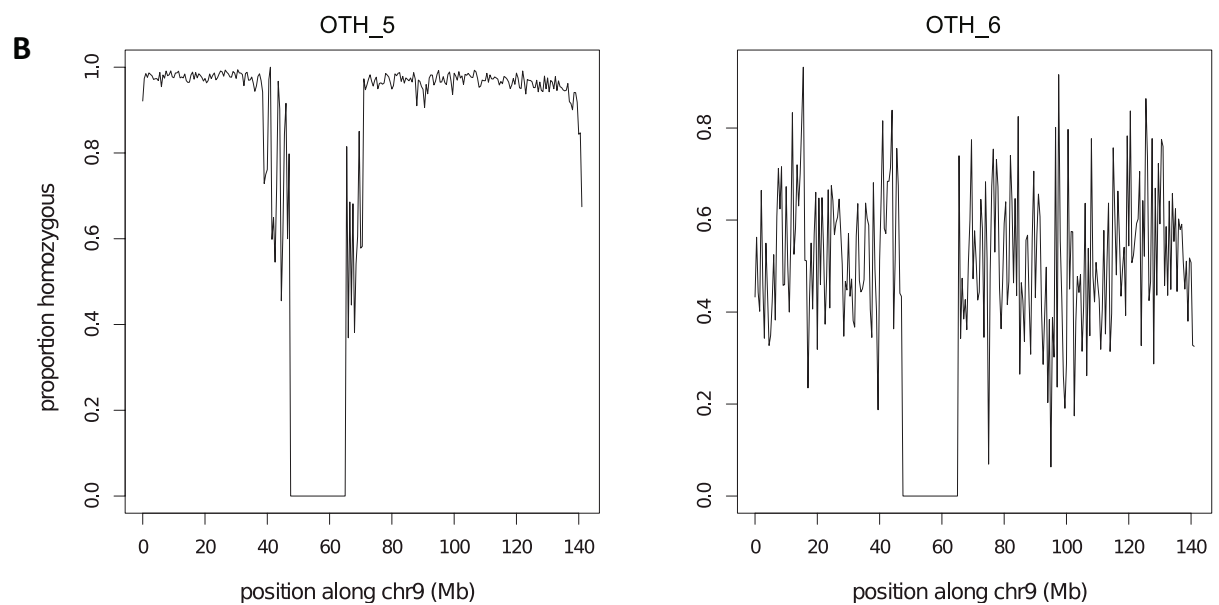
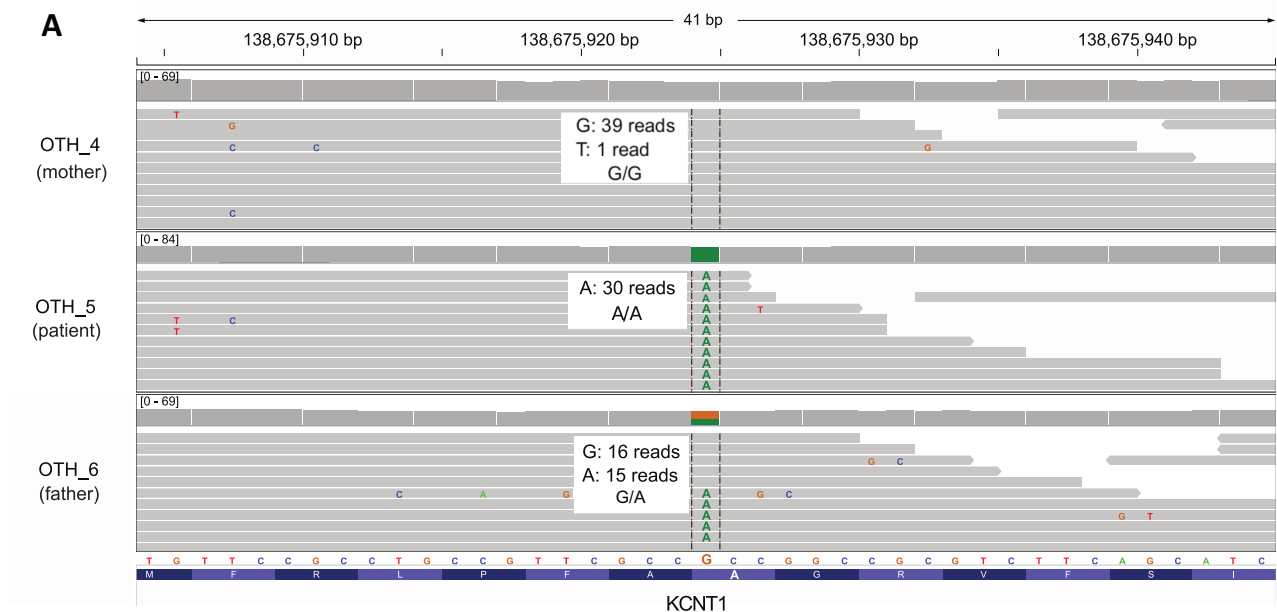
cytosolic domains have previously been reported in patients with intractable epilepsies (Shi *et al.*, 2011), including a girl who had symptoms that were “similar but distinct from Ohtahara syndrome” (Ogiwara *et al.*, 2009). These previous reports provide strong evidence that these *de novo* mutations are responsible for OS in these children.

2.3.1.3 Patient 2: *KCNT1*

Patient 2 had very severe early-onset epilepsy, an EEG consistent with OS, and profound developmental delay. His father’s maternal aunt, grandmother and nephew had had childhood idiopathic epilepsy, but it was unknown whether this was related to the proband’s condition. Patient 2 did not have any compelling *de novo* mutations. However, low chromosomal heterozygosity and detection of multiple Mendelian errors (Figure 2.1) suggested that he had paternal isodisomy for chromosome 9. A SNP array subsequently confirmed that he was completely homozygous on this chromosome but had normal copy number.

This finding prompted two new alternative hypotheses: that the patient’s symptoms were due to aberrant expression of an imprinted gene on chromosome 9, or that there was a recessive pathogenic mutation on this chromosome that had become homozygous as a result of the isodisomy. There was no evidence in the literature for imprinted genes on this chromosome that had plausible links to epilepsy. We therefore scanned the patient’s chromosome 9 for rare homozygous variants that might

Figure 2.1 (*following page*): **Discovery of paternal isodisomy for chromosome 9 in Patient 2.** a) Initially, Mendelian errors were noted on this chromosome. This example illustrates the *KCNT1* mutation: the mother is homozygous reference, the father heterozygous and the child homozygous for the mutant allele. Grey bars represent individual sequencing reads from the sample indicated on the left, and coloured letters divergences from the reference sequence. Only a subset of reads are shown, sorted to aid visualisation. The grey pile-up along the top indicates the sequence coverage. b) WGS data revealed complete homozygosity across chromosomes 9. The plots show the average proportion of variants that are homozygous in 500kb windows across chromosomes. There is an apparent drop in homozygosity around the centromere, owing to the sequence being inaccessible to short read alignment. c) The isodisomy was confirmed using a CytoSNP12 array, which showed that the whole chromosome was homozygous (top plot: B allele frequency), despite normal copy number (bottom plot: LogR ratio).



be pathogenic. The only plausible candidate was a novel nonsynonymous variant in *KCNT1* at 9q34.3 that disrupted a highly conserved alanine residue in the intracellular C-terminal domain, NM_020822:c.G2896A:p.A966T.

KCNT1 encodes the Na⁺-activated K⁺ channel known as Slack, which is very widely expressed throughout the brain (Bhattacharjee *et al.*, 2002). While our project was under way, two papers were published reporting that dominant mutations in *KCNT1* cause autosomal dominant nocturnal frontal lobe epilepsy (ADNFLE) (Heron *et al.*, 2012) and malignant migrating partial seizures of infancy (MMPSI) (Barcia *et al.*, 2012; McTague *et al.*, 2013). Interestingly, one of the MMPSI patients with a *KCNT1* mutation was described as having a subtle burst-suppression EEG (McTague *et al.*, 2013). Distinct from this patient, however, our Patient 2 did not have migrating seizures, and had a clear burst-suppression EEG pattern.

The A966 residue in *KCNT1* is completely conserved across all vertebrates for which genome sequences are available. It seemed very likely that this mutation was causal in Patient 2 but, in order to investigate its functional consequences, we enlisted collaborators, the Kaczmarek group, to conduct electrophysiological experiments. They created a mutant construct with the rat version of *KCNT1*, encoding the Slack protein with the mutation at position 945, equivalent to position 966 in the human protein. They expressed this A945T construct and the wild-type (WT) version in *Xenopus laevis* oocytes, and measured channel activity by performing two-electrode voltage clamping experiments. These experiments showed that activity of the A945T mutant channel was increased significantly by a factor of 13 relative to the WT at +60 mV (Figure 2.2A,B). The amounts and integrity of the WT and A945T cRNA used in these experiments appeared similar, just 8% higher in the mutant than the WT (Figure 2.2C). However, the A945T mutant channel was consistently more active than the WT channel at each of the positive membrane potentials tested (Figure 2.2D).

This mutation had a much more profound effect than the two previously characterised MMPSI-causing mutations, R409Q and A913T (Barcia *et al.*, 2012). Whereas channel activity was increased by a factor of 3 at +60 mV in the two MMPSI mutant channels, the corresponding increase was 13-fold for the A945T mutation. This suggests that the patterns of neuronal firing that produce the distinctive EEG patterns characteristic of these different disorders can vary as a function of KCNT1 channel

activity, and can furthermore influence the nature, severity and onset of the seizures. Since we published our case (Martin *et al.*, 2014), there have been other reports of *de novo KCNT1* mutations in patients several kinds of epilepsy (Ohba *et al.*, 2015; Møller *et al.*, 2015). One mutation reported by Møller *et al.* (2015) was incompletely penetrant, which may explain why the father of our Patient 2, who was heterozygous for the mutation, had no epileptic phenotype.

2.3.1.4 Patient 4: *PIGQ*

Patient 4, who was of West African origin, had severe EOE with a burst-suppression EEG, consistent with OS, as well as hypotonia. Since the first three OS patients analysed had presumably pathogenic mutations in sodium and potassium channel genes, it seemed likely that Patient 4 would have a similar mutation.

We first considered two compound heterozygous, nonsynonymous SNVs identified in *SCN4A*. One (D952N) was novel and the other (M1808I) had a frequency 0.002 in 1000 Genomes. The channel encoded by *SCN4A*, Nav1.4, is primarily known as a skeletal muscle channel, and mutations in it have been shown to cause various muscle diseases (Raja Rayan and Hanna, 2010). *SCN4A* is expressed in the brain (Allen Brain Atlas), so it seemed plausible that these compound heterozygous mutations could be responsible for the seizures in this patient, as well as his severe hypotonia. However, the patient’s unaffected sibling was subsequently found to be a compound heterozygote for the same *SCN4A* variants, so they were deemed benign.

The second candidate mutation considered in Patient 4 was a *de novo* nonsynonymous mutation at a highly conserved position in *C5ORF42*. Compound heterozygous mutations in this gene were recently reported to cause Joubert Syndrome (JS) (Srour *et al.*, 2012), a recessive disorder characterized by a distinct mid-hindbrain malformation, hypotonia, ocular-motor apraxia and hyperpnea². JS patients occasionally suffer from seizures (Bachmann-Gagescu *et al.*, 2012), but since Patient 4 did not have the diagnostic “molar tooth sign” on the MRI, characteristic of JS, or any breathing abnormalities, and we could not find a second mutation in *C5ORF42* that would suggest recessive inheritance, we concluded that this was not a case of misdiagnosis and that the mutation was not the cause of OS in this patient.

²Hyperpnea refers to increased depth of breathing due to metabolic demands.

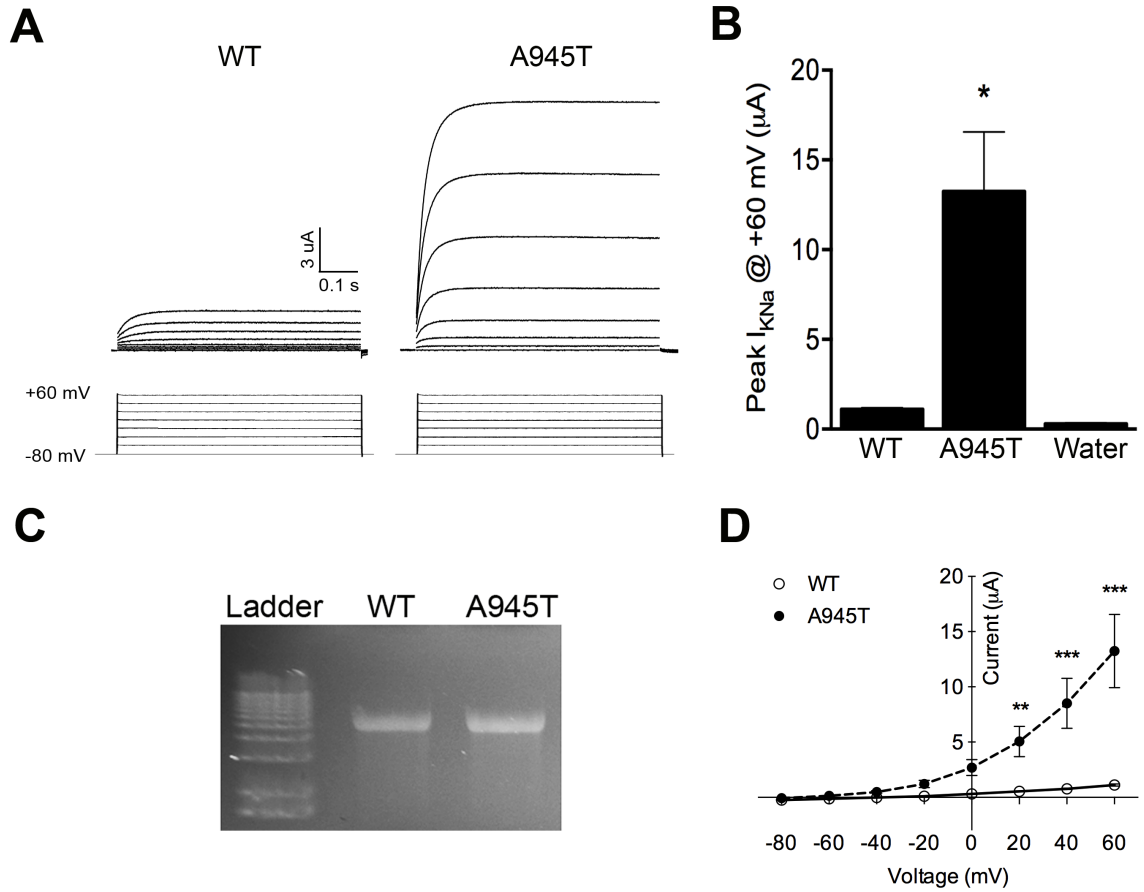


Figure 2.2: **Electrophysiological and channel expression analysis of *KCNT1* mutation found in Patient 2.** This figure was made by Grace Kim from the Kaczmarek group, who conducted the experiments. The WT or A945T mutant Slack channel was expressed in *Xenopus laevis* oocytes, and two-electrode voltage clamping (TEVC) performed. (A) A representative trace of current activity recorded from an oocyte expressing WT or A945T is shown on top, with the voltage-clamping protocol displayed underneath. (B) Averaged quantitation of the peak current is compared at +60mV ($p < 0.001$, $n=5$, Student's t-test; representative of three independent experiments). (C) The quality of RNA injected into *Xenopus* oocytes was checked on a 1% formaldehyde agarose gel. (D) Current-voltage relations for the WT or A945T channels. Channel activity as measured at peak current amplitude and normalized to the value at +60 mV is plotted against voltage (** $p < 0.01$, *** $p < 0.001$, $n = 5$, 2-way ANOVA, Sidak's multiple comparisons test).

Although the patient was reported to be non-consanguineous, we noticed several extended homozygous regions in his genome. Within a 2Mb homozygous region on chr16, he had a novel homozygous single nucleotide variant that was predicted to disrupt the highly conserved splice acceptor site of exon 3 of the *PIGQ* gene: NM_004204:exon3:c.690-2A>G (Figure 2.3). Both parents were heterozygous and two unaffected siblings were either heterozygous or homozygous for the reference allele. *PIGQ*, formerly called *GPII*, encodes a subunit of an N-acetylglucosaminyltransferase that catalyzes the first step in the biosynthesis of glycosylphosphatidyl inositol (GPI), a glycolipid that is attached to the C-terminus of some proteins (Watanabe *et al.*, 1998). The *PIGQ* protein stabilizes the enzyme complex (Hong *et al.*, 1999). A nonsense mutation in the X-linked *PIGA* gene, which encodes another subunit of this enzyme, was recently reported to cause a lethal disorder characterized by multiple congenital abnormalities, structural brain malformations, joint contractures and neonatal seizures with a burst-suppression EEG (Johnston *et al.*, 2012). Recessive mutations in other GPI synthesis genes cause clinically heterogeneous syndromes, all of which involve seizures (Maydan *et al.*, 2011; Almeida *et al.*, 2006; Krawitz *et al.*, 2010; Thompson *et al.*, 2012). Thus, this *PIGQ* mutation seemed a very plausible candidate for causing OS in Patient 4.

The patient died before this mutation was discovered so it was not possible to obtain samples to test the effect of this homozygous mutation on *PIGQ* activity. However, analysis of the RNA extracted from his parents' blood revealed two *PIGQ* transcripts, one consistent with the reference annotation and another with a deletion of exon 3, as expected given that they were heterozygous for the variant at the splice acceptor site for this exon (Figure 2.4). Since exon 3 falls immediately before the catalytic domain of *PIGQ*, the mutation seemed likely to abrogate the function of the enzyme and lead to a reduction in GPI synthesis, as was seen for the nonsense mutation in *PIGA* (Johnston *et al.*, 2012). The parents were tested for abnormalities in serum alkaline phosphatase levels and in expression of CD59 on red blood cells, which are typical signs of impaired GPI synthesis (Johnston *et al.*, 2012; Maydan *et al.*, 2011; Krawitz *et al.*, 2010, 2012), but none were found. This is not entirely surprising since heterozygous carriers of other *PIG* gene mutations had normal CD59 levels (Almeida *et al.*, 2006).

The skipping of exon 3 causes an in-frame deletion of 44 amino acids. Collaborators tested whether this abnormality affected *PIGQ* function in vitro, by transfecting

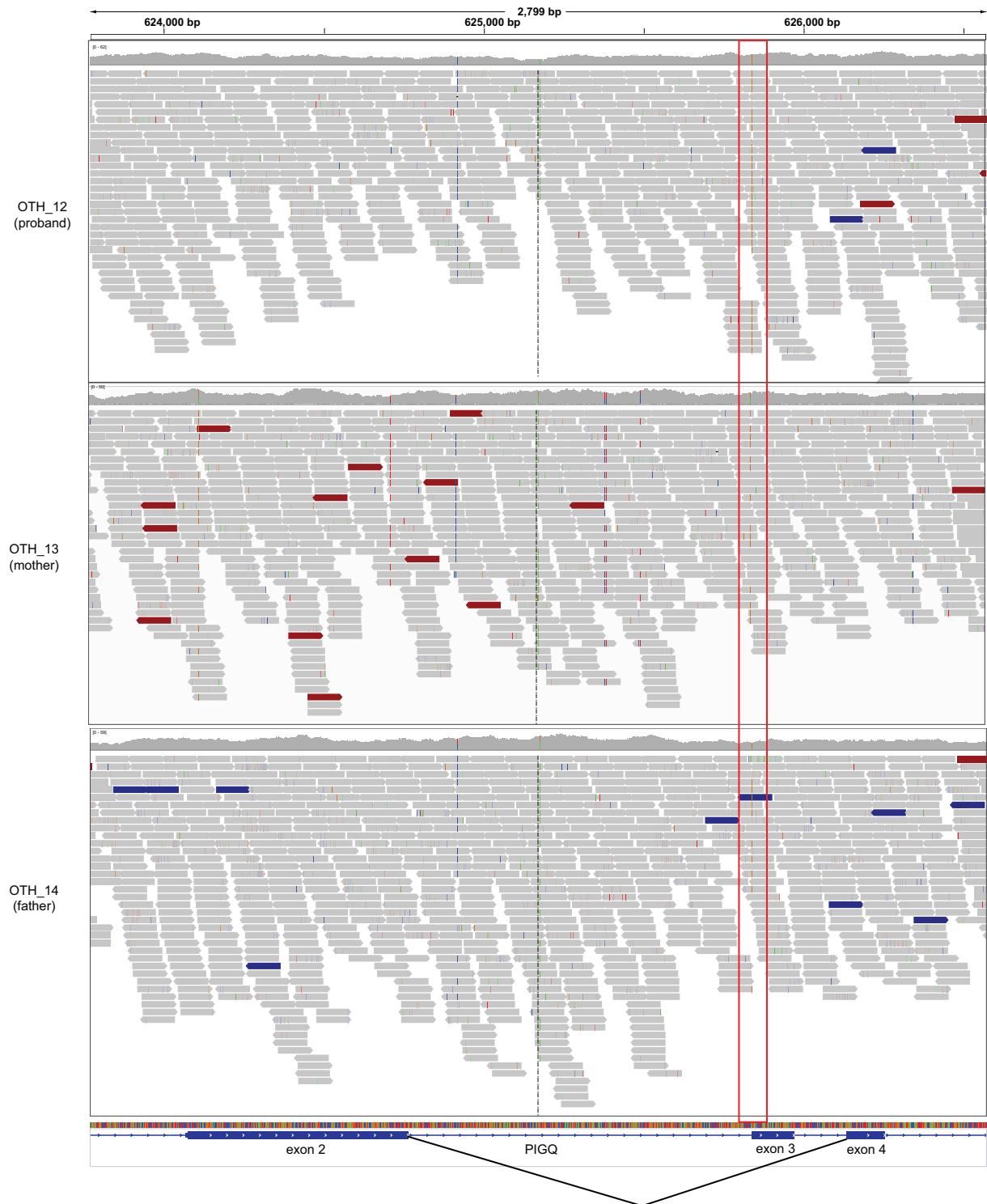


Figure 2.3: **Patient 4 has a novel homozygous variant in a splice site of *PIGQ*.** The variant (NM.004204:exon3:c.690-2A>G) disrupts the splice acceptor site preceding exon 3 (red box) by altering the consensus “AG” sequence to “GG”. It was thus expected to cause skipping of this exon, as indicated by the black lines at the bottom. Grey bars represent individual sequencing reads from the sample indicated on the left, and coloured lines divergences from the reference sequence, with the base quality denoted by the colour intensity. The grey pile-up along the top indicates the sequence coverage.

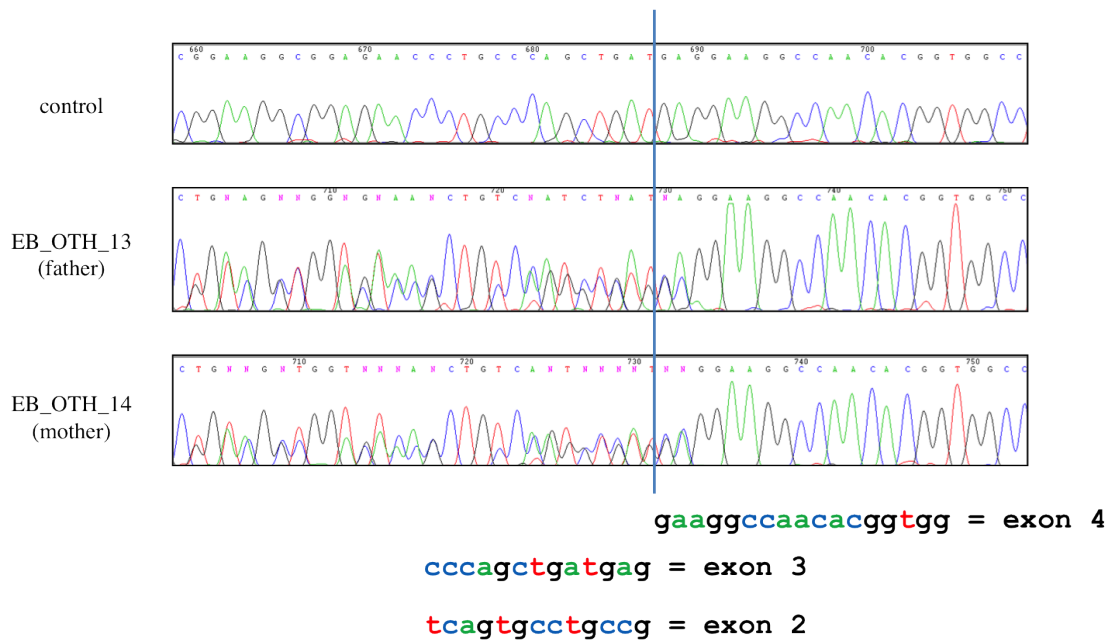


Figure 2.4: **Confirmation of exon skipping in the parents of Patient 4, heterozygous carriers of the *PIGQ* mutation.** RNA was extracted from their blood and converted to cDNA. Sanger sequencing confirmed that one of the two transcripts was missing exon 3. The coloured trace indicates the Sanger read and the vertical blue line the break between the exons. Note the superimposition of the two different sequences to the left of the line, corresponding to the isoform with either exon 2 or exon 3 preceding exon 4. The exonic sequences are indicated underneath, with the colours corresponding to those in the Sanger trace.

human *PIGQ* cDNA either with or without exon 3 into *PIGQ*-deficient cell lines. They demonstrated that the *PIGQ* protein lacking the 44 amino acids had some functional activity but was unstable, and so GPI synthesis was impaired (for more details, see Martin *et al.* (2014)). A recent report of a second patient with a homozygous *PIGQ* mutation who had intractable seizures, developmental delay, and optic atrophy (Alazami *et al.*, 2015) makes it seem highly likely that this mutation is pathogenic in Patient 4.

Further work is needed to determine how impaired GPI synthesis leads to epilepsy, since over 150 proteins have been reported to have GPI anchors (Fujita and Kinoshita, 2012). Several of these play important roles in neural development and function (McKean and Niswander, 2012; Labasque and Faivre-Sarrailh, 2010; Borrie *et al.*, 2012) and would presumably be good targets for further studies in model organisms.

2.3.1.5 Patient 5: *CSNK1G1*

Patient 5 had severe epilepsy, microcephaly and developmental delay. The most likely causal variant was a *de novo* nonsynonymous variant at a highly conserved site in *CSNK1G1* (NM_022048: c.C688T:p.R230W). This gene encodes casein kinase 1 (CK1), gamma 1, a serine/threonine kinase expressed in many tissues including brain. CK1 regulates the phosphorylation of NMDA receptors and is thus important for synaptic transmission (Chergui *et al.*, 2005). It also plays a key role in Wnt signalling (Davidson *et al.*, 2005), which is crucial for multiple embryonic processes including neuronal differentiation (Hirabayashi *et al.*, 2004) and craniofacial development (Fossat *et al.*, 2012). There is also evidence that *CSNK1G2* is associated with febrile seizures (Yinan *et al.*, 2004). Thus, this variant in *CSNK1G1* is a good candidate for causing the severe early-onset epilepsy in Patient 5 but genetic validation studies did not find any additional patients (see below) and further work would be needed to definitively establish it as causal.

2.3.1.6 Patient 6: *CBL*

Patient 6 had severe tonic epilepsy, microcephaly, and developmental delay. He was found to have two *de novo* mutations in the *CBL* gene that were 93bp from each other³. Of the two mutations, one was synonymous (NM_005188:c.C1320T:p.G440G)

³A single Illumina read suggested these were in *cis* and this was subsequently confirmed using allele-specific PCR and Sanger sequencing (data not shown).

and the other predicted to affect splicing (NM_005188:exon9:c.1228-1G>A).

The *CBL* gene encodes a ubiquitously expressed adaptor molecule and ubiquitin ligase that regulates the Ras/MAPK pathway (Swaminathan and Tsygankov, 2006). The gene is primarily recognized as a tumor suppressor (Kales *et al.*, 2010), but germline mutations in it and other genes involved in the Ras/MAPK pathway can also cause various developmental disorders collectively known as the neuro-cardio-facial-cutaneous (NCFC) syndromes or RAS-opathies (Denayer and Legius, 2007). Splicing assays conducted by collaborators showed that the splice site mutation did indeed lead to skipping of exon 9 (Martin *et al.*, 2014), which encodes the C-terminal part of the RING finger domain that is responsible for Cbl’s ubiquitin ligase activity. Heterozygous missense mutations in the RING domain of *CBL* have been reported to cause facial, cutaneous and cardiac abnormalities, hypotonia and developmental delay (Martinelli *et al.*, 2010), as well as microcephaly and a predisposition to juvenile myelomonocytic leukemia (JMML) (Perez *et al.*, 2010; Niemeyer *et al.*, 2010). Curiously, Niemeyer *et al.* (2010) reported two JMML patients with an almost identical splice site mutation to that in Patient 6. One patient had developmental delay and reduced growth but neither was reported to have epilepsy. Niemeyer *et al.* found that their mutation (1228-2A>G) gave rise to some transcripts with premature stop codons, as well as the transcript with the in-frame deletion of exon 9 that was observed in Patient 6 (Δ exon 9). The retention of the LZ dimerization domain in the Δ exon 9, but not the 1228-2A>G mutant with the premature stop codon, may account for the different consequences, since one would anticipate that Δ exon 9 Cbl would retain its adaptor function in multiple cellular signalling pathways, even if its ubiquitin ligase activity were ablated.

Although mutations in *CBL* have not previously been reported to cause epilepsy, mutations in *KRAS* and *BRAF*, in the same pathway, have been reported in two boys with refractory epilepsy and cardio-facial-cutaneous syndrome (CFC) (Adachi *et al.*, 2012). Following the finding of the *CBL* mutation, Patient 6 was reviewed by several clinical geneticists, who noted hypopigmented skin lesions and a history of congenital heart disease. However, they did not think that this patient fitted phenotypically into NCFC syndromes. Nonetheless, since *CBL* mutations are notorious for their variable phenotypes and incomplete penetrance, even in the case of the recurrent Y371H mutation (Perez *et al.*, 2010), this mutation still seems a good candidate in Patient 6. We did not find any other likely pathogenic *CBL* mutations in the larger

panel of epilepsy patients (see below), so it may be that a) *CBL* mutations only rarely manifest themselves primarily as epilepsy, b) this mutation is not causal, or c) there are additional modifier variants in this patient.

2.3.1.7 Mutation screening in other epileptic encephalopathy cases

To try to confirm causality of the *KCNT1*, *PIGQ*, *CBL* and *CSNK1G1* mutations, we asked collaborators to screen these genes in other cohorts. The first cohort comprised about 500 epileptic encephalopathy⁴ cases from Australia, which were screened using a targeted resequencing approach (Carvill *et al.*, 2013). It included four cases of OS, five of early myoclonic encephalopathy (EME; a syndrome which shares features with OS (Djukic *et al.*, 2006)), and thirty-eight of early-onset epileptic encephalopathy (defined as onset within the first 3 months of life). The second cohort comprised thirteen other cases of OS and EME from a UK cohort, which were Sanger sequenced. There were no coding variants that would fit a recessive model in *KCNT1* and *PIGQ*, or a *de novo* model in *CBL* and *CSNK1G1*, as was observed in the WGS500 patients. This likely reflects the genetic heterogeneity in severe EOE, and is not surprising given the relatively small number of patients screened who had similar phenotypes to those described here (total 16 OS, 6 EME).

2.3.1.8 Family 7: uncertain cause

The two daughters in Family 7 had severe mental retardation and epileptic encephalopathy. The parents were initially reported to be normal, and thus the simple and compound heterozygous models were considered most plausible. We also looked for shared *de novo* mutations. There were no coding variants that fitted these models using the filters described in Section 2.2.3, nor any large CNVs. We thus considered rare noncoding variants that fitted the inheritance model. There were no shared rare homozygous noncoding variants, but there was one gene of potential interest, *IGF11*, that had compound heterozygous mutations in regions annotated as regulatory by Ensembl. This gene encodes insulin-like growth factor-1 receptor, and coding mutations in it have been reported to cause severe growth defects and microcephaly (Walenkamp *et al.*, 2006; Raile *et al.*, 2006). IGF1R is one of the receptors that is ubiquitinated by Cbl (Sehat *et al.*, 2008), encoded by the gene mutated in Patient 6. Conceivably, these compound heterozygous variants could each lead to misregulation of *IGF1R* expression, leading to microcephaly and disrupting brain development

⁴Epileptic encephalopathy manifests as intractable seizures that lead to progressive cognitive and neuropsychological decline.

(Joseph D’Ercole and Ye, 2008) in these girls. However, the fact that the children with *IGF1R* coding mutations did not have epilepsy, and that our patients are not abnormally small, argues against this hypothesis.

Further discussion with the clinical geneticist revealed that that the patients’ mother had mild learning disability (LD), suggesting that she may, in fact, have a less severe form of their disease. Analysis of the data under a dominant model uncovered a rare nonsynonymous variant in *KCNQ1*. Mutations in this potassium channel gene are known to cause long QT syndrome (LQTS) (Bokil *et al.*, 2010) and this same mutation has previously been seen in a LQTS patient (Choi *et al.*, 2004). The clinician requested an electrocardiogram of the mother following this finding. Some LQTS mutations in *KCNQ1* can reportedly cause seizures (Goldman *et al.*, 2009), but the fact that such severe learning disability is not commonly observed in patients with these mutations suggests that this one is probably not the cause of these girls’ condition.

Assuming the mother does indeed have a milder form of the phenotype, one might imagine a two-hit model in which the girls both inherited a heterozygous dominant mutation from their mother and a modifying mutation from their father in a different gene. There were no rare heterozygous variants in ion channel genes that would fit this model, but there could be such variants in other genes. However, if this possibility were considered, the search space would be enormous and the likelihood of pinpointing the causal mutations very low. Another possible explanation is that the causal mutation is an intermediate-sized duplication, inversion or translocation, since our sensitivity to detect such variants in the WGS500 pipeline was probably quite low. The uncertainty about the mother’s phenotype in this family renders the analysis particularly challenging.

2.3.2 Putatively novel syndromes

One of the most successful uses of WGS and WES has been to help diagnose patients whose condition does not appear to match any known syndromes (Need *et al.*, 2012; Ku *et al.*, 2012). In some cases this has led to the realisation that the patient in fact has a clinically recognised syndrome presenting in an unusual way (e.g. Schrader *et al.*, 2011). In others, a new disease gene has been discovered (e.g. Worthey *et al.*,

2011). We analysed two such cases as part of WGS500.

2.3.2.1 Mental retardation and disturbed sleep-wake cycle

The first case was a family in which the two sons had severe physical and mental retardation (MR). They were born to healthy, nonconsanguineous parents with no family history of developmental disability. The boys could not run until they were in their teens, and in their early twenties, when this project was begun, they could still only speak a few words. Many genes have been implicated in mental retardation (reviewed in Ropers (2010)), but a large proportion of cases remain undiagnosed even in large-scale WGS or WES efforts (Gilissen *et al.*, 2014; Wright *et al.*, 2015). However, in addition to their low cognitive ability, these boys have an additional and remarkable phenotype: a very disturbed sleep-wake cycle. They alternate 24- to 72-hour periods of continuous intense activity with 24- to 72-hour periods of deep sleep. There are no reports in the literature of this condition, although there are some genes linked to milder sleep disorders: polymorphisms in the serotonin pathway (Xu *et al.*, 2014) have been associated with obstructive sleep apnea syndrome, and variants in circadian genes have been linked with non-24-hour sleep-wake cycles (Kripke *et al.*, 2014). The novelty and severity of this condition made it a good candidate for WGS, since it was hoped that our findings might assist with diagnosis or at least lead to better understanding of the pathogenesis.

There was only one variant in the WGS data that met the criteria described in Section 2.2.3 for any of the possible inheritance models (simple recessive, compound recessive, X-linked or shared *de novo*). This was an X-linked variant in a splice site of the *CHM* gene, which both boys inherited from their mother. This was a novel variant where we first discovered it, but it has since been seen in 17 males in the Exome Aggregation Consortium (ExAC) ($N = 53,653$ chromosomes), from which individuals with severe paediatric disease have been excluded, so it is very unlikely to be causal.

WES data were available for the quartet and so these were scanned for variants that might have been missed in the WGS. The only plausible one was a variant in the *GRIA3* gene that did fit an X-linked model but that been erroneously called as heterozygous in the WGS data because of a single read containing the reference base (Figure 2.5). This had been missed because, in searching for candidate variants,

we ignored those called as heterozygous on the X chromosome in males, which were almost all due to sequencing or PCR errors. In this case, low-level contamination of the sequencing libraries for the individuals in this family seems a likely explanation (Figure 2.6).

This variant was a novel nonsynonymous mutation in the *GRIA3* gene, which encodes GluR3, a subunit of an AMPA-type ionotropic glutamate receptor. Mutations in this gene have previously been reported to cause a milder form of MR (Gecz *et al.*, 1999; Chiyonobu *et al.*, 2007; Wu *et al.*, 2007). Some MR patients with *GRIA3* missense mutations had similar physical features to our patients: poor muscle bulk, muscle weakness, hyporeflexia and asthenic habitus (Wu *et al.*, 2007). There are 83 patients recorded in the DECIPHER database (Swaminathan *et al.*, 2012) with a deletion or duplication disrupting *GRIA3*, and 51 of these have a recorded phenotype, mostly intellectual disability and/or global developmental delay. Interestingly, one of the patients is described as having autism, intellectual disability and sleep disturbance. We attempted to contact this patient’s clinician via DECIPHER to ask for more details about the phenotype, but were unsuccessful.

Our mutation, A653T, was not seen in ~86,000 chromosomes in ExAC, nor were any other coding mutations observed in the same exon. The mutation disrupts a residue in the pore region of GluR3 which is highly conserved both across species and in all other glutamate receptor genes (Figure 2.7A). A653 plays an important role in occluding the ion channel. The neighbouring amino acid in the GluR δ 2 subunit is mutated in the “Lurcher” mouse (Zuo *et al.*, 1997) (Figure 2.7A), which has been extensively studied as a model of neurodegeneration (Vogel *et al.*, 2007). The Lurcher mutation is also an alanine (A) to threonine (T) mutation and it results in the channel spontaneously opening due to sterical clashes. This causes excitotoxicity and apoptotic death of Purkinje cells, which control motor coordination, resulting in the lurching gait of these mice. It seemed likely that the A653T mutation has a very similar effect in the GluR3 channel, and that this has led to degeneration of the parts of the brain in which GluR3 is expressed. Notably, GluR3 is the predominant AMPA receptor subunit in the mouse thalamus (Wang *et al.*, 2011a), which is known to play a role in circadian rhythm control (Colavito *et al.*, 2015), and in Syrian hamsters, the light-dark cycle affects the expression pattern of GluR3 in the suprachiasmatic nucleus, also involved in the sleep-wake cycle (Chambille, 1999). There is some evidence that GluR3 regulates sleep patterns in mice (Steenland *et al.*, 2008) and SNPs

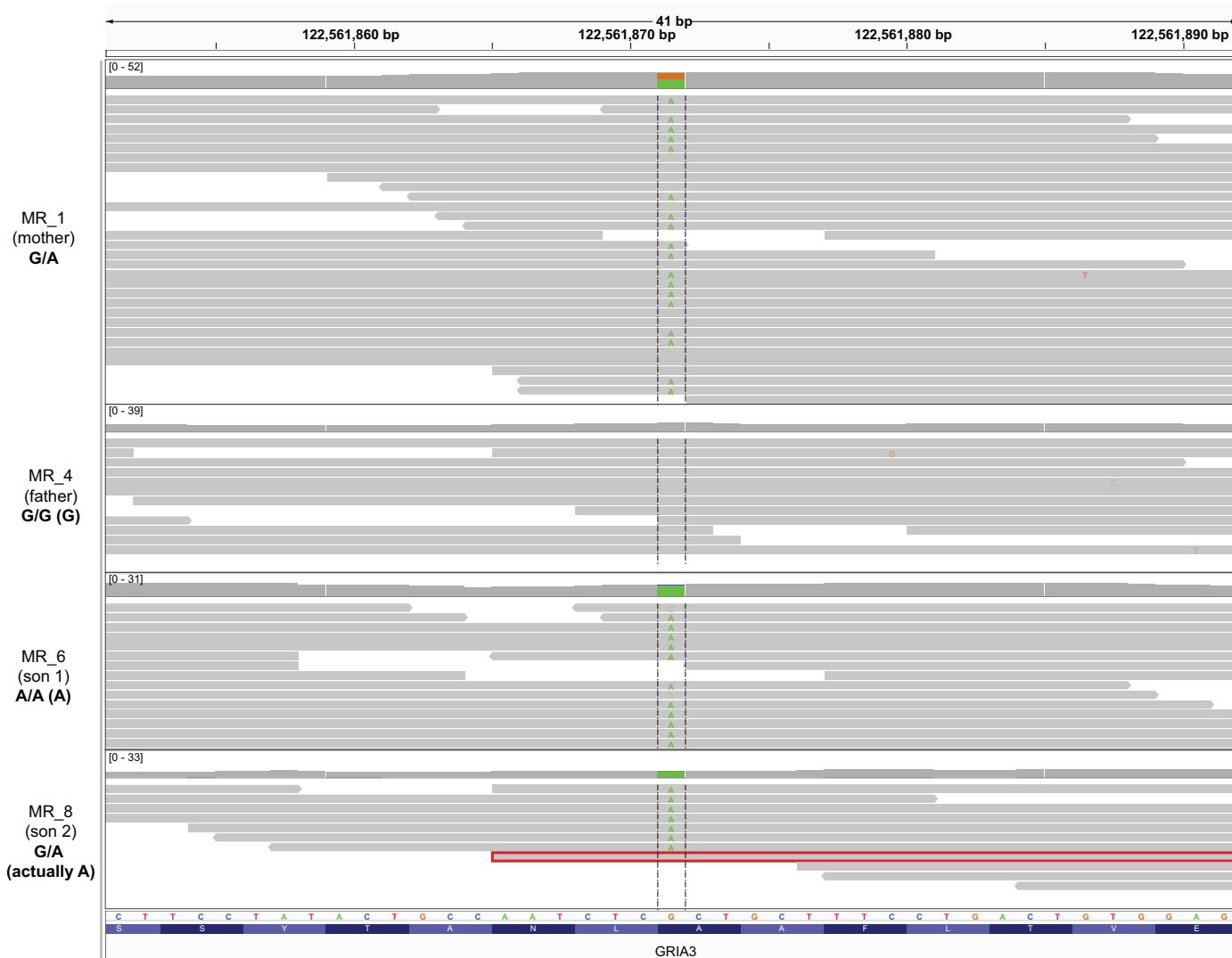


Figure 2.5: **Misgenotyping of the X-linked *GRIA3* variant in the WGS data.** The genotypes called by Platypus are indicated in bold under the sample IDs on the lefthand side. (Note that two alleles are called for the males, although they are hemizygous.) In the WGS data, the variant was called as heterozygous (G/A) in one of the affected sons, MR_8, due to the presence of a single read that contained the reference base (G; red box around read). The mother, father and other affected son were correctly called as heterozygous (G/A), hemizygous for the reference allele (G) and hemizygous for the alternate allele (A) respectively.

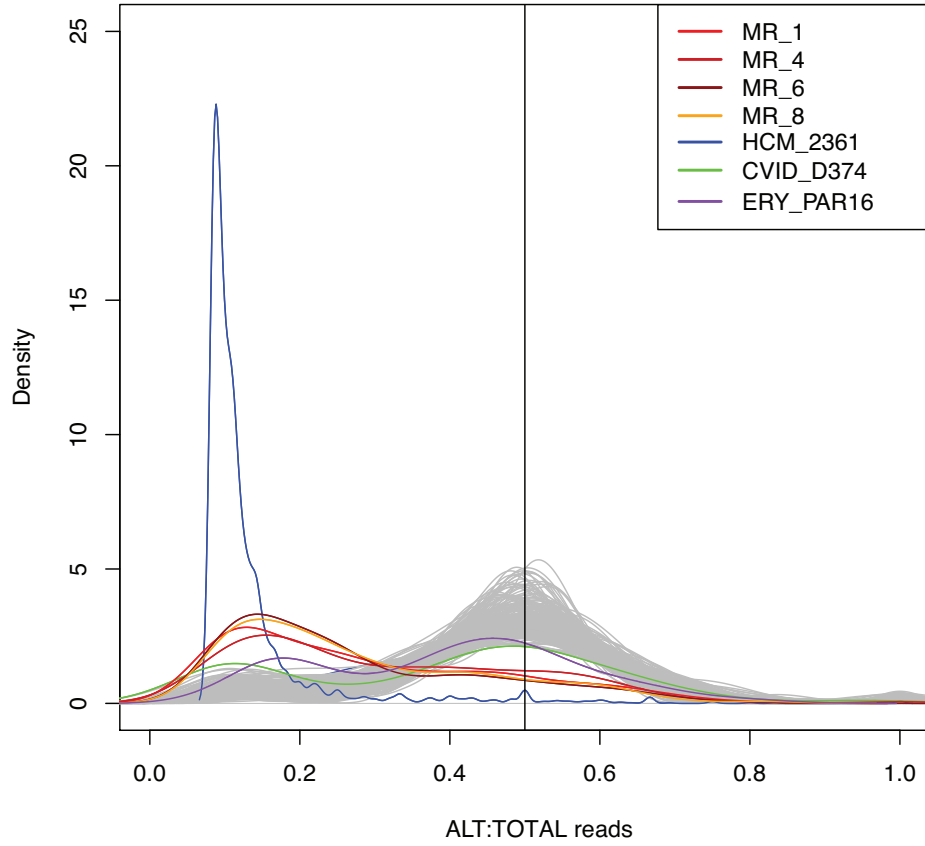


Figure 2.6: **Contamination in the DNA samples from the MR family.** This plot shows the distribution, for each sample in WGS500, of the ratio of the number of reads with the alternate allele (ALT) to the total number of reads (TOTAL), for novel heterozygous variants. We expect the mean to be 0.5. Individuals with a mean < 0.4 , including the individuals from the MR family, are shown with coloured lines. These are likely to have sample contamination which leads to a larger number of heterozygous calls for which there are few ALT reads.

in *GRIA3* were found to be associated with sleep duration in a candidate gene study of 3147 Finnish women (Utge *et al.*, 2011) (Bonferroni corrected p-value=0.0052).

Collaborators conducted electrophysiological experiments to test the effect of the A653T mutation on the channel, by overexpressing both the mutant and wildtype version of the gene in HEK293 cells, which do not normally express *GRIA3*. Contrary to the initial hypothesis, they found that the mutant showed ~10-fold reduced efficacy in response to the partial agonist kainate compared to the wild-type channel, suggesting that the additional hydrophobic contacts mediated by the threonine keep the gate more tightly closed (Figure 2.7) (data not shown). However, demonstrating an effect at the level of the channel does not prove that the A653T mutation is responsible for our patients' phenotype. Thus, the Transgenic Core at the Wellcome Trust Centre for Human Genetics made a knock-in mouse line with this mutation. Collaborators tested for abnormalities in the endogenous circadian rhythms in hemizygous male mice kept under different light conditions. No strong phenotypic difference were observed between mutant and wild-type B6 mice. However, there was some evidence that the mutant mice did not tend to "nap" during the night as much as the wild-type mice, and that they displayed slight period-lengthening when kept in constant light, that are being investigated further.

The lack of an obvious circadian phenotype in these *Gria3* A653T mice could have many possible explanations. One could be that the GluR3 subunit is expressed in different tissues in humans and mice, and so is only important for circadian rhythm regulation in humans. It may be that, in mice, the effect of the mutation is very subtle and can only be detected electrophysiologically by not phenotypically. Notably, Steenland *et al.* (2008) did not observe circadian rhythm disturbances in *Gria3*-knockout mice, but they did find EEG abnormalities during sleep. Collaborators are currently carrying out field recordings on brain slices from our *Gria3* A653T mice, to try to detect electrophysiological changes in the thalamus.

It is possible that the *GRIA3* mutation accounts for the intellectual disability seen in our patients but not the sleep phenotype. However, if it were responsible for the severe cognitive deficits, one might expect to see behavioural abnormalities in the mice, but these are not obvious, although they might be detected by comprehensive behavioural phenotyping. There may be a second mutation in the family that modifies the effect of the *GRIA3* mutation, but since this could be present in one of the

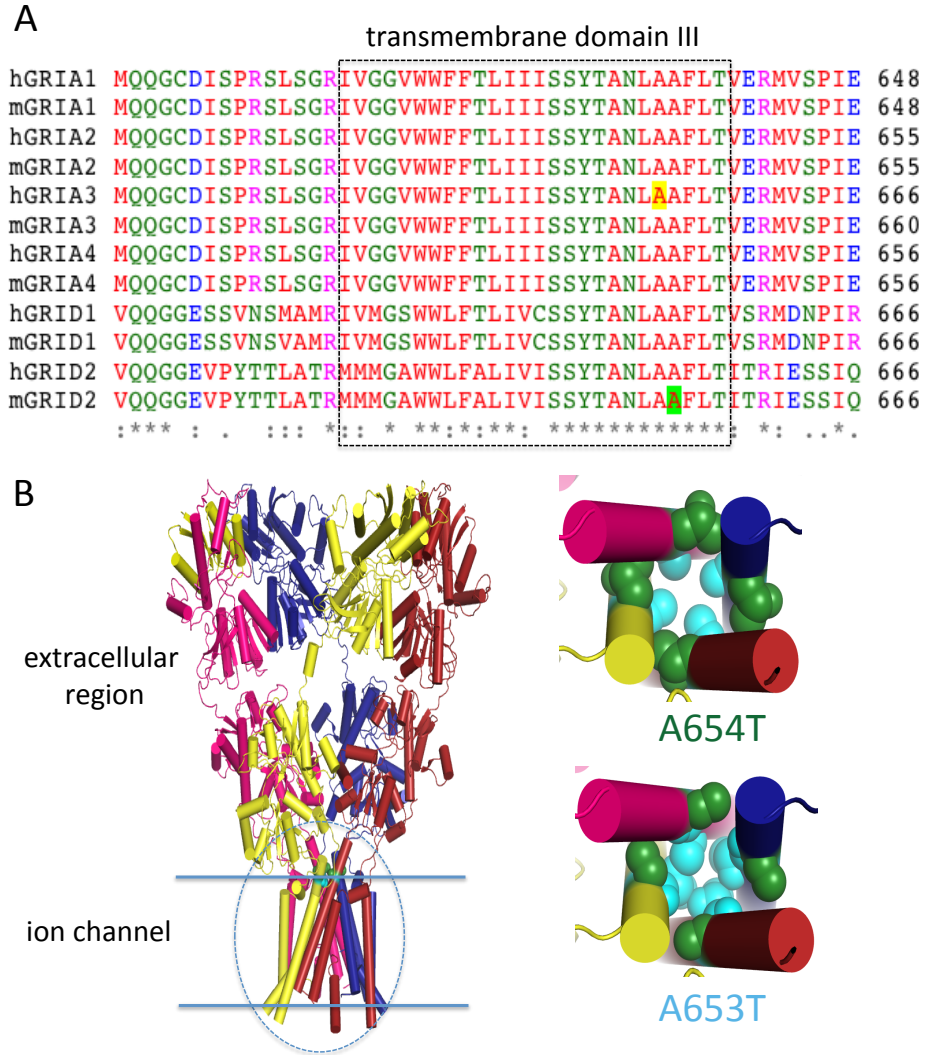


Figure 2.7: The A653T *GRIA3* mutation mimics the Lurcher mutation. A) Alignments of parts of different human (h) and mouse (m) glutamate receptor proteins. The mutated alanine (highlighted in yellow) falls within a highly conserved transmembrane domain that forms part of the glutamate receptor (GluR) pore. The neighbouring alanine residue in GRID2 (highlighted in green), encoding a different glutamate receptor subunit, GluD2, is mutated to threonine in the Lurcher mouse (A654T). B) Left: structure of the GluR2 receptor, with the transmembrane ion channel indicated. Right: structure of the ion channel, viewed from above. The cyan and green spheres represent residues 653 and 654 respectively. Threonine (T) is a bulkier amino acid than alanine (A), and the A654T mutation produces steric hindrance within the pore, causing the receptor to be constitutively open and leading to excitotoxicity. Due to its predicted structural similarity, it was initially hypothesized that the A653T mutation would have the same effect, but in fact, electrophysiology experiments showed that instead it stabilised the channel in the closed conformation. These protein structure figures were made by R. Aricescu.

parents as well, there are likely to be many candidates. Crossing the A653T mutation into a different mouse background may reveal a phenotype, if there is a modifier masking its effect in the B6 strain. Finally, it is also possible that the *GRIA3* mutation is not pathogenic at all, and that the true causal mutation has been missed. It could be a coding mutation that was not found owing to poor coverage of the exon, the contamination that originally caused us to miss the *GRIA3* mutation in the WGS data, or an error in genotyping. Alternatively, it might be noncoding, in which case, there will likely be thousands of candidates. Finally, it is also possible that the causal mutation is a CNV or structural variant that was not called by the WGS500 pipeline. Thus, we are revisiting the WGS and WES data to search for new candidates. We anticipate this will be complicated by the sample contamination, so we may have to obtain new DNA samples.

2.3.2.2 Spastic paraplegia, learning disability and pseudopseudohypoparathyroidism

The second case with a putatively novel syndrome was a consanguineous family of Pakistani ancestry (Figure 2.8) in which four individuals, two sisters and their uncle and aunt, had learning disability, spastic paraplegia (weakness and stiffness of the leg muscles) and pseudopseudohypoparathyroidism (resistance to parathyroid hormone but without biochemical abnormalities (Levine, 2012)). Over 60 genes have been linked to spastic paraplegia, and these play roles in a variety of pathways, including mitochondrial membrane transport, lipid synthesis and metabolism, DNA double-stranded break repair, axon development, and stabilisation of the myelin sheath (Lo Giudice *et al.*, 2014). Many types of hereditary spastic paraplegia (HSP) also involve cognitive impairment. However, pseudopseudohypoparathyroidism has not been reported to occur in HSP patients. Thus, it was suspected that this family had a new syndrome.

There were multiple large regions of homozygosity (i.e. greater than 1Mb) in the genomes of the four affected individuals but none in LDS_30, suggesting that this woman's parents were not consanguineous and that this may have been a case of nonpaternity. Analysis of the SNP chip data with MERLIN confirmed that LDS_30 was only the half sister of LDS_28 and LDS_29. Thus, her genotype data were not useful for the exclusion of candidate recessive variants.

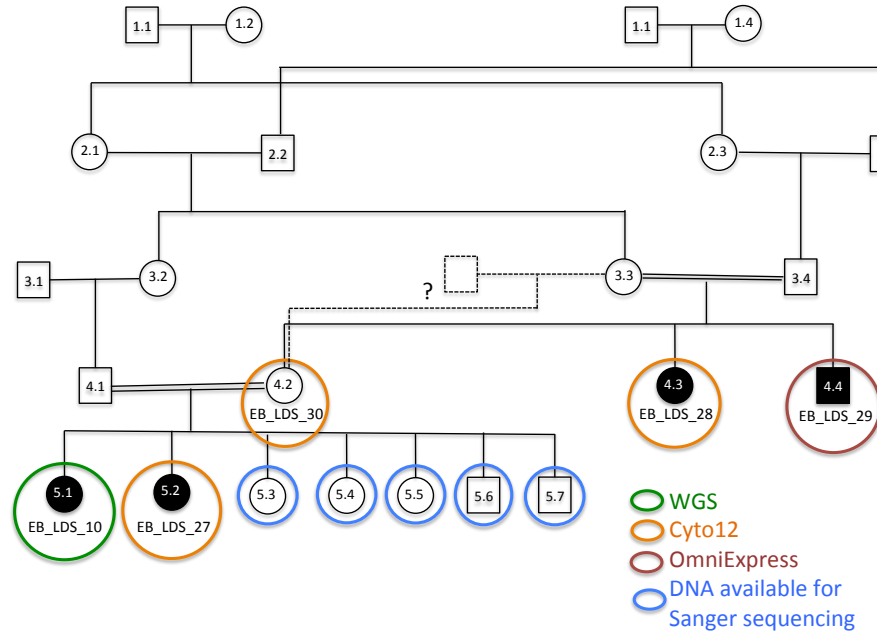


Figure 2.8: **Pedigree of the family with learning disability, spastic paraplegia and pseudopseudohypoparathyroidism.** Affected individuals are indicated in black. The coloured circles indicate the kind of data available on each individual. The dotted lines denote a suspected nonpaternity event.

Linkage analysis revealed a single significant ($\text{LOD} > 3$) linkage peak (Figure 2.9): a 12.7Mb region on chr1:40921145-53713336 that had a LOD score of 3.01 and was homozygous in all affected individuals. This region contained 145 protein-coding genes, and was interrogated for rare simple recessive variants in LDS_10. There was only one coding variant that met the frequency filters described in Section 2.2.3 and that disrupted a well-conserved position. This was a nonsense variant, NM_199342:c.C82T:p.Q28X, in *CCDC23* (coiled-coil domain containing 23). The variant was not found in a cohort of 4350 healthy British Pakistanis and Bangladeshis who had been exome-sequenced (Vagheesh Narasimhan, personal communication), and nor was any other rare homozygous loss-of-function variant. Only one loss-of-function variant was seen in *CCDC23* (Q53X) in 60,706 samples in ExAC, and this was only seen in a single individual, who was heterozygous.

CCDC23 encodes a short protein of only 66 amino acids that is highly conserved across mammals and is known to play a role in angiogenesis (Suzuki *et al.*, 2010). The variant segregated with the phenotype in all family members for whom DNA was available. There was no evidence in the literature to connect *CCDC23* with the patients' symptoms and it is reportedly not expressed in the mouse brain (Allen Brain

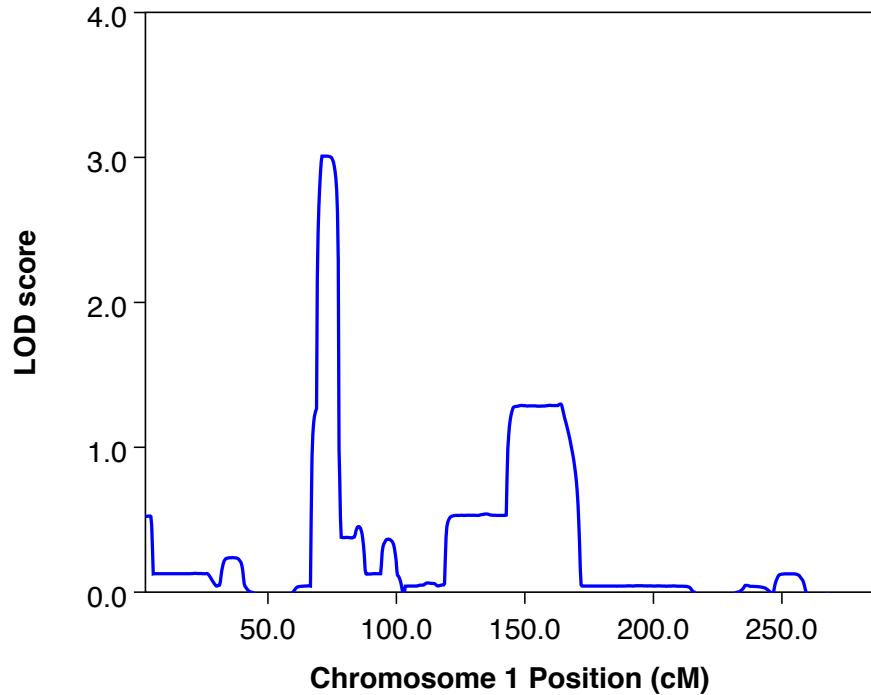


Figure 2.9: **Linkage analysis in the LDS family.** The plot indicates the LOD (log of odds) score across chromosome 1. None of the other chromosomes had any linkage peaks with $\text{LOD} > 2$.

Atlas; (Suzuki *et al.*, 2010)). Nonetheless, collaborators tested the effect of knocking down this gene in zebrafish with morpholinos. They found no dose-dependent differences between morphants and controls in any of the brain structures or in the peripheral neurons. This does not preclude *CCDC23* from being causative. It may be that the morpholino used for knock-down was inefficient, or that the right phenotypes were not examined. However, experiments were terminated since the existing evidence in favour of causality was not strong enough to merit the effort required to complete these.

We searched DECIPHER (Swaminathan *et al.*, 2012) for patients with CNVs overlapping *CCDC23* and found three deletions and one duplication, each of which encompassed 29 or more genes. The three deletions had a phenotype recorded, all involving brain abnormalities and either intellectual disability or global developmental delay. For two of the patients, the inheritance of the deletion was unknown, but in the third, it was *de novo*. This deletion, at chr1:42,269,545-44,007,321, encompassed 40 genes. The patient was reported to have intellectual disability, microcephaly and

spastic diplegia⁵, a syndrome that bares some resemblance to the one in our family. Since the *de novo* 1.74Mb deletion in the DECIPHER patient is entirely contained within our 12.7Mb linkage region, it seems likely that the syndrome is caused by mutations disrupting the same gene within that 1.74Mb region. The mutation is clearly recessive in our family, but in the DECIPHER patient, it seems to be dominant (*de novo*).

If our patient and the DECIPHER patient do have the same syndrome (which is unclear), there are several possible explanations. We obtained DNA from this patient and Sanger-sequenced *CCDC23* to check if the undeleted copy had a loss-of-function mutation, but found none. An alternative explanation is that our family has a point mutation affecting the expression of a particular gene which, when homozygous, has the same effect as the total deletion of one copy of the gene in the DECIPHER patient.

We scanned the linkage region for rare homozygous noncoding variants that might affect the expression of relevant genes. This revealed a SNV in a putative regulatory region in intron 1 of the *ELOVL1* gene that segregated with the phenotype in the family. *ELOVL1* (elongation of very long chain fatty acids (VLCFA)-1) is one of the genes deleted in the DECIPHER patient. It may be involved in the pathogenesis of adrenoleukodystrophy, a disorder of white matter caused by overproduction of VLCFAs, which leads to damage to the brain, adrenal glands and peripheral nervous system (Kemp *et al.*, 2012) and which sometimes involves or even presents as spastic paraplegia (Shaw-Smith *et al.*, 2004). It seemed possibly that this variant increased *ELOVL1* expression and thus VLCFA synthesis. However, quantitative real-time PCR found no evidence for differences in expression of this gene in blood between LDS_10, who was homozygous, and LDS_30, who was heterozygous. This assay is very noisy, and so without further samples (including controls matched for age) we cannot exclude the possibility that this gene is aberrantly expressed in the blood or in other tissues of affected individuals, and that this contributes to the phenotype.

Genome-wide searches (see Section 2.2.2) and manual inspection of the linked region in the Integrated Genomics Viewer (IGV) (Thorvaldsdottir *et al.*, 2012) revealed no rare structural variants. A compound heterozygous model was also considered in

⁵There is some debate about the difference between spastic diplegia and spastic paraplegia (Colver and Sethumadhavan, 2003), but generally, the latter is used to describe spasticity than only affects the legs.

the three regions that were IBD2 between all affected individuals but not homozygous (totalling 10Mb), but no good candidates were found. It is possible that, since the family is consanguineous, there are actually several recessive pathogenic mutations segregating, each responsible for a different part of the phenotype, but again, no further likely candidates were found. Thus, more work is needed in this family to determine the causal mutation. We have registered the phenotype and genes in the linkage region on GeneMatcher (Sobreira *et al.*, 2015) to try to find other families with the same syndrome.

2.4 Discussion

In this chapter we have described several case studies of WGS applied to Mendelian neurological diseases. The results are summarised in Table 2.2. Given that, in seven out of nine families, the putatively causal mutations are in coding regions or splice sites, this study would suggest that WES is probably a more efficient and cost-effective option than WGS for studying families with Mendelian diseases. Even if the causal mutations in families 7 and 9 eventually prove to be noncoding, it would still have been cheaper to do WES on all families first (27 samples) and then do WGS on the unsolved families (5 samples), than to do WGS on everyone initially (£37,750 vs. £94,500, assuming a cost of £3000 per sample for WGS and £750 for WES⁶). That said, at the moment, it is very difficult to pinpoint pathogenic noncoding mutations for rare disorders. It is notable that, when Gilissen *et al.* (2014) sequenced the whole genomes of fifty trios in which the children had intellectual disability (ID), having failed to find a causal mutation initially with WES, they found no likely causal non-coding mutations in ENCODE-annotated elements around known ID genes.

The benefits of the higher coverage in WES are emphasised by our analysis of Family 8, in which a single erroneous read (likely due to contamination) in the WGS data caused us to miss the causal variant. On the other hand, this error could have been avoided by altering the parameters used in the variant-calling algorithm or flagging heterozygous X-chromosome calls in males. Another argument commonly used in favour of WGS over WES is that whole-exome capture platforms introduce biases and can miss some exons (Asan *et al.*, 2011). In the WGS500 paper (Taylor *et al.*, 2015), we estimated that 15% of the 33 causal variants identified in the study (all but

⁶These were the costs for bulk sequencing at the Oxford Genomics Centre at the time.

two of which were coding) would likely have been missed by WES in an independent study of 141 exomes (compared to 0.5% in WGS), if at least six reads are required to detect a novel heterozygous variant, three supporting the alternative allele. Combined with the fact that the detection of structural variants and CNVs is much more reliable in WGS than WES data, although the current methods are far from perfect, WGS is undoubtedly the preferable method in the absence of budget constraints.

This research has led to surprising discoveries about the aetiology of these diseases. (1) The finding of a recessive variant becoming homozygous due to isodisomy in Patient 2 was completely unexpected, and prompted us to check for extended homozygous regions and Mendelian errors in the other WGS500 samples. Since we made this discovery, several large WES studies have reported uniparental disomy (UPD) as a rare cause of Mendelian disorders: Yang *et al.* (2014) found five UPD cases out of 2000 patients sequenced in the Baylor study⁷, and Wright *et al.* (2015) reported six events out of 1,057 DDD trios. (2) Two of the candidate genes found, *GRIA3* and *CBL*, were known to cause diseases with some similarities to those of our patients. The remarkable discovery of the *GRIA3* mutation that resembles the extensively studied “Lurcher” mouse, if established to be definitely causal, is likely to prompt further research into its role in circadian rhythm regulation. *CBL* mutations are notorious for their variable expressivity (Martinelli *et al.*, 2010; Perez *et al.*, 2010), and if this mutation is confirmed as pathogenic, this will expand the spectrum of phenotypes caused by Ras/MAPK pathway mutations. (3) *PIGQ* is an example of a gene which would previously not have been considered, since the patient did not have the congenital abnormalities common in children with other PIG gene mutations, but which became a very plausible candidate after the publication of a paper about a mutation in the gene encoding its binding partner, *PIGA* (Johnston *et al.*, 2012). However, despite these successes, our experience indicates that, even when data from multiple relatives are available and the candidate region is quite small (e.g. Family 9), it can still be very challenging to pinpoint the causal mutation.

An important outcome of this project has been a better awareness of the current difficulties in using WGS for clinical diagnosis. In two of the nine families (1 and 3), the mutation was identified quickly and we were sufficiently confident of

⁷One of these UPD cases was a child with amyotrophic lateral sclerosis who was found to have paternal isodisomy on chromosome 9. This led to the discovery of a homozygous pathogenic mutation in *SIGMAR1*.

Table 2.2: **Summary of results from case studies.** OS: Ohtahara syndrome, NSEO: nonsyndromic early-onset epilepsy; MR: mental retardation; LD: learning disability; MMPSI: malignant migrating partial seizures of infancy; ADNFLE: autosomal dominant nocturnal frontal lobe epilepsy; NCFC: neuro-cardio-facial-cutaneous syndrome; PPHPT: pseudopseudohypoparathyroidism.

Family	Phenotype	Gene	Inheritance mechanism	Presumed causal mutation	Evidence for pathogenicity
1	OS	<i>KCNQ2</i>	de novo	NM_004518:c.C827T:p.T276I	KCNQ2 previously implicated in OS
2	OS	<i>KCNT1</i>	recessive; UPD9	NM_020822:c.G2896A:p.A966T	<i>KCNT1</i> previously implicated in MMPSI and ADNFLE; electrophysiology demonstrated effect on channel current
3	OS	<i>SCN2A</i>	de novo	NM_001040143:c.A5558G:p.H1853R	<i>SCN2A</i> previously implicated in OS
4	OS	<i>PIGQ</i>	simple recessive	NM_004204:exon3:c.690-2A>G	Binding partner <i>PIGA</i> implicated in similar syndrome; mutation leads to exon skipping and reduced GPI synthesis
5	NSEO	<i>CSNK1G1?</i>	de novo	NM_022048:c.C688T:p.R230W	<i>CSNK1G1</i> involved in synaptic transmission
6	NSEO	<i>CBL?</i>	de novo	NM_005188:exon9:c.1228-1G>A	<i>CBL</i> implicated in NCFC; mutation leads to exon skipping
7	NSEO	unsolved	probably simple recessive	N/A	N/A
8	MR, circadian rhythm disturbance	<i>GRIA3?</i>	X-linked recessive	NM_000828:c.G1957A:p.A653T	<i>GRIA3</i> implicated in MR; electrophysiology showed mutation alters channel function; possible subtle phenotype in mouse model; functional work ongoing
9	LD, spastic paraplegia, PPHPT	unsolved	presumed recessive	unsolved	N/A

its pathogenicity to be able to report this back to the families almost immediately. Learning that the causal mutation was *de novo* and thus that the recurrence risk was low had important psychosocial benefits for the parents, even though the finding does not suggest any better courses of treatment at the moment. In contrast, the rest of the families required more extensive analysis, as well as multiple follow-up appointments to reassess their condition and to obtain further samples from other relatives. In some cases, the clinicians were reluctant to participate in this, believing that it was inappropriate for what they considered a purely clinical project. Thus, the boundaries between diagnostic testing and research need to be better defined in future projects of this kind.

It is evident that the mutations identified in clinical WGS studies like this will often require further work to confirm their pathogenicity, beyond what is possible in a purely clinical context. Sometimes, relatively straightforward functional work will suffice because the gene or closely related genes are known to cause the same or similar diseases, and they have direct links to the pathology. For example, for *KCNT1*, which had already been implicated in other types of epilepsy, we were able to collaborate with electrophysiologists who tested the effects of the mutation on the function of the encoded potassium channel. However, for the genes with less direct roles on the organ or pathway of interest, proving that the identified mutations are truly causal would be extremely onerous. For example, even though the *PIGQ* mutation in Patient 4 was shown to impair GPI synthesis, uncovering the precise mechanism by which this leads to epilepsy would require extensive further work. Similarly, even if one demonstrated that the *CBL* mutation in Patient 6 altered Ras/MAPK signaling, showing that this disrupted brain development would require many additional experiments, probably in mouse models. The Deciphering Developmental Disorders Study (2015) employed an effective and relatively high-throughput functional validation method, knocking down thirty-two candidate developmental disorder genes in zebrafish. They found that twenty-one of these altered development, of which eleven showed phenotypic similarity to the human patients and were thus classed as strong candidates. This approach shows considerable promise for further studies, although it obviously cannot be employed in a purely clinical setting.

Identifying additional patients with mutations in the same genes is a complementary method of validation, but is challenging for genetically heterogeneous disorders. A good example of this is a study by Allen *et al.* (2013), published while

our project was underway, who sequenced the exomes of 264 children with epileptic encephalopathies (infantile spasms and Lennox-Gastaut syndrome) and their parents. Because of their relatively large sample size, they were able to show statistical evidence implicating two new genes, *GABRB3* and *ALG13*, which had a greater-than-expected number of *de novo* mutations. It is anticipated that larger sample sizes and meta-analysis of multiple studies together will allow more genes to be discovered in the same way. If, as some expect (Visscher and Gibson, 2013), millions of clinical genomes are sequenced in the next few years, this will greatly increase the chance of finding mutations in the same gene in multiple similar patients, and thus will alleviate some of the need to validate candidates using functional studies.

2.5 Acknowledgements

All laboratory work mentioned in this chapter was conducted by others. Alistair Pagnamenta did the qRT-PCR of *ELOVL1*, as well as all of the Sanger sequencing. He and Karl Hudspeth did the splicing assays in *PIGQ* and *CBL*. Kevin Leyden at the John Radcliffe Hospital did the flow cytometry to test the CD59 levels for the parents of Patient 4. Grace Kim at Yale did the electrophysiology experiments for *KCNT1*, supervised by Len Kaczmarek. Taroh Kinoshita and Yoshiko Murakami did the experiments on GPI synthesis to test the *PIGQ* mutation. Radu Aricescu made the structure of the GluR3 protein shown in Figure 2.7. Ondrej Cais, in Ingo Gregor's group in Cambridge, did the electrophysiology on GluR3. Laurence Brown in Russell Foster's group did the phenotyping of the *Gria3* mutant mice, with help from Sibah Hasan and Stuart Peirson. Kavita Praveen and Erica Davis, in Niko Katsanis' group, did the knockdown experiments on *CCDC23* in zebrafish.

Chapter 3

General lessons from the WGS500 project

3.1 Introduction

WGS500 is one of the largest clinical whole-genome sequencing datasets, and is unusual in terms of the range of diseases that it covers. In addition to allowing the identification of mutations causing disease in specific families and thus learning more about aetiology, as described in the previous chapter, it also provides an opportunity to compare the utility of different approaches for finding such mutations. In this chapter, rather than focusing on specific disease projects, we address some high-level questions about the merits of particular analysis strategies. These analyses were included in the WGS500 umbrella paper (Taylor, Martin *et al.* 2015). Acknowledgements are given at the end of this chapter.

As reviewed in Chapter 1, since the first discovery of a disease-causing mutation by exome sequencing in 2009 (Choi *et al.*, 2009), hundreds of studies have applied WES/WGS to various disorders. These studies have typically used a filter-based method for prioritising variants based on functionality, frequency, sharing between affected individuals and/or inheritance information. The search is normally limited to coding variants, since these are thought likely to have the most severe functional consequence. Indeed, most studies have used exome rather than genome sequencing and are thus restricted to coding regions. Variants at conserved genomic positions are often prioritised, since conservation across species probably indicates functional importance. There are several algorithms for predicting the consequences of variants based on conservation and the biochemical properties of the amino acid substitution, such as PolyPhen (Adzhubei *et al.*, 2010), phyloP (Pollard *et al.*, 2010) and SIFT

(Ng and Henikoff, 2003). Some studies cite scores for their best candidate variant as support for its causality (e.g. Arboleda *et al.*, 2012; Riazuddin *et al.*, 2012; Doyle *et al.*, 2012), and, in some projects in WGS500, these scores were used to prioritize variants in the absence of other information. However, the results of these predictive methods are poorly correlated and most have only modest sensitivity (0.6-0.8) and specificity (e.g. SIFT: 0.5) (Thusberg *et al.*, 2011; Li *et al.*, 2013b), although combining the output from multiple tools can improve performance (Li *et al.*, 2010).

Another key criterion applied in medical sequencing studies, and in Chapter 2, is the frequency of the variant. Early studies sometimes filtered out variants that were present in dbSNP, but this can lead to the exclusion of the causal variant if it has been seen before, such as a variant in *GJB2* that causes recessive deafness and that has a frequency of 2% in Europeans (Gasparini *et al.*, 2000). The 1000 Genomes Project and NHLBI Exome Sequencing Project (ESP) ($n \sim 6,500$ exomes) datasets are most frequently used as control sets, and have proved very valuable for reducing the number of candidates. However, these have some disadvantages:

- The very low-coverage whole-genome sequencing used in 1000 Genomes can miss variants, and the power to impute rare variants is low.
- The exome capture technologies used in the ESP (and on some 1000 Genomes samples) produce low coverage of some exons, so variants may be missed, particularly if they are near the edge of exons, in splice junctions.
- A large fraction of the individuals in the ESP have heart, lung or blood conditions, and since the publicly available variant counts are not stratified by disease status, this cohort is not appropriate for screening out variants found in patients with such diseases.
- Rare variants show considerable population stratification (Mathieson and McVean, 2012) and thus are best filtered against controls individuals chosen to have as similar ancestry as possible to the cases. Although the 1000 Genomes Project includes individuals of several different ancestries, it contains only a relatively small sample (100-200) of individuals from each population. The ESP includes only European and African Americans, albeit in larger samples. Some studies combat this by sequencing or genotyping a set of controls with ancestry matched closely to their cases, but usually the cost prevents their doing a large number

(Zhang *et al.*, 2012; Aldahmesh *et al.*, 2012; Wang *et al.*, 2011b).

- Each study will produce a set of spurious variant calls as a result of the combination of sequencing technology, mapping and calling pipelines. These are not likely to be present in publicly available samples that have been through a different pipeline, and so locally sequenced controls that have been processed in the same way as the cases are more useful for ruling out technical errors.

In addition to filtering on function and frequency, often both unaffected and affected members of the same family are sequenced so that variants can be filtered based on inheritance. For sporadic cases, a common design is to sequence a trio and search for *de novo* or recessive variants (e.g. Pagnamenta *et al.*, 2012a; Riviere *et al.*, 2012; O’Roak *et al.*, 2012b; Deciphering Developmental Disorders Study, 2015), and, for families with several affected children, to sequence a quartet (e.g. Roach *et al.*, 2010). Many disease-causing variants have been identified in consanguineous families by first identifying shared homozygous regions using SNP chip data, and then applying WES (Schossig *et al.*, 2012; Najmabadi *et al.*, 2011). These strategies were employed in the case studies described in Chapter 2. Other studies have carried out linkage analysis using SNPs extracted from the WES data, in order to eliminate a large fraction of the candidate variants (Smith *et al.*, 2011; Park *et al.*, 2013; Guergueltcheva *et al.*, 2012).

In contrast, many studies have sequenced a cohort of unrelated affected individuals and searched for genes that carry likely pathogenic variants in multiple patients. In some cases, such as *MYH3* in Freeman-Sheldon syndrome (Ng *et al.*, 2009), or *KCNT1* in malignant migrating partial seizures in infancy (Barcia *et al.*, 2012), this strategy has been successful. However, for some very genetically heterogeneous diseases such as autism, only a handful of genes are seen with multiple hits even in a large cohort (e.g. hundreds or even thousands of patients) (O’Roak *et al.*, 2012b; Iossifov *et al.*, 2012; Sanders *et al.*, 2012; Neale *et al.*, 2012), although pathway analysis can point to potentially important underlying biology. For example, Iossifov and colleagues demonstrated a significant enrichment of targets of the Fragile X Mental Retardation Protein amongst the genes containing coding mutations in autism cases (Iossifov *et al.*, 2012). Other WES studies have focused their initial analyses on genes that have already been reported for the relevant disorder (Veeramah *et al.*, 2013),

or that are functionally related to known genes (Reis *et al.*, 2013). An important question is whether this is an efficient prioritisation strategy, or whether it is better to filter first based on family information or presence of variants in the same gene in multiple unrelated individuals, and then evaluate whether a particular gene is likely to be relevant to the phenotype.

In this chapter, we will use data from the WGS500 project to evaluate different strategies for pinpointing the mutations causing putatively monogenic Mendelian diseases. We will first ask how much is gained by the consideration of parental information (Section 3.3.1), and by the use of locally-sequenced samples as controls (3.3.2). Next we will investigate the utility of candidate gene lists by examining the distribution of potentially deleterious coding variants across samples (3.3.3). Finally, we will consider the burden of putative regulatory variants (3.3.4) and then conclude the chapter with a discussion of the implications of the findings from these analyses.

Table 3.1: **WGS500 samples and projects included in the analysis in this chapter.** Multiple related individuals are counted as one set. M: presumed monogenic; I: immunological. * Note that one individual, HCM_2361, was excluded due to suspected sample contamination, and thus final number of samples analysed in this chapter was 198.

Type	Disease	Number of samples	Number of independent sets
I	Common Variable Immunodeficiency	35	35
I	Inflammatory bowel disease	13	13
I	Multiple sclerosis	6	3
I	Systemic lupus erythematosus	10	10
M	Acquired erythrocytosis	3	1
M	Asplenia	1	1
M	Cerebellar Ataxia	1	1
M	Congenital dyserythropoetic anaemia type I	1	1
M	Congenital myasthenic syndrome	5	5
M	Craniosynostosis	12	4
M	Dilated cardiomyopathy	3	2
M	Dystonia	1	1
M	Erythrocytosis	10	6
M	Familial juvenile hyperuricemic nephropathy	8	8
M	Familial tubulo-interstitial nephropathy	2	1
M	Hypertrophic Cardiomyopathy	12*	9
M	Inherited hypoparathyroidism	7	5
M	Learning disability, spastic paraparesis, hypoparathyroidism	1	1
M	Left-Ventricular Non-Compaction	2	1
M	Lipodystrophy	4	
M	Long QT syndrome/Hypertrophic Cardiomyopathy	2	1
M	Long QT syndrome	1	1
M	Mental Retardation	4	1
M	Multiple Adenoma	20	16
M	Ohtahara syndrome	22	7
M	Repolarisation abnormalities	2	1
M	Saethre-Chotzen syndrome	8	6
M	Tuberculosis	2	1
M	Variable cardiomyopathy/long QT syndrome	1	1
Total		199*	143

3.2 Methods

3.2.1 Samples

The results described in this chapter are based on 198 non-cancer samples from the WGS500 freeze 4 dataset (release date: 7 August, 2012). These comprise 143 independent sets (where a pair or group of related individuals is counted as one set) and cover twenty-nine different diseases, including four immunological disorders and twenty-five putatively monogenic syndromes (Table 3.1). The mapping and variant calling were conducted by other people in the WGS500 consortium. Briefly, the reads were mapped with Stampy (Lunter and Goodson, 2011) and variants called with Platypus (Rimmer *et al.*, 2014). Both normal and tumour samples from the cancer projects were excluded, as was an individual from the cardiomyopathy project (HCM_2361). This individual had a very large number of novel heterozygous calls for which the distribution of the fraction of reads with the alternate allele was shifted markedly towards lower values, making us suspect sample contamination (see Figure 2.6 in Chapter 2).

3.2.2 Variant annotation for overarching analysis

Originally, ANNOVAR (Wang *et al.*, 2010) was used to annotate RefSeq transcripts in all WGS500 samples, but analysis by Davis McCarthy found that an in-house tool, VariantAnno, similar to the Ensembl Variant Effect Predictor, had a more inclusive definition of splicing variants, and that annotation of all transcript isoforms increased the number of variants ascribed important functional consequences (McCarthy *et al.*, 2014). Thus, analyses in this chapter are based on VariantAnno annotations of all Ensembl V65 transcripts (see Table 3.3 for a comparison with ANNOVAR). We also annotated each variant with various conservation scores: GERP (Davydov *et al.*, 2010), phyloP (Pollard *et al.*, 2010) and phastCons (Siepel *et al.*, 2005). A variant was considered conserved if it had a GERP or phyloP score greater than 2 or a phastCons score greater than 0.95, or was in a GERP constrained element or a phastCons constrained element. We used the Ensembl Variant Effect Predictor (<http://www.ensembl.org/info/docs/variation/vep/index.html>) to add information about regulatory regions from the Ensembl V65 Regulatory Build (http://www.ensembl.org/info/docs/funcgen/regulatory_build.html), which is curated using experimental data about chromatin state, and transcription factor binding sites from Jasparr (<http://jaspar.genereg.net/>).

3.2.3 Searching for variants in candidate genes

For each disease investigated (craniosynostosis (CRS), mental retardation (MR) and early-onset epilepsy (EOE)), we compiled three lists of candidate genes based on searches of the Human Gene Mutation Database (HGMD) (Stenson *et al.*, 2009), the Mammalian Protein-Protein Interaction Database (MIPS) (Pagel *et al.*, 2005), the Gene Ontology project (GO) (Ashburner *et al.*, 2000) and the Kyoto Encyclopedia of Genes and Genomes (KEGG; <http://www.genome.jp/kegg/>). Tier 1 contains genes that are known to cause the disease of interest, tier 2 contains genes that interact with tier 1 genes or that have been linked to related disorders, and tier 3 contains genes involved in relevant biological pathways. The precise search terms used for EOE and CRS are detailed in Table 3.2. The list of X-linked mental retardation genes was compiled by searching HGMD for “mental retardation” and “intellectual disability” and then restricting to chromosome X.

Table 3.2: Search terms used to define candidate genes for early-onset epilepsy (EOE) and craniosynostosis (CRS).

Disease	Tier	Database	Search term	Number of genes
EOE	1	HGMD	Ohtahara syndrome (disease/phenotype)	1
			epileptic encephalopathy (disease/phenotype)	10
			early infantile epileptic encephalopathy (disease/phenotype)	4
	2	HGMD	epilepsy (disease/phenotype)	83
			seizure (all fields)	1
	3	MIPS	gene names from tier 1	9
		HGMD	ion channel (all fields)	240
			ion channel (gene description)	134
			ion channel (gene ontology)	226
			brain development (gene ontology)	78
CRS	1		brain development (all fields)	98
		GO	ion channel complex	200
			brain development	335
		HGMD	craniosynostosis (all fields)	11
			Saethre-Chotzen (all fields)	2
	2	MIPS	gene names from tier 1	4
		HGMD	microcephaly (disease/phenotype)	47
			macrocephaly (disease/phenotype)	10
		GO	embryonic cranial skeleton morphogenesis	32
			neural crest formation	38
			floor plate development	9
	3	KEGG	neural tube formation	26
			neural tube patterning	24
			MAPK signaling	259
			Hedgehog pathway	51
			Wnt pathway	151
			axon guidance	129

3.3 Results

3.3.1 Use of parental information

A popular design for WES and WGS studies on Mendelian syndromes has been to sequence multiple family members and use the inheritance information to filter candidate variants (Bamshad *et al.*, 2011; Roach *et al.*, 2010). There were fourteen families in WGS500 from which both the healthy parents and their affected offspring were sequenced. An important question is the extent to which including the parental information helped to reduce the number of candidates. The following analysis is based on six trios and one quartet with early-onset epilepsy (EOE) (described in Chapter 2), a quartet with mental retardation and disturbed circadian rhythm (MR) (also described in Chapter 2), three trios with craniosynostosis (CRS) and two trios with Saethre-Chotzen syndrome (SC), a type of CRS. It also includes an erythrocytosis (ERY) trio in which both the mother and daughter were affected; although dominant inheritance was suspected here, the trio was treated in this analysis as though the mother were unaffected.

Firstly, let us consider a *de novo* dominant model for the twelve trios but suppose we do not have parental information. We are looking for novel heterozygous coding variants in the trio children (Table 3.3). These are variants that are absent from dbSNP, 1000 Genomes, ESP and from all of the individuals in WGS500 that are not related to the proband of interest. There are, on average, 189 such variants per child (standard deviation, SD = 125; range 109-508) that are confidently called as heterozygous (genotype log likelihood ratio (GLLR) < -5) (top row of Table 3.3). If we then apply the condition that both parents must be confidently called as homozygous (GLLR < -5), thus filtering for *de novo* variants, this average is reduced by 90% to 19.3 (SD = 11; range 6-36). Manual inspection of the reads suggested that many of these putative *de novo* variants were actually inherited but had been missed in the raw calls from Platypus (Rimmer *et al.*, 2014). Thus, joint calling was performed for each trio, such that the observation of a variant in one of the individuals reduced the required threshold for calling it in the others. Based on these joint calls, the average number of putative *de novo* coding variants, is reduced by a further 86% to 2.75 per trio (SD=1.76; range 0-5). Thus, using the parental information eliminates 99% of putative *de novo* coding candidates.

Table 3.3: **The number of putative *de novo* protein-altering variants in trio children under different filtering strategies.** When segregation is ignored, this is simply the number of novel heterozygous coding variants as defined by absence from the control datasets indicated. Parental genotypes (either raw calls or jointly called genotypes) were checked to determine whether the variants were likely to be *de novo*. Annotations were done either with VariantAnno or ANNOVAR, with both sets of results shown here for comparison. All results presented in the text, and other tables and figures, are from the VariantAnno annotations.

Annotation tool	Segregation	Control datasets	CRS 4103	CRS 4447	CRS 4917	SC 2930	SC 4659	OTH 2	OTH 5	OTH 7	OTH 12	OTH 15	OTH 18	ERY PAR07c
VariantAnno	ignored	1000G+ESP+WGS500	131	113	117	120	119	109	165	132	232	508	372	151
		1000G+ESP	316	338	357	331	318	301	368	348	496	813	630	396
		WGS500	181	163	164	172	171	159	203	179	391	569	436	206
	<i>de novo</i> raw	1000G+ESP+WGS500	17	9	18	11	6	12	11	33	36	35	12	32
		1000G+ESP	79	94	93	77	63	80	68	100	126	142	86	124
	WGS500		22	12	19	18	9	15	14	38	55	41	15	38
VariantAnno	<i>de novo</i> joint	1000G+ESP+WGS500	3	1	1	2	0	4	1	5	5	3	3	5
		1000G+ESP	7	7	3	2	2	11	1	6	8	7	11	7
		WGS500	3	1	1	2	0	4	1	5	5	3	3	5
	ignored	1000G+ESP+WGS500	90	81	88	97	94	76	113	101	164	383	275	125
		1000G+ESP	201	225	237	222	175	187	201	234	316	576	445	264
		WGS500	130	121	126	135	134	112	142	142	270	433	319	165
ANNOVAR	<i>de novo</i> raw	1000G+ESP+WGS500	12	5	14	7	6	9	3	26	25	30	11	24
		1000G+ESP	48	66	60	49	28	54	24	75	78	101	62	79
		WGS500	16	5	12	10	7	9	5	29	37	34	11	30
	<i>de novo</i> joint	1000G+ESP+WGS500	1	1	1	2	0	4	1	4	3	3	3	5
		1000G+ESP	3	5	3	2	2	9	1	5	3	7	10	7
		WGS500	1	1	1	2	0	4	1	4	3	3	3	5

Let us next investigate a simple recessive model, but assume we do not have parental information (Table 3.4). Our putative causal mutations will be rare homozygous coding variants and we will set the following frequency cutoffs: 0.5% in 1000 Genomes and ESP (corresponding to an expected homozygous frequency of 1 in 40,000), and 0 homozygotes and 2 heterozygotes amongst the unrelated WGS500 samples. We will consider all trio children except the two with an abnormally large number of rare homozygous variants described in Chapter 2: OTH_5 (who has complete isodisomy for chromosome 9) and OTH_12 (whose parents are consanguineous). For these ten children, the average number of rare homozygous coding variants is 2.9 (SD = 0.4; range 1-7). If we then require that both the parents be called as heterozygous in the joint calls, such that the variant follows a simple recessive model, the average is 0.4 (SD = 0.5; range 0-1). It was somewhat surprising that so many (86%) of these homozygous calls failed the Mendelian inheritance check. Manual inspection revealed that these tended to be in low-complexity regions which are prone to spurious or incorrect calls, or have very low coverage. For the two quartets, there were no rare coding variants that were homozygous in both children and heterozygous in both parents, and hence that would fit a simple recessive model.

We can now do the same analysis for the X-linked model (Table 3.5). There are an average of 5.6 rare hemizygous coding variants on chromosome X in the male probands (SD = 3.6; range 3-14). All of these are heterozygous in the mother. Thus, the use of maternal genotype information does not help reduce the number of X-linked candidates. There is only one X chromosomal coding variant that, apparently, does not fit the X-linked model: this is a variant in the MR family that was miscalled as heterozygous in one of the affected sons (in fact, this was the *GRIA3* variant described in Chapter 2, probably due to contamination of that sample).

The fourth plausible model for these Mendelian families would be one of compound heterozygous inheritance, in which the affected children have inherited two deleterious heterozygous variants in the same gene, one from each parent (Table 3.6). In the absence of parental information, we can only look for genes with at least two rare heterozygous coding variants. Applying the same frequency thresholds as for the simple recessive model, there are an average of 23 genes per trio child that fit this criterion (SD = 20; range 8-78). Requiring that at least one of the variants be inherited from each parent reduces this by 83% to an average of 6 genes per trio child

Table 3.4: **The number of putative simple recessive protein-altering variants (VariantAnno annotation) in trio and quartet children under different filtering strategies.** When segregation is ignored, this is simply the number of rare homozygous coding variants as defined by their frequency in the control datasets indicated. In order to classify the variant as recessive, the parents were both required to be heterozygous (in the raw calls) and, in families with two affected children, they both had to share the same homozygous variant (OTH 21 and 22; MR 6 and 8). Note that OTH 5 has an abnormal number of rare homozygous variants due to isodisomy for one chromosome, and OTH 12 due to consanguinity. These two samples were excluded from the averages given in the text.

Segregation	Control Datasets	CRS 4447	CRS 4103	CRS 4917	SC 4659	SC 2930	OTH	OTH	OTH	OTH	OTH	OTH	OTH	OTH	OTH	MR	MR	ERY PAR07c
ignored	1000G+ESP+WGS500	1	2	5	7	1	2	18	4	24	2	2	3	1	3	5	3	
	1000G+ESP	219	215	230	281	238	206	253	219	237	238	236	239	261	202	195	242	
	WGS500	2	3	7	8	1	2	20	5	126	10	2	4	4	3	10	3	
recessive	1000G+ESP+WGS500	1	1	0	0	0	1	0	0	16	0	1	0	0	0	0	0	
	1000G+ESP	23	21	30	20	31	16	34	28	45	18	41	6	5	5	4	30	
	WGS500	2	1	2	0	0	1	0	1	88	6	1	0	0	0	0	0	

Table 3.5: **The number of putative X-linked recessive protein-altering variants (VariantAnno annotation) in male trio and quartet children under different filtering strategies.** When segregation is ignored, this is simply the number of rare hemizygous coding variants as defined by their frequency in the control datasets indicated. In order to classify the variant as X-linked recessive, the mother was required to be heterozygous (in the raw calls) and, in the family with two affected boys, they both had to share the same hemizygous variant (MR_6 and MR_8). Since the top candidate causal variant in the MR quartet was originally missed because it was erroneously called as heterozygous in MR_8, we also show the counts obtained if heterozygous calls on the X are considered.

Allowing heterozygous genotype	Inheritance	Control datasets	CRS 4447	CRS 4917	SC 2930	OTH 2	OTH 5	OTH 12	MR 6	MR 8
No	ignored	1000G+ESP+WGS500	3	4	7	5	5	14	4	3
		1000G+ESP	13	11	12	10	14	22	11	10
		WGS500	3	6	11	5	5	30	6	5
	XLR	1000G+ESP+WGS500	3	4	7	5	5	14	3	3
		1000G+ESP	12	11	12	10	12	21	9	9
		WGS500	3	6	11	5	5	30	5	5
Yes	ignored	1000G+ESP+WGS500	5	4	7	5	9	14	18	19
		1000G+ESP	34	26	24	23	32	34	38	46
		WGS500	5	6	11	5	9	30	20	21
	XLR	1000G+ESP+WGS500	3	4	7	5	9	14	12	6
		1000G+ESP	22	20	24	15	24	27	26	21
		WGS500	3	6	11	5	9	30	14	8

(SD = 7; range 1-19).

So in summary, for most of the inheritance models, using parental information substantially decreases the number of candidate variants.

3.3.2 Screening against WGS500

A standard approach in medical next-generation sequencing projects is to filter variants based on their frequency in publicly available datasets such as 1000 Genomes and the NHLBI ESP. Sequencing local controls can add value firstly if their ancestry is well matched to those of the cases (e.g. Zhang *et al.*, 2012; Aldahmesh *et al.*, 2012; Wang *et al.*, 2011b), and secondly because this allows reduction of technical artefacts specific to the sequencing and analysis pipeline. For each individual disease project in WGS500, we routinely filtered out variants that were seen at greater than some threshold frequency in these public datasets or in other WGS500 samples. In the case of a severe Mendelian disorder, this seems appropriate since the causal mutation is presumed to be highly penetrant and therefore should be absent from or, for recessive mutations, very rare and only heterozygous in, individuals who do not have that disease. It is pertinent to ask how much was gained by screening against WGS500 in addition to 1000 Genomes and ESP.

Let us first pretend that the causal mutation in each of the 198 WGS500 individuals is a novel heterozygous coding variant. If novelty is defined by absence from dbSNP, 1000 Genomes and ESP, there are, on average, 384 such variants in each individual (SD = 107; range 253-813) (Figure 3.1; Table 3.3). If we also filter out variants seen in other unrelated WGS500 samples, the mean number of novel heterozygous coding variants is 163 (SD = 89; range 84-508). Thus, including information from WGS500 reduces the number of candidate variants under this model by 58% across all samples. This reduction is greater for the 173 Caucasians than for the 25 non-Caucasians (62% vs. 44%), which we discuss further below.

Suppose now that we imagine the causal variant is inherited by a simple recessive model, but do not have parental information (Figure 3.2; Table 3.4). We are looking for rare homozygous coding variants, and will set an allele frequency cutoff of 0.5%. If we filter based on the frequency in 1000 Genomes and ESP alone, each person has an average of 240 such variants (SD = 28; range 184-349), excluding individuals with

Table 3.6: **The number of genes with putative compound recessive protein-altering variants (VariantAnno annotation) in trio and quartet children under different filtering strategies.** When segregation is ignored, this is simply the number of genes with two rare heterozygous coding variants as defined by their frequency in the control datasets indicated. In order to classify the variants as compound recessive, each of them was required to have been inherited from a different parent (in the raw calls) and, in families with two affected children, they both had to share the same variants (OTH_21 and OTH_22; MR_6 and MR_8).

Inheritance	Control datasets	CRS 4103	CRS 4447	CRS 4917	SC 2930	SC 4659	OTH 0007	OTH 12	OTH 15	OTH 18	OTH 0021	OTH 0022	OTH 2	OTH 5	MR 6	MR 8	ERY PAR07c
ignored	1000G+ESP+WGS500	18	17	8	17	11	16	78	48	30	15	19	15	13	80	66	16
	1000G+ESP	81	88	93	96	80	89	155	142	101	95	102	79	80	165	141	99
	WGS500	26	24	15	21	17	20	411	79	42	32	34	18	20	83	72	22
compound recessive	1000G+ESP+WGS500	3	2	1	5	2	2	19	14	18	1	1	2	3	2	2	1
	1000G+ESP	8	5	6	15	6	6	28	22	25	8	4	7	8	4	3	13
	WGS500	5	5	4	6	2	2	113	21	21	1	1	4	6	2	2	2

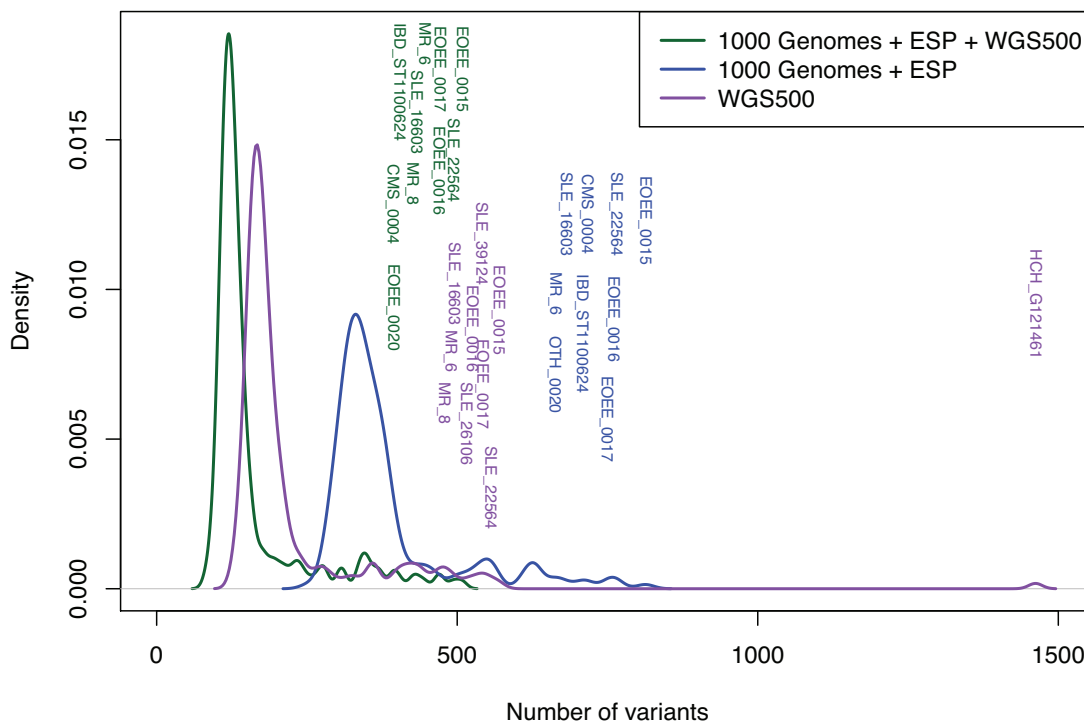


Figure 3.1: **Density plot of the distribution of the number of novel heterozygous coding variants across all individuals.** Novelty was defined as absence from the control datasets indicated, and VariantAnno was used for annotation. The individuals in the top 5th percentile are shown; all of these are known to have African or South Asian ancestry except for MR_6 and MR_8, which we suspect have some sample contamination (see Figure 2.6 in Chapter 2).

large regions of homozygosity due to consanguinity or uniparental disomy. If we also filter variants based on their frequency in unrelated WGS500 samples (0 homozygotes, ≤ 2 heterozygotes), the average number of variants meeting these criteria is only 6 (SD = 10; range 0-67). This constitutes a 98% reduction, and is very similar for Caucasians and non-Caucasians. Thus, having our own cohort of control samples sequenced using the same technology and processed through the same pipeline allowed us to substantially reduce the number of candidate variants.

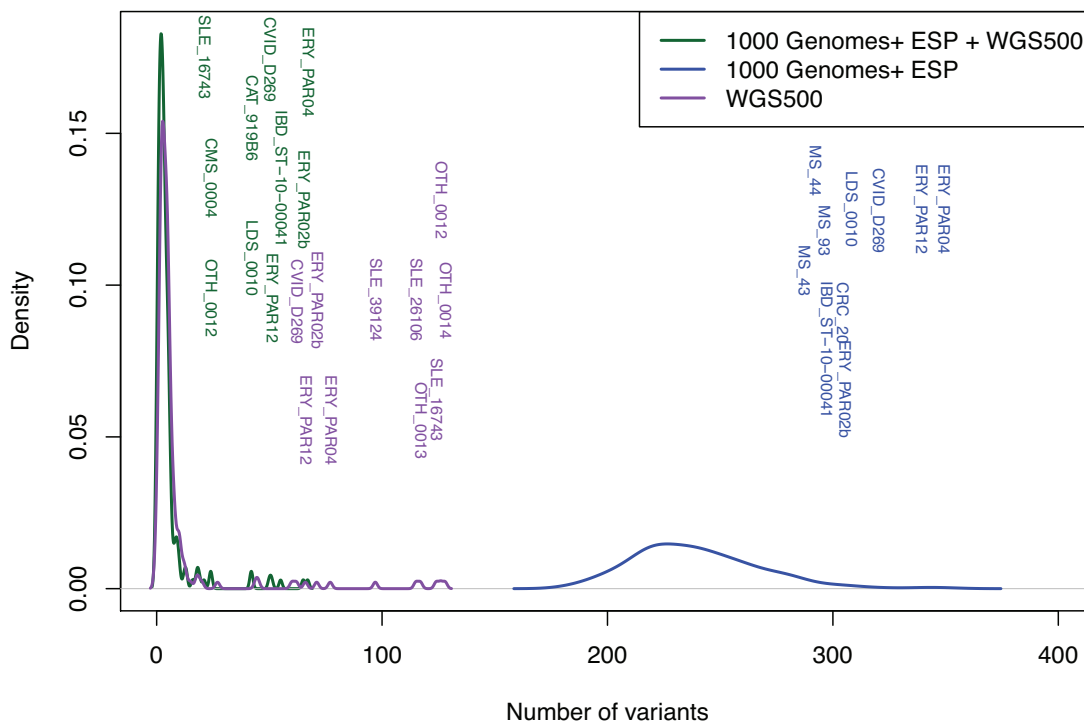


Figure 3.2: **Density plot of the distribution of the number of rare homozygous coding variants across all individuals.** Rarity was defined by an allele frequency less than 0.005 in the indicated control datasets, and VariantAnno was used for annotation. The individuals in the top 5th percentile are shown; all of the outliers for the most stringent filtering strategy (1000 Genomes+ESP+WGS500) are known to have consanguineous parents.

3.3.3 Utility of candidate gene lists

In WGS500, two complementary strategies were used to find putative causal variants for the Mendelian diseases. For family-based projects such as those described in Chapter 2, we first screened for variants that would fit appropriate inheritance models, and then investigated the published literature on those genes to determine which were good candidates. For projects in which little or no family data were available, lists of candidate genes were composed for the relevant disease, and these genes were screened for rare, potentially deleterious variants in the proband(s). Extra family information was then sought to check whether these variants segregated. This latter approach has been the standard one used in the clinic to test for mutations in known genes for a disease, but often parental data are not available and so, if a

rare and potentially deleterious variant is found, it is often deemed to be likely causal.

In WGS500, both of these approaches produced plausible candidates and, in some cases, further functional work and/or screening of other patients has shown that these are definitely or highly likely to be causal. The examples of the EOE trios and the MR quartet, described in Chapter 1, demonstrated how powerful the “inheritance first” strategy can be. In this section, we investigate the value of using candidate gene lists compiled in a semi-automated manner.

We chose as test cases the Mendelian disease projects mentioned above in Section 3.3.1: EOE, MR, and CRS (treating CRS and SC together). For each disease, we made three lists of candidate genes based on high-throughput searches of several public databases (see Section 3.2.3 and Table 3.2). In general, tier 1 contains genes that are recorded in the Human Gene Mutation Database (HGMD) (Stenson *et al.*, 2009) as causing the disease of interest, tier 2 contains genes that interact with tier 1 genes (as defined in the MIPS database (Pagel *et al.*, 2005)) or that are listed in HGMD as causing related diseases (e.g. for EOE, “epilepsy” was considered a related disease), and tier 3 contains genes that are involved in biological pathways known to be involved in pathogenesis.

We next screened for coding variants in these genes that would appear to fit a *de novo* dominant (DN), simple recessive (SR), X-linked (XL), or compound heterozygous (CH) model in the absence of parental information, across the 198 WGS500 samples; these are referred to as “pseudo-candidates”. That is, we looked for novel heterozygous variants (DN model), very rare homozygous/hemizygous variants (SR/XL models) or genes with two very rare heterozygous variants (CH model; phase was ignored since it was unknown), filtering against 1000 Genomes, ESP and unrelated WGS500 samples.

Figure 3.3 shows the distribution of the number of variants meeting these criteria for each of the EOE gene lists. Clearly, many individuals have rare or novel variants in the EOE candidate genes, although most of them do not have EOE. For example, four individuals have a heterozygous pseudo-candidate in a tier 1 EOE gene, but none of them have severe epilepsy; they presented with multiple sclerosis, familial juvenile hyperuricemic nephropathy, inherited hypoparathyroidism and asplenia respectively. Fifty-six percent of all individuals have at least one heterozygous pseudo-candidate

in a gene implicated in epilepsy (tier 2), and 87% in a gene involved in brain development or function (tier 3). These numbers are lower for the SR and CH models, but there are nonetheless still many controls who have suggestively pathogenic variants in genes that are good candidates for a disease they do not have.

Similar results were found for CRS (Figure 3.4), in which 8% of samples have at least one heterozygous pseudo-candidate in the 11 known genes, but none of these individuals had been ascertained for this phenotype. Even more strikingly, 40% of males have at least one pseudo-candidate in a gene that had been previously reported for X-linked mental retardation (Figure 3.5), including the two brothers who were referred for this condition for whom this variant, in *GRIA3*, is thought likely to be causal (see Chapter 2). This result emphasises how important it is to have further information about the variant (demonstrating, at the very least, its biochemical effect on the protein, and ideally also recapitulating the phenotype in an *in vivo* model) or other patients with mutations in the same gene, in order to prove causality.

A second point to note from Figures 3.3 and 3.4 is that the probands generally have several novel or rare coding variants in candidate genes for their disease that are not considered to be causal because they do not fit the correct inheritance model. There was no significant enrichment (using Fisher’s exact test) for pseudo-candidates in any of the tiers for either heterozygous or homozygous candidates among the EOE or CRS patients compared to controls. Conversely, of the twelve families with Mendelian diseases for which we think we know the causal mutation (Table 3.7), only four of these mutations are in candidate genes for that disease that were found using high-throughput database searches.

3.3.4 Regulatory variants

Until this point, we have focused exclusively on coding variants, as has been standard for NGS studies of Mendelian disease, since these are probably most likely to be causal and also easiest to interpret. A major advantage of WGS500 over many other high-coverage medical sequencing datasets is that it includes noncoding as well as exonic variants. Although we only found two cases in which the best candidate for a causal variant was noncoding (Taylor *et al.*, 2015), WGS500 nonetheless provides a useful resource in which to examine the extent of regulatory variation across a large number of clinical samples. Determining whether a variant is likely to effect gene

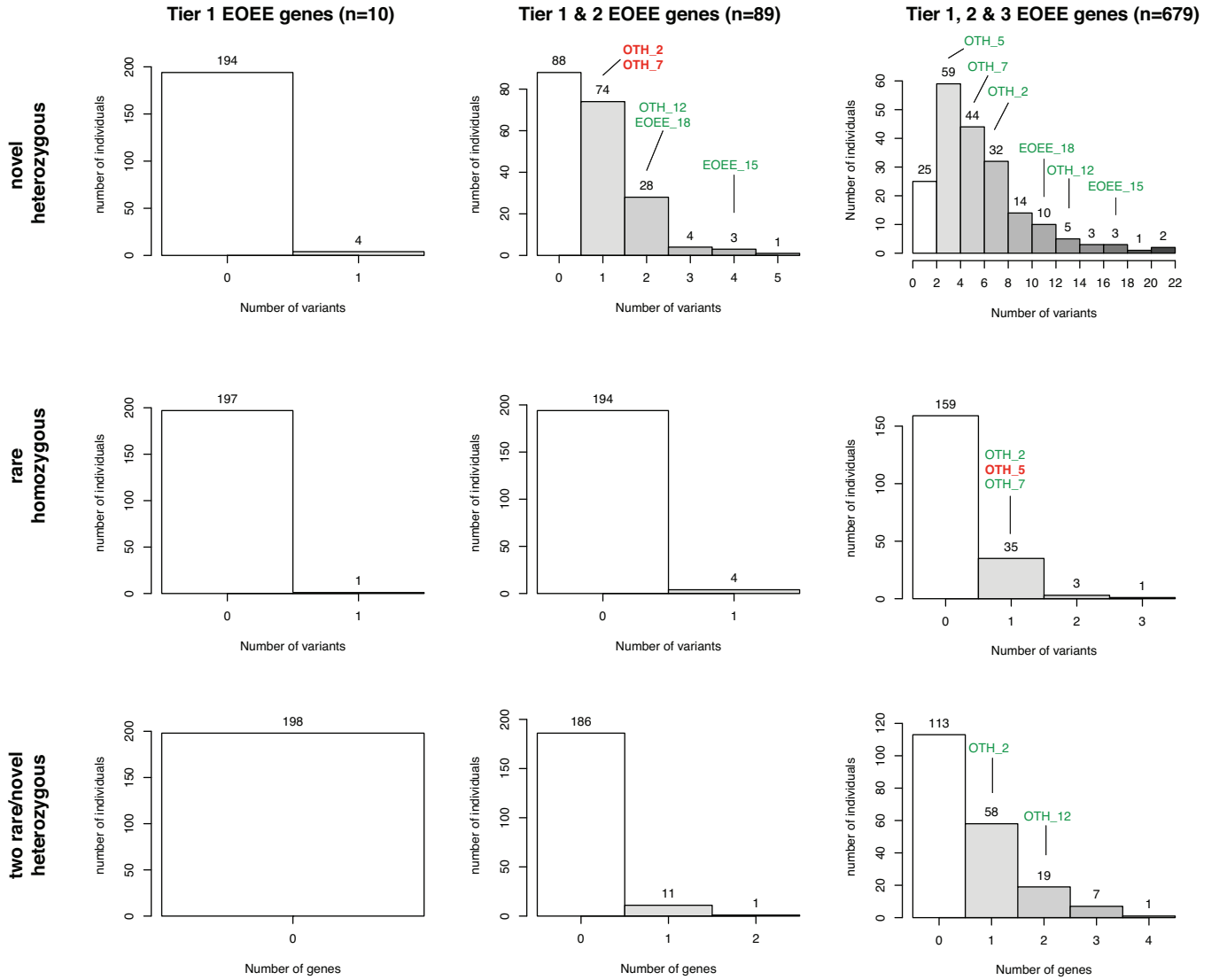


Figure 3.3: **Variant burden in candidate genes for early-onset epilepsy.** Histograms of the number of potentially pathogenic (conserved) coding variants in different candidate gene sets for EOE across all 198 WGS500 freeze 4 samples, under different inheritance models. Top row: novel heterozygous (*de novo* dominant model); middle row: rare homozygous (simple recessive model); bottom row: two rare/novel heterozygous (compound heterozygous model). The candidate genes were chosen by high-throughput searches (see Section 3.2.3). If found in a patient with this disease, these variants would be good candidates for pathogenicity, but in fact, only the indicated individuals are affected. Sample names in green text indicate that the variant is not likely to be pathogenic (since it does not fit a plausible inheritance model or is less functionally compelling than another candidate); red text indicates that the variant is thought to be causal.

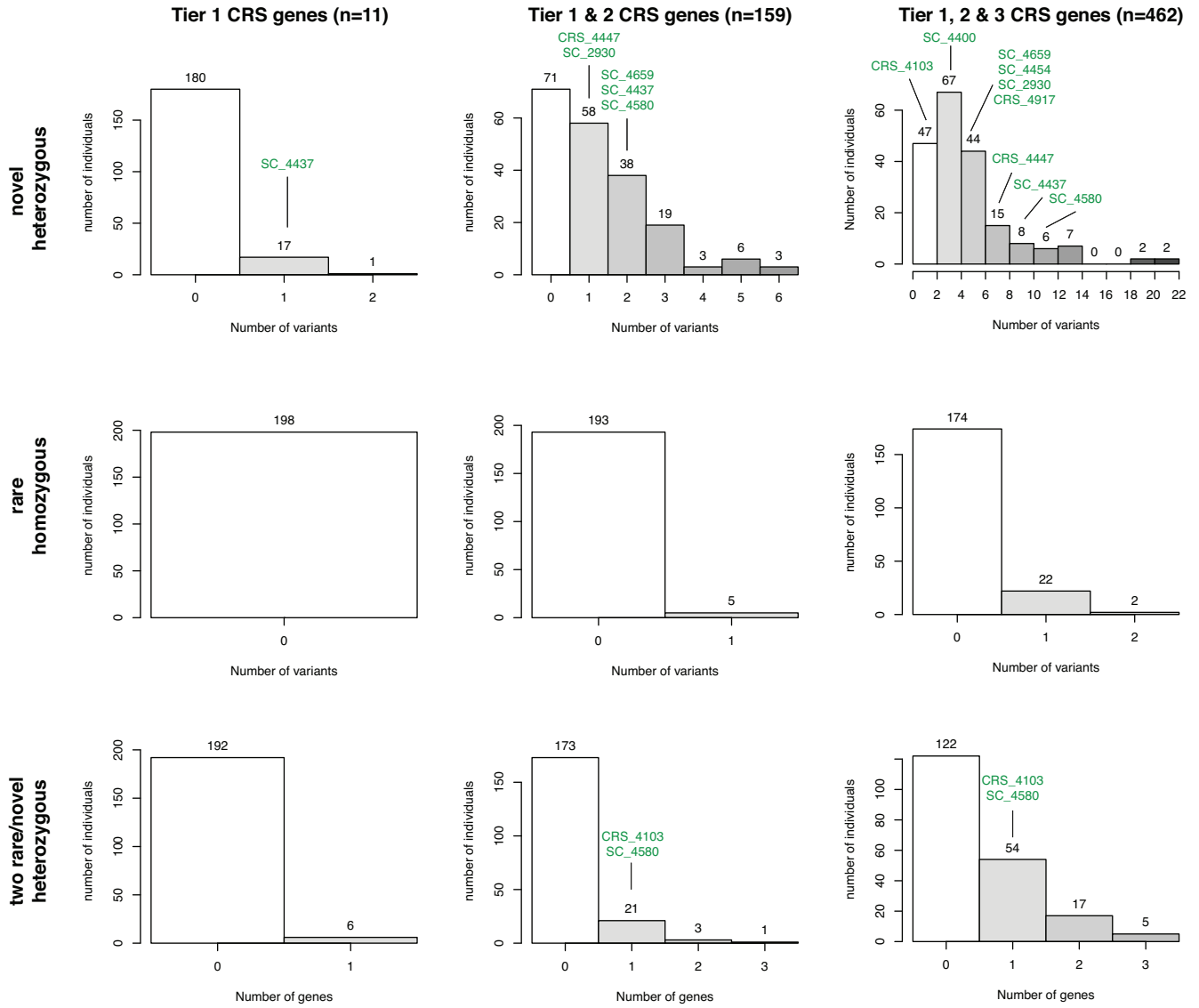


Figure 3.4: **Variant burden in candidate genes for craniosynostosis.** Histograms of the number of potentially pathogenic, conserved coding variants in different candidate gene sets for CRS across all WGS500 freeze 2 samples, under different inheritance models. SC: Saethre-Chotzen syndrome. Note that the CRS/SC patients include five from the trios mentioned in Section 3.3.1.

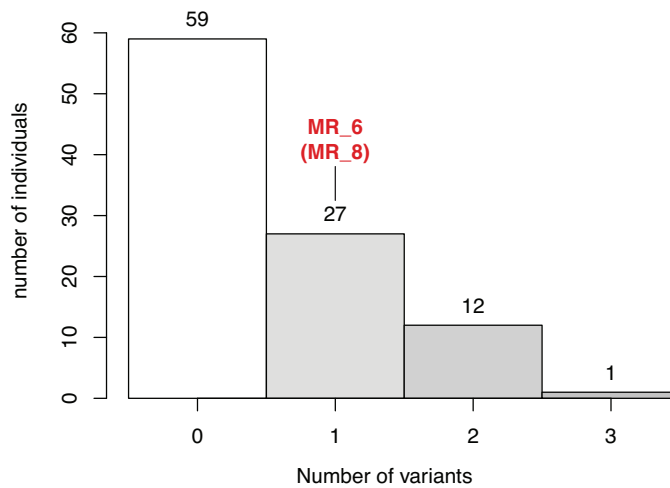


Figure 3.5: **Variant burden in candidate genes for X-linked mental retardation.** Histograms of the number of potentially pathogenic, conserved coding variants in known XLMR genes (see Section 3.2.3) across all male WGS500 freeze 4 samples. If found in a patient with intellectual disability, these variants would be good candidates for pathogenicity, but only the indicated individuals (described in Chapter 2) were ascertained for this phenotype. MR.8 is indicated in brackets because this variant, the one in *GRIA3*, was actually called as heterozygous in this individual, possibly due to sample contamination (Figure 2.6).

Table 3.7: **The variants thought to be causal in the EOE and CRS trios and the MR quartet.** The positions given are for hg19. The mutations are given in the format: reference allele>alternate allele.

Sample	Position	Mutation	Gene	Consequence	Inheritance	Candidate gene
CRS_4103	chr17:2290734	G>A	<i>MNT</i>	nonsynonymous	<i>de novo</i>	no
CRS_4447	chr3:147131157	C>A	<i>ZIC1</i>	nonsense	<i>de novo</i>	yes, but for EIEE and MR tier 3
CRS_4917	chr7:97873936	G>A	<i>TECPR1</i>	nonsynonymous	<i>de novo</i>	no
SC_4659	chrX:53674333	C>T	<i>HUWE1</i>	nonsynonymous	<i>de novo</i>	yes, but for MR tier 1
SC_2930	chr11:63883876	C>T	<i>FLRT1</i>	nonsynonymous	<i>de novo</i>	no
OTH_0007	chr2:166245874	A>G	<i>SCN2A</i>	nonsynonymous	<i>de novo</i>	tier 2
OTH_12	chr16:625837	A>G	<i>PIGQ</i>	essential splice site	recessive	no
OTH_15	chr15:64499785	G>A	<i>CSNK1G1?</i>	nonsynonymous	<i>de novo</i>	no
OTH_18	chr11:119149219	G>A	<i>CBL?</i>	essential splice site	<i>de novo</i>	no
OTH_2	chr20:62071051	G>A	<i>KCNQ2</i>	nonsynonymous	<i>de novo</i>	tier 2
OTH_5	chr9:138675924	G>A	<i>KCNT1</i>	nonsynonymous	recessive - UPD9	tier 3
MR_6, MR_8	chrX:122561871	G>A	<i>GRIA3?</i>	nonsynonymous	X-linked	tier 1

expression is still an open problem. Some groups have used ENCODE RNA-seq data and ChIP-seq data on transcription factor (TF) binding and chromatin marks to identify regulatory regions (Ward and Kellis, 2012; Ritchie *et al.*, 2014). We decided to use regions from the Ensembl Regulatory Build, which is based on integration of data on DNase1 hypersensitivity sites, histone modifications, and polymerase and TF binding. We focused on variants at conserved positions in regulatory regions within 50kb of genes.

Figure 3.6 shows the distribution of the number of these putative regulatory variants across the WGS500 samples (red line); the mean for novel heterozygous variants was 610 (SD=313; range 298-1889), almost four times as many as the equivalent number for coding variants (Figure 3.1; mean=163). Interestingly, the number of variants does not substantially change if one considers only regulatory regions within 5kb of genes (Figure 3.6, black line), since regulatory regions tend to be close to genes anyway. Notably, many individuals have pseudo-candidates in regulatory regions close to candidate genes for EOE, CRS and MR (Figure 3.7). There are a few that fit relevant inheritance patterns in the probands for those diseases that could potentially be interesting candidates, but the top coding candidates in each of those families seemed more compelling (Table 3.7), since one can more easily predict their functional consequences.

To further hone in on variants that were most likely to alter gene expression, we narrowed the criteria to include only those variants that were within transcription factor binding sites (TFBSs) defined by Jaspar position weight matrices (PWMs). There were considerably fewer of these (Figure 3.8), an average of 5.6 per sample (SD=5.7; range 0-18). All except one of these could be eliminated as candidates for the EOE, MR and CRS families once parental information was considered.

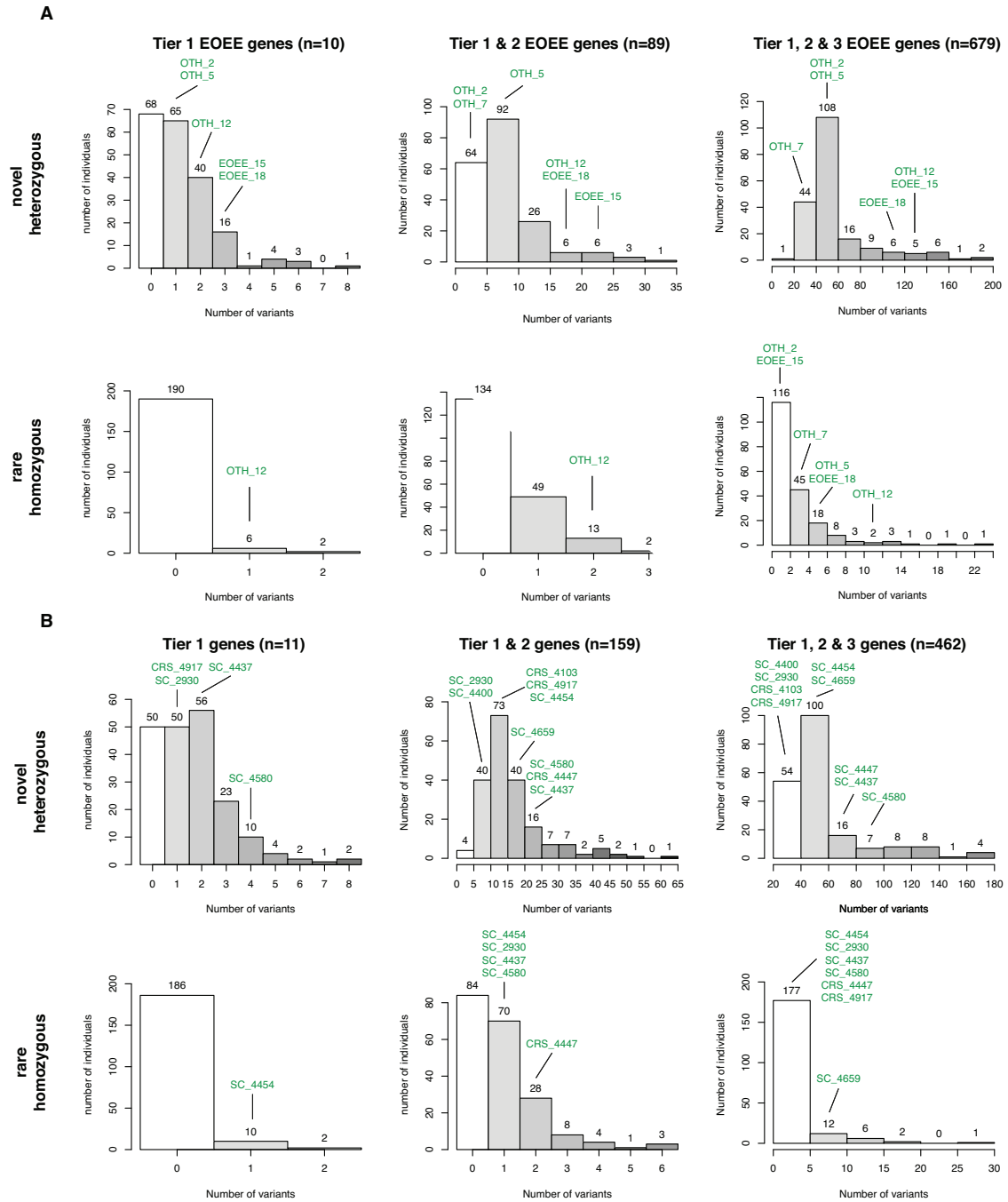


Figure 3.7: The burden of putative regulatory variants around candidate genes for EOE and CRS. As for Figures 3.3 and 3.4, but for variants at conserved positions in regulatory regions within 50kb of a gene.

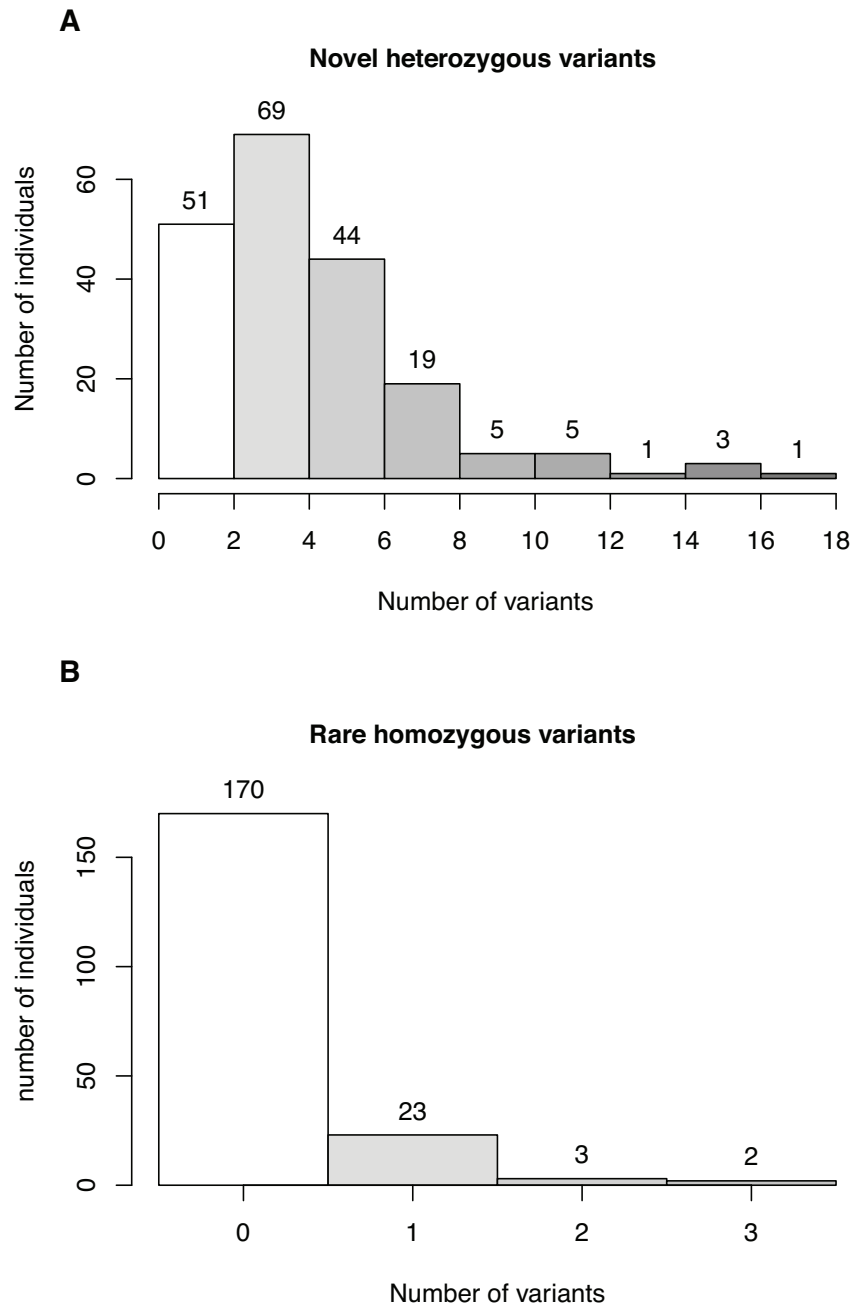


Figure 3.8: **The burden of variants likely to disrupt transcription factor binding sites (TFBSs).** (A) and (B) show the distributions of the number of novel heterozygous and rare homozygous variants respectively that alter conserved positions in TFBSs (see Section 3.2.2) within 50kb of a gene.

3.4 Discussion

In this chapter, we examined the merits of different analysis strategies in the WGS500 project. Our experience, and those of other large WES studies (Yang *et al.*, 2014; Beaulieu *et al.*, 2014; Wright *et al.*, 2015; Deciphering Developmental Disorders Study, 2015), should be useful in designing future clinical sequencing projects such as Genomics England. The focus was on Mendelian disease projects, since these are the most tractable, and because we had several test cases of families in which the causal mutation was known with reasonable confidence. However, some of the points illustrated by these analyses, such as the importance of having controls sequenced in-house, and the abundance of potentially regulatory variants, are relevant to WGS studies of complex disease too.

A key component of the strategies employed in Chapter 2 to the EOE and MR families was the use of parental information to filter for variants that would fit relevant inheritance models. In this chapter, we demonstrated that this approach substantially reduced the number of candidates: the number of coding variants to consider decreased by 99%, 86% and 83% for the *de novo*, simple recessive and compound recessive models respectively. This is similar to results from the DDD study: Wright *et al.* (2015) reported that each proband had an average of 30 rare coding variants overlapping known developmental disorder genes, but this was reduced to an average of 1 per proband (range 0-13) when inheritance information was used. Somewhat at odds with these findings, Beaulieu *et al.* (2014) reported that in the FORGE Canada study, the parental data was not necessary for the detection of known disease-causing genes in genetically heterogeneous conditions because, in the majority of cases, a single candidate was evident before even considering the parents. However, as they pointed out, implicating novel genes containing *de novo* variants is difficult without parental information.

We were unsurprised by the large reduction in candidates we observed under the *de novo* model when using parental information, since humans are known to have a large number of very rare variants due to recent population size expansion (Fu *et al.*, 2013), and although many of these will be private to families, and hence appear as singletons when filtering against unrelated people, we should only expect one or two *de novo* coding variants. The reduction in the number of compound recessive candidates reflects the current difficulty in phasing variants using NGS data. As read

lengths increase, the chance of a single read overlapping both variants will be higher, and so it may not be necessary to have parental information in order to determine whether two variants in the same gene are on the same or different chromosomes. The X-linked model was the only one for which using the parental data added no information, since variants called hemizygous in males were almost always heterozygous in their mothers. Although one would expect 1-2 *de novo* variants per male on chromosome X, these will generally be noncoding. For this reason, other WGS studies of X-linked disorders have sequenced only the proband, or several probands, and used SNP chip data to check inheritance (Johnston *et al.*, 2012; Tarpey *et al.*, 2009); however, this approach cannot be applied to small pedigrees in which the inheritance pattern is unclear, such as the MR quartet with just two affected brothers.

The WGS500 project was designed to include multiple patients with different diseases partly on the premise that it would be useful to use individuals from the large cohort as controls for each specific disease project, as long as they did not have related conditions. Filtering against ~200 WGS500 samples eliminated a large proportion of the candidate variants that remained after screening against 1000 Genomes and ESP: 58% of novel heterozygous and 98% of rare homozygous variants. Initially, it seemed surprising that this was so helpful for homozygous variants, since it was assumed that homozygous calls were less likely to be spurious. However, when parental information was incorporated for the ten trios, it emerged that 88% of the rare homozygous variants (defined by having a frequency < 0.005 in 1000 Genomes and ESP alone) were Mendelian errors. Of the variants that were correctly inherited as simple recessive (mean 26 per trio if 1000 Genomes and ESP were used for filtering; second bottom row of Table 3.4), many were in low complexity regions that are difficult to call, and 37% were not in Hardy-Weinberg equilibrium ($p < 0.05$), suggesting genotyping errors. Thus, most of these variants were probably artifacts of Stampy mapping (Lunter and Goodson, 2011) and Platypus calling (Rimmer *et al.*, 2014) that were eliminated by screening against other WGS500 samples. This emphasises how important it is that the variant calls are produced in the same way for both cases and controls; relying on publicly available datasets will not remove common artifacts specific to the sequencing technology or mapping and calling pipeline.

The benefits of filtering against other WGS500 samples were greater for the European samples than for those from Africa and Asia. The variants that were eliminated by this step were presumably either a) sequencing or calling errors specific to

our dataset or b) real variants absent from public databases, which are dominated by Caucasian samples. The non-Caucasian samples presumably have more variants which have not been seen before but which are real. Since there are so few individuals with Asian or African ancestry in WGS500, this filtering step did not help as much to eliminate these variants as it did for the Caucasians. Thus, ideally, medical sequencing projects should try to include a large number of samples with similar ancestry, to avoid having to rely on public databases.

In contrast to the “inheritance first” strategy we used to find the causal mutations in the EOE and MR families (Chapter 2), other projects in WGS500 focused first on variants in candidate genes for the relevant disease, and then sought family information to check segregation. The efficacy of this latter approach depends on the burden of potentially pathogenic variants in the healthy population. How much weight should one put on the discovery of a novel missense variant at a conserved position in a known epilepsy gene in an Ohtahara syndrome patient, given the distribution of such variants across individuals without this disorder? In examining the distribution of such variants in candidate genes for EOE, CRS and MR (Figures 3.3 and 3.4), we saw that it was common to find pseudo-candidates in individuals who did not have the disease in question and that, without parental information, this candidate gene approach would lead to a substantial number of false positives. The frequency of such pseudo-candidates should be of particular concern to the field, since, at least when the WGS500 study was starting, it was still quite common practise in clinical genetics to screen a panel of candidate genes for mutations, and then look no further if a plausible variant is found, particularly if family information was unavailable. The results presented here and elsewhere (e.g. Bell *et al.*, 2011; Norton *et al.*, 2012; Xue *et al.*, 2012; Groth *et al.*, 2015) emphasise that this often happens even in unaffected individuals, and thus, caution must be taken in interpreting such variants when found in patients.

The abundance of rare coding variants in known XLMR genes in males is sobering. If these variants were detected in an individual with MR, they would be considered prime candidates. However, since most have completely unrelated phenotypes, this merely represents the considerable background variation in these genes that has previously been noted (Tarpey *et al.*, 2009). Indeed, for this reason, Wright *et al.* (2015) decided not to consider any inherited missense variants on the X chromosome in

the DDD project, except if the patient had a family history of the disorder consistent with X-linked inheritance or if the variant was present in HGMD or the Leiden Open Variation Database. In investigating the X-linked *GRIA3* variant described in the previous chapter, we were careful to seek further bioinformatic and functional evidence of pathogenicity, but we will not be confident of its causality unless other patients with *GRIA3* mutations and circadian rhythm disruption are found, or unless a robust phenotype is found in the knock-in mouse.

It is also important to emphasise that this candidate gene-based approach leads to false negatives too (Table 3.7). For most of the EOE and CRS trios, the most likely causal genes were only identified as good candidates through reading the recent literature about their biochemical functions and about related genes. These would have been missed if the search had been restricted to candidate gene lists automatically generated using databases such as HGMD and GO. On the other hand, if we had used OMIM to construct our gene lists (particularly for tier 1) using Human Phenotype Ontology terms (Köhler *et al.*, 2014), we may not have had such a problem with false negatives, since this database tends to be more up-to-date, although false positives would still have been a problem. The best approach is undoubtedly to use family information first, if available, then to consider candidate variants (fitting the right inheritance pattern) in a well-curated list of candidate genes, as well as applying sensitive and specific text-mining algorithms to query the literature for information on the remaining genes. The laborious approach employed for most projects in WGS500, of investigating manually every variant that was sufficiently rare and fitted the right inheritance model, will simply not be feasible in large-scale clinical sequencing projects such as Genomics England.

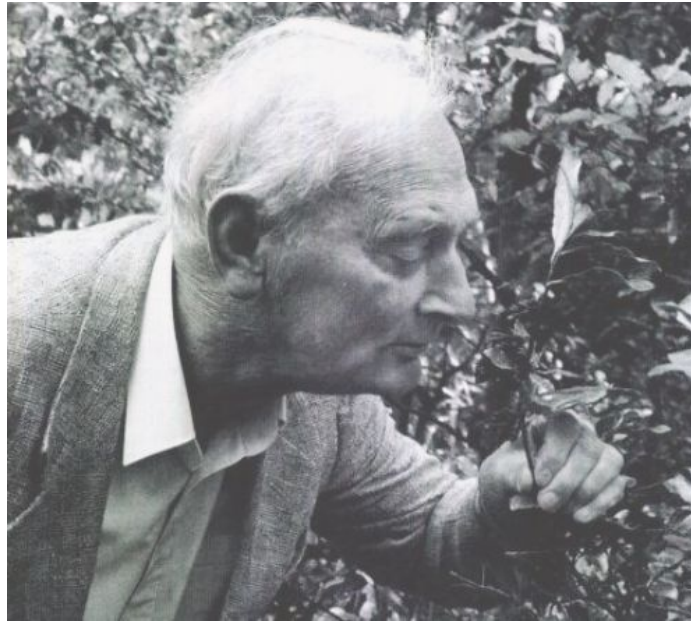
For the monogenic disease projects described in Chapter 2, we looked only at coding variants, except in the few cases in which no good coding candidates were found (LDS_10 and the EOE quartet). For studies of more complex disease and for unsolved Mendelian diseases, it would be useful to be able to better hone in on potentially pathogenic regulatory variants. One way to do this could be to use the set of regulatory regions defined by Ensembl, based primarily on ChIP-seq from ENCODE. We found that all individuals in WGS500 had hundreds of rare variants at conserved positions within these regions (Figure 3.6), and most had several such variants close to genes that are known to cause a disease they do not have (Figure 3.7). Thus, inheritance information in large pedigrees, extensive functional work and/or statistical

evidence (burden analyses) will be necessary to demonstrate causality. Incorporating ENCODE data from the relevant cell lines (if known) may help to reduce the number of candidates to consider. The approach of Ritchie *et al.* (2014) for prioritising noncoding variants based on factors such as conservation, DNase I hypersensitivity, and certain histone modifications may be helpful.

Following many reports of incorrect assignments of causality (Bell *et al.*, 2011; Norton *et al.*, 2012; Xue *et al.*, 2012), MacArthur *et al.* (2014) recently published a set of guidelines for establishing the pathogenicity of variants. They placed particular emphasis on the importance of reporting statistical evidence for segregation or association of each variant. These guidelines have not yet been widely adopted, although increasingly, studies are using statistical tests as an important criterion for demonstrating causality. For example, the DDD study tested for an enrichment of *de novo* variants on a per-gene basis, using estimates of gene-specific mutation rates (Samocha *et al.*, 2014) to construct the null distribution of the number expected by chance (Deciphering Developmental Disorders Study, 2015), and other groups have implicated genes in heart defects on the basis of their being enriched for rare variants in a cohorts of cases *versus* controls from other studies (Al Turki *et al.*, 2014; D'Alessandro *et al.*, 2015). It is expected that the guidelines suggested by MacArthur *et al.* (2014), when applied to larger and larger cohorts, should accelerate the discovery of new Mendelian disease genes.

3.5 Acknowledgements

Gil McVean and Peter Donnelly contributed substantially to the design of analyses in this chapter. Andrew Wilkie and Steve Twigg provided guidance on the search terms to use for craniosynostosis. The Genomics Core at WTCHG did the read mapping, generated the variant calls, and did some of the annotation. Peter Humburg assisted with running VariantAnno. I also thank Stefano Lise for useful discussions.



To Nick,
When your
turn comes, do it
better.
Dad. 1975

My grandfather, Peter Gordon Martin (1923-1994), wrote this dedication to my father in his book on the chromosomes of Australian mammals (Hayman and Martin, 1974). He was an inspirational teacher and an influential pioneer of evolutionary genetics in Australia. I dedicate this platypus chapter to him.

Chapter 4

Whole-genome sequencing of the platypus

4.1 Introduction

4.1.1 Natural history of the platypus

As we saw in Chapter 1, next-generation sequencing technologies have greatly facilitated studies into the diversity and population structure of non-model organisms, including chimpanzees (Auton *et al.*, 2012; Leffler *et al.*, 2013), other primates (Xue *et al.*, 2015; Prado-Martinez *et al.*, 2013) and pigs (Ai *et al.*, 2015). They have also spearheaded the *de novo* assembly of many new genomes, leading to a greater understanding of the biology of various species, including marine mammals (Foote *et al.*, 2015), fish (Wang *et al.*, 2015) and birds (Jarvis *et al.*, 2014). We decided to undertake population sequencing of the platypus (*Ornithorhynchus anatinus*), a semi-aquatic mammal with a strange mixture of mammalian and reptilian characteristics that lives in the freshwater rivers and lakes of eastern Australia and Tasmania (Figure 4.1). Platypus¹ are monotremes, mammals that lay eggs. The other monotremes, echidnas, which live throughout Australia and in Papua New Guinea, share some features with the platypus (e.g. milk, eggs) but appear phenotypically very different, resembling hedgehogs and porcupines. A recent study using fossil and molecular data estimated that the monotremes, or Prototheria, diverged about 166 million years ago (MYA) from the other mammalian lineage, the Theria, which split about 148MYA into the marsupials (Metatheria) and placentals (Eutheria) (Bininda-Emonds *et al.*,

¹“Platypus” is a latinised form of the Greek *πλατυς* (platus, flat) and *πους* (pous, foot). Properly, the plural should be “platypodes”, although “platypus” and “platypuses” are frequently used. In this thesis, we shall use “platypus” as both the singular and plural.

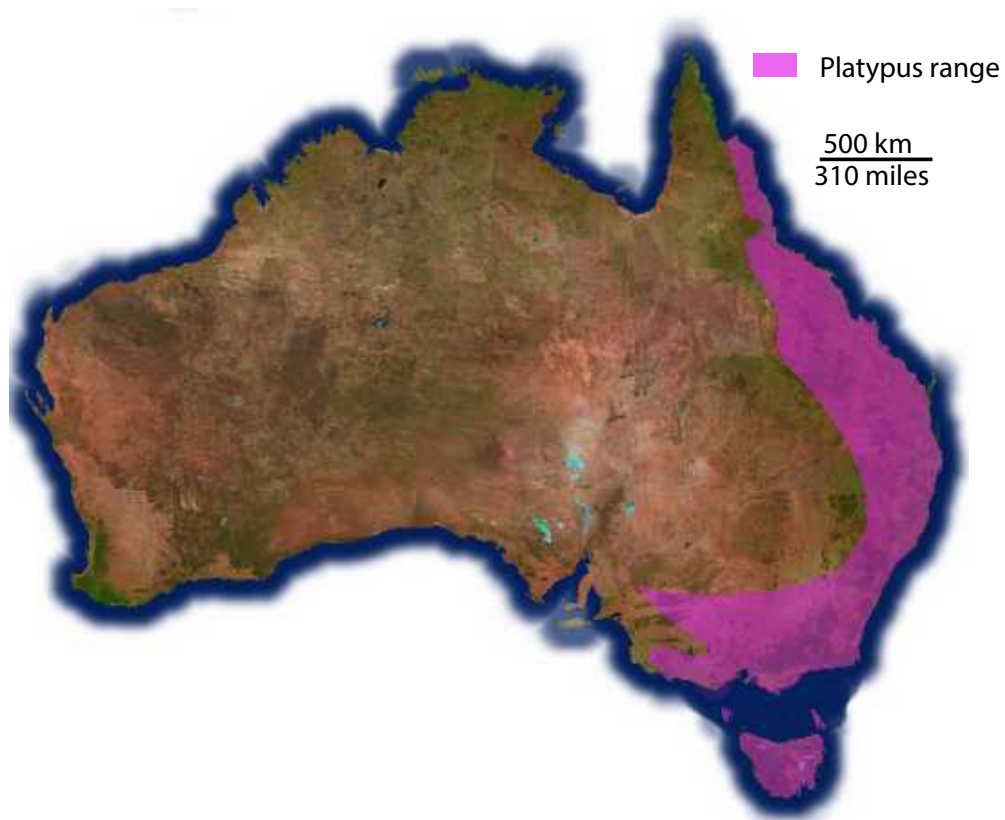


Figure 4.1: **The range of the platypus.** This map was obtained from <http://www.wildlife.org.au/wildlife/speciesprofile/mammals/platypus.html>.

2007) (Figure 4.2). Thus, the monotremes may be informative about the most recent common ancestor (MRCA) of all mammals. For example, a trait shared between monotremes and birds, if it has the same genetic basis, probably existed in the MRCA of all mammals.

In addition to laying eggs, platypus have an unusual set of characteristics. These have been described in detail by Grant (2007), so we will just summarise them briefly here. Platypus have fur, webbed feet, and a large bill that contains electroreceptors used for sensing their prey (Figure 4.3). They feed on water-dwelling insects and their larvae, shrimps and tadpoles. Male platypus have a venomous spur which is thought to be used for fighting other males during the mating season. They live in burrows on the banks of rivers, where the females lay their eggs. They typically lay two small eggs about 11mm in diameter, which develop *in utero* for about 28 days before being incubated externally between the tail and abdomen for 10 days. These eggs are meroblastic, meaning that the ovum splits incompletely, leaving the yolk

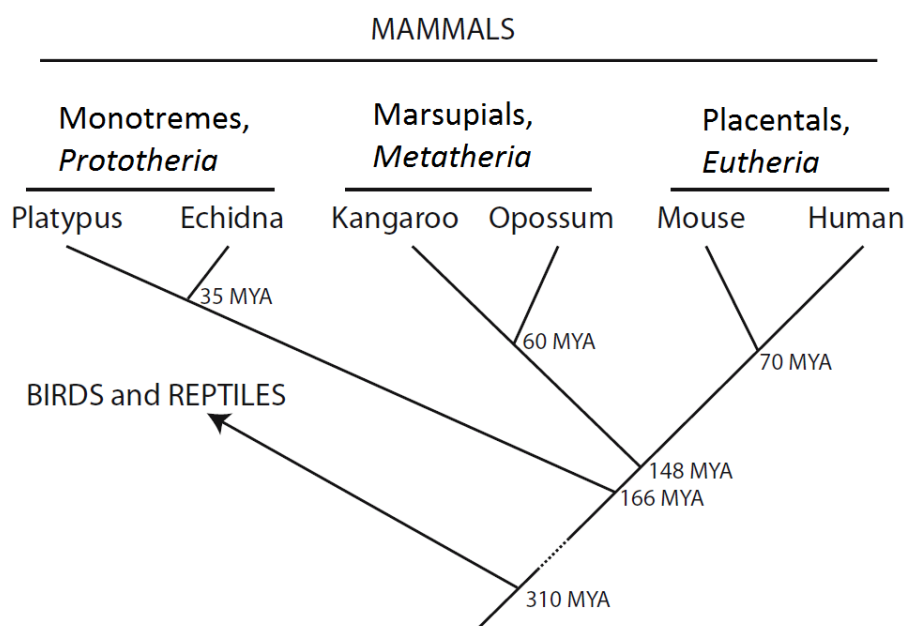


Figure 4.2: **Phylogenetic tree of mammals and other vertebrates.** This figure is taken directly from McMillan *et al.* (2007), and the divergence dates are from Bininda-Emonds *et al.* (2007).

continuous with the egg’s cytoplasm. This is similar to the eggs of birds and reptiles, but differs from those of mammals, which go through holoblastic cleavage i.e. after fertilisation, the ovum splits into multiple divisible daughter cells. Once hatched, the blind and hairless young are fed by their mother’s milk, which is secreted directly through pores in the skin, since female platypus have no teats. Platypus have been reported to live up to 20 years in the wild, and 21 years in captivity.

4.1.2 The platypus genome project

Early work on the platypus genome focused on determining the number of chromosomes (Hayman and Martin, 1974). This was debated for several years until Bick and Sharman (1975) established that both males and females had $2n = 52$, but that several pairs were not exact homologs. This eventually turned out to be due to the numerous sex chromosome pairs, discussed below. Warren *et al.* (2008) produced the first platypus genome assembly using established whole-genome shotgun methods. The reference individual was a female named Glennie, from the Barnard River in New South Wales (NSW). Warren *et al.* generated $\sim 6X$ coverage in paired-end reads of average length $\sim 700\text{bp}$ from whole-genome libraries of plasmids, fosmids and bacterial artificial chromosomes (BAC), which had insert sizes of 1.5-10kb, $\sim 40\text{kb}$ and $\sim 150\text{kb}$, respectively. These were assembled into contigs and then into scaffolds using PCAP² (Huang *et al.*, 2006). The resultant assembly contained 1.84Gb of sequence (excluding about 0.16Gb of gaps), which is about 80% of the estimated total genome size, $\sim 2.3\text{Gb}$. The assembly was further scaffolded by aligning the BAC-end sequences to a fingerprint map made by digesting the BAC library with a restriction enzyme and overlapping the BACs on the basis of the resulting fragment pattern. Lastly, about 437Mb of the sequence was ordered and oriented along twenty of the chromosomes using fluorescent *in situ* hybridisation (FISH).

The final assembly, called ornAna1, had 201,523 scaffolds, many of which are actually unscaffolded contigs. The N50 length was 983,898bp and the N50 number was 39,590, meaning that 50% of the sequence was contained in 39,590 scaffolds greater than 983,898bp. This was considerably more fragmented than the chimpanzee and chicken assemblies, which were produced using similar methods and had similar coverage and read quality. Warren *et al.* did not have a satisfactory explanation for

²A contig is a contiguous sequence made by overlapping DNA segments. A scaffold consists of one or more oriented contigs connected together, separated by gaps.



Figure 4.3: **The platypus.** Note the bill and fur in panel A, and the webbed feet in panel B. Photos obtained from www.zoo.org.au/healesville/animals/platypus and www.animalfactguide.com/animal-facts/platypus/.

this. The average GC content in the assembled platypus genome is 45.5%, rather higher than the average 41% in eutherian and chicken genomes. There was evidence that the platypus reads were being prematurely truncated in regions of higher GC content, which may have contributed to the low contiguity. Another factor could be the higher interspersed repeat content (Table 4.A.1). The most common repeats were copies of the 5kb long-interspersed-element (LINE2) and of the SINE with which it co-evolves (mammalian-wide interspersed repeat, MIR, Mon-1 in monotremes), which each comprise about 20% of the genome.

Warren *et al.* predicted 18,527 protein-coding genes in ornAna1, of which 82% have orthologues in chicken, dog, mouse, human and opossum. Analysis of these genes gave clues as to the underlying basis of various platypus characteristics. For example, a large expansion of the vomeronasal receptor type 1 family, completely absent from chicken and anole lizard, may have facilitated pheromonal communication or the detection of water-soluble odorants during foraging. The venom genes seem to have evolved independently in platypus and reptiles from the same gene families, including the β -defensins. Platypus have four genes encoding proteins of the zona pellucida (egg envelope) in humans, and two seen only in birds, fish and amphibians. Like those in other mammals, the platypus casein genes, which encodes milk proteins, are tightly clustered in the genome, with the addition of a duplicated β -casein gene. Thus, the genome of the platypus has some features in common with mammals, some with sauropsids (reptiles and birds), and some unique to this species.

4.1.3 Previous work on platypus population genetics

The original genome paper shed light on several aspects of platypus biology but included only a limited analysis of inter-individual variation and population structure based on 57 polymorphic retrotransposon loci. Subsequently, several other studies have investigated population structure at fine and broad scales, using microsatellites or mitochondrial DNA (mtDNA).

Kolomyjec *et al.* (2009) genotyped 12 polymorphic microsatellites in 120 samples, predominantly from the adjacent Shoalhaven and Hawkesbury River systems near Sydney (see Figure 4.7). They found greater variance between than within the river systems, and reported that 13 of the 120 individuals appeared to be first-generation

migrants from one river system to the other³. The data supported an isolation-by-distance scenario between the two unconnected river systems, which implies that there must be overland dispersal of platypus, consistent with previous reports from recapture and radio-tracking data (Grant, 2007). Of note, during normal conditions, the Shoalhaven and Hawkesbury systems are only $\sim 1.5\text{km}$ apart at their closest point, and so terrestrial migration is probably relatively easy.

Gongora *et al.* (2012) sequenced part of the mitochondrial control region and of the cytochrome b gene in 284 samples from 22 river basins across the whole platypus range, along the east coast from Tasmania to North Queensland (from 16°S to 43°S). They found that the samples fell into two major clades, corresponding to mainland Australia and Tasmania, and that the mainland samples split into two subclades, corresponding to Northeastern Queensland+Southern Queensland/Northern Murray-Darling and Southern Murray-Darling/New South Wales+Victoria (see Figure 4.7 for a map). The Bayesian phylogenetic tree presented in Gongora *et al.* (2012) is reproduced in Figure 4.4 for reference. As shown in this figure, the time to the MRCA (TMRCA) of the major platypus clades was estimated from the mtDNA control region to be 0.94MYA (95% credible interval 0.39–1.52MYA). The estimate from the third codon sites of the cytochrome b gene was slightly more recent (0.7MYA, 95% credible interval 0.1–1.46MYA), but both estimates are long before Tasmania diverged from the mainland as a result of rising sea levels ($\sim 12,000$ years ago).

Furlan *et al.* (2013) analysed 13 microsatellite loci and two mitochondrial genes in 752 platypus samples from New South Wales, Victoria and Tasmania. They found very little genetic differentiation among 183 individuals from different parts of Tasmania but substantial structure between populations on the mainland, and they presented evidence that suggested more frequent overland travel in Tasmania than in Victoria, possibly due to the wetter, cooler climate and higher density of rivers. This corroborates previous reports of a higher frequency of overland movement in Tasmanian platypus, with dispersal events of greater than 20km (Connolly *et al.*, 1998; Otley *et al.*, 2000). By comparing the maternally-inherited mtDNA markers to the nuclear microsatellites, Furlan *et al.* (2013) found evidence for male-biased dispersal, which led to patterns of higher genetic similarity between females within a short river

³This was determined using methods which first calculate the likelihood that an individual comes from a particular population, and then identify individuals with a high probability of being misassigned, who are presumed to be migrants (Piry *et al.*, 2004).

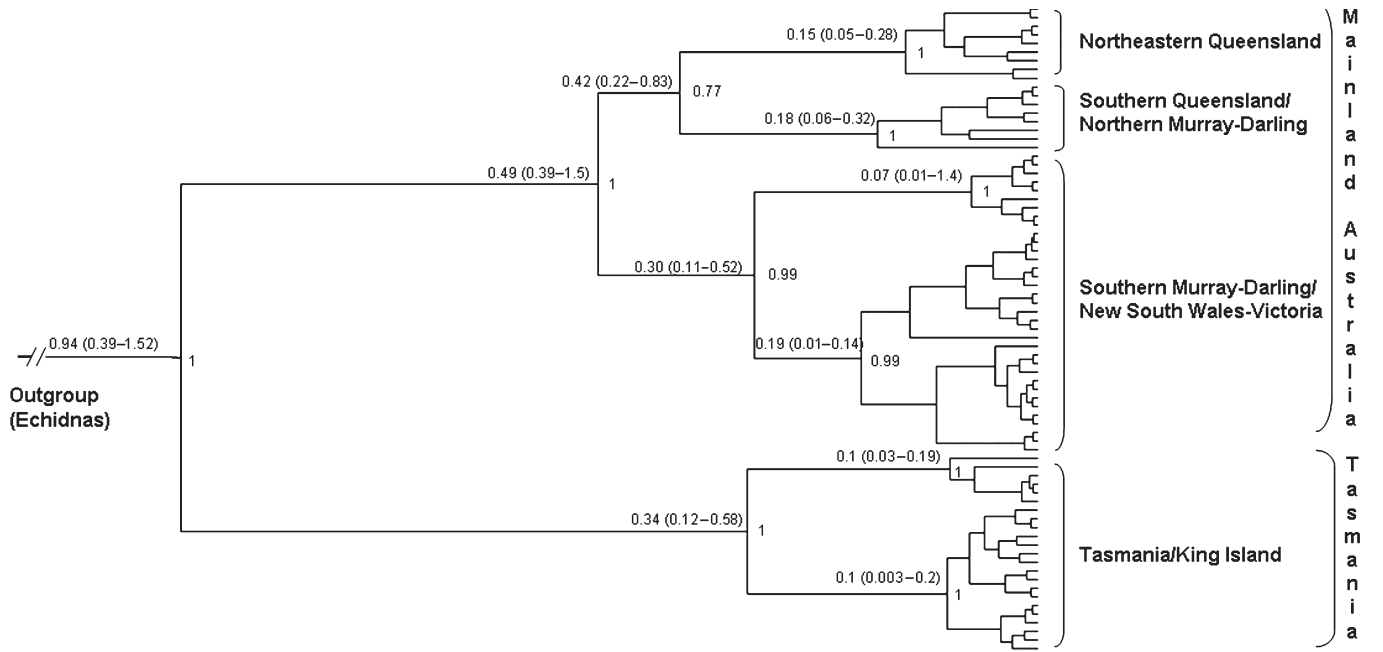


Figure 4.4: **Bayesian phylogenetic tree based on mtDNA from 74 platypus samples, taken from Gongora *et al.* (2012).** Divergence time estimates (in MYA) are shown on the branches, with 95% credibility intervals in brackets, and posterior probabilities after the nodes. The prior distribution for the age of the root node (the platypus-echidna split time) was modelled as a normal distribution with a mean of 32MYA and a standard deviation of 7.14MY. This choice was based on a recent molecular estimate which placed the divergence time at between 19 and 38MYA (Phillips *et al.*, 2009).

distance ($\leq 1.4\text{km}$) than between males. Finally, they also reported no genetic differentiation at the measured loci between Victorian platypus either side of the Great Dividing Range, a range of mountains (maximum altitude 2,228 meters) that runs $\sim 3,500\text{km}$ along the eastern coastline of Australia (shown as a grey transparent line in Figure 4.7). This was somewhat surprising since it differs from what was seen in other aquatic and semi-aquatic species, and since the steep slopes and absence of connecting waterways were anticipated to restrict platypus dispersal.

These papers have provided an important glimpse at platypus population structure and differentiation, but we anticipated that more could be learned from whole-genome sequencing data, which contains considerably more information than the small number of microsatellites and the mtDNA examined in previous studies.

4.1.4 Platypus sex chromosomes

Perhaps the most remarkable aspect of platypus biology is their bizarre sex chromosome system. Before describing this, we will give a brief overview of the sex chromosomes in other species and how they are thought to have formed (for a more thorough review, see Charlesworth *et al.*, 2005; Graves, 2006). In therians, females have two copies of the X chromosome, and males have one X and one Y⁴. The Y chromosomes are typically much smaller than their X chromosome partners (called “gametologs”), but they share homology within the pseudoautosomal regions, and, to some extent, between coding genes in the X-specific and Y-specific regions. The equivalent is true for the Z and W chromosomes in birds, except that the females are heterogametic, with a ZW karyotype. Heteromorphic sex chromosomes are thought to have evolved independently in multiple lineages, but by a similar mechanism. Various lines of evidence support the theory that they evolved from a pair of homologous autosomes after one of the pair acquired a sex-determining (SD) locus, or two interacting SD loci, leading to suppression of recombination (Nei, 1969). This reduces the efficiency of selection, leading to the accumulation of deleterious mutations which precipitates gene decay, build-up of repeats, and eventual divergence between the members of the pair. If the recombination suppression happened at different times, perhaps through

⁴There are a few exceptions to this. Mole voles have XX females and X0 males (i.e. with one X but no Y) (Just *et al.*, 2007), and both the salt flat mouse (*Salinomys delicatus*) (Lanzone *et al.*, 2011) and the marsupial potoroo (*Potorus tridactylus*) (Sharman and Barber, 1952) have independently evolved a second Y chromosome, so the females are XX and the males XY1Y2.

separate inversion events, this can lead to regions of the X and Y (or Z and W in birds) having quite different levels of divergence. These are known as “strata”, and several of them have been identified in placentals (Lahn and Page, 1999; Sandstedt and Tucker, 2004), of which the oldest contains the sex-determining gene *SRY*. In many species, one or more regions of the sex chromosomes retain homology and continue to recombine, possibly because recombination is essential for proper segregation. These pseudoautosomal regions (PARs) can contract as a result of the expansion of the non-recombining region through the creation of new strata (Otto *et al.*, 2011).

Early karyotyping work suggested that the platypus had an XY system like most placentals and marsupials. Murtagh (1977) reported that there were an additional four unpaired elements which, in male meiosis, formed a chain of ten chromosomes along with the sex chromosomes and two pairs of small autosomes. Watson *et al.* (1990) mapped the orthologs of eight genes present on the human and marsupial X to the putative platypus X using radioactive *in situ* hybridisation, and thus concluded that the monotreme and therian X chromosomes were derived from a common ancestor.

However, the story became more complicated when Grutzner *et al.* (2004) carried out chromosome painting in male metaphase cells using paints derived from flow sorted chromosomes, and found that, in fact, there were five different male-specific chromosomes (named Y chromosomes), and five different chromosomes present in one copy in males and two copies in females (X chromosomes). These were observed to form a multivalent chain in male meiosis, alternating X1Y1X2Y2X3Y3X4Y4X5Y5, with connections through nine pseudoautosomal regions (PARs) (Figure 4.5) to ensure segregation into X1X2X3X4X5-bearing and Y1Y2Y3Y4Y5-bearing sperm. Grutzner *et al.* reported that X1 (the chromosome previously reported to have homology to the human X) and Y1 were at one end of the chain, and shared a PAR comprising a whole chromosome arm, suggesting that this pair was at an early stage of differentiation from an autosomal to fully sex chromosomal pair. In contrast, Y5, at the other end of the chain was tiny, and any homology with X5 was too small to be detected by chromosome painting, indicating that Y5 had been through a long period of degeneration, as explained earlier. Thus, Grutzner *et al.* proposed that the chain had started with the X5Y5 pair, and progressed through a series of reciprocal translocations involving exchanges with autosomes to the X1Y1 end. The platypus chain is very similar to that of the echidna, the main difference being that echidna has a

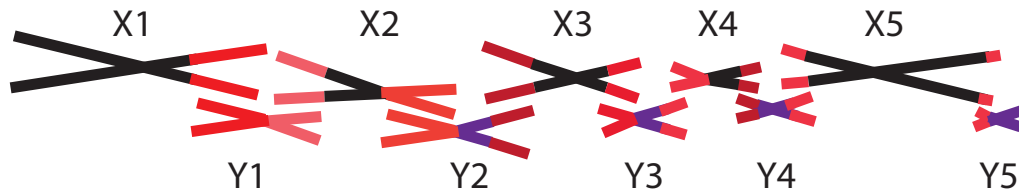


Figure 4.5: **Schematic of the chain of the ten sex chromosomes in platypus male meiosis.** Unpaired X-specific regions are indicated in black, unpaired Y-specific regions (also known as male-specific Y, MSY) in purple, and the nine different pseudoautosomal regions (PARs) in different shades of red/pink. Note that Y5 is the most degenerated, indicating that the X5Y5 pair is the oldest, whereas the youngest pair, X1Y1, still share a large region of homology. In females, the five pairs of X chromosomes each synapse like normal autosomes.

fusion of Y3 and Y5 (Rens *et al.*, 2007), and so it must be at least 35 million years old (Bininda-Emonds *et al.*, 2007). A similar translocation chain is seen in some termite species (Syren and Luykx, 1977; Martins and Mesa, 1995) and in evening primroses (Cleland, 1972). Grutzner *et al.* showed that the *DMRT1* gene, the dosage-sensitive sex-determining gene located on the bird Z chromosome, was present on the platypus X5. Thus, they posited that this chromosome had homology to the bird Z, but that *DMRT1* was unlikely to be the sex-determining gene in platypus, since the dosage would be opposite to that in birds. Platypus were found to lack *SRY*, so it was unclear how their sex was determined.

Veyrunes *et al.* (2008) took 32 BACs that mapped to female X chromosomes in the FISH experiments conducted as part of the platypus genome project (Warren *et al.*, 2008), and hybridised these to male mitotic metaphase chromosomes. They used the BAC end sequences to determine which of the scaffolds in *ornAna1* they came from, and identified the genes on those scaffolds in order to uncover homology with the human and chicken genome. The 32 X-chromosomal BACs were found to contain genes with orthologs only on human autosomes (primarily 2, 3, 4, 5, 6, 9 and 19), and the majority of the orthology with chicken was to the Z chromosome, mostly on X5 and X3 (see Figure 4.6). Twelve other BACs containing genes orthologous to the X conserved region⁵ in therians mapped to platypus chromosome 6, including six genes that Watson *et al.* (1990) had reported as being on X1. It is possible that, in

⁵This is the region, comprising two-thirds of human chromosome X, which is shared with marsupial chromosome X, and therefore presumed to be ancestral.

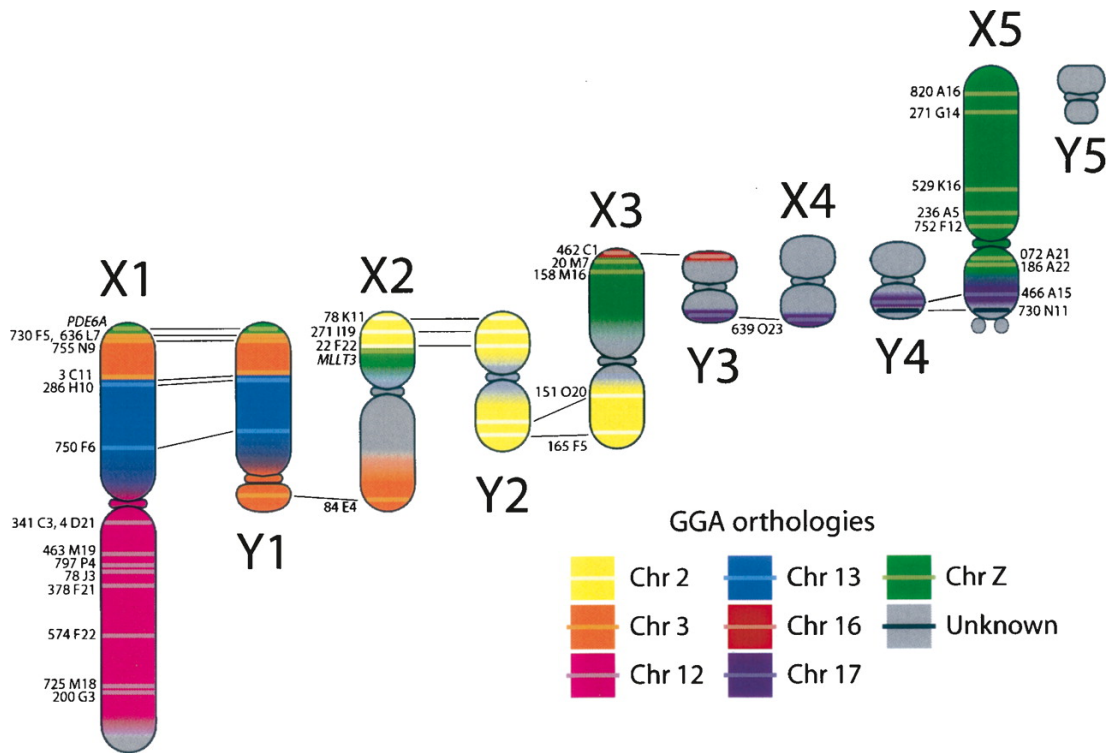


Figure 4.6: **Homology of the platypus sex chromosomes with the chicken genome.** This figure is taken directly from Veyrunes *et al.* (2008), and is based on the BAC mapping experiments they and others conducted. Homology with chicken (GGA) is represented by different colours, and pseudoautosomal regions are connected by lines. The small numbers indicate the BACs used.

the radioactive *in situ* hybridisation experiments with human cDNA probes in the earlier study, Giemsa-stained chromosomes 6 and X1 were misidentified, or that the hybridisation of human clones onto divergent platypus chromosomes was unreliable. Veyrunes *et al.* (2008) therefore concluded that the XY system in placentals and marsupials must have evolved after the prototherian-therian divergence but before the placental-marsupial divergence, so between 166 and 145MYA.

Since the platypus reference genome was made from a female, the Y chromosomes are missing from ornAna1, and thus, these have been less well studied than the X chromosomes. In a recent paper, Cortez *et al.* (2014) investigated the evolution of Y chromosomes in 15 mammals, including platypus, using high-throughput genome and transcriptome sequencing. For the species without Y chromosome sequences (that is, all except human, chimpanzee, macaque, dog and cat), they applied a subtraction approach that assembles transcripts using male-specific RNAseq reads not mapped to the female reference, and then confirms their Y identity by comparing the mapping

of male and female WGS data to the assembled transcripts. Through this method, Cortez *et al.* identified 25 platypus Y genes (18 of which were also found in echidna) and their corresponding gametologs on the X chromosomes. They grouped genes into strata by clustering them on the basis of d_s (synonymous site divergence) and phylogenetic analysis. From this they concluded that there were six potential Y chromosome strata in monotremes, five of which were formed before the platypus-echidna divergence. Levels of sequence identity between the X and Y copies of these genes are found in Table 4.1. One of the genes in the oldest stratum, S1, is *AMH* (anti-Mullerian hormone), known to play an important role in sex determination cascades across multiple vertebrates. In placentals, *AMH* acts two steps downstream of *SRY* to block the development of female reproductive organs. *AMH* therefore seems to be a good candidate for the sex-determining gene in monotremes. Consistent with this, Cortez *et al.* found that the Y copy is on Y5, although the X copy seems to have moved to X1.

4.1.5 Aims of the platypus population sequencing project

In order to learn more about different aspects of the biology of this remarkable species, we sequenced the genomes of 58 platypus samples from Queensland (QLD), New South Wales (NSW) and Tasmania (TAS) (Figure 4.7; Table 4.2). We sampled between one and eighteen individuals from each of seventeen waterways across most of the platypus range, excluding Victoria only because no high-quality DNA samples were available. We included the reference individual (Glennie, sample N725 in this chapter), a father-daughter pair obtained from Taronga Zoo (N735 and N736), who are originally from the Fish River in the Macquarie basin in NSW, and a duplicate sample (N703 and N749) from Dirran Creek, North QLD, which we intended to use for quality control. The aims of this project were severalfold. We wanted to investigate platypus population structure and history in greater detail than had been possible with the microsatellites and mtDNA data in earlier studies. Additionally, we wanted to use the population genetic data to study historical recombination rates and patterns of natural selection (not described in this chapter). We also aimed to learn more about the evolution of the sex chromosomes. All of these analyses rely upon variant calls made against the reference genome, and so we devote some of this chapter to describing reference genome improvements (Section 4.3.1) and quality control of the SNP calls (Section 4.3.2). We finish by presenting preliminary findings about genetic diversity and population structure (Section 4.3.3) and the sex chromosomes (Section

Table 4.1: Density of X-Y divergences in the Y chromosome genes identified by Cortez *et al.* (2014).

Stratum	Gene	Alignment length excluding gaps (bp)	Number of X-Y divergences	Mean distance between X-Y divergences (bp)	Percentage identity
S1	<i>AKAP8L</i>	579	115	5.0	80.1%
	<i>AMH</i>	1506	504	3.0	66.5%
	<i>FEM1C</i>	969	243	4.0	74.9%
	<i>MED26</i>	1555	505	3.1	67.5%
	<i>PPP4R2</i>	390	150	2.6	61.5%
	<i>THAP11</i>	608	152	4.0	75.0%
all S1 genes		5607	1669	3.4	70.2%
S2	<i>KHSRP</i>	1519	257	5.9	83.1%
	<i>RFX1</i>	1165	292	4.0	74.9%
	<i>STK11</i>	647	140	4.6	78.4%
	all S2 genes	3331	689	4.8	79.3%
S3	<i>ARF4</i>	534	63	8.5	88.2%
	<i>CTCF</i>	747	82	9.1	89.0%
	<i>MAU2</i>	1455	204	7.1	86.0%
	all S3 genes	2736	349	7.8	87.2%
S4	<i>DAZAP1</i>	279	47	5.9	83.2%
	<i>HNRNPA0</i>	66	7	9.4	89.4%
	<i>JUND</i>	338	35	9.7	89.6%
	<i>PPP1R10</i>	387	59	6.6	84.8%
all S4 genes		1070	148	7.2	86.2%
S5	<i>HNRNPK</i>	1174	100	11.7	91.5%
	<i>PRPF4B</i>	2304	201	11.5	91.3%
	<i>SSR1</i>	648	55	11.8	91.5%
	all S5 genes	4126	356	11.6	91.4%
S6	<i>EDF1</i>	357	9	39.7	97.5%
	<i>SDHA</i>	1904	124	15.4	93.5%
	all S6 genes	2261	133	17.0	94.1%

4.3.4).

Many people have contributed to this project. Jaime Gongora and Frank Grutzner provided the samples, some of which were originally collected by Tasman Daish, Steven Kolomyjec and Tom Grant. Tasman Daish and Portia Westall prepared the DNA samples. Rory Bowden, Simon Engledow, Paolo Piazza, Cath Green and Mariateresa di Cesare assisted with sample preparation and/or sequencing. Julie Hussin did the *de novo* assembly (Section 4.2.2.4). Liz Batty carried out the mapping, variant calling and coverage analysis jointly with me. I did the genome reassembly, quality control of variant calling, and downstream analysis of population structure and the sex chromosomes. Liz Batty, Julie Hussin, Jaime Gongora and Tom Grant assisted with the interpretation of results presented in this chapter, as did Peter Donnelly, who led the project.

Table 4.2: **Platypus samples sequenced.** The samples in blue font are in the unrelated set used for the population genetic analyses (see Section 4.2.5). The final column indicates how the samples were divided into five groups, and in parentheses, into four more general groups, on the basis of results presented in Section 4.3.3. Some individuals were excluded from these groups because there were insufficient samples from that cluster on the Principal Component Analysis shown in Figure 4.18. These are indicated by N/A. M: male, F: female; NQLD: North Queensland; NNSW: North New South Wales; CNSW: Central New South Wales; TAS: Tasmania.

Sample	Sex	Sampling Location	Group (larger group)
N701	M	Dirran Creek	NQLD (NQLD)
N702	M		
N703	F		
N704	F		
N748	M		
N749	F		
N750	F		
N711	M	Rifle Creek, Julatten	
N751	F	Barron River	
N745	F	Running River	N/A (N/A)
N746	F	Broken River	N/A (N/A)
N747	M		
N753	F	Carnarvon Creek	N/A (N/A)
N730	M	Gwydir River	N/A (NNSW)
N707	M	Barnard River	Barnard (NNSW)
N713	M		
N714	M		
N715	M		
N716	M		
N717	M		
N718	F		
N719	F		
N720	M		
N721	M		
N722	F		
N723	M		
N724	M		
N725	F		
N735	M	Fish River (Taronga Zoo)	Wingecarribee (CNSW)
N736	F		
N732	F	Wingecarribee River	
N733	M		
N744	F		
N759	F		

Continued on next page

Table 4.2 Continued

Sample	Sex	Sampling Location	Group (larger group)
N712	F	Shoalhaven River	Shoalhaven (CNSW)
N726	F		
N727	F		
N728	M		
N729	M		
N731	M		
N734	F		
N737	F		
N738	M		
N739	M		
N740	F		
N741	F		
N742	F		
N743	F		
N752	M		
N757	M		
N760	F		
N761	M		
N705	M	Launceston	TAS (TAS)
N710	M	Mangalore	
N754	M	South Esk River	
N755	M	Brumby's Creek	
N756	F	Brumby's Lake	
N758	M	Upper Derwent River	

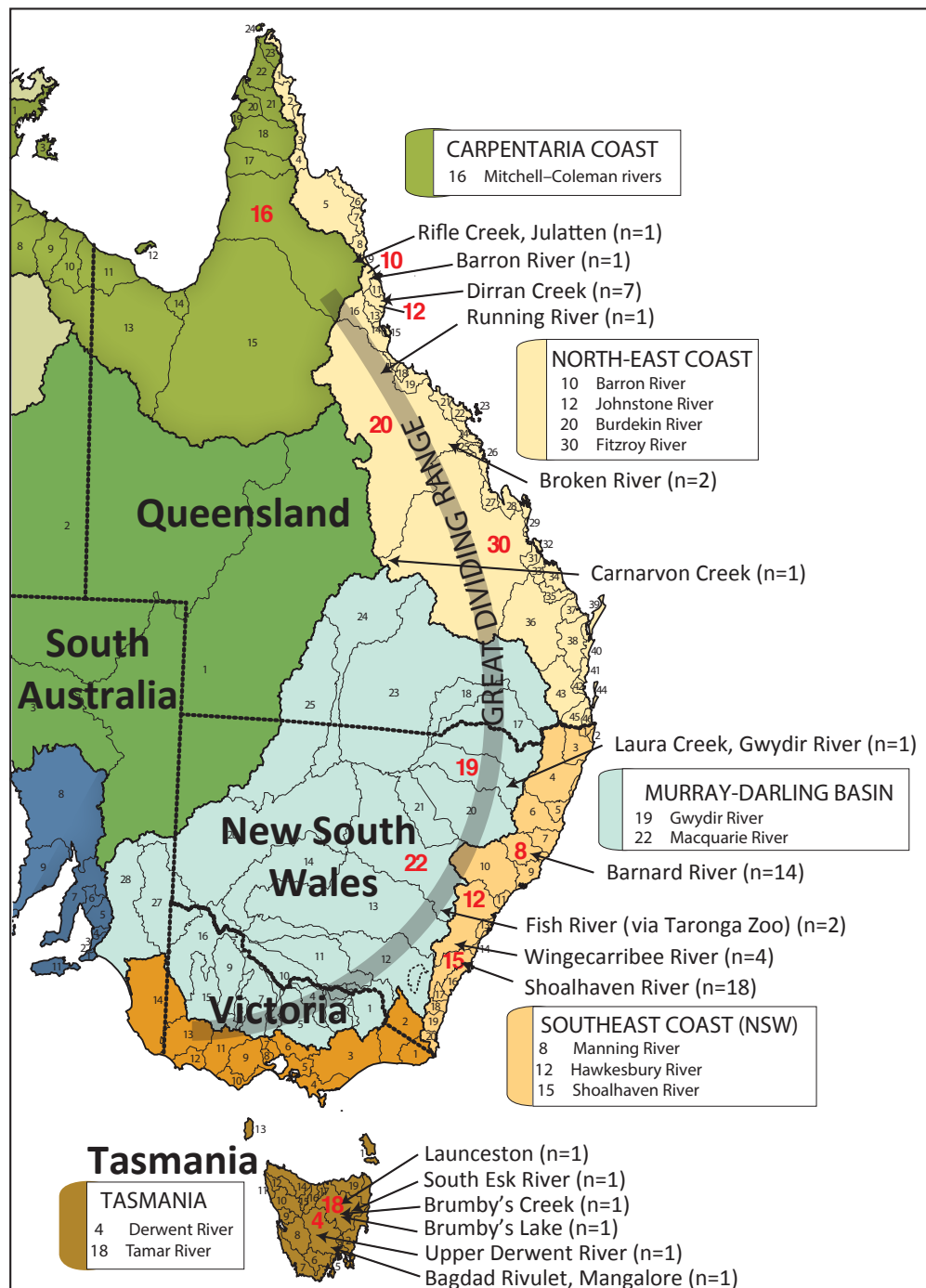


Figure 4.7: Map showing our sampling locations within river regions in eastern and southeastern Australia. The waterways our samples come from are indicated by arrows, with sample sizes in brackets after the river name. The specific catchments these rivers fall into are labelled with large red numbers and are named in the boxes corresponding to the drainage division. The transparent grey line represents the Great Dividing Range. This map is adapted from one obtained from the Australian Bureau of Meteorology (www.bom.gov.au/water/geofabric/documents/BOM002_Map_Poster_A3_Web.pdf).

4.2 Methods

4.2.1 Samples and sequencing

Collaborators in Australia provided us with DNA samples from 61 platypus, extracted from toe webbing, spleen, liver or cell lines. These were sequenced in four tranches. Details of the library preparation and paired-end (PE) sequencing are shown in Table 4.3. Three samples were discarded because they were found to have very uneven coverage. For the remaining 58 samples, mean coverage ranged from 13.1X to 21.6X, with a median of about 15X. Mean coverage and insert size per sample are shown in Figure 4.8 and Table 4.A.2.

4.2.2 Improving the reference genome

As described below in Section 4.3.1, we encountered various problems with the platypus reference genome, *ornAna1*, so decided to improve it using a variety of technologies.

4.2.2.1 Mate pair data

The Genomics Core at the WTCCHG produced four mate pair libraries from a single male sample, N721. These were constructed using the Nextera mate pair protocol from Illumina, with custom indexes, and were sequenced on the HiSeq 2500 Rapid (150bp, PE). Three libraries, with mean insert sizes 5.9kb, 8kb and 10.5kb, were made using the gel-plus protocol. The fourth library was constructed using a gel-free protocol, which requires less input DNA but gives a broader distribution of insert sizes. This library had a mean insert size of 2.5kb with a standard deviation of 0.9kb.

NextClip (Leggett *et al.*, 2014) was run to remove the junction adaptors from reads⁶, and to identify and remove reads that did not contain an adaptor, since they may have been paired-end sequences that had slipped through the biotin enrichment step in library preparation. Reads which were shorter than 50bp after adaptor removal

⁶In the Nextera protocol, a biotinylated adaptor is attached to both ends of each DNA fragment. The fragments are then circularised, joining the two junction adaptors. Next, the circularised DNA is fragmented and biotin enrichment is used to pull down the fragments containing an adaptor that marks the junction. These fragments are sequenced from both ends, reading inwards towards and sometimes through the junction adaptors. The adaptor sequences must be removed before aligning the reads to the reference genome.

Table 4.3: Library preparation and sequencing.

Tranche	Date	Samples	Library preparation	Indexes	Sequencer	Read length	Location
1	August 2012	N701-N709 ^a	Nextera DNA library preparation kit	Nextera index kit	HiSeq 2000	100bp	WTCHG
2	June 2013	N710-N712	standard Illumina protocol, specifics unknown		HiSeq 2000	100bp	Sydney
3	June 2014	N713-N725	NEBNext; no amplification ^b ; 100ng starting material; 1 min fragmentation with Covaris; bead size selected	TruSeq adaptors (barcoded)	HiSeq 2500	150bp	WTCHG
4	February 2015	N726-N761	Wafergen Apollo 324 NGS Library Prep system; 100ng starting DNA; 1 min fragmentation with Covaris	TruSeq adaptors (barcoded)	HiSeq 2500	150bp	WTCHG

^aSamples N706, N708 and N709 were found to have very uneven coverage (data not shown), thought to be due to low DNA quality, so these were discarded from all analyses.

^bThe libraries for samples N713-N725 were originally prepared using the NEB Ultra DNA Library Prep Kit. However, after the first round of sequencing, some samples were found to have extremely unequal coverage that seemed to be due to differential GC bias during the PCR. Thus, new libraries were produced without the amplification step.

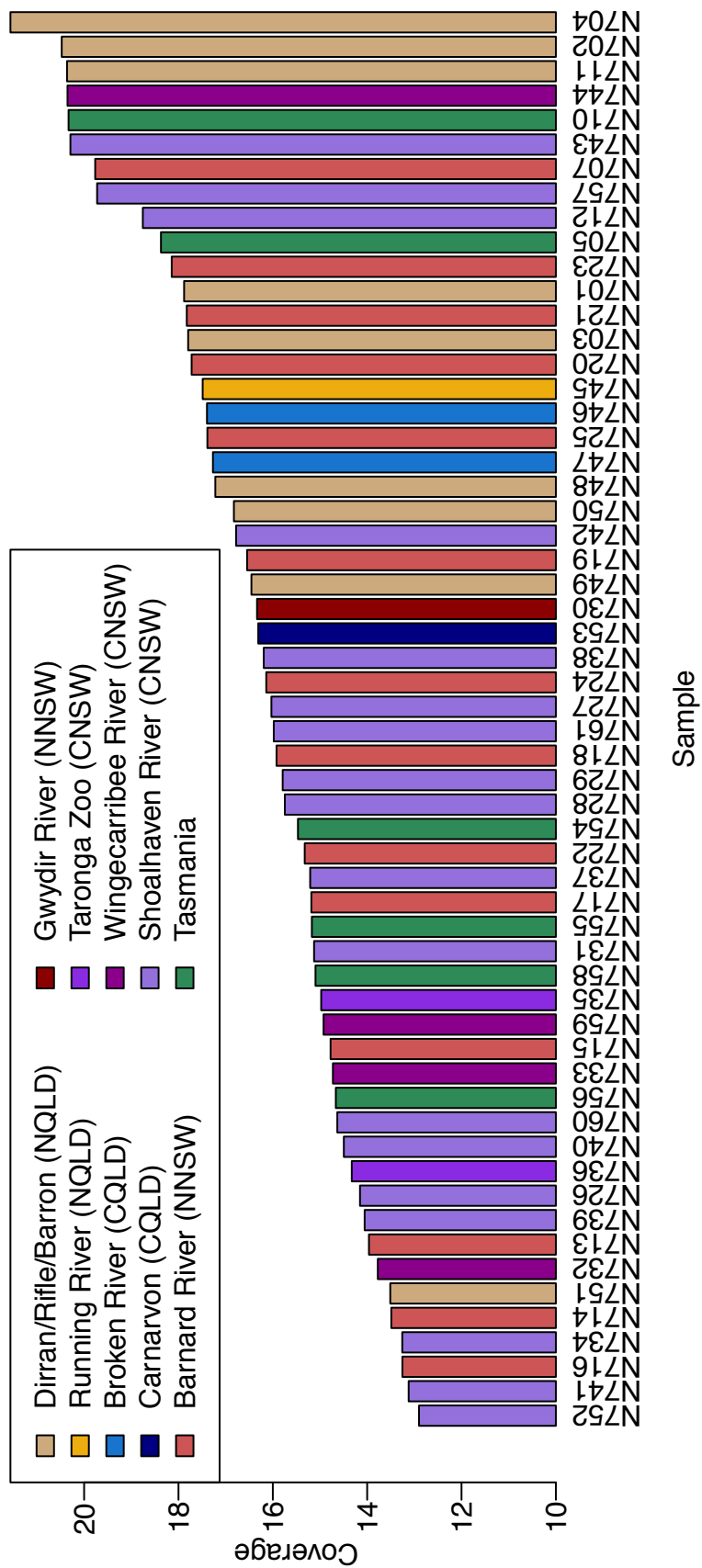


Figure 4.8: Average genome-wide coverage per sample. This is calculated using reads with $\text{MAPQ} > 0$.

were discarded, since these were more likely to map spuriously downstream. The gel-plus libraries were found to have a very high proportion of duplicate read pairs (Table 4.A.3), and so only the gel-free library was used in the scaffolding described below.

4.2.2.2 BACs and fosmids

Wes Warren (Washington University, St Louis) provided us with the end sequences of BACs and fosmids from the original platypus genome assembly. The reads had a mean length of 788bp with a standard deviation (s.d.) of 138bp. The mean insert size was 28.68kb (s.d. 10.4kb) for the fosmids and 126.8kb (s.d. 54.9kb) for the BACs (see Figure 4.A.1).

4.2.2.3 Moleculo data

We also obtained long reads from sample N721, produced using Illumina’s Moleculo technology. In this protocol, genomic DNA is fragmented to about 10kb, and the fragments are amplified, sheared, marked with a unique barcode, and sequenced. The short reads originating from each molecule are then assembled into synthetic long reads. We obtained $\sim 1X$ coverage in reads over 7kb, and about $\sim 4X$ coverage in reads over 0.5kb (Figure 4.A.2).

4.2.2.4 *De novo* assembly

We assembled the Illumina paired-end data *de novo* using ABySS (Simpson *et al.*, 2009). Three different sets of samples were used as input: 1) three males with the highest coverage available at the time (F3M: N701, N702 and N705), 2) three females with the highest coverage (F3F: N703, N704, N712) and 3) the male used for the mate pair and Moleculo sequencing (N721). The ABySS parameters used were as follows: $n=3$, $m=20$, $s=100$, and $k=64, 72$ or 80 . The *de novo* assemblies produced were considerably more fragmented than the reference genome, but the contigs from six different assemblies (F3M with $k=64, 72$ or 80 , F3F with $k=64$ or 72 , and N721 with $k=72$) were used for gap filling, as described in the next section.

4.2.2.5 Rescaffolding and gap filling

Our procedure for rescaffolding and gap filling is illustrated in Figure 4.9. We first broke the platypus reference genome, *ornAna1*, into its original contigs and used Stampy (Lunter and Goodson, 2011) to map the mate pairs and the paired-end reads

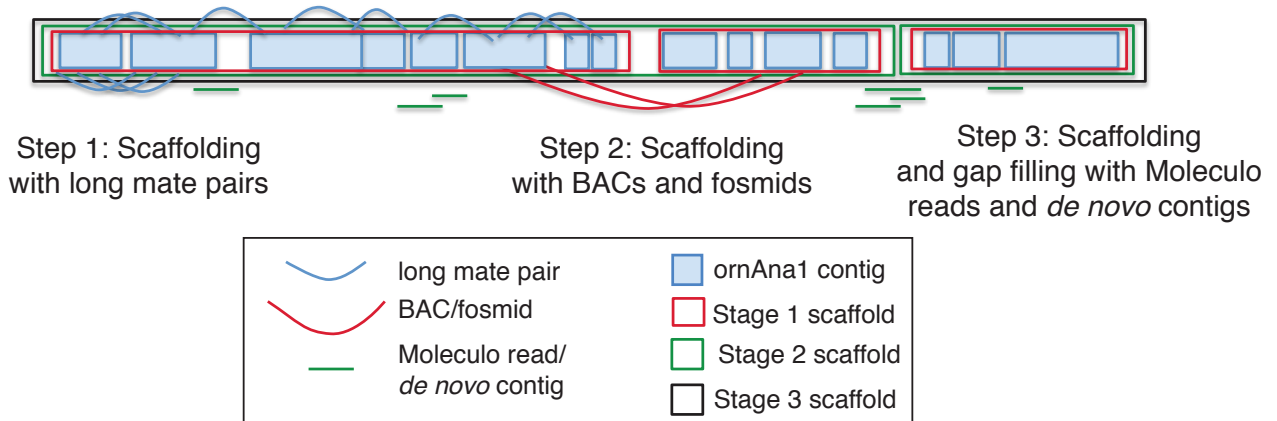


Figure 4.9: **Rescaffolding and gap filling procedure to improve the reference genome.**

from the first 22 samples to these⁷. To minimise downstream errors resulting from incorrectly mapped reads, we retained only read pairs that mapped with a mapping quality (MAPQ) greater than 30, and from which at least 50bp of each 100bp or 150bp read mapped outside an annotated repeat. (We lifted over the coordinates of the repeats identified by RepeatMasker in the ornAna1 scaffolds to their constituent contigs.)

Next we used SSPACE v3.0 (Boetzer *et al.*, 2011) to join the contigs into new scaffolds using the information from the paired-end and mate pair libraries (Figure 4.9). This algorithm works by first identifying putative pairs of neighbouring contigs such that the distance between them satisfies the range set by the user. We set the range to be between the 1st and 99th percentile of the distribution of insert sizes for each library separately. SSPACE then builds scaffolds iteratively by combining contigs if a minimum number of read pairs, k , supports the connection. We used $k = 5$. The scaffolds produced by this process are henceforth referred to as Stage 1 scaffolds.

We then mapped the BACs and fosmids to these Stage 1 scaffolds using bwa mem (Li, 2013) and retained the best alignment for each read. This left us with 0.06X and 0.1X coverage by BACs and fosmids respectively (3.4X and 2.3X if we consider the average number of read pairs spanning each position in the genome). We then ran SSPACE a second time to produce Stage 2 scaffolds, permitting insert sizes of

⁷Only 22 samples were available at the time this rescaffolding was done.

1-50kb for the fosmids and over 50kb for the BACs. We set $k = 2$, a lower threshold than was used with the mate pairs and paired-end reads, since the coverage was lower.

Our final step was to incorporate the Molecuro reads for gap filling. We decided to supplement these with the *de novo*-assembled contigs of length greater than 500bp, treating these as long reads. PBJelly (English *et al.*, 2012) was used to fill gaps within the Stage 2 scaffolds and do further scaffolding to produce Stage 3 scaffolds. This algorithm first aligns long reads to the input scaffolds using BLASR (Chaisson and Tesler, 2012). It then identifies reads that map across or at the edge of gaps, and uses these to close the gap or extend its flanks. We specified that all gaps should be considered (using `-minGap==1`). SSPACE inserts a gap of length 1bp if the mean insert size of the read pairs indicate that two contigs should be merged but no overlap is found. To avoid PBJelly classing these as “overfilled” because the sequences that extend the flanks do not have an identifiable overlap but have a total length greater than the predicted gap size, we arbitrarily set the lengths of such gaps to be 200bp. The BLASR parameters used for mapping the reads were as follows: `-minMatch 8 -sdpTupleSize 8 -minPctIdentity 95 -bestn 1 -nCandidates 10 -maxScore -500 -nproc 5 -noSplitSubreads -maxMatch 30 -maxAnchorsPerPosition 50`.

To assign the Stage 3 scaffolds to chromosomes as far as possible, we took the end sequences of BACs that had been placed on chromosomes using FISH in the original genome project, and mapped these to the new scaffolds using `bwa mem` (Li, 2013).

4.2.3 Mapping

The Stage 3 scaffolds produced by PBJelly were used as a new reference genome that we will call `ornAna1.2`, on which all other analyses described in this chapter were carried out. We mapped the paired-end data from all 58 samples to this reference using Stampy (Lunter and Goodson, 2011). Duplicates were removed using Picard (<http://picard.sourceforge.net>).

4.2.4 Variant calling

Variant calling was performed on the scaffolds longer than 50kb (67.7% of the genome) using three different callers, each one applied to all 58 samples jointly.

We used the GATK UnifiedGenotyper (DePristo *et al.*, 2011) to call SNPs and indels separately, and then used the VariantAnnotator to annotate SNPs that failed a set of filters based on coverage, read mapping quality, proximity to indels (≤ 20 bp from an indel), proximity to other SNPs (≥ 3 SNPs within ≤ 10 bp), and other metrics (shown in Table 4.A.4). Where appropriate, these filters were adjusted to deal with X-linked scaffolds (see Section 4.2.7). SNPs that failed any of these filters were removed. The GATK SNP sets before and after filtering are described henceforth as “GATK raw” and “GATK PASS”. For some purposes, we used a further filtered set in which the genotypes in the GATK PASS set were set to missing if the genotype quality (GQ) was less than 20 (“GATK PASS with GQ filter”). The selection of this threshold is described in Section 4.3.2.

To produce a more stringent set of SNPs, we used CORTEX, a graph-based approach that avoids some problems inherent in reference genome-based methods, such as spurious heterozygous SNPs due to underrepresented repeats (Iqbal *et al.*, 2012). CORTEX requires a considerable amount of memory to run; we used a server with 2Tb of RAM. The final step in the CORTEX pipeline involves mapping the sequence immediately 5’ of each variant to the reference genome to identify its position and determine the reference and alternate alleles. In doing so, it flags variants with low mapping quality or with more than two alleles, variants that overlap another variant, and variants at which none of the alleles match the reference. We removed SNPs that failed any of these filters, or that were identified as being sequencing errors or being due to repeats. The remaining SNPs are referred to henceforth as the “CORTEX PASS” set.

The analysis in Section 4.3.2 suggested that, at SNPs called by both GATK and CORTEX, the GATK genotypes were more accurate. We thus produced a stringent callset by using the SelectVariants module from GATK to extract the GATK PASS SNPs that were also called by CORTEX, taking the GATK genotypes at these sites, and setting those with GQ less than 20 to missing. We removed any SNPs at which the reference individual, N725, was called as 1/1 (homozygous for the alternate allele).

For comparison, we also called variants with Platypus v0.8.1 (Rimmer *et al.*, 2014). We retained only SNPs that passed all of the in-built filters, which are similar to those

used for GATK (Table 4.A.4), to make a “Platypus PASS” callset. We took the intersection between the Platypus PASS SNPs and the CORTEX PASS SNPs, retained the Platypus genotypes, and set those with GQ less than 20 to missing.

4.2.5 Identifying relatives

To identify relatives among the sampled individuals, we ran KING (Manichaikul *et al.*, 2010), which uses the number of SNPs that are identical-by-state 0, 1 or 2 to estimate the kinship coefficient. For each population (North Queensland, Barnard, Shoalhaven, Wingecarribee+Fish, and Tasmania) we removed SNPs from the GATK/CORTEX intersection set with minor allele frequency (MAF) < 0.05 across that population, and then ran KING-homo, which assumes a homogenous population. For several of the population genetic analyses described below (F_{ST} and nucleotide diversity calculations; STRUCTURE), we removed one individual from each relative pair (up to and including third-degree relatives), leaving 42 individuals (shown in blue in Table 4.2). Specifically, we removed N713, N719, N720 and N721 from the Barnard River; N703 and its duplicate, N749, from Dirran Creek; N727, N734, N741, N742, N757 and N760 from the Shoalhaven River; N732 and N744 from the Wingecarribee River; N736 from the Fish River (Taronga Zoo); and N756 from Brumby’s Lake in Tasmania.

4.2.6 Population genetic analyses

All these analyses were restricted to scaffolds longer than 50kb which were unambiguously classed as autosomal or pseudoautosomal (see Section 4.2.7 below), and had no putative Y-homologous regions (Section 4.2.8). Unless otherwise specified, we used the GATK/CORTEX intersection callset with the GQ filter.

We used EIGENSOFT v6.0.1 (Patterson *et al.*, 2006) to estimate F_{ST} between populations⁸ and to run a Principal Components Analysis (PCA). We ran STRUCTURE (Pritchard *et al.*, 2000) with the admixture model for $k = 1, 2, 3, \dots, 10$. For these analyses, we took the GATK/CORTEX intersection callset with the GQ filter and removed SNPs with MAF < 0.05 or over 10% missingness across all 42 unrelated samples, leaving us with 1,045,508 SNPs. We also created an LD-pruned SNP set by using PLINK (Purcell *et al.*, 2007) to thin the SNPs on each scaffold. Specifically, for windows of 50 SNPs, we removed one of each pair of SNPs if the r^2 between them

⁸This uses Hudson’s estimator, following the recommendations of Bhatia *et al.* (2013).

was greater than 0.1, and then shifted the window 5 SNPs forward (PLINK option: `--indep-pairwise 50 5 0.1`). This left us with 58,784 SNPs.

We ran PSMC to investigate historical population sizes (Li and Durbin, 2011). For this we used the GATK PASS callset, since we could determine how much of the sequence was truly “callable” by applying the same filters (i.e. those based on coverage, mapping quality etc. described in Section 4.2.4 and Table 4.A.4) to all positions along scaffolds. PSMC operates on 100bp bins which are classed either as containing at least one heterozygous SNP, containing no heterozygous SNPs, or having too much missing data. We tried using different cutoffs for missingness ($> 10\%$, 50% or 90% of bases per bin callable) and also restricting the analysis to scaffolds of different sizes ($> 50\text{kb}$, 500kb , 1Mb , or 2Mb) (Figure 4.A.21). For scaling the output, we used for the mutation rate $\mu = 1.2 \times 10^{-8}$ /bp/generation, taken from estimates in humans (Scally and Durbin, 2012), and, for the generation time, $g = 10$ years, taken from Furlan *et al.* (2012).

4.2.7 Identifying X-linked scaffolds

To identify scaffolds that were X-linked, we compared the coverage between males and females. First, for each sample, we calculated the mean coverage across each contig (excluding reads with $\text{MAPQ} = 0$) within the 6159 scaffolds longer than 50kb , and then divided this by the mean genome-wide coverage for that sample to give “scaled coverage”. We then took the average scaled coverage across males and across females and calculated their ratio for each contig. The distribution of this ratio was clearly bimodal (Figure 4.10). We classed contigs as X-linked if they had a male-to-female scaled coverage ratio less than 0.7 and if the p-value from a t-test of the difference in scaled coverage between males and females was less than 0.01. If the ratio was greater than 0.7, we classed the contig as (pseudo)autosomal, since we could not distinguish PARs and autosomes based on coverage. Then, for each scaffold, if all of the constituent contigs had been classed as X-linked, we classed the scaffold as X-linked, and similarly for (pseudo)autosomal scaffolds.

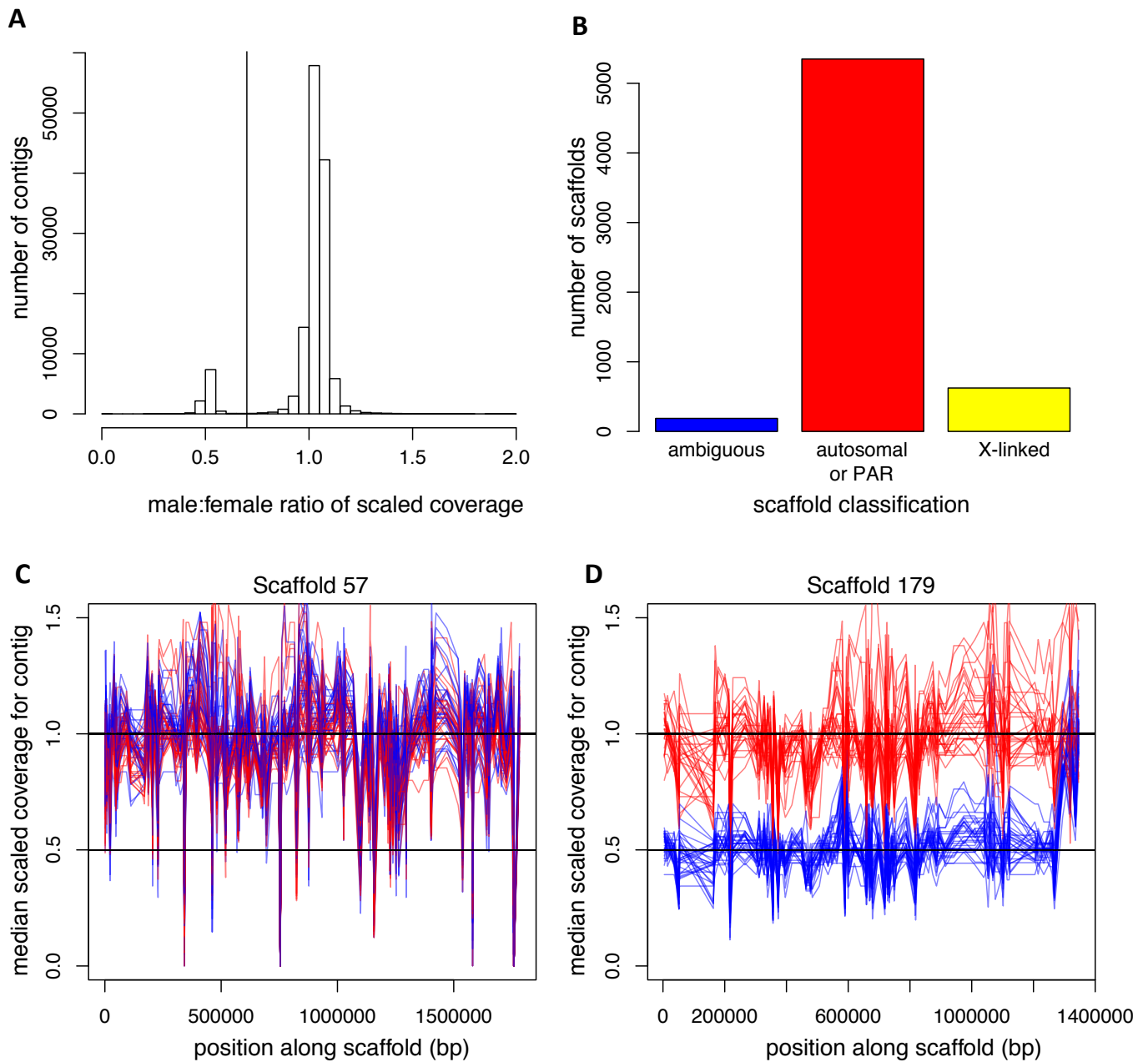


Figure 4.10: Assigning scaffolds as autosomal or X-linked. A) The distribution of the ratio of scaled coverage per contig in males to females was clearly bimodal. Contigs for which this ratio was less than 0.7 and that had a p-value from a t-test of the difference in scaled coverage between males and females less than 0.01 were assigned as X-linked. B) Scaffolds were then assigned as X-linked, autosomal (or PAR), or ambiguous on the basis of their constituent contigs. C) and D) show the median scaled coverage per contig across two scaffolds, with red lines representing females and blue lines representing males. Scaffold 57 was classified as autosomal (or PAR). Scaffold 179 was X-linked for most of the scaffold but transitioned into a region with equal coverage in males and females in the last ~100kb, suggesting it contained a PAR boundary (see Section 4.3.4).

4.2.8 Identifying regions homologous to the Y chromosomes

While examining putatively X-linked scaffolds, we detected regions in which all males had dense heterozygous SNPs at which all females were homozygous reference. We inferred that these were due to reads from the Y chromosomes, which are missing from the reference, mapping to the homologous region on an X chromosome. To detect these regions genome-wide, we applied the following heuristic algorithm to the GATK raw callset⁹:

1. Specify the set of samples required to have heterozygous genotypes (Set 1 - see definition of sets below) and the set of samples required to have homozygous reference genotypes (Set 2).
2. For each scaffold longer than 50kb:
 - Identify any SNPs that are heterozygous in all samples in Set 1 and homozygous reference in all samples in Set 2 (“X-Y divergences”).
 - Move along the scaffold in windows of 50 SNPs, incrementing by one SNP at a time; if $> 25\%$ of SNPs in the window are putative X-Y divergences, flag that window.
 - Merge any overlapping windows, and refine their boundaries such that the outermost SNPs are X-Y divergences, to produce putative “Y homologous regions” (YHRs).

We initially ran this algorithm to detect SNPs that were heterozygous in all males and homozygous reference in all females. Then, to allow for the possibility of polymorphism between or within populations, we ran it using the following sets of samples as Set 1 *versus* all other samples as Set 2: all except n males, where n ranges from 1 to 10; Barnard males; Barnard + Gwydir males; Shoalhaven males; Shoalhaven + Wingecarribee + Fish males; North Queensland males; Queensland males; mainland males; Queensland + North New South Wales males; New South Wales males; Shoalhaven + Wingecarribee + Fish + Tasmania males. We did not use the Tasmanian males alone because it seemed likely that this would lead to spurious classification of

⁹We used this less stringent callset since many of these apparent heterozygous SNPs were removed by our filtering procedures due to being high density or due to the reads’ low mapping quality.

regions as YHRs simply because there was a particular haplotype segregating in Tasmania for which the males (five of the six samples) all happened to be heterozygous.

Having identified possible YHRs using this algorithm, we looked at the scaffolds containing them in the Integrated Genomics Viewer (IGV) (Thorvaldsdottir *et al.*, 2012), and manually refined the boundaries of PAR/autosomal regions, X-specific regions, and YHRs.

4.2.9 Gene identification and synteny

We took the Ensembl gene annotations from ornAna1 and used BLAST to align the genes' exons to the new ornAna1.2 scaffolds. We determined the best BLAST hit for each exon, and then identified the genes from which all or nearly all exons were aligned on the same scaffold in the correct order. To try to place scaffolds of interest on chromosomes using synteny, we took the genes on those scaffolds and extracted their predicted orthologs in chicken and human using Ensembl.

4.3 Results

4.3.1 Improving the platypus reference genome

In the early stages of this project, we detected several types of likely errors in the platypus reference genome, ornAna1 (Warren *et al.*, 2008). Figure 4.11 shows one of these: a change in the male-to-female coverage ratio along a scaffold assigned to chromosome 14, suggesting a misassembly or incorrect chromosome assignment in the FISH data from the original genome project. Another indication of errors was that, when we started investigating fine-scale recombination rates¹⁰, we observed many peaks in rate which, when scrutinised closely, were found to be caused by scaffolding errors that produced an apparent breakdown in linkage disequilibrium over short distances (Figure 4.12). The fact that the genome assembly was so fragmented also impeded many analyses. Thus, we decided to try to improve it.

We suspected that the majority of errors occurred during the scaffolding stage, rather than in making the contigs, and our observations in the early recombination

¹⁰Analyses of fine-scale recombination are ongoing and will not be presented in this thesis.

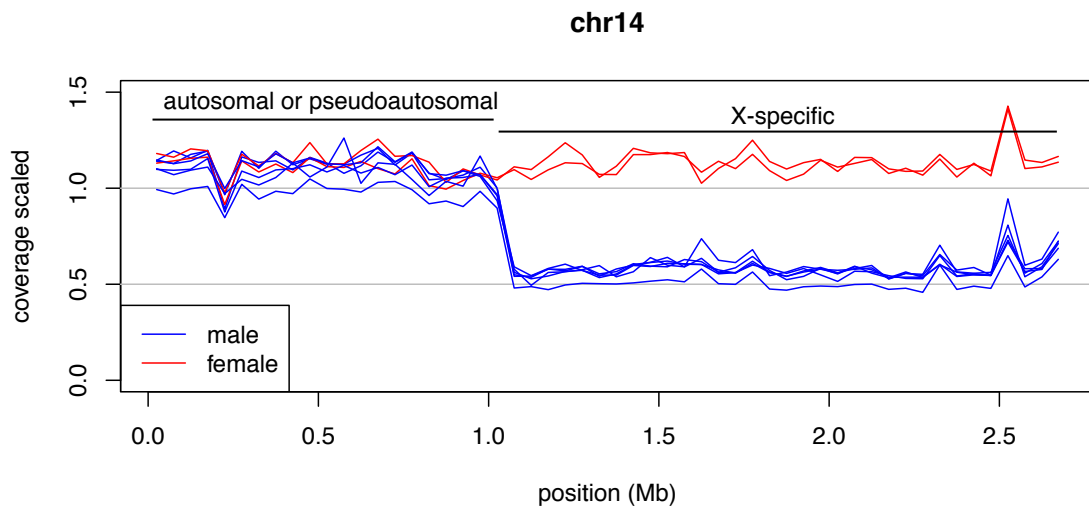


Figure 4.11: **Coverage for the first nine samples along a scaffold assigned to chromosome 14 in ornAna1.** The drop in male relative to female coverage at around 1Mb suggests either a misassembly or misassignment. The coverage was scaled as described in Section 4.2.7. Note that it appears slightly higher than 1 in pseudoautosomal/autosomal regions and slightly higher than 0.5 for males in X-linked regions because the scaling is done using a genome-wide average that included positions that have Ns (i.e. gaps), so this average is slightly underestimated.

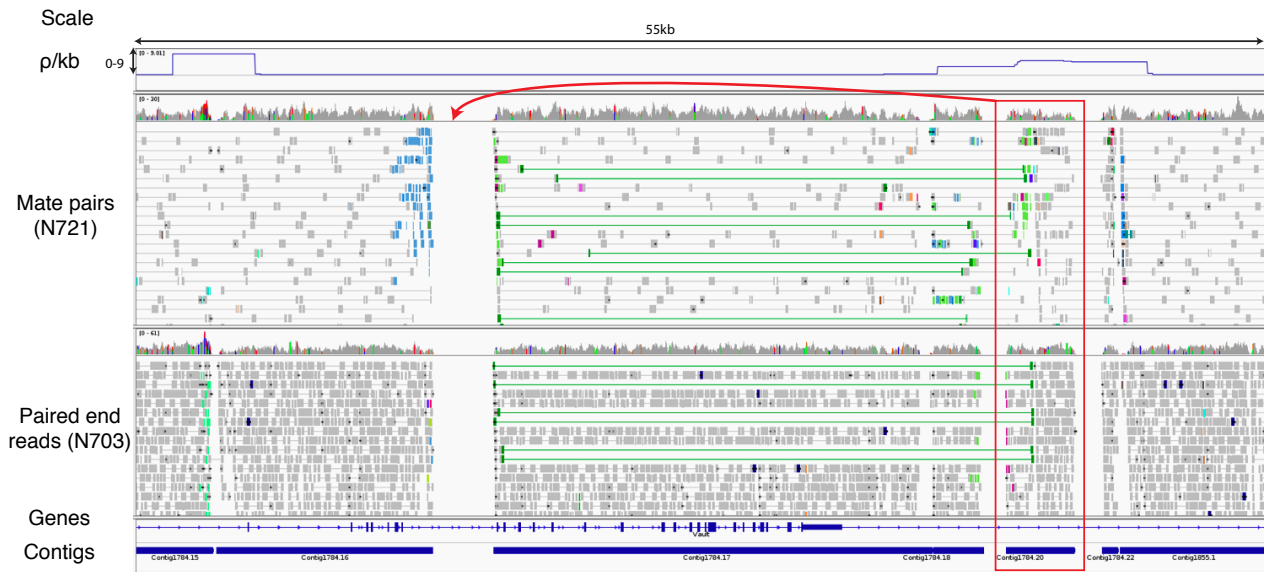


Figure 4.12: **Example of a spurious recombination hotspot caused by a mis-assembly in ornAna1.** The line in the top track indicates the recombination rate in ρ /kb estimated with LDHAT (McVean *et al.*, 2004). The next two tracks show sequencing reads from the mate pair library and a paired-end library, represented as grey bars if the insert size is within the expected range, green bars (connected with long green lines) if the read mapping indicates an intra-chromosomal translocation, or bars in colours other than green if one of the pair maps to a different scaffold. The green read pairs indicate that one contig (in the red box) has been inserted in the wrong place, instead of in the position indicated by the red arrow. This has caused an apparent breakdown in linkage disequilibrium, leading to increases in the recombination rate. The pileup along the top of the read tracks indicates coverage, with coloured lines indicating differences from the reference sequence.

analyses supported this. Therefore, we broke the original scaffolds into their constituent contigs (i.e. at the gaps), rescaffolded these using paired-end reads, mate pairs, BACs and fosmids, and then filled gaps and did further scaffolding with long Molecuro reads and *de novo* assembled contigs (method described in Section 4.2.2.5 and illustrated in Figure 4.9). We tried to be stringent while doing the rescaffolding by ignoring paired-end and mate pair reads that mapped primarily inside repeats, since these could lead to our making incorrect joins between contigs¹¹. We refer to the resultant assembly as ornAna1.2.

After this rescaffolding, the distribution of scaffold lengths was similar to that in the original assembly, but with fewer scaffolds over 1Mb (Figure 4.13; Figure 4.A.3). Eighty of the scaffolds, totalling 47.7Mb, could be assigned to chromosomes using the FISH data. We managed to fill a substantial fraction of intra-scaffold gaps using Molecuro reads and *de novo* contigs (Table 4.4). The filled gaps were substantially enriched for repeat elements, particularly LINEs, compared to the rest of the genome. Whereas in ornAna1, 20% of the sequence was composed of LINEs, this proportion was 31% for sequence in filled intra-scaffold gaps and 51% for sequence in filled inter-scaffold gaps (Table 4.A.5). This is consistent with the finding in the original genome paper that contigs tended to end in LINEs and SINEs, presumably due to their higher GC content and the difficulty in assembling over repeats.

Table 4.4: Number of gaps filled using long Molecuro reads and *de novo* contigs. In the last step of our rescaffolding pipeline described in Section 4.2.2.4, PBJelly (English *et al.*, 2012) was run on the Stage 2 scaffolds to try to fill gaps and join together scaffolds. As a result of this, gaps were either filled, had one or both flanks extended, or were “overfilled”, meaning that the sequences extending the flanks did not have an identifiable overlap but had a total length greater than the predicted gap size. Some scaffolds were extended but not joined to other scaffolds (rightmost column).

result	intra-scaffold gap (% of all such gaps)	inter-scaffold gap	unjoined scaffold
gap filled	36,728 (18.5%)	26,232	NA
one flank extended	20,704 (10.4%)	629	49,253
both flanks extended	11,766 (5.9%)	849	NA
gap overfilled	11,312 (5.7%)	NA	NA

¹¹We detected some spurious joins when using less stringent criteria for scaffolding.

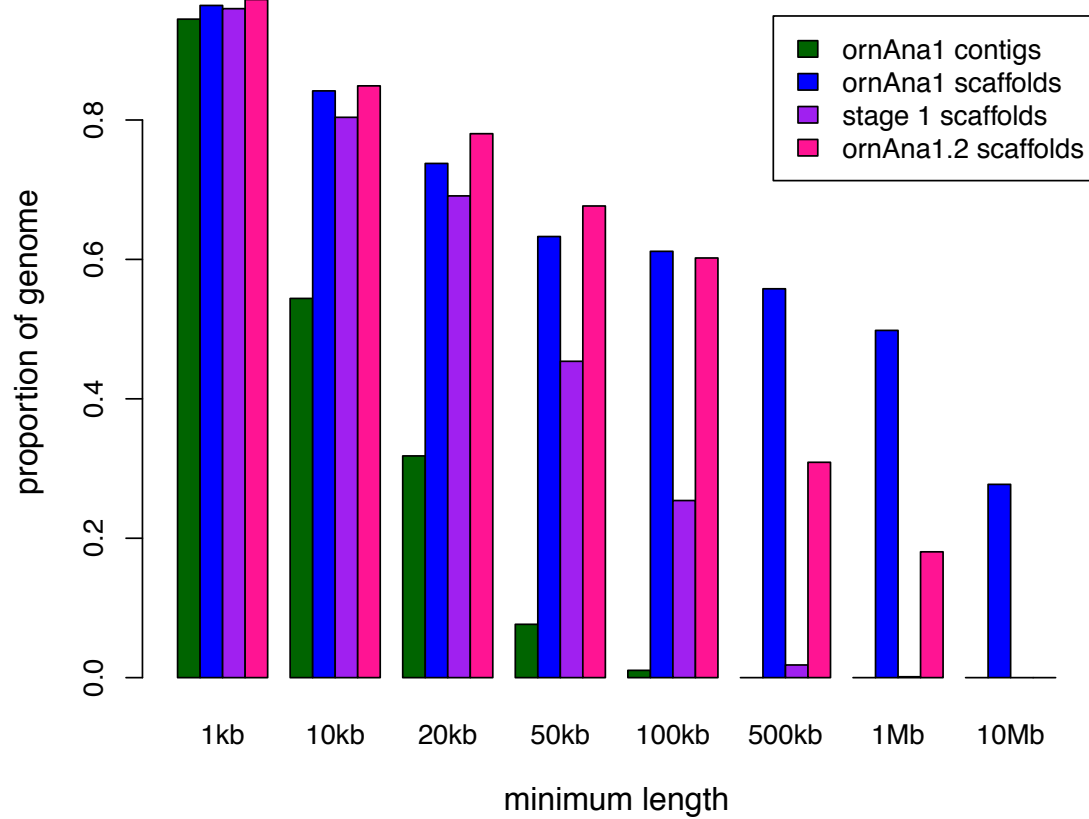


Figure 4.13: **Proportion of genome in contigs or scaffolds over different lengths.** See Section 4.2.2.4 for a description of the rescaffolding. Stage 2 scaffolds are not shown because their length distribution was almost identical to that of the final ornAna1.2 (Stage 3) scaffolds.

Although we failed to produce more long scaffolds through our reassembly process, we did remove some of the scaffolding errors present in ornAna1. We were able to assess this using inheritance patterns in a family quartet (two parents + two offspring) which we found amongst the samples from the Shoalhaven River¹². Specifically, we counted changes in the pattern of inheritance in the two offspring (“switches”), which represent either crossovers or switch errors. The density of these depends on the number of informative SNPs (i.e. SNPs at which one parent is heterozygous and the other homozygous). The mean number of switches per informative SNP was 25% lower in ornAna1.2 than in ornAna1, indicating that the reassembly had eliminated some of the errors (Table 4.A.6). However, the error rate in ornAna1.2 is still quite high, and so we are trying to make further improvements to the genome using PacBio long read data and also a HiC-based method from Dovetail Genomics (Putnam *et al.*, 2015). At the time of writing, these data are not yet available. Thus, the results in this thesis will be based on the ornAna1.2 assembly.

4.3.2 Variant calling in population samples

We ran three different variant callers on our 58 samples: GATK (DePristo *et al.*, 2011), Platypus (Rimmer *et al.*, 2014) and CORTEX (Iqbal *et al.*, 2012), as described in Section 4.2.4. Analysis with KING indicated that we had sequenced a duplicate sample (N703 = N749), as well as a whole family quartet (Figure 4.A.4; Table 4.A.7). We could therefore use these samples, along with the father-daughter pair from the Fish River (via Taronga Zoo), for quality control (QC) on our different callsets. For the analyses in this chapter, we used only SNPs and discarded indel calls, since these were more likely to be erroneous and were not necessary for the analyses presented here.

With the CORTEX PASS calls alone, the rate of discordant genotypes between N703 and its duplicate sample N749 was unacceptably high (1.8%) (Table 4.5), as was the Mendelian error rate (1.5% in the quartet) (Table 4.6). The Mendelian error rate was also high for the GATK PASS calls (1%), but better for the Platypus PASS calls (0.3%). We found that taking the SNPs called by both GATK and CORTEX

¹²Although we did not intend to collect relatives, perhaps, in hindsight, it is not surprising that we picked some up, given the dense sampling in the Shoalhaven River. This is discussed further in Section 4.4.

and retaining the GATK genotypes at those SNPs greatly improved both the genotype concordance between the duplicate samples and the Mendelian error rate. The same was true when we took the Platypus/CORTEX intersection. When we set any genotypes with GQ less than 20 to missing, this further improved the QC metrics in the GATK/CORTEX and Platypus/CORTEX intersection. Applying a GQ threshold to the CORTEX genotypes was not satisfactory, since GQ was found to be highly correlated with genotype (Figure 4.A.5), so setting a GQ threshold of, for example, 10, would lead to the exclusion of half of the homozygous calls, but a much smaller proportion of heterozygous calls. In contrast, with GATK and Platypus, GQ was not so strongly associated with genotype (Figures 4.A.6 and 4.A.7), and setting a GQ threshold of 20 eliminated about 80% of the discordant genotypes between the duplicate samples in the GATK/CORTEX and Platypus/GATK intersections (Table 4.5), without removing a large number of genotypes that were likely correct.

The discordant calls are dominated by SNPs at which N703 is called as heterozygous and N749 is called as homozygous for either the reference or alternate allele, or at which N749 is called as heterozygous and N703 is called as homozygous for either the reference or alternate allele (IBS1 calls in Table 4.5). We can get a sense of the cause of errors in each callset by looking at the coverage and the fraction of reads carrying the alternate allele at SNPs with concordant *versus* discordant genotypes in the duplicate samples. Examining the discordant calls, we see that in the CORTEX calls, the sample called as heterozygous tends to have higher coverage than the sample called as homozygous (Figure 4.A.8), and the fraction of reads carrying the alternate allele is centred around 0.5, as expected (Figure 4.A.9). This suggests that, at those SNPs, the heterozygous genotype is the correct one, and CORTEX should have set the other sample's genotype to missing rather than call it as homozygous. In contrast, in the Platypus and GATK calls, the heterozygous genotype at these discordant SNPs tends to have lower GQ (Figures 4.A.6 and 4.A.7), lower coverage (Figures 4.A.10 and 4.A.11), and a skewed alternate allele fraction (Figures 4.A.12 and 4.A.13), suggesting that the heterozygous call is incorrect.

A subset of the Mendelian errors presented in Table 4.6 are *de novo* mutations. Table 4.7 shows the number of putative *de novo* mutations for the two offspring in the quartet. For the less stringent callsets, many of these are likely genotyping errors, but a subset are probably real. We expect that those seen in the most stringent callsets are more likely to be truly *de novo*. For population genetic analyses, it would

Table 4.5: **Rate of discordant genotypes between duplicate samples for different callsets.** The error rate is the proportion of SNPs called that had different genotypes in N703 *versus* N749. IBS: identity-by-state; IBS0: no alleles called in common in both samples i.e. one sample was called 0/0 (homozygous for the reference allele) and the other 1/1 (homozygous for the alternate allele); IBS1: one allele was called in both samples i.e. one sample was 0/1 (heterozygous) and the other was 0/0 or 1/1. Only SNPs not called as missing in either sample were included.

callset	Total Number of SNPs		Error rate per SNP		
	IBS0	IBS1	Total	IBS0	IBS1
CORTEX PASS	2,613	42,401	2,398,862	1.09×10^{-3}	1.77×10^{-2}
CORTEX PASS with GQ < 10 filtered	0	603	601,888	0	1.00×10^{-3}
GATK PASS	2,245	195,130	9,976,074	2.25×10^{-4}	1.96×10^{-2}
GATK/CORTEX intersection	0	7,372	2,362,242	0	3.12×10^{-3}
GATK/CORTEX intersection with GQ < 20 filtered	0	917	2,335,575	0	3.93×10^{-4}
Platypus PASS	112	64,787	8,413,819	1.33×10^{-5}	7.70×10^{-3}
Platypus/CORTEX intersection	3	5,452	2,278,649	1.32×10^{-6}	2.39×10^{-3}
Platypus/CORTEX intersection with GQ < 20 filtered	0	1,125	2,226,432	0	5.05×10^{-4}

Table 4.6: **Rate of Mendelian errors for different callsets.** The error rate is the proportion of called SNPs that were Mendelian errors. This was evaluated in the two trios from the Shoalhaven quartet (N742-N712-N752 and N757-N712-N752), and the Fish River (Taronga) duo (N736-N735). Note that in the duo, the only Mendelian errors that can be detected are those SNPs at which the father is 0/0 and the daughter is 1/1, or *vice versa*. The rightmost column shows the average number of SNPs considered across the trios and duo. SNPs were only considered if they had a nonmissing genotype in all three/two individuals.

SNP callset	N742-N712-N752			N757-N712-N752			Quartet average		N736-N735		N. SNPs
	N. errors	Error rate		N. errors	Error rate		Error rate		N. errors	Error rate	
CORTEX PASS	41,510	1.76×10^{-2}		30,025	1.27×10^{-2}		1.51×10^{-2}		12,884	5.45×10^{-3}	2,365,048
GATK PASS	104,887	1.05×10^{-2}		94,951	9.53×10^{-3}		1.00×10^{-2}		12,300	1.24×10^{-3}	9,961,349
GATK/CORTEX intersection	3,606	1.55×10^{-3}		3,367	1.45×10^{-3}		1.50×10^{-3}		615	2.64×10^{-4}	2,329,245
GATK/CORTEX intersection with GQ filter	564	2.64×10^{-4}		596	2.77×10^{-4}		2.70×10^{-4}		170	8.05×10^{-5}	2,133,550
Platypus PASS	24,841	2.97×10^{-3}		26,166	3.12×10^{-3}		3.05×10^{-3}		4,777	5.71×10^{-4}	8,372,262
Platypus/CORTEX intersection	1,991	8.76×10^{-4}		2,038	8.96×10^{-4}		8.86×10^{-4}		495	2.18×10^{-4}	2,274,197
Platypus/CORTEX intersection with GQ filter	458	2.07×10^{-4}		520	2.34×10^{-4}		2.21×10^{-4}		101	4.56×10^{-5}	2,215,677

be useful to calculate the mutation rate, but to do this, we need to know what proportion of the genome (or, in this case, of the autosomal scaffolds over 50kb) is actually callable. This is not straightforward to determine for the callsets involving CORTEX or Platypus, but for the GATK-only callsets, we can determine this by simply applying the same filters (see Section 4.2.4 and Table 4.A.4) to all positions in the genome. Using the GATK PASS calls with GQ filter ($GQ > 20$), and ignoring the putative *de novos* seen in both offspring, which are likely spurious, we estimate:

$$\begin{aligned}\mu &\approx \frac{\text{N. putative } de \text{ novos for N742} + \text{N. putative } de \text{ novos for N757}}{\text{N. sites callable in N742 \& parents} + \text{N. sites callable in N757 \& parents}} \\ &\approx \frac{66 + 103}{872184858 + 883117446} \approx 9.6 \times 10^{-8} / \text{bp/generation}.\end{aligned}$$

The true mutation rate in this quartet is probably somewhat lower than this, since, as we noted earlier, GATK, when used alone, seems to over-call heterozygous sites. If we use the GATK/CORTEX intersection callset with the GQ filter but generously assume the callable region (the denominator) is the same as used for the GATK PASS calls with GQ filter, we estimate that $\mu = 1.2 \times 10^{-8} / \text{bp/generation}$, which is exactly the estimate given for humans by Scally and Durbin (2012), who took an average of the number of *de novo* mutations found in several trio sequencing studies. Since the callable region (the denominator) is, in reality, much reduced with CORTEX, the true mutation rate for this quartet is probably somewhere between 1.2×10^{-8} and $9.6 \times 10^{-8} / \text{bp/generation}$. We intend to try to validate the putative *de novo* mutations from the less stringent callset (GATK PASS calls with the GQ filter) using an independent technology, Sequenom, to give us a more accurate estimate, but the data are not yet available.

Overall, the GATK/CORTEX and Platypus/CORTEX intersections with the GQ filter had similar rates of error. We used the GATK/CORTEX intersection with GQ filter for the analysis in this chapter, except when otherwise noted.

Surprisingly, using this stringent callset, a large number of SNPs (65,542) were called as homozygous for the alternate allele (1/1) in N725 (Glennie), the reference individual. SNPs that were called 1/1 in all samples had already been removed during filtering¹³, so 1/1 calls remaining in N725 cannot be simply attributed to errors in the reference. We examined about 10 of these SNPs in IGV and could not identify

¹³Of the 16,762,328 SNPs in the GATK raw SNP set, 193,992 (1.15%) were fixed for 1/1 in all samples (presumably reference errors) and 677,611 were 1/1 in N725 (4.04%).

Table 4.7: **Number of putative *de novo* mutations in different callsets.** These have been filtered to remove any SNPs at which either parent had one or more reads containing the alternate allele, or which were seen in any of the other 57 samples. The rightmost column shows the number of SNPs that were heterozygous in both offspring and homozygous reference in all other samples. These could be due to mutations mosaic in a parent, present at low frequency or not at all in the sampled tissue, but they are more likely technical artefacts (i.e. the alternate allele was missed in the parent it was inherited from).

SNP callset	N742 only	N757 only	both
CORTEX PASS	143	179	112
GATK PASS	102	161	53
GATK PASS with GQ filter	66	103	21
GATK/CORTEX intersection	9	18	3
GATK/CORTEX intersection with GQ filter	6	15	1
Platypus PASS	62	83	26
Platypus/CORTEX intersection	8	17	3
Platypus/CORTEX intersection with GQ filter	7	17	3

any problems with them. We checked whether the SNPs called as 1/1 in N725 tended to fall between contigs, in sequence that had been filled in with Moleculo reads from N721 or with the contigs assembled *de novo* from other samples, but only 1.6% of them were within 100bp of the end of a contig. These SNPs tended to have higher alternate allele frequencies than SNPs called as 0/0 (homozygous for the reference allele) or 0/1 (heterozygous) in N725 (Figure 4.14A). They had a very similar distribution of coverage and of GQ to the 0/0 SNPs (Figure 4.14B,C), speaking against the idea that they were just under-called heterozygous SNPs. Only 2,627 of the 65,642 (4.0%) had any reads with the reference allele (compared to 55,913 of the 2,080,056 0/0 SNPs (2.7%)). The reads supporting the alternate allele had a shift towards lower MAPQ than those supporting the reference allele (Figure 4.14D). There was a small but significant enrichment of 1/1 SNPs in repeats (odds ratio = 1.025, $p=0.007$ in Fisher's exact test). However, none of these metrics could explain a substantial proportion of these SNPs. We think it is possible they are due to reads mapping in collapsed paralogues or repeats that have not been annotated by RepeatMasker, but we find it difficult to see why this would not result in 0/1 rather than 1/1 calls, and why the SNPs do not have systematically higher coverage. The 1/1 SNPs are slightly more clustered than random SNPs (Figure 4.15). They also show an unusual pattern of substitutions (Figure 4.16). Whereas the stringent SNPs set as a whole

is dominated by C-to-T substitutions, particularly at CpG sites, presumably due to the spontaneous deamination of methylated cytosines, the SNPs 1/1 in Glennie are depleted for these and enriched for T-to-C substitutions. The significance of this is not clear. Thus, we do not have a satisfactory explanation for the large number of 1/1 calls in Glennie, but since, by the metrics described earlier, this callset seemed to have high quality genotypes, we decided to proceed with other analyses, from which we excluded these suspect SNPs.

One might be concerned that variation in coverage between samples (Figure 4.8) would lead to artifactual results due to differences in power to call SNPs. There is clearly an inverse correlation between the level of coverage and proportion of missing SNP calls, but there is no strong relationship between coverage and the proportion of SNPs called heterozygous or homozygous for the alternate allele (Figure 4.17). This suggests that the coverage levels and filtering criteria chosen were sufficient to avoid biases in SNP calling, and so this is unlikely to have a substantial effect on results.

4.3.3 Population structure, diversity and history

For the analyses in this section (unless otherwise noted), we reduced our 58 samples to a subset of 42 unrelated samples, removing one from each relative pair in Table 4.A.7 so as to maximise the number of samples. The samples retained are indicated in blue font in Table 4.2. We first ran a PCA to summarise the genetic variation. Figure 4.18 shows the first two principal components (PCs) when all SNPs are used, after filtering on MAF and missingness, as described in Section 4.2.6. The first PC separates the Tasmanian from the mainland samples, and accounts for 34.5% of the variation, while the second separates the mainland samples on a north-south axis and accounts for 12.4%. If we prune the SNPs based on LD, the first two PCs are exchanged (Figure 4.A.14) and together account for only $14.6\% + 13\% = 27.6\%$ of the variation. This is due to the removal of many SNPs which are fixed for the alternate allele in Tasmania *versus* the mainland and are therefore in perfect LD with each other.

In order to explore differences between populations, we divided the samples into five groups based on a combination of these PCA results and their known geographical proximity to one another. These five groups, and four more general groups, are

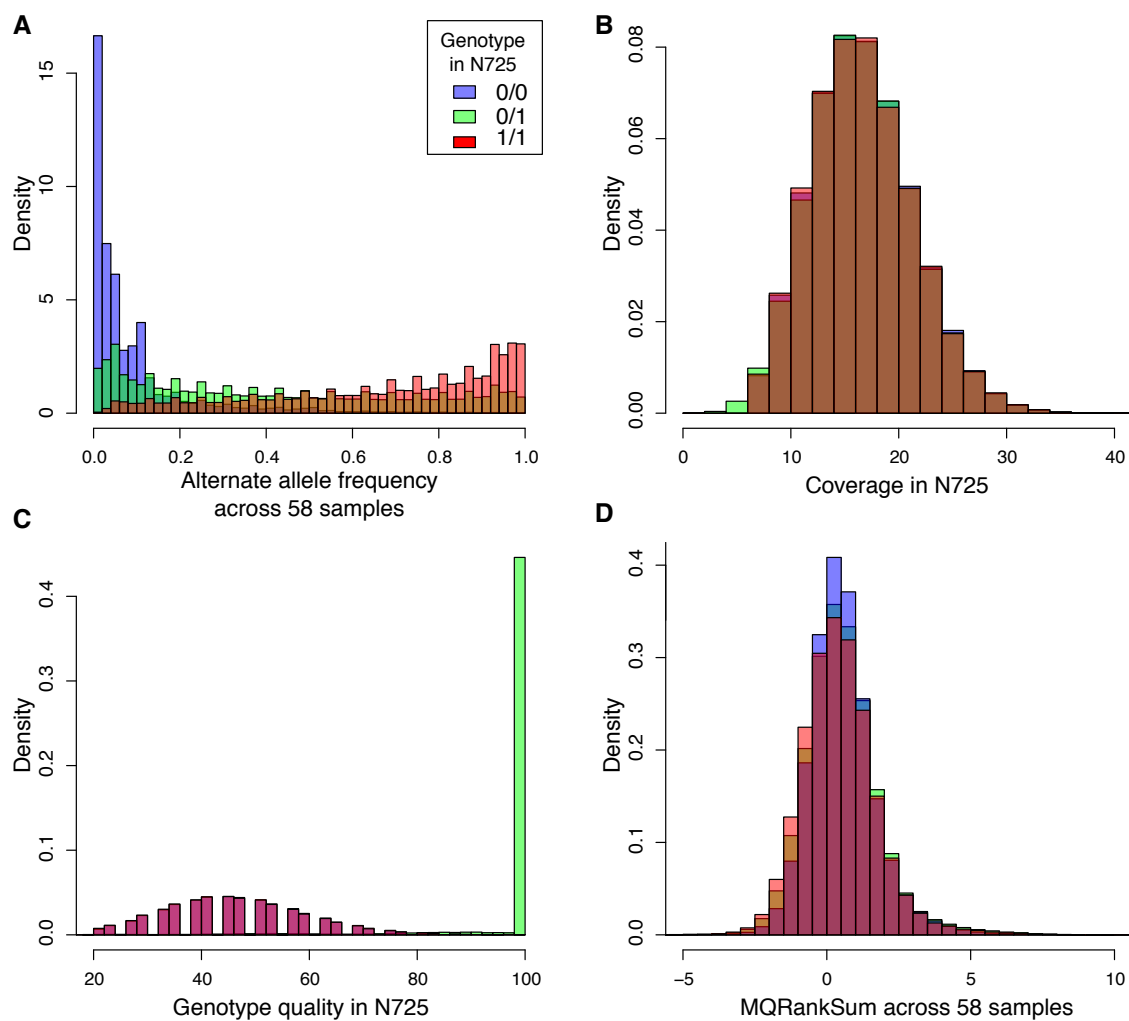


Figure 4.14: **Distribution of frequency and quality metrics for SNP calls in N725 (reference individual).** MQRankSum: Z-score from Wilcoxon rank sum test of comparing MAPQ for reads with the alternate *versus* reference allele. A negative value indicates that the reads supporting the alternate allele have lower MAPQ than those supporting the reference allele.

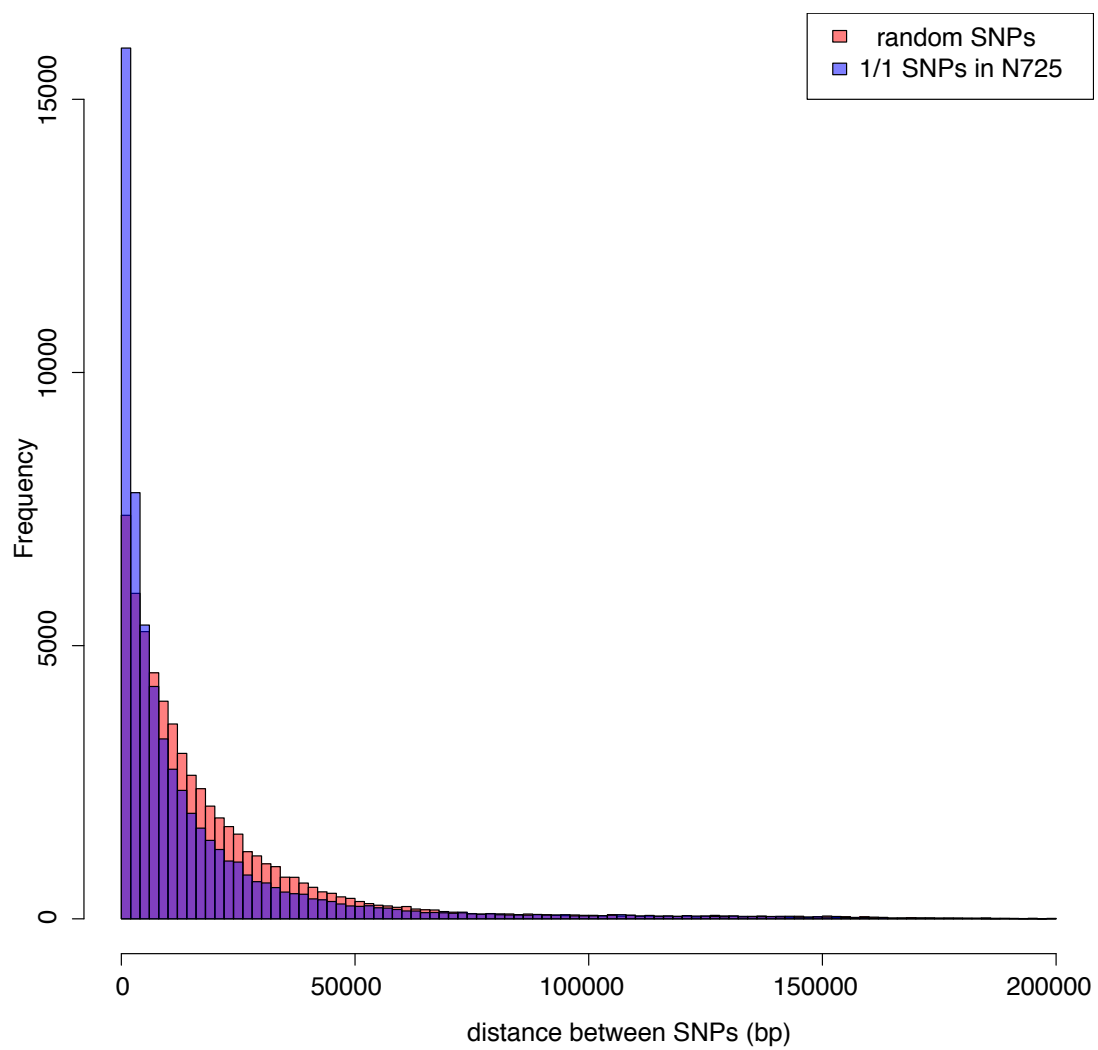


Figure 4.15: Distribution of distances between SNPs 1/1 in N725 *versus* random SNPs.

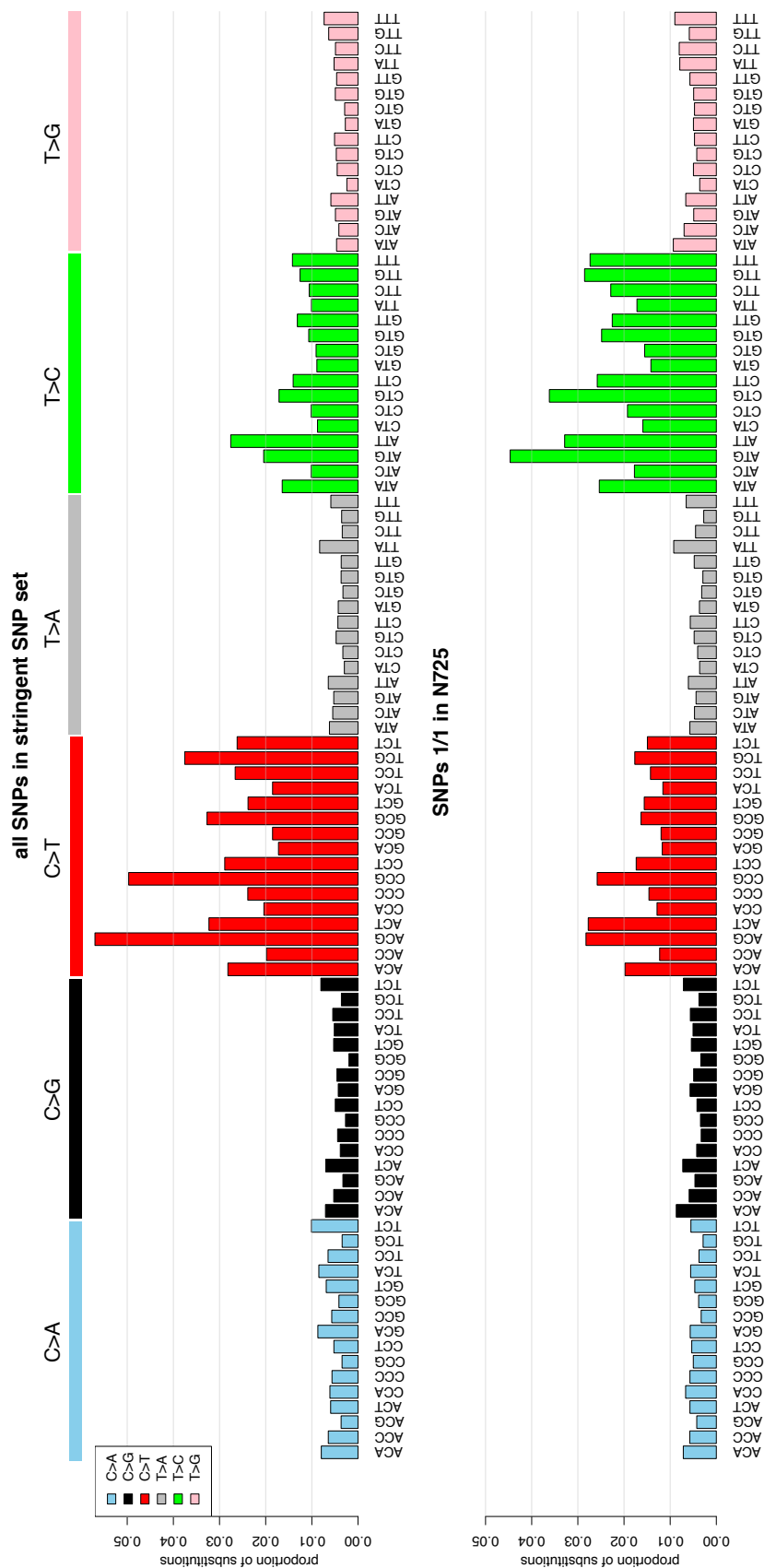


Figure 4.16: **Sequence context of nucleotide substitutions for SNPs 1/1 in N725 versus random SNPs.** The SNP substitutions are shown with the pyrimidine base first, with the 3bp surrounding the mutated base indicated by the labels under the bars. The y-axis indicated the proportion of all SNPs.

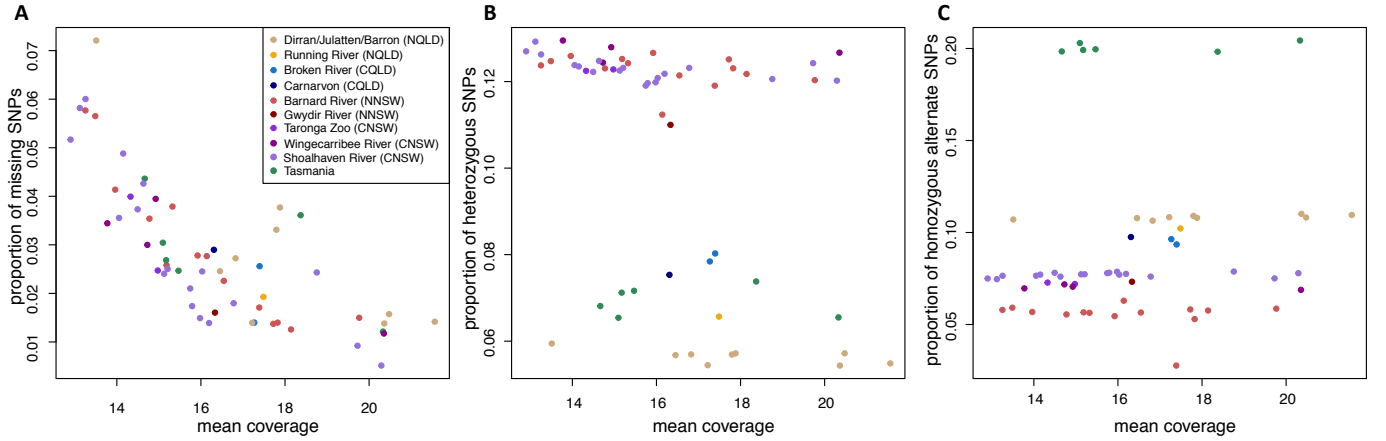


Figure 4.17: **Effect of coverage on SNP missingness and genotypes.** These plots show proportion of SNPs in the stringent SNP set that are missing (A), heterozygous (B) or homozygous for the alternate allele (C) *versus* coverage for each sample. The samples are coloured by population.

indicated in Table 4.2. The groups show different levels of nucleotide diversity¹⁴, π (Nei and Li, 1979), ranging from 7.04×10^{-4} in the North QLD population to about 1.3×10^{-3} in the Barnard and Shoalhaven rivers (Table 4.8). A large proportion of the SNPs segregating in each population are only polymorphic in that population (Figure 4.19): 33% of those in North QLD, 29% in North NSW, 37% in Central NSW and 74% in Tasmania.

There was high F_{ST} between the different population groups, with the highest value between Tasmania and North QLD (0.792) and the lowest between the Wingecarribee and North NSW populations (0.085) (Table 4.9). The F_{ST} values were slightly lower when we pruned SNPs based on LD, presumably because of the removal of many fixed differences between populations. There were a large number of fixed differences between both the Tasmanian and North QLD populations and the reference (Figure 4.A.15): 15.2% of the ~ 2.3 million SNPs segregating in the fifty-eight samples were fixed for the alternate allele in the five unrelated Tasmanian samples, and 4.1% were fixed in the seven unrelated North QLD samples.

We applied STRUCTURE (Pritchard *et al.*, 2000), a Bayesian model-based clustering algorithm, to identify subpopulations and assign the samples to them without using any prior information. Figure 4.20 shows the results from an admixture model

¹⁴The nucleotide diversity is the average number of nucleotide differences per site between any two DNA sequences chosen randomly from a sample population.

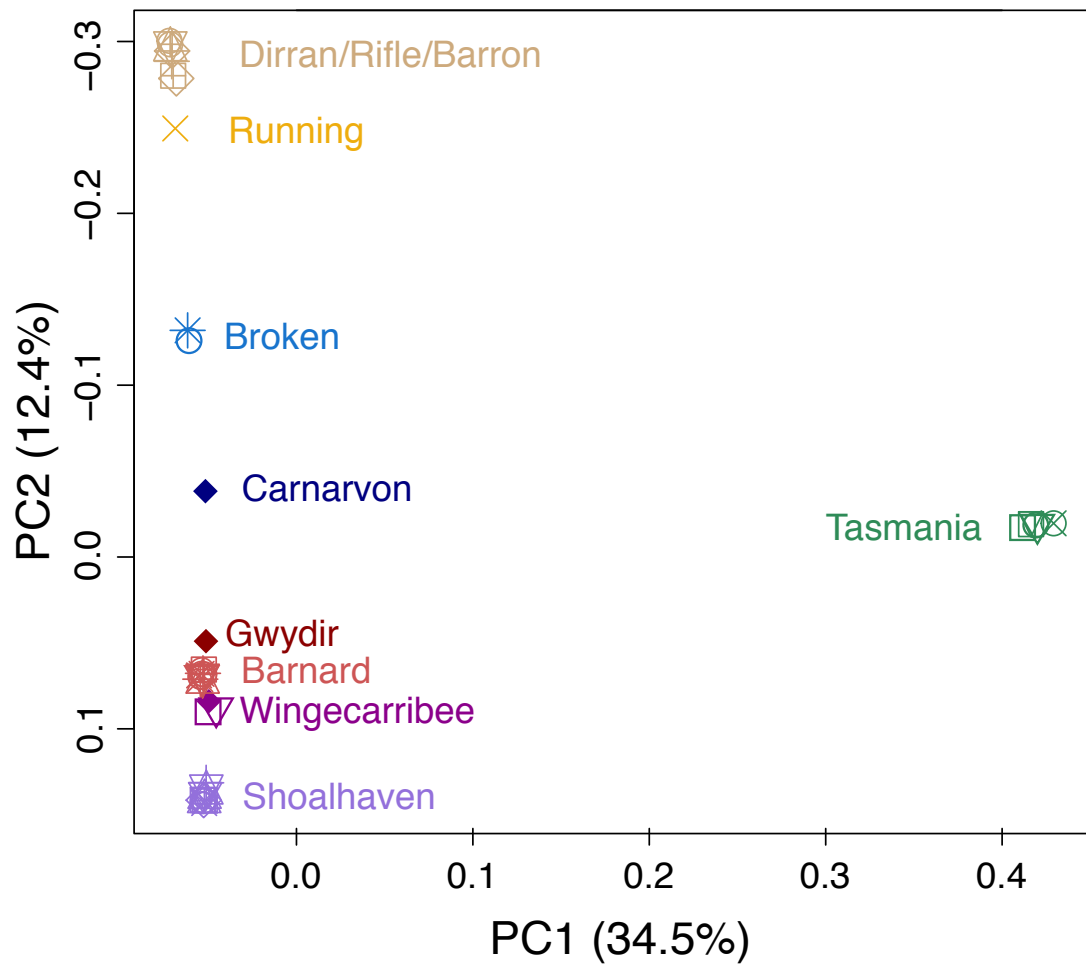


Figure 4.18: **Principal component analysis of 42 unrelated samples.** The first two principal components are shown, with the proportion of variance accounted for indicated in parentheses. The populations are labelled in the same colour as the corresponding points.

Table 4.8: **Nucleotide diversity (π) in different populations.** Note that the number of callable bases was only determined accurately for the GATK PASS SNPs. The same number was used to calculate π for the GATK/CORTEX intersection with GQ filter, although the true number of callable bases is probably much lower for this callset, and thus π has been underestimated. Note that North NSW is Barnard + Gwydir, and Central NSW is Shoalhaven + Wingecarribee + Fish Rivers.

Population	N. samples	$\pi \times 1000$	
		GATK PASS	GATK/CORTEX intersection with GQ filter
North QLD	7	0.704	0.131
Barnard	10	1.370	0.273
North NSW	11	1.386	0.277
Shoalhaven	12	1.342	0.273
Central NSW	15	1.398	0.284
TAS	5	0.796	0.162
all	42	1.674	0.348

Table 4.9: F_{ST} **between different populations.** The black numbers below the diagonal are calculated using SNPs before LD pruning, and the blue ones above the diagonal are calculated after LD pruning (see Section 4.2.6).

	North QLD	North NSW	Wingecarribee	Shoalhaven	TAS
North QLD	-	0.337	0.337	0.371	0.682
North NSW	0.388	-	0.078	0.128	0.472
Wingecarribee	0.394	0.085	-	0.081	0.461
Shoalhaven	0.427	0.147	0.09	-	0.48
TAS	0.792	0.658	0.652	0.659	-

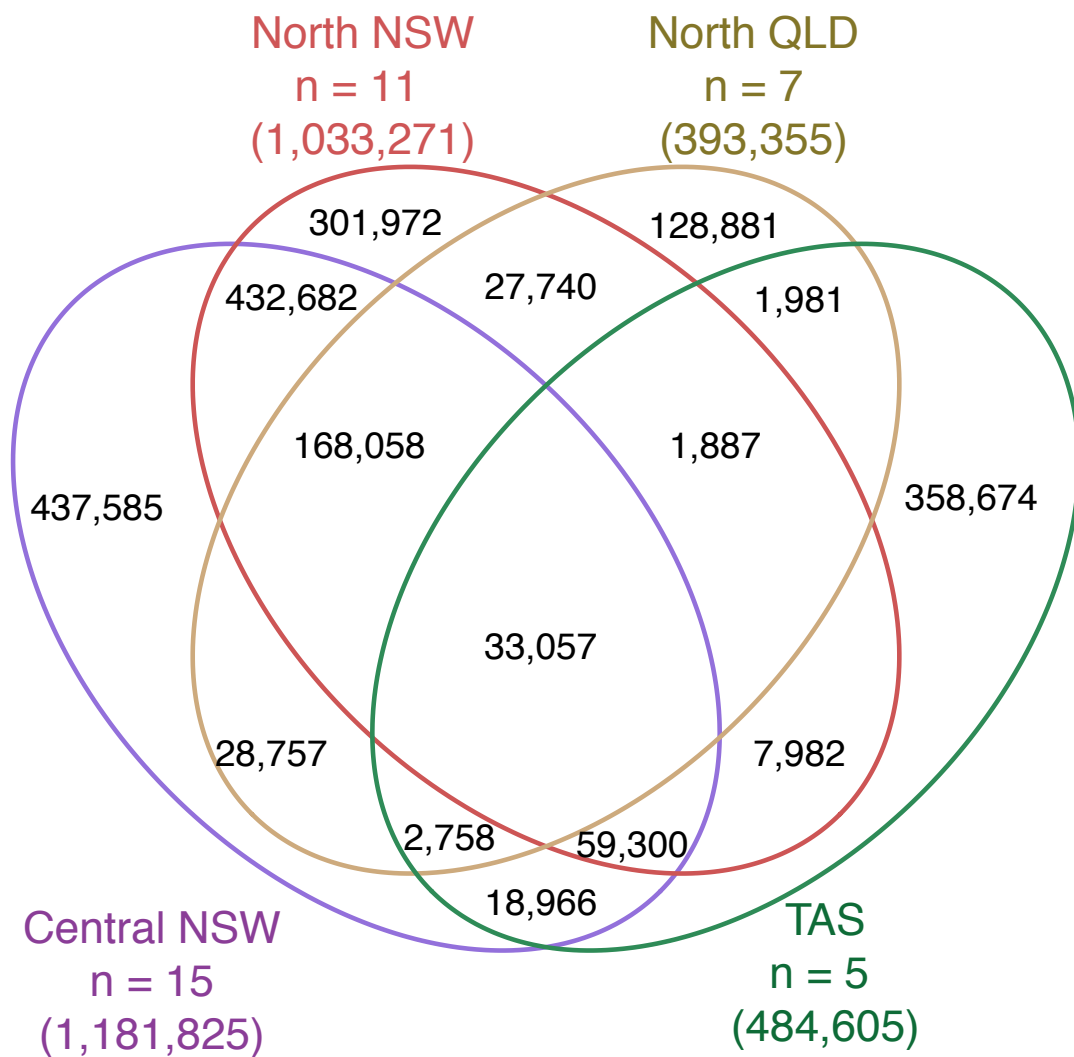


Figure 4.19: **Venn diagram showing number of SNPs segregating in different populations.** The numbers in parentheses are the total numbers of segregating SNPs in the unrelated samples from that population, out of the ~2.3 million SNPs segregating in the 58 samples.

in which individuals are allowed to have membership in more than one of the K populations, or clusters. For $K = 2$, the clusters are anchored by North Queensland and Tasmania, with the Central QLD and NSW populations inferred to be a mixture of these two. The next cluster at $K = 3$ corresponds to these Central QLD and NSW individuals, and at $K = 4$, the North NSW individuals are delineated (Barnard and Gwydir), along with individuals from the Fish and Wingecarribee Rivers. The additional cluster at $K = 5$ contributes the majority of the ancestry of the Central QLD individuals, as well as a small amount to the samples from the Running River (originally classed as North QLD on the basis of the PCA results, although it should perhaps be classed as Central QLD since it is from the same river system as the Broken River individuals - see Figure 4.7), the Gwydir River (North NSW), and the Fish and Wingecarribee Rivers (both Central NSW). We tried running STRUCTURE with $K = 6, \dots, 10$ but found that the proportion of ancestry assigned to the additional clusters was always very low in all samples, and the posterior mean was close to 0, so, although slightly higher log likelihoods were observed for some runs at $K = 8$, we think that the simpler model with $K = 5$ represents the data well.

The most striking point in this figure is that the Wingecarribee and Fish River samples look more similar to the Gwydir and Barnard samples than they do to the Shoalhaven samples, despite being geographically closer to Shoalhaven (see Figure 4.7). This is consistent with the clustering seen on the PCA. Additionally, we see some similarity between the samples from the Broken River and from Carnarvon Creek, despite these waterways being several hundred kilometres apart and on different sides of the Great Dividing Range. We discuss possible explanations for the observed similarities in Section 4.4.

To explore the demographic history of the different populations, we applied the Pairwise Sequentially Markovian Coalescent (PSMC) method of Li and Durbin (2011) described in Chapter 1 to each of the 57 unique samples. Figure 4.21 shows estimates of N_e from each sample at a series of time intervals. These estimates were scaled using a mutation rate, μ , of 1.2×10^{-8} /bp/generation, the estimate given by Scally and Durbin (2012) for humans (which also happens to be the same as the conservative estimate from the platypus quartet)¹⁵ and a generation time of $g = 10$ years,

¹⁵Arguably, we should use the neutral substitution rate, which is likely to be lower than the *de novo* mutation rate since some *de novo* mutations will be eliminated by purifying selection. On the other hand, the substitution rate might be expected to change as N_e changes, due to different efficacy of selection. Thus, for the analysis here, we use an estimate of the *de novo* mutation rate,

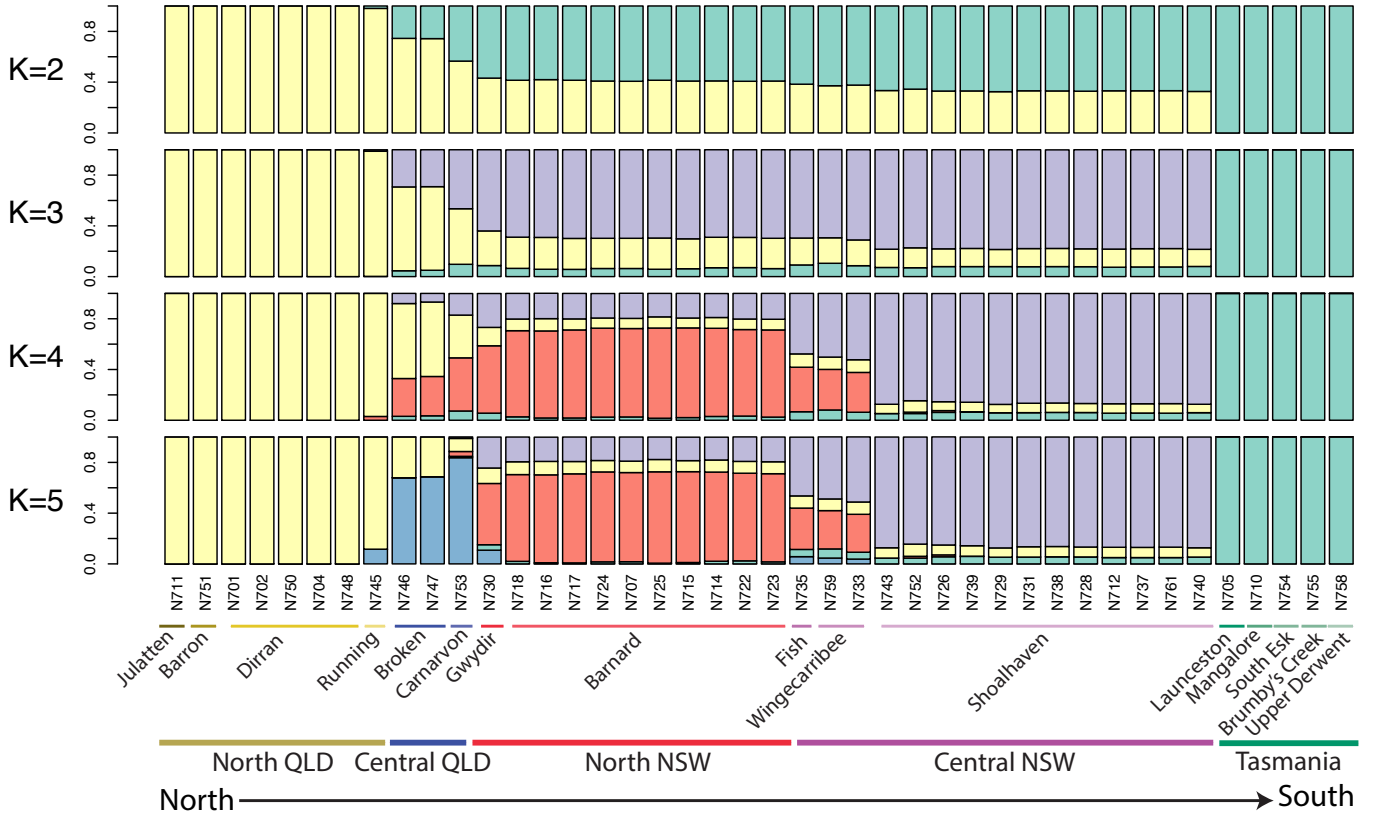


Figure 4.20: **Population structure inferred from 42 unrelated individuals using STRUCTURE.** Each individual is represented by a vertical bar partitioned into K coloured segments that represent the individual's estimated membership fractions in K clusters. Ten STRUCTURE runs at each K produced very similar results, and so the run with the highest likelihood is shown. The Broken River and Carnarvon samples are labeled as Central QLD and the Running River sample as North QLD, even though these samples did not form part of large clusters on the PCA and were excluded from the groupings used in Tables 4.8 and 4.9 and Figures 4.19 and 4.A.15.

from Furlan *et al.* (2012). The μ parameter affects the scaling on both the X- and Y-axes (representing time and N_e respectively), and g affects the X-axis. Although the μ value we have used may differ from the true value by several fold, and g may be off by several years, changing these parameters will affect estimates from all samples equally (assuming all populations have the same true μ and g) (Figures 4.A.16 and 4.A.17). Thus, although we should not draw any firm conclusions about actual dates from the PSMC results, we can still infer historical differences between populations.

The first point to note from Figure 4.21 is that individuals from the same river show strikingly similar trajectories, supporting the precision of the N_e estimates. The samples from North and Central NSW all look very similar (Figure 4.A.18), suggesting that these may have been part of the same population within the timeframe accessible to the method¹⁶. However, we do note that although samples from the same population would be expected show the same N_e trajectory, having the same N_e does not necessarily mean that samples are from the same population. The North QLD samples from Dirran Creek, Rifle Creek and the Barron River appear to have been through a strong bottleneck (Figure 4.A.19), consistent with the low nucleotide diversity observed (Table 4.8). In contrast, the sample from Running River shows a different trajectory without the bottleneck (Figure 4.A.19), more similar to those of the Central QLD samples. This is consistent with the small amount of shared ancestry observed between this sample and the Central QLD individuals using STRUCTURE (blue component in Figure 4.20). The samples from different parts of Tasmania all have very similar N_e trajectories (Figure 4.A.20), with the exception of the most recent timepoint (until 20,000 years ago with the μ and g parameters used in this figure), which is known to be noisy due to the low number of such recent coalescences between two sequences (Li and Durbin, 2011). This suggests that the populations in the Tamar and Derwent River basins, which these samples come from, diverged more recently, and indeed, there may still be gene flow between them.

which we think it is unlikely to have changed substantially through the history of modern platypus.

¹⁶There are insufficient coalescences between two samples in the very recent and very ancient past to draw robust inferences in these time periods.

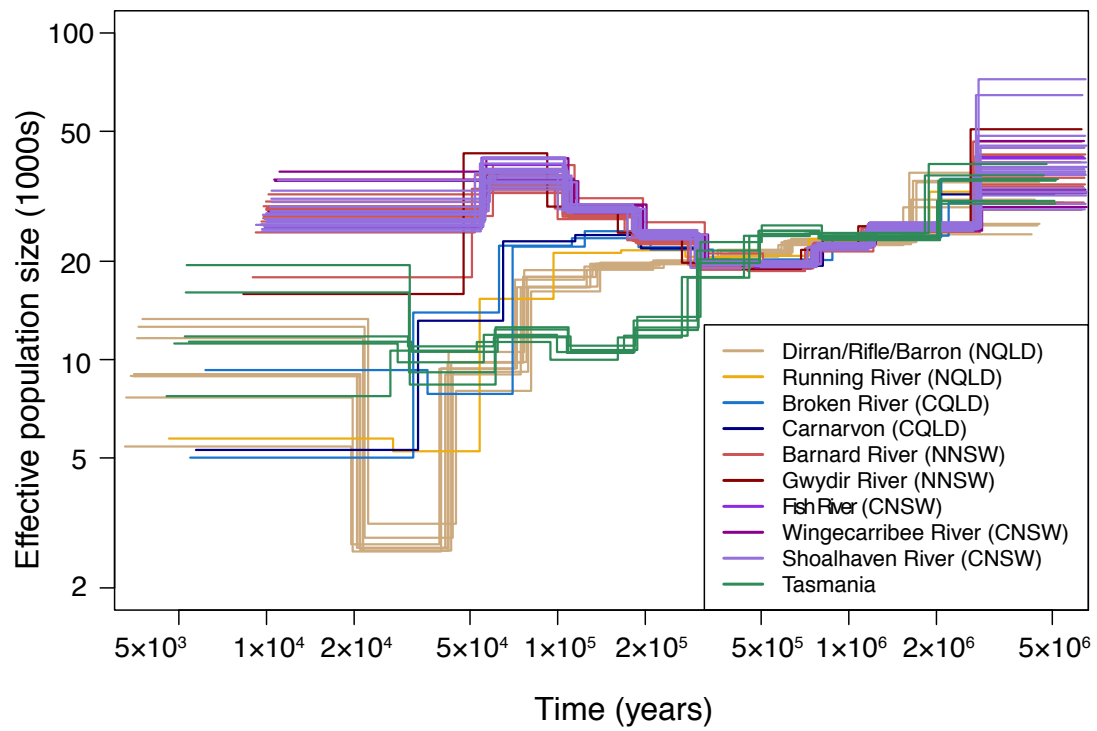


Figure 4.21: **Historical effective population sizes inferred using PSMC.** These were scaled using $g = 10$ and $\mu = 1.2 \times 10^{-8}$.

4.3.4 Sex chromosomes

In the original genome assembly, four scaffolds had been assigned to X chromosomes, specifically to chrX1 (45.5Mb scaffold), chrX2 (5.6Mb), chrX3 (6.0Mb) and chrX5 (27.8Mb). To identify X-linked scaffolds in our new assembly, we compared male and female coverage as described in Section 4.2.7. Amongst the scaffolds over 50kb, we found 624 unambiguously X-linked scaffolds on which the males had half the coverage of the females, and 5348 scaffolds with equal coverage in males and females, which are either autosomal or pseudoautosomal. There were an additional 187 scaffolds which contained a mixture of apparently X-linked and PAR/autosomal contigs. We term these “mosaic scaffolds”. Many of these are probably due to misassemblies, and indeed, we did find some inter-contig joins within these scaffolds that were only weakly supported by the mate pairs, BACs or fosmids.

Upon closer investigation, we found that thirteen of these mosaic scaffolds had regions of close homology to a Y chromosome (see an example in Figure 4.22). We implemented a heuristic algorithm to detect these Y homologous regions (YHRs) across all scaffolds (Section 4.2.8), and found them on thirty-two of the scaffolds longer than 50kb. These thirty-two scaffolds, which are described in Table 4.A.8, had a total length of ~17.4Mb of which ~2.9Mb was classified as YHR.

Based on the coverage, twenty of the thirty-two YHR-containing scaffolds appeared to come from an X-specific region, but had YHRs across some or all of the scaffold, in which the male and female coverage was approximately equal when the putative chrY reads were included. Five scaffolds (Scaffolds 40, 621, 2901, 5515 and 10588) appeared fully (pseudo)autosomal, apart from some YHRs ranging from ~400bp to ~15kb long, in which the male coverage was about ~1.5 times higher than in females. These are probably regions on an autosome or within a PAR which have been duplicated onto a Y chromosome and then diverged.

On five scaffolds (Scaffolds 0¹⁷, 101, 368, 1480, 2660), the coverage suggested that one end was (pseudo)autosomal and the other end was X-specific, with one or more YHRs either right at or close to the boundary, within the X-specific region (XSR). We scrutinised the inter-contig joins on these boundaries carefully. Those over the boundary on Scaffolds 101 and 2660 seemed likely to be spurious, but we found that

¹⁷The numbering of the scaffolds produced by PBJelly begins at 0, and we retain this labelling to avoid future confusion.



Figure 4.22: **An example of aligned reads indicating a Y-homologous region on Scaffold 0.** The grey bars indicate mapped reads (outlined in red if their mate is not mapped; connected by a black horizontal line if they have deletions in them), and vertical coloured lines within them are divergences from the reference. These divergences are also displayed as coloured bars in the pileup along the top of each block of reads, which indicates coverage. Heterozygous variants appear as two different colours. Note the dense heterozygous SNPs in the male sample, absent from the female. These are presumed to be divergences between this X-chromosomal scaffold and its Y gametolog.

those on Scaffolds 0 (Figure 4.23), 368 and 1480 were well supported by the mate pairs (MPs) and matched, more or less, the order of contigs in ornAna1. It therefore seems likely that these three scaffolds contain *bona fide* PAR boundaries, where the PAR transitions into the XSR, and that the XSR proximal to the boundary comes from a stratum that formed relatively recently, and thus still has substantial homology with the Y gametolog.

Table 4.A.9 shows the levels of X-Y divergence in the YHRs we found, and Table 4.1 shows the levels in the genes reported by Cortez *et al.* (2014). On Scaffold 368, the level of X-Y divergence in the YHR at the putative PAR boundary was similar to the genes assigned to Cortez’s Stratum 6 (1 per $\sim 18.8\text{bp}$)¹⁸. On Scaffold 1480, it was much lower (1 per $\sim 55.5\text{bp}$). The YHRs at 1,337-1,351kb on Scaffold 179 and at the PAR boundary on Scaffold 0 had the lowest level of X-Y divergence: 1 per $\sim 146\text{bp}$ and 1 per $\sim 120\text{bp}$ respectively.

We attempted to determine which chromosomes these scaffolds came from using the FISH data from the original genome paper, combined with synteny information from chicken and human (Section 4.2.9; Table 5.A.10). The FISH data allowed us to place the longest scaffold, Scaffold 0 ($\sim 6.4\text{Mb}$), in the PAR of chrY2 and either chrX2p or chrX3p (these could not be distinguished using FISH). Based on the observations described above, Scaffold 0 seems likely to span the boundary of this PAR with the neighbouring XSR on either chrX2 or chrX3. Consistent with this, and with Veyrunes *et al.* (2008) (Figure 4.6), we found synteny between Scaffold 0 and chicken chr2. The synteny information allowed us to place, tentatively, five additional YHR-containing scaffolds after the PAR boundary in Scaffold 0, within the XSR: Scaffolds 3642, 481, 1937, 8700 and 2552. These six scaffolds, together, are syntenic with the region at $\sim 64.7\text{-}73.4\text{Mb}$ on chicken chr2. The YHRs on Scaffolds 3642, 481, 1937, 8700 and 2552 had similar levels of X-Y divergence (all 1 per $\sim 11\text{-}16\text{bp}$), all much denser than the YHRs on Scaffold 0, suggesting that they may form part of the same stratum, which is older than the stratum (or strata - see below) on Scaffold 0. In fact, Scaffold 2552 and 3642 each contained a gene whose Y gametolog was found by Cortez *et al.* (2014), which they assigned to Stratum 5: *PRPF4B* on Scaffold 3642 and *SSR1* on Scaffold 2552.

¹⁸However, note that levels of X-Y divergence within genes are probably lower than in noncoding sequence.

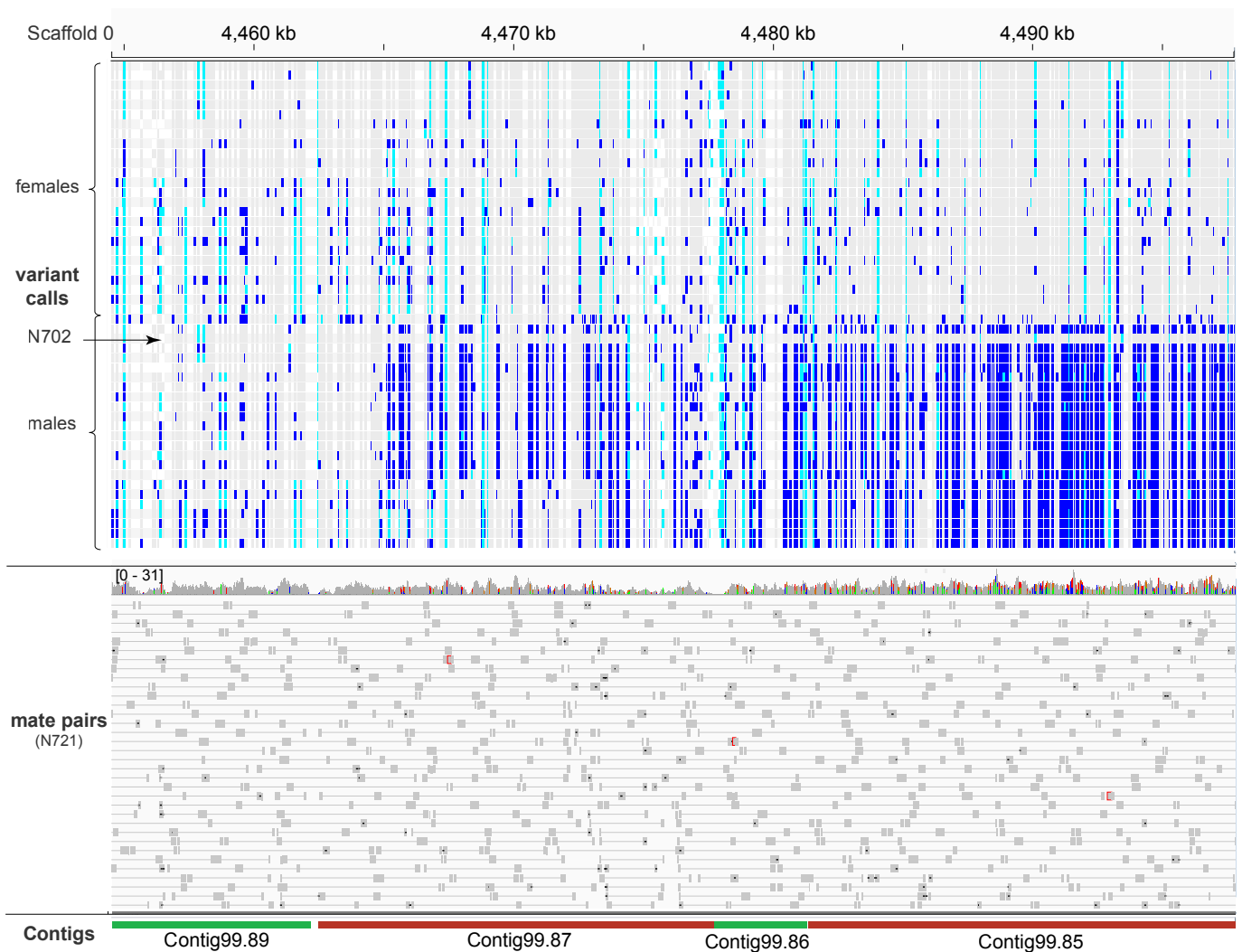


Figure 4.23: **The putative PAR boundary on Scaffold 0.** The top part of the figure shows the variant calls (GATK raw callset), with small dark blue and cyan bars representing SNPs that are heterozygous or homozygous non-reference respectively. Each row represents one sample, with females on the top and males underneath. Notice the transition into a region of dense heterozygous SNPs in all males (except N702, as discussed below). The coverage in all samples is approximately the same across this whole region, but in males, about half the reads on the right hand side are presumed to come from a Y chromosome. Thus, this is likely to be a PAR boundary, with PAR on the lefthand side, and X-specific sequence on the righthand side that still has close homology with its Y counterpart. The bottom part of the figure shows the mate pair reads (see legend to Figure 4.22 for an explanation), with contigs shown as alternating green and orange bars beneath them. The assembly appears to be correct here, since there are properly paired reads across the whole region.

Scaffold 40 is also syntenic with a different part of chicken chr2 (~6.2-8Mb), and Scaffold 179 contains several genes whose orthologs are in different regions of this chromosome, including *SDHA*, whose Y gametolog was assigned to Stratum 6 in Cortez *et al.* (2014). Another Stratum 6 gene, *EDF1*, is located within the YHR we detected on Scaffold 2833, which is syntenic with chicken chr17. Thus, based on this information, it is not clear whether Scaffolds 40, 179 or 2833, or Stratum 6, are on the same platypus chromosome as Scaffold 0 and (presumably) as Stratum 5.

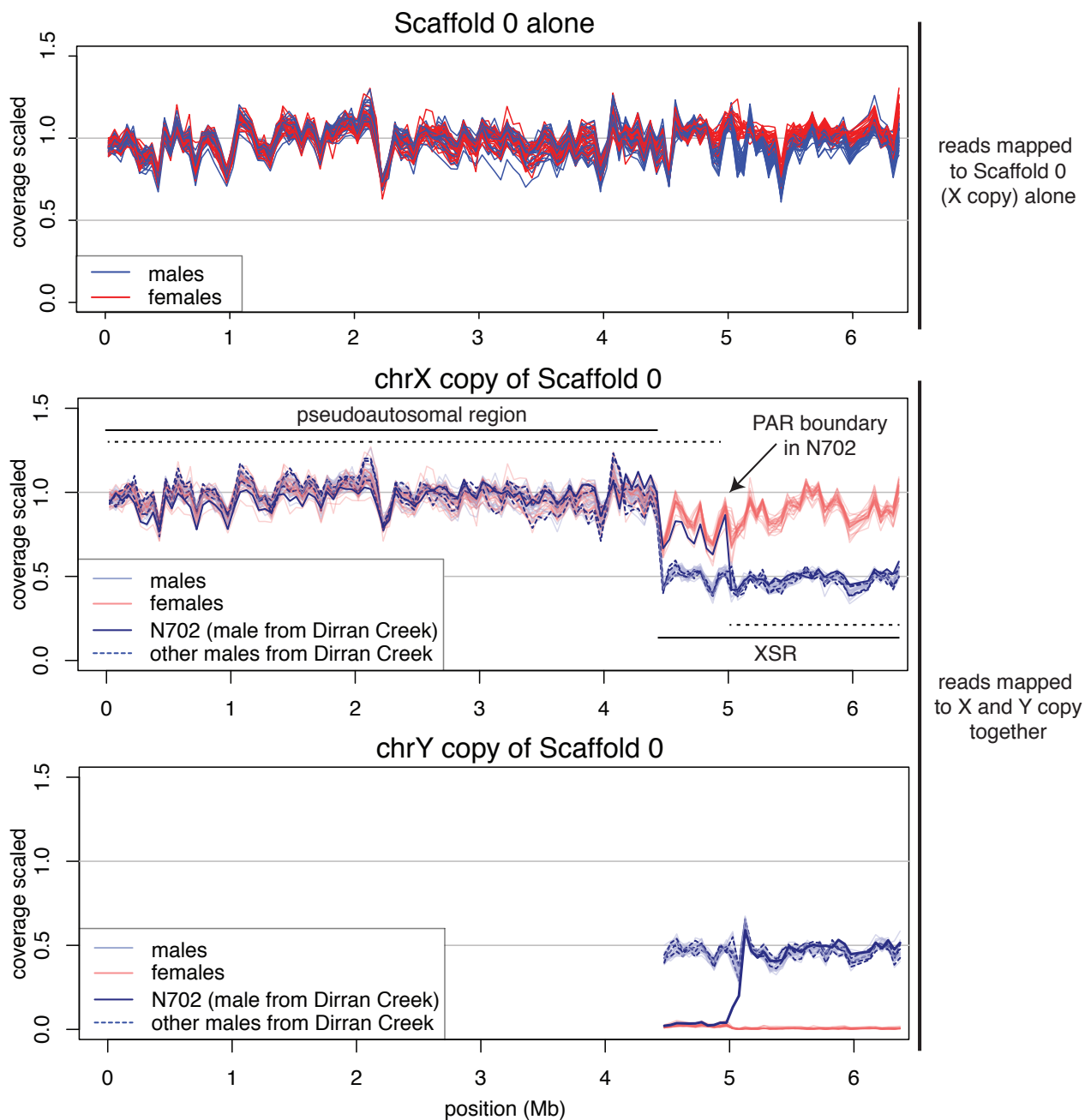
The pseudo(autosomal) part of Scaffold 101 is syntenic with chicken chr4 and also human chr4. The ~50kb YHR at the end of this scaffold (which may have been erroneously added, as noted in Table 4.A.8), contains a gene (ENSOANG00000031547) whose ortholog is on chicken chr13 and on human chr5. Within the same 2Mb region on chicken chr13/human chr5 are the orthologs of genes from Scaffolds 4505, 5038, 2839 and 11479, and of one of two genes on Scaffold 2660 (*SMAD5*, falling in the YHR). This suggests that these scaffolds may be on the same platypus chromosome, possibly chrX1, since the short arm of this chromosome (most of which is in the PAR with chrY1) was shown to share homology with chicken chr13 by Veyrunes *et al.* (2008) (Figure 4.6). However, if these scaffolds were lined up in the order suggested by the synteny with chicken chr13, there would be multiple switches between XSRs and PARs. Thus, it seems likely that there have been multiple rearrangements with respect to the chicken genome, and Scaffolds 101, 2660, 2839, 4505, 5038, and 11479 probably do not all fall on the same platypus chromosome. Alternatively, a single chromosome may contain several segments of recombining PAR flanked on both sides by non-recombining X-specific sequence.

Across a ~600kb region (~4.5-5.1Mb) within the YHR just to the right of the PAR boundary in Scaffold 0, sample N702 was homozygous reference at all the positions where all other males were heterozygous (see Figure 4.23). Thus, he appeared to have the “female haplotype”. We initially suspected that he simply had a deletion of the homologous region on the Y chromosome. To check this, we created a decoy sequence for this region by substituting in the alternate allele at all the male-specific heterozygous SNPs, and then realigned all male and female reads to the decoy sequence (the “Y copy”) and Scaffold 0 (the “X copy”) (Figure 4.24). Using this method, we showed that N702 was diploid for the X copy in the region ~4.5-5.1Mb, and haploid for both the X and the Y copy from 5.1Mb onwards. While the level of divergence seen between the X and Y in the YHRs suggests that recombination must

have been suppressed between them for a long time, it appears that this male has inherited the product of a crossover between the X and Y that occurred at around 5.1Mb on Scaffold 0 (likely to have been between either chrX2 or chrX3 and chrY2, based on the FISH results described above). This putatively recombinant haplotype is segregating with the river population, since other males from Dirran Creek were haploid for both the X and Y copy in the whole region from ~ 4.5 Mb onwards. It is possible that this male actually has the ancestral longer PAR and that all other males have the derived version (the X and Y haplotypes), but given the frequency, the hypothesis of a recent crossover seems more likely.

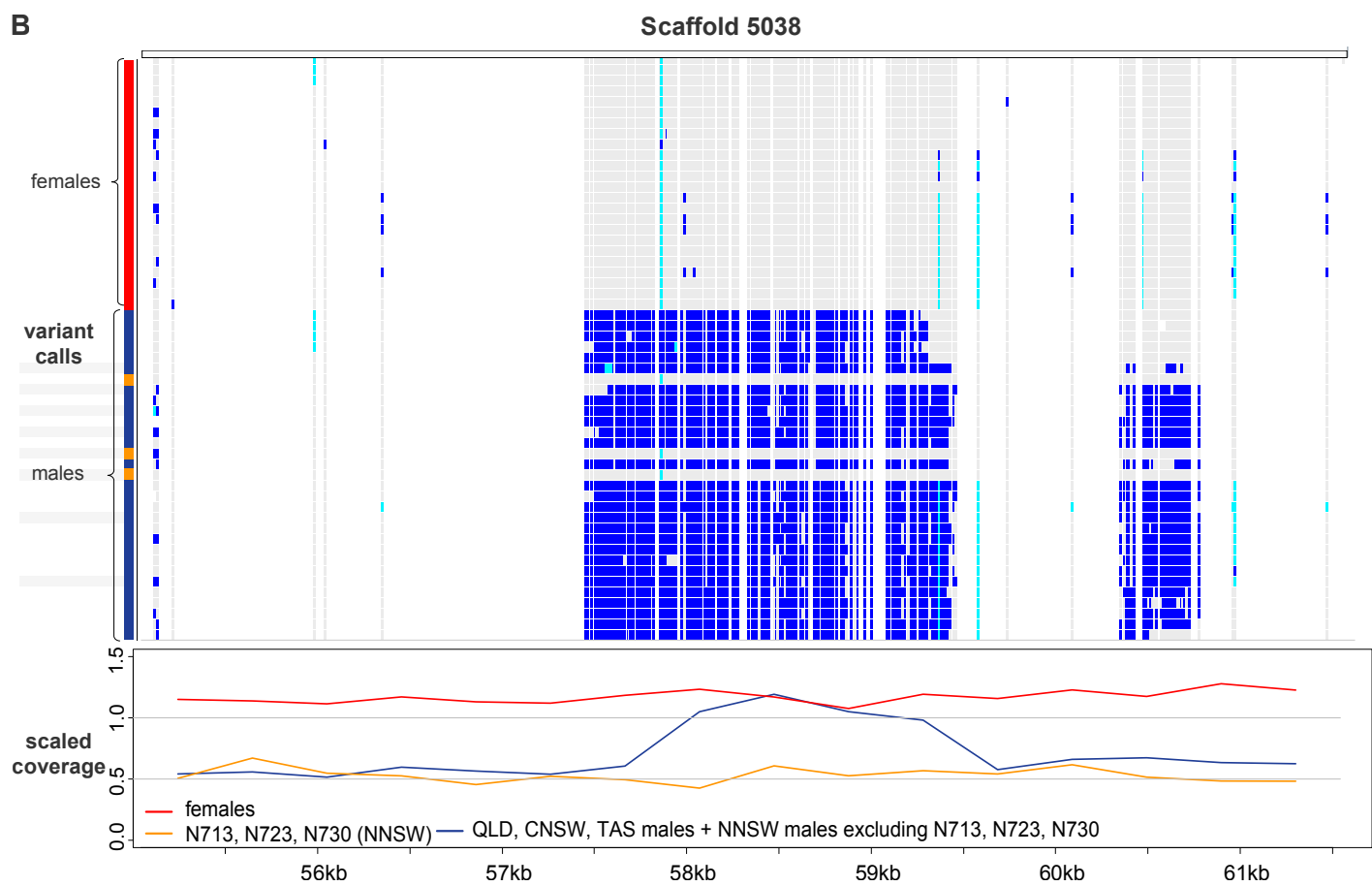
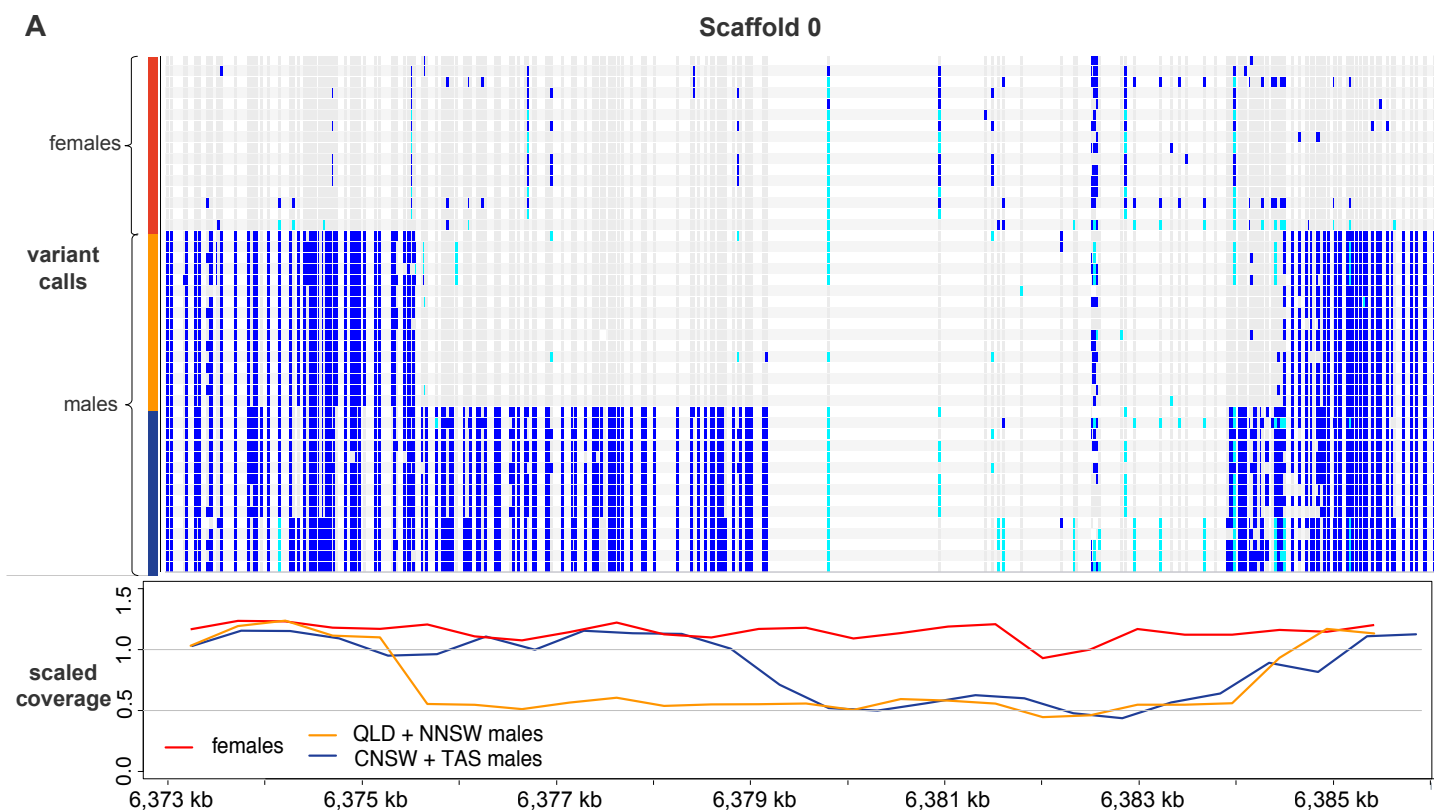
There were several other smaller parts of YHRs similar to this, in which there was a highly divergent Y haplotype that was seen in some but not all males. These were between ~ 1.5 kb and 10kb long, so much smaller than the 600kb region noted in N702. We initially wondered if they might be due to gene conversion events replacing small pieces of Y with the corresponding sequences from the X, which has been documented in primates (Trombetta *et al.*, 2014). However, a recent paper in humans suggested that gene conversion tract lengths are generally between 100bp and 1kb (Williams *et al.*, 2015), and so the regions we observed are probably too large for this. Another simpler explanation could be that these apparently polymorphic YHRs were due to segregating deletions on a Y chromosome, as we had initially suspected for the 600kb region on Scaffold 0. Figure 4.25 shows two examples where the coverage is apparently consistent with a deletion but not with a gene conversion (which would lead to diploid copy number in the region of the X lacking the YHR). However, the coverage pattern observed could also be consistent with the existence of multiple very different

Figure 4.24 (*following page*): **Examining coverage on the X and Y copies of Scaffold 0 indicates a polymorphic PAR boundary.** Red and blue lines represent the scaled coverage for female and male samples respectively, when the reads are mapped to Scaffold 0 alone (X copy; top plot) or to Scaffold 0 (X copy; middle plot) and the decoy sequence (Y copy; bottom plot) together. In the middle and bottom plots, N702 is shown with the dark blue solid line and other males from the same river with dark blue dotted lines. The black horizontal lines in the middle plot indicate the PAR and XSR, with the dotted line representing N702 and the solid line indicating all other samples. Note that coverage in the PAR (first ~ 4.5 Mb of the scaffold) is only shown in the plot of the X copy, although there is actually one copy of this sequence on the X and one on the Y. The female coverage is slightly below 1 within the XSR on the middle plot because we have removed reads with mapping quality 0 (i.e. which map equally well to the Y copy).



segregating Y haplotypes, one or more of which are too diverged from the X for the reads to align with the default substitution rate parameter (0.001) in Stampy (Lunter and Goodson, 2011). It may be possible to detect Y sequences which have diverged further from the X by increasing this substitution rate parameter while mapping the reads. In future analyses, we intend to compare this approach with *de novo* assembly of the Y chromosomes.

Figure 4.25 (*following page*): **Examples of inter- and intra-population polymorphism in Y-homologous regions, likely due to Y chromosomal deletions.** In each panel, the top part indicates the SNP genotypes, as described in the legend to Figure 4.23, with the coloured bars on the left indicating different groups of samples. The plot along the bottom shows the average scaled coverage along the scaffold (the X copy) for these groups. Panel A shows a region on Scaffold 0 where the males from QLD and North NSW (orange bar) lack the dense heterozygous SNPs seen in the Central NSW and TAS males (around 6,375.5-6,379kb and 6,383.8-6,384.4kb). The coverage plot suggests that the QLD and North NSW have only haploid coverage in the region 6,375.5-6,384.4kb, which implies either that they carry a deletion of the homologous copy on the Y, or that their Y haplotype is so far diverged from the X at this point that the reads cannot align. Panel B shows a similar example on Scaffold 5038, where the polymorphism occurs within the NNSW population. Most males show dense heterozygous SNPs and have diploid coverage in the region 57.5-59.3kb, but N713, N723 and N730 lack these SNPs and show haploid coverage on the X copy, indicating that no chrY reads are mapping here.



4.4 Discussion

In this chapter, we have described the first population sequencing of the platypus. The early analyses presented provide insights into the population structure and levels of diversity in this species, as well as their remarkable and apparently still-evolving sex chromosomes.

Throughout this project, problems with the reference genome have been a major challenge. Our reassembly has fixed some of these, as indicated by the reduction in switch error rate (Table 4.A.6), but clearly further major improvements are needed. Firstly, *ornAna1.2* still contains a large number of scaffolds, the vast majority of which are unassigned to chromosomes. Secondly, there are evidently still many spurious joins between contigs, since the switch error rate, though improved compared to *ornAna1*, remains high. Thirdly, there were 187 scaffolds that appeared to contain a mixture of X-specific and (pseudo)autosomal contigs, most of which cannot be explained by YHRs, although a few may contain PAR boundaries, perhaps from the older sex chromosome pairs. Scrutinising some of these cases suggested that errors had been caused by BAC or fosmid reads that landed within repeats. We had only removed short reads (mate pairs and paired-end reads) that landed within repeats, since the BAC and fosmid reads were ~ 800 bp long, so we presumed they would be more likely to span unique sequence as well. However, we should be more stringent about this in the next round of assembly, which we hope will be greatly improved by the inclusion of long PacBio reads and by the HiC-based method from Dovetail Genomics. In hindsight, we probably should have generated PacBio data early in the project, but because of its high error rate and cost at the time, we instead chose mate pair sequencing and Molecule. However, due to the difficulty in making libraries (for the mate pairs), and the cost (for Molecule), we did not generate sufficient coverage of either to make a substantial difference to the contiguity. Additionally, it is likely that the short-read assembly in the Molecule protocol broke down in the same places that the original assembly did i.e. in long, frequently interspersed repeats.

In the analyses presented in Sections 4.3.2 and 4.3.3, with the exception of the PSMC results, the relative positions of the SNPs analysed are not important. Thus, we think aberrantly placed contigs have probably had a minimal effect on results, although there were likely some genotyping errors made in regions with collapsed repeats. These would not have been detected by our QC if they did not cause Mendelian

errors or have discordant genotypes between N703 and N749. However, for analyses of recombination, which are of particular interest to us, we definitely need to improve the contiguity and accuracy of the assembly.

Little is known about the extent to which platypus disperse over their lifetimes. Our data shed some light on this, since we detected many pairs of relatives within the same river, including a whole quartet from the Shoalhaven River, which, fortuitously, allowed us to check the Mendelian error and switch errors rates and estimate the mutation rate. The mother, her sister and her two offspring were collected in a pool at a junction about 2km downstream of the point where the father was sampled, which was in Jerrabattgulla Creek, a small tributary of the Shoalhaven. Unfortunately, because of difficulties in determining the age of platypus, it is not possible to tell whether the two offspring were born as dizygotic twins (common in platypus) or at different times, which would indicate that their parents mated in more than one year. The fact that these samples were all collected within a small distance of each other over a relatively short timeframe (< 3 years) is interesting since previously there was no evidence for families staying within a single stretch of stream (Grant, 2007). It suggests that perhaps we should not be surprised at having sampled so many pairs of second- and third-degree relatives as well (Table 4.A.7). Since it appears to be common to pick up related individuals when sampling at the same location across several years, it is likely that previous population genetic studies on the platypus (Kolomyjec *et al.*, 2009; Gongora *et al.*, 2012; Furlan *et al.*, 2013) included relatives but did not detect these due to the small number of markers used.

Our population genetic analyses confirm the strong structure previously reported in the species (Gongora *et al.*, 2012; Furlan *et al.*, 2013), with the Tasmania-mainland differences being the most pronounced. Although the platypus is considered to be a single species across its whole range, there is some phenotypic variability between the different regions, which may have a genetic basis. For example, platypus in the south are substantially bigger than those in the north (Grant and Temple-Smith, 1983), and they tend to breed later in the year (Grant, 2007). One might speculate that at least part of the reason zoos have found it difficult to breed platypus in captivity is that they have been trying to cross individuals from different populations, and there are pre- or post-zygotic barriers to reproduction. Further analysis of our data may reveal regions of the genome which show particularly strong differences between populations, possibly due to selection, but establishing whether these are responsible

for phenotypic differences or reproductive incompatibilities is likely to be very challenging.

Consistent with Furlan *et al.* (2013), we saw no evidence for structure within the Tasmanian population, but this could simply be because of our small sample size (6 individuals). Furlan *et al.* also reported that there was little or no genetic differentiation in mtDNA and microsatellite markers between platypus either side of the Great Dividing Range in Victoria, and our STRUCTURE results (Figure 4.20), based on genome-wide SNP data, suggest that this is also the case in Queensland. We observed similarity between platypus from the Broken River, near the Queensland coast, and Carnarvon Creek, which is about 500km further inland to the southeast, on the other side of the Great Dividing Range, although within the same drainage basin (Figure 4.7).

We were surprised to note the similarities in the STRUCTURE results between the Wingecarribee/Fish River and North NSW samples, particularly the one from the Gwydir River, given that the platypus from the Shoalhaven River do not share the red component seen in Figure 4.20. The Wingecarribee individuals come from the Hawkesbury River basin which is directly adjacent to the Shoalhaven River basin, and several hundred kilometers further south of the Manning River basin which contains the Barnard River. These three river basins form part of the South East Coast drainage division, and flow towards the sea. In contrast, the Gwydir River flows west from the Great Dividing Range, and forms part of the Murray-Darling basin (Figure 4.7), as does the Fish River (in the Macquarie catchment), although it is geographically very close to the Wingecarribee River. Interestingly, Kolomyjec *et al.* (2009) also reported lower F_{ST} between the Hawkesbury and Gwydir platypus populations than between Hawkesbury and Shoalhaven (0.02 *versus* 0.05)¹⁹. Consistent with our findings, Gongora *et al.* (2012) noted that their mtDNA data suggested paraphyletic²⁰ relationships between individuals from some drainage divisions. There are several possible explanations for this. Although the Gwydir and Manning River basins are much

¹⁹These numbers are rather lower than ours: we estimated $F_{ST} = 0.085$ for North NSW *versus* Wingecarribee/Fish, 0.09 for Shoalhaven *versus* Wingecarribee/Fish, and 0.147 for North NSW *versus* Shoalhaven. This difference could be because we used a different F_{ST} estimator from Kolomyjec *et al.* (2009) or a different method of averaging F_{ST} over loci (Bhatia *et al.*, 2013), or because their estimates were based on just twelve microsatellite markers, which might have undergone repeat mutations since the populations split.

²⁰A paraphyletic group consists of all the descendants of the group's MRCA except for a small subset. A monophyletic group consists of an ancestral species/population and all its descendants.

too far from the Hawkesbury basin for platypus to have travelled directly overland, it is possible that, over many generations, they traversed progressively through the barriers between those rivers. Alternatively, it may be that individuals can easily cross overland from the west-flowing Macquarie catchment of the Murray-Darling basin into the east-flowing Hawkesbury catchment, thus connecting the latter with catchments in North NSW. There may also have been connections between those river systems in the past that have been altered due to climate or geological changes. The similarity between the sample from Rifle Creek, which is part of the west-flowing Mitchell River basin, and those from east-flowing Dirran Creek and Barron River, could possibly be explained by reasonably frequent overland travel, maybe even within a generation, since, although they are part of different river systems, the sampling locations are all within 100km and the climate is much wetter.

The PSMC results provide a glimpse of the ancient demographic histories of the different populations. One might be concerned that the short length of the scaffolds and any misplaced contigs would somehow affect these results by altering the apparent length of a block of heterozygous SNPs at a particular density. However, setting different cutoffs for the scaffold length did not seem to substantially change the results (Figure 4.A.21). Even if it did, this is likely to affect all individuals in the same way, and so comparisons between them would still be possible. One might anticipate that samples with lower coverage would have more false negative SNP calls, which would have the same effect as a lower mutation rate (Figure 4.A.16). Indeed, due to the vagaries of the sequencing tranches, the Shoalhaven samples tended to have lower coverage than the North QLD samples (Figure 4.8). However, since the Shoalhaven population was still estimated to have about twice the nucleotide diversity of the North QLD population (Table 4.8), and since we effectively accounted for callability by setting sites with low coverage across all samples or with low genotype quality in a particular sample (which reflects the coverage) to missing, we suspect this difference in coverage had a minimal impact. Additionally, it is encouraging to see that individuals from the same river had such similar N_e estimates, since this suggests that any batch effects resulting from the samples having been sequenced in different tranches (Tables 4.3 and 4.A.2) is fairly minimal.

PSMC revealed a strong bottleneck in the samples from Rifle Creek, Dirran Creek and the Barron River (Figure 4.A.19), consistent with their low nucleotide diversity, followed by a more recent increase in N_e . Other species in tropical QLD went through

a bottleneck during the Last Glacial Maximum (LGT) 19-26.5KYA, when the climate was particularly dry (Oza *et al.*, 2012). Uncertainty in the mutation rate and generation time make it difficult to say whether we are observing the same event, but a rate similar to the stringent *de novo* mutation rate we estimated in the quartet would suggest we are. The Carnarvon and Broken River samples from Central QLD also appear to have been through a period of population decline, from which they did not recover, while the Running River sample is intermediate between them and the other North QLD samples. This likely reflects the historical and current paucity of suitable habitat for platypus between North (tropical) and Central QLD. In water-dependent animals, the hot and dry “Burdekin gap” (named for the Burdekin River) is known to act as a barrier to genetic exchange (James and Moritz, 2000; Schauble and Moritz, 2001). Indeed, Carnarvon Gorge is an oasis in the semi-arid heart of Central QLD, with a climate considerable wetter and colder than the surrounding terrain, making it the only suitable habitat for platypus for hundreds of kilometres, consistent with the low recent N_e . In contrast to the QLD samples, the NSW individuals appeared to show relatively little fluctuation in N_e over the time period accessible to the method, suggesting that the climate and water availability in those areas has been favourable over this time.

Although having the same N_e at a particular timepoint does not necessarily mean that two individuals were in the same ancestral population, PSMC results can be used used to infer split times of different populations by noting where the N_e estimates start to diverge (Li and Durbin, 2011; Nadachowska-Brzyska *et al.*, 2013; Thomas *et al.*, 2015). It is a little difficult to determine this in Figure 4.21, but one might tentatively infer that the Tasmania-mainland split occurred between 0.8 and 2MYA (with the μ and g used in this figure), which overlaps with the mtDNA-based estimate of Gongora *et al.* (2012). It is unclear from Figure 4.21 whether the Rifle Creek/Dirran Creek/Broken River population diverged from the others at the same time as the Tasmania population, or slightly later, around 300,000 years ago, but it seems that the Broken River/Running River/Carnarvon Creek population split off from the NSW populations after that, about 200,000 years ago. These estimates are not consistent with the topology of the tree inferred from mtDNA by Gongora *et al.* (2012) (Figure 4.4). This may be because the tree is genuinely different for the mitochondrial *versus* the nuclear genome, because the mtDNA tree was incorrect, or because our interpretation of the N_e estimates is incorrect, and different populations just happened to have had similar N_e at the same time, although they had already

diverged. Further reference genome improvements will hopefully allow us to phase the SNP calls and run methods such as MSMC (Schiffels and Durbin, 2014), the extension of PSMC for multiple sequences, to estimate more recent N_e and divergence times between river basins.

Perhaps the most novel finding in this project so far is the presence of what seem to be relatively recent strata on the sex chromosomes. These were not detected by Cortez *et al.* (2014) because of their approach of assembling reads that did not map to the female reference. In contrast, we detected the YHRs because the X-Y divergence is sufficiently low in these regions that the chrY reads still map to the reference X copy. Some of the genes Cortez *et al.* reported do indeed fall within YHRs we discovered: *DAZAP1* from Stratum 4 (Scaffold 17598:7,055-33,182), *PRPF4B* and *SSR1* from Stratum 5 (Scaffold 3,642:41547-65,008 and Scaffold 2552:219,234-229,751 respectively), and *EDF1* and *SDHA* from Stratum 6 (Scaffold 2833:31,733-42,172 and Scaffold 179:575,140-597,510 respectively). However, the YHRs on Scaffold 0 and at Scaffold 179:1,336,872-1,350,588, have levels of divergence between 3- and 8.5-times lower than that of the youngest stratum (S6) reported by Cortez *et al.* (see Tables 4.1 and 5.A.9). Further work will include trying to date these strata. It is possible the youngest are the result of a recent transposition from an X to a Y chromosome, like the X-transposed region on the human male-specific Y, which was transposed about 3-4 million years ago and has retained 99% identity with the X (Skaletsky *et al.*, 2003). The genes reported by Cortez *et al.* show levels of X-Y divergence similar to the genes in the X-degenerate region on the human Y, which have between 60% and 96% nucleotide sequence identity with their X gametologs (Skaletsky *et al.*, 2003).

Homology with the chicken and human genomes allowed us to try to assign scaffolds containing YHRs to particular platypus chromosomes. Comparing levels of X-Y divergence provided clues as to which YHRs form part of the same stratum. ENSOANG00000001201, a gene on Scaffold 481, which may possibly be in Stratum 5 from Cortez *et al.*, along with Scaffolds 3642 and 2552, encodes an uncharacterised protein that is orthologous to four chromodomain proteins within palindromic regions on the human Y (*CDY1*, *CDY1B*, *CDY2A*, and *CDY2B*). There is some evidence that deletion of *CDY1* and *CDY1B* is associated with spermatogenic failure (Choi *et al.*, 2012; Ghorbel *et al.*, 2014; Sen *et al.*, 2015). It is tempting to speculate that the Y gametolog of ENSOANG00000001201 might also be important for sperm production

in the platypus.

The regions of X-Y divergence we have found are presumably due to recombination suppression, but the cause of this is unknown. It is likely to be due to inversions, but we cannot identify these without good assemblies of the X and Y chromosomes. We have attempted to assemble the Y chromosomes using unmapped male reads, and managed to reconstruct most of the genes reported by Cortez *et al.*²¹, but the assembled scaffolds were mainly very short. Y chromosomes have been challenging to assemble in other mammals due to the palindromic regions and high repeat content (Skaletsky *et al.*, 2003; Hughes *et al.*, 2010, 2012; Soh *et al.*, 2014; Li *et al.*, 2013a). This may well turn out to be the case for platypus, particularly for the older Ys. We hope that long PacBio reads will help, but we still anticipate difficulties. Assembling the platypus X and Y chromosomes together is likely to be difficult in regions of high homology. An alternative, or parallel approach, might be to try mapping the male reads to the female reference using Stampy with a higher substitution rate parameter, to identify regions of higher sequence divergence than the YHRs we have already detected²², and then assembling these reads alone. We have also considered flow-sorting the Y chromosomes in order to sequence them individually, but a recent attempt to do this for the gorilla and orangutan Y chromosomes has produced only very fragmented assemblies (Marta Tomaszekiewicz, personal communication).

Recombination suppression might be possible even in the absence of inversions if there is strong sexually antagonistic (SA) selection acting, that is, selection that works in opposite directions in males and females (Otto *et al.*, 2011). SA polymorphisms tend to accumulate in the PAR, close to the boundary with the X-specific or Y-specific region, since the partial sex linkage allows the frequencies in males and females to diverge, and this favours further suppression of recombination. It is tempting to speculate that the region on Scaffold 0, where one North QLD male appears to have inherited the product of a recent recombination event between the X and the Y (Figure 4.25), might have diverged between them as a result of SA selection, rather than an inversion (although see next paragraph). We do see many SNPs with nominally significant sex differences in allele frequency ($p < 0.05$ in Fisher's exact test) along the whole ~ 4.5 Mb PAR on Scaffold 0, which is a characteristic signature of

²¹Data not shown. This was done by Julie Hussin.

²²We noticed that shorter YHRs tended to have higher levels of X-Y divergence, probably because the reads either side are too diverged to align.

balancing selection due to sexual antagonism (Otto *et al.*, 2011). However, such a signature can also arise very close to the PAR boundary simply due to the close sex linkage, in the absence of SA selection. Kirkpatrick and Guerrero (2014) showed that these associations with sex disappear within a population recombination rate value of $\rho < 5$ in the absence of sexual antagonism. Future work on recombination in our data should allow us to determine to what physical distance this corresponds. However, the observed sex differences in allele frequency may simply be an artefact of having insufficiently accounted for population structure²³, and we should also be conscious of multiple testing.

There are several possible mechanisms that could have produced the polymorphic PAR boundary on Scaffold 0 (Figure 4.24). If the putative young stratum in that region formed due to an inversion, recombination may still have been possible through an inversion loop (Boyd *et al.*, 2005; Torgasheva *et al.*, 2013). It is likely that the level of divergence between the X and Y copies in the region ~ 4.5 -5.1Mb is sufficiently low (~ 1 per 120bp, on average) that crossovers can still occur in the absence of an inversion, since recombination happens normally in hybrids between mouse subspecies with about 1% divergence (Paigen *et al.*, 2008; Keane *et al.*, 2011). It may not be coincidental that the sequence identity in the region ~ 4.5 -5.1Mb is higher than in the region from ~ 5.1 Mb onwards, which has 1 X-Y divergence per ~ 50 bp. It is not known what level of sequence identity is required for successful recombination between homologous chromosomes, but in mouse hybridoma cells, at least 500-1000bp of homology seem to be needed (Mundia *et al.*, 2014). There is an instance of the PAR being extended in humans through an insertional translocation of X chromosomal sequence onto the Y, mediated by nonallelic homologous recombination between flanking long terminal repeat elements (Mensah *et al.*, 2014). However, this produced a duplication of a region on the X, which is not observed in N702. A polymorphic PAR boundary has been noted in the papaya (Lappin *et al.*, 2015) (16% of wild individuals surveyed had a 100-130kb expansion of the non-recombining region), and in *Silene latifolia* (Qiu *et al.*, 2015), both of which have young sex chromosomes. In neither case is the genome well characterised enough to speculate as to the process that formed these polymorphic PARs.

²³To try to account for this, we simply sampled equal numbers of unrelated males and females from each of the five population groups.

The most frequently asked and probably most important question about the platypus sex chromosomes is what advantage might be conferred by this chain of ten. Indeed, one might expect this complicated system to have more room for error. However, the segregation seems to be well controlled, since, out of 284 sperm, Grutzner *et al.* (2004) did not observe any with both X and Y chromosomes. Possibly relevant are the findings of Dohm *et al.* (2007), who reported that parts of the major histocompatibility complex (MHC) are in the PARs on X3/Y3 and Y4/X5. It might be advantageous for these genes to be in a PAR if it has higher diversity (expected if there is higher recombination (Begun and Aquadro, 1992; Spencer *et al.*, 2006)), or if they are closely linked to the sex-specific region and are therefore more likely to be heterozygous in males, conferring immunity to more pathogens. However, it seems highly improbable that this would be a strong enough force to drive the emergence of a whole chain of chromosomes with multiple PARs, several of which are so long they probably do not have an increased recombination rate conferring higher diversity anyway. An alternative explanation could be mechanistic; perhaps some molecular process in the monotreme ancestor facilitated multiple reciprocal translocations between chromosomes, and there was not sufficiently strong selection against this. Improving the genome assembly and annotating the genes on the sex chromosomes may provide further clues as to how the chain formed, where the strata are located, and what the advantage of this complicated system may be.

Appendix

4.A Appendices

4.A.1 Supplementary Figures

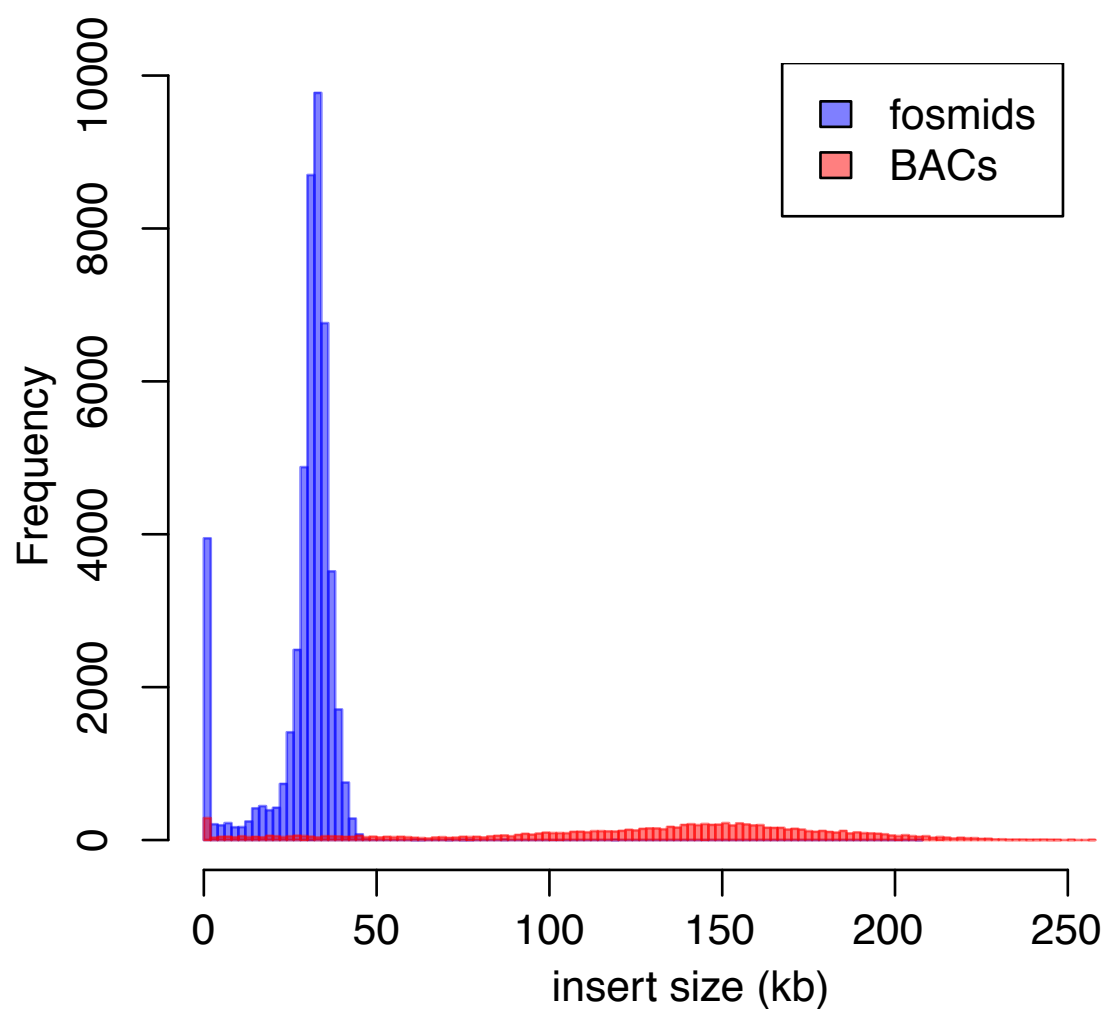


Figure 4.A.1: **Distribution of insert sizes for BAC and fosmid libraries.** This was determined by mapping the BAC and fosmid ends to ornAna1 using bwa mem (Li, 2013), and taking the pairs of which both reads mapped to the same scaffold.

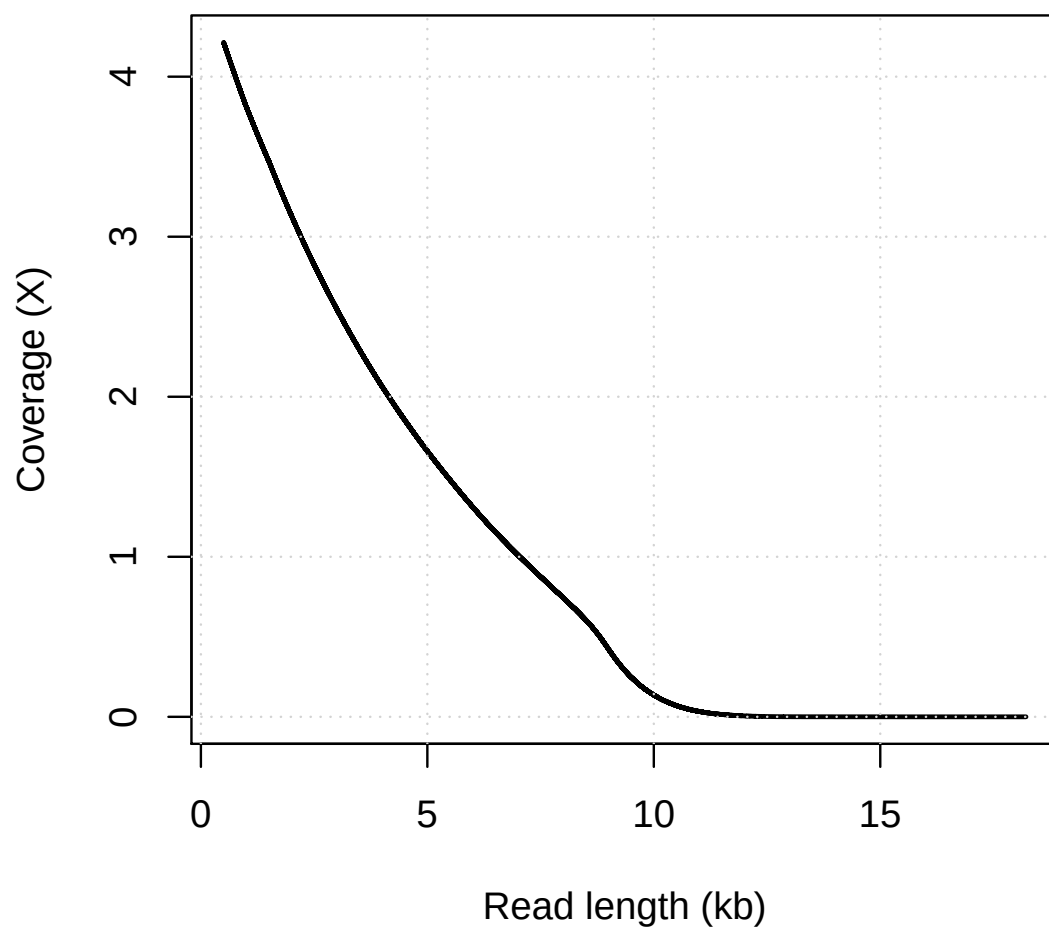


Figure 4.A.2: **Plot showing the distribution of coverage by Molecule reads.** The Y-axis indicates the coverage of the genome in reads over the length specified on the X-axis.

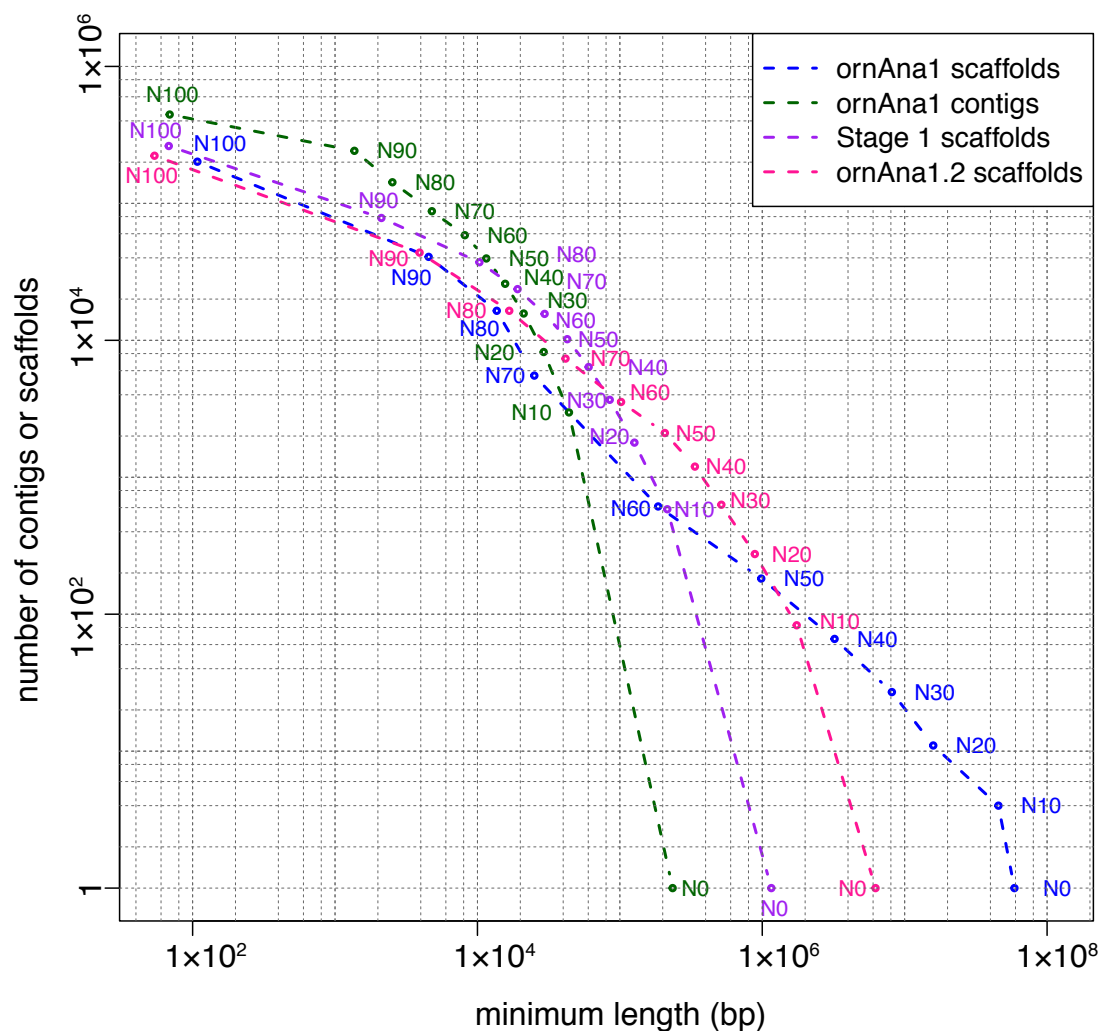


Figure 4.A.3: **Plot showing distribution of lengths of contigs and scaffolds from the original and new assemblies.** The labels on the points indicate NX values, which represent the length of the shortest scaffold (or contig) such that the total length of all scaffolds (or contigs) larger than that comprises X% of the genome. The X-axis indicates the NX length and the Y-axis indicates the number of scaffolds with length larger than that. The distributions for the contigs and scaffolds in the original assembly, ornAna1, are shown, along with those for the Stage 1 scaffolds and final scaffolds in the new assembly, ornAna1.2 (Stage 3 scaffolds). Stage 2 scaffolds are not shown because their length distribution was almost identical to that of the final ornAna1.2 scaffolds.

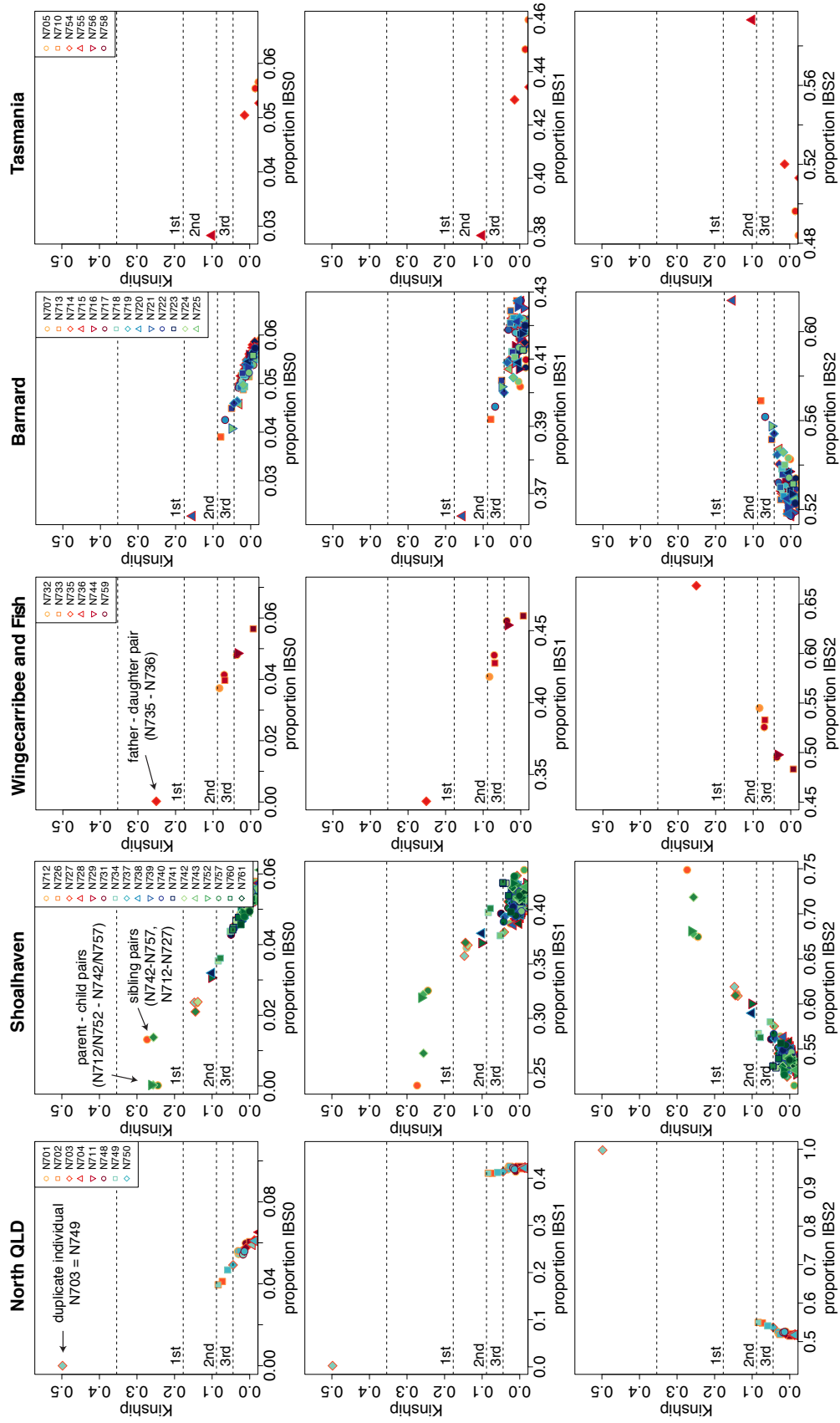


Figure 4.A.4: **Kinship estimates from KING.** These plots show the kinship coefficients estimated with KING *versus* the proportion of SNPs with identity-by-state (IBS) 0, 1 or 2. The legends indicate the colour and shape corresponding the point from each individual. The border colour and shape indicates the first individual in a pair, and the fill colour indicates the second individual. The estimates are also displayed in Table 4.A.7. The dotted lines and labels (1st, 2nd, 3rd) indicate the ranges of kinship coefficients which correspond to 1st, 2nd or 3rd degree relatives. Pairs of 1st degree relatives and a duplicate sample are indicated.

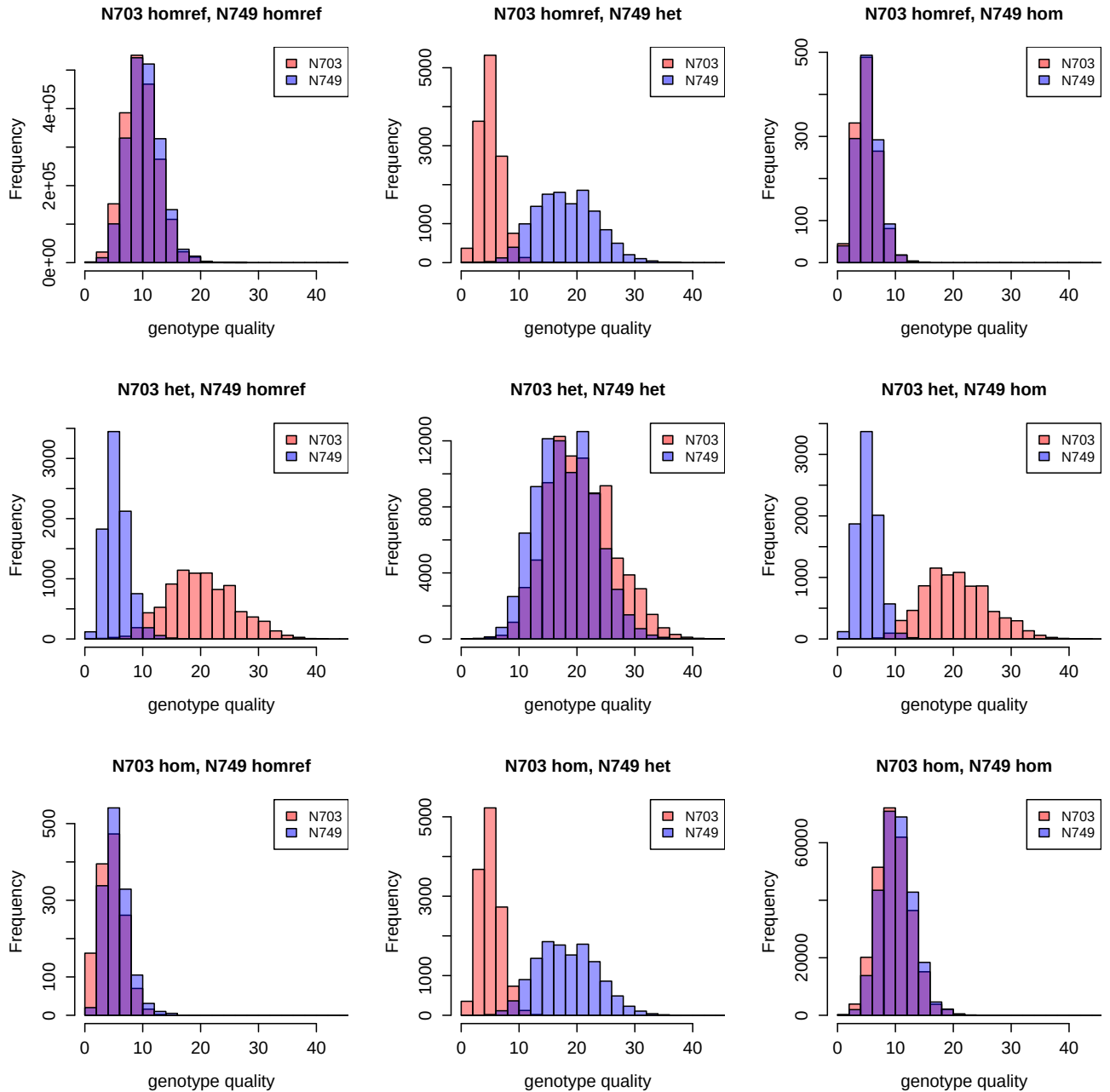


Figure 4.A.5: **Distribution of genotype qualities for CORTEX PASS SNPs in the duplicate samples.** Each histogram shows the distribution of genotype qualities (GQs) for SNPs at which N703 and N749 had the genotypes specified in the plot title. (homref: homozygous for the reference allele; het: heterozygous; hom: homozygous for the alternate allele.) The plots along the diagonal are concordant genotypes, and those off the diagonal are discordant. Note that hom and homalt genotypes primarily have GQ less than 20, with a median of about 10, whereas het genotypes typically have GQ between 10 and 35, with a median of about 20. Thus, although setting a GQ threshold of 10 would eliminate most of the discordant genotypes, it would also eliminate about half of all hom or homref calls.

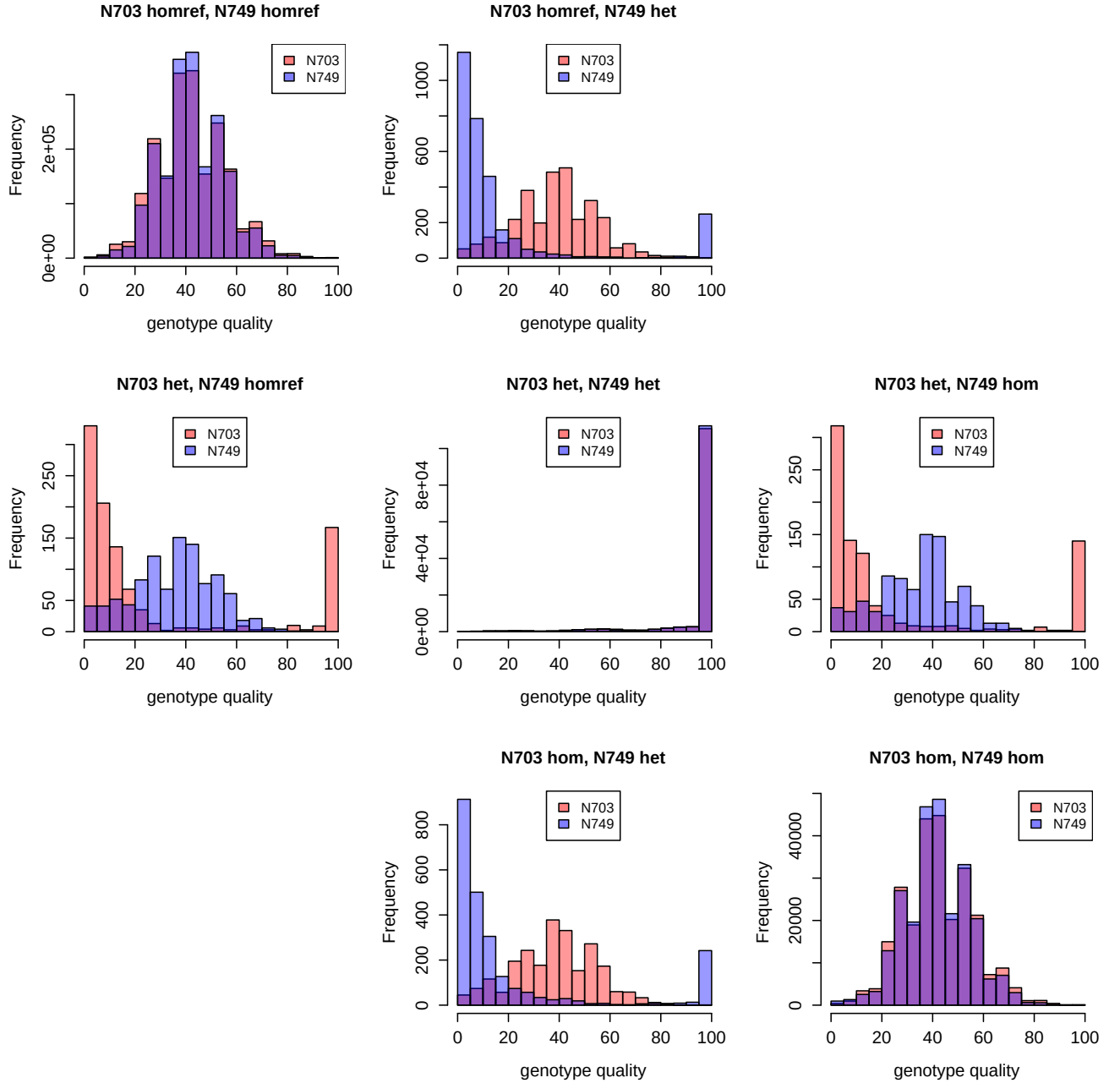


Figure 4.A.6: **Distribution of genotype qualities for GATK/CORTEX intersection SNPs in the duplicate samples.** As for Figure 4.A.5, but for the GATK/CORTEX intersection callset. Blank spaces have been left if there were no SNPs in a particular genotype category, to facilitate comparison with the other figures. Genotype qualities (GQs) of 100 or more are set to 99 by default. Note that imposing a GQ threshold of 20 would eliminate a substantial proportion of the discordant genotypes without having an excessively biased effect on the number of hom or homalt calls remaining relative to het calls, in contrast to CORTEX.

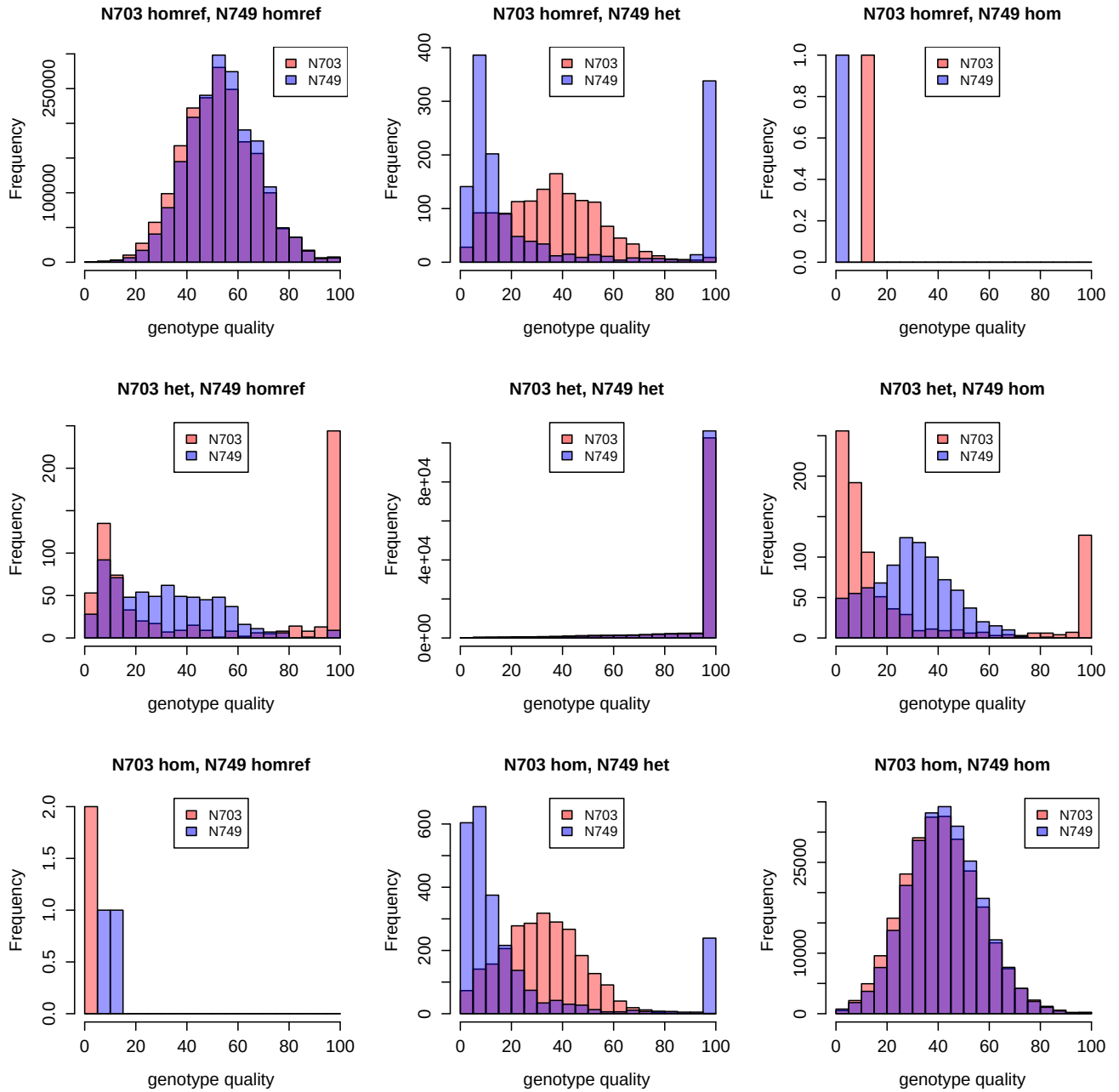


Figure 4.A.7: Distribution of genotype qualities for Platypus/CORTEX intersection SNPs in the duplicate samples. See legend to Figure 4.A.6.

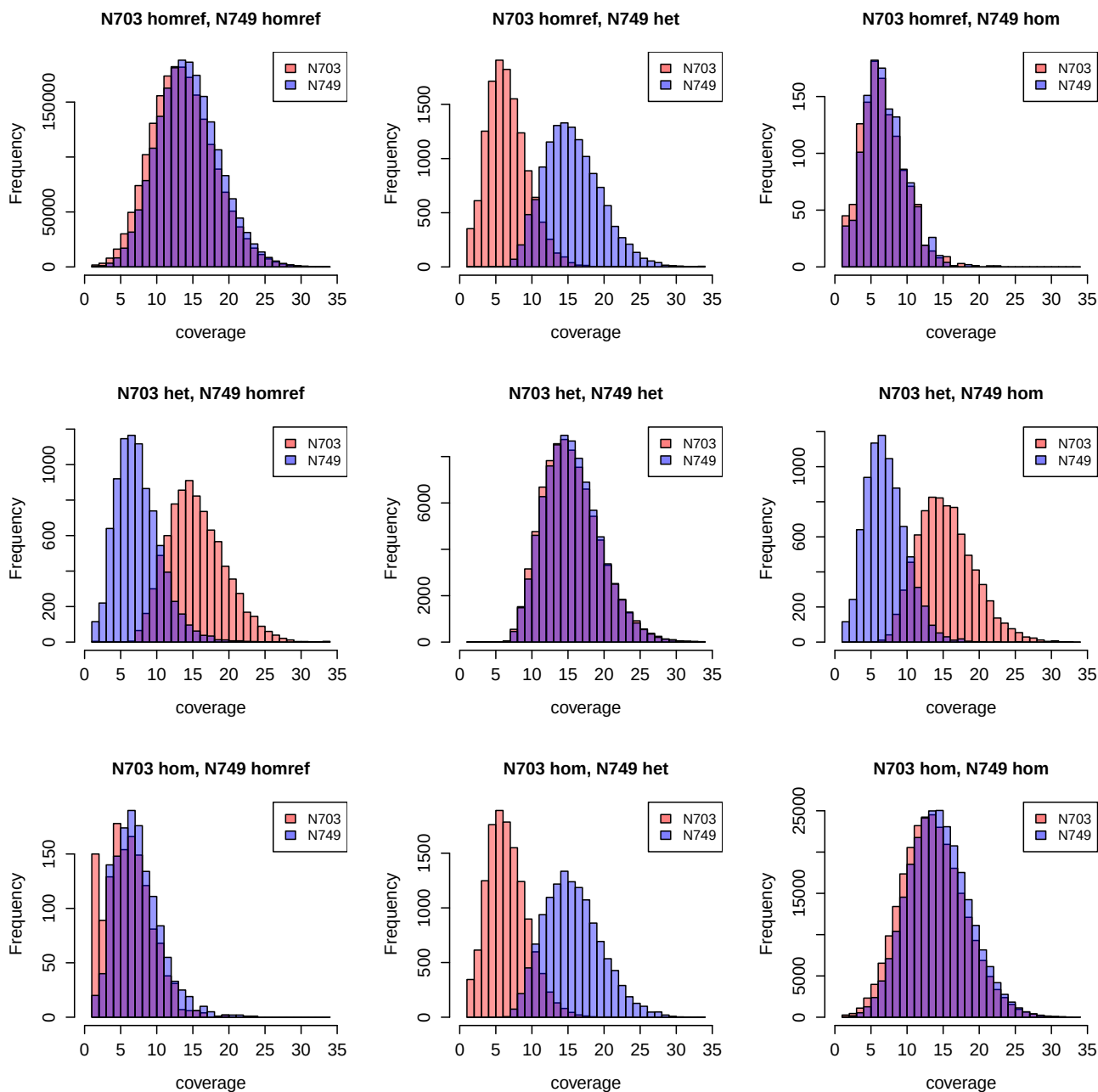


Figure 4.A.8: Distribution of coverage at CORTEX PASS SNPs in the duplicate samples. See legend to Figure 4.A.5.

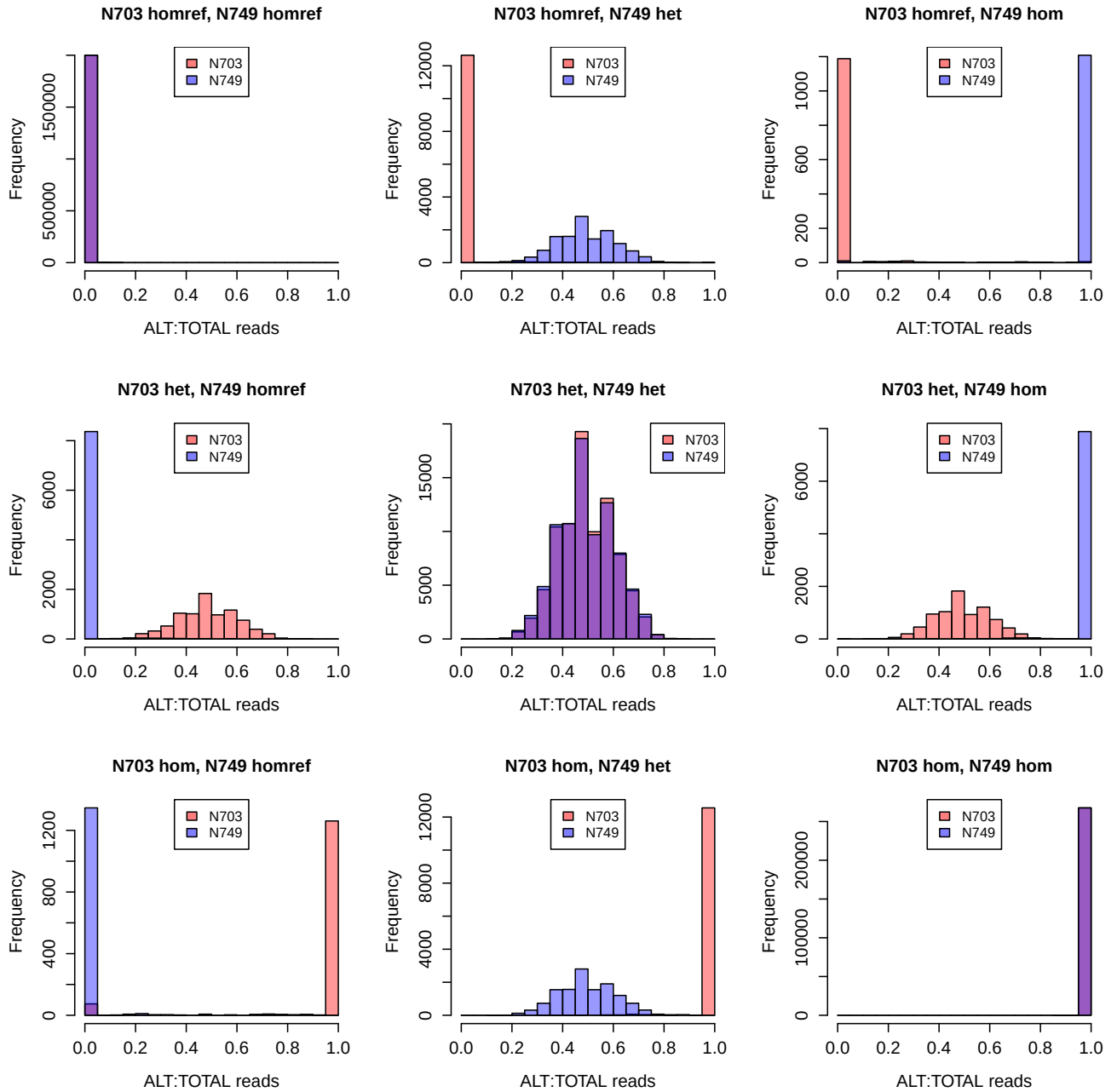


Figure 4.A.9: **Distribution of the fraction of reads carrying the alternate allele at CORTEX PASS SNPs in the duplicate samples.** See legend to Figure 4.A.5. ALT:TOTAL ratio - the ratio of the number of reads carrying the alternate allele to the total number of reads.

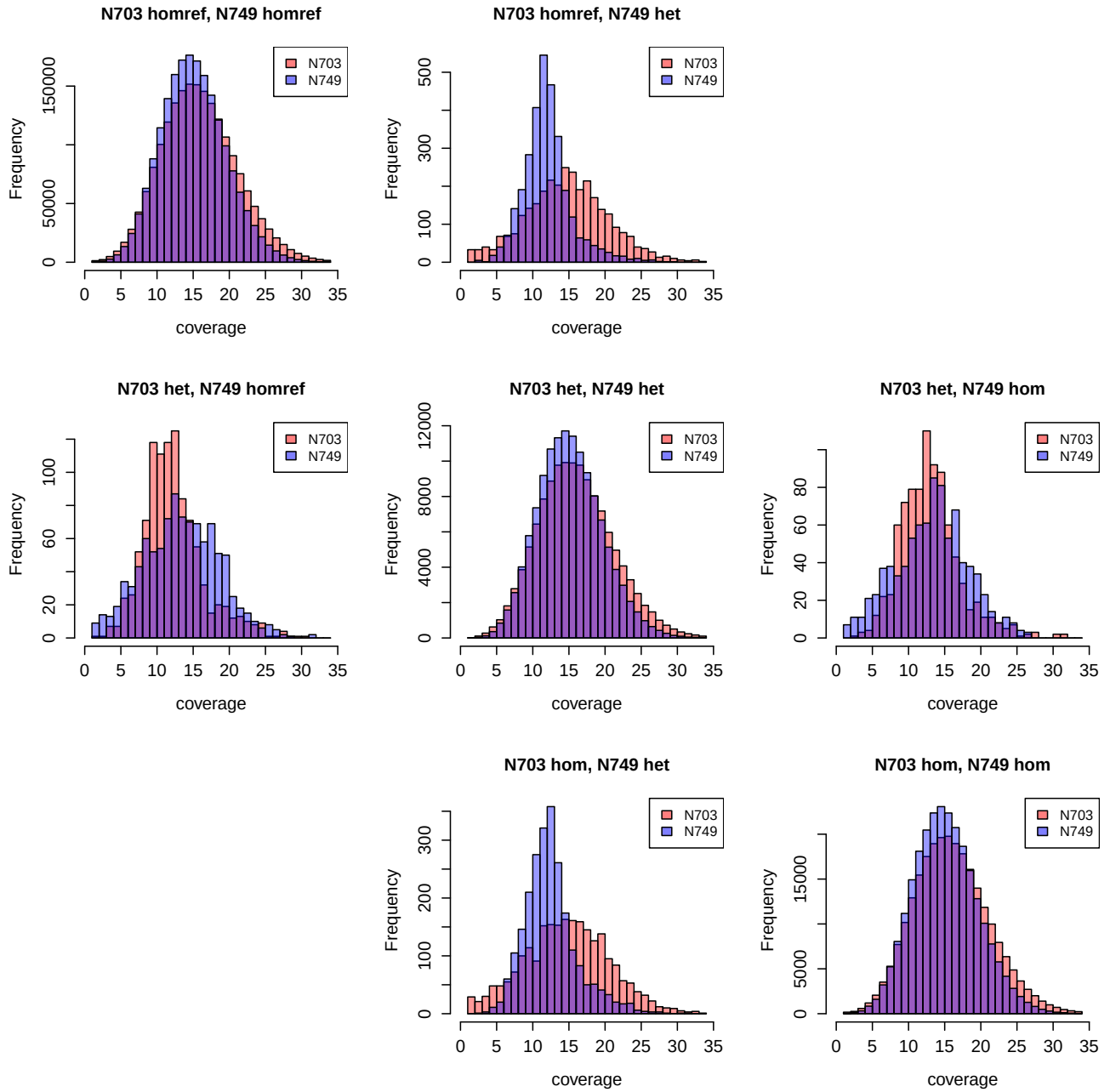


Figure 4.A.10: **Distribution of coverage at GATK/CORTEX intersection SNPs in the duplicate samples.** See legend to Figure 4.A.8. Blank spaces have been left if there were no SNPs in a particular genotype category, to facilitate comparison with the other figures.

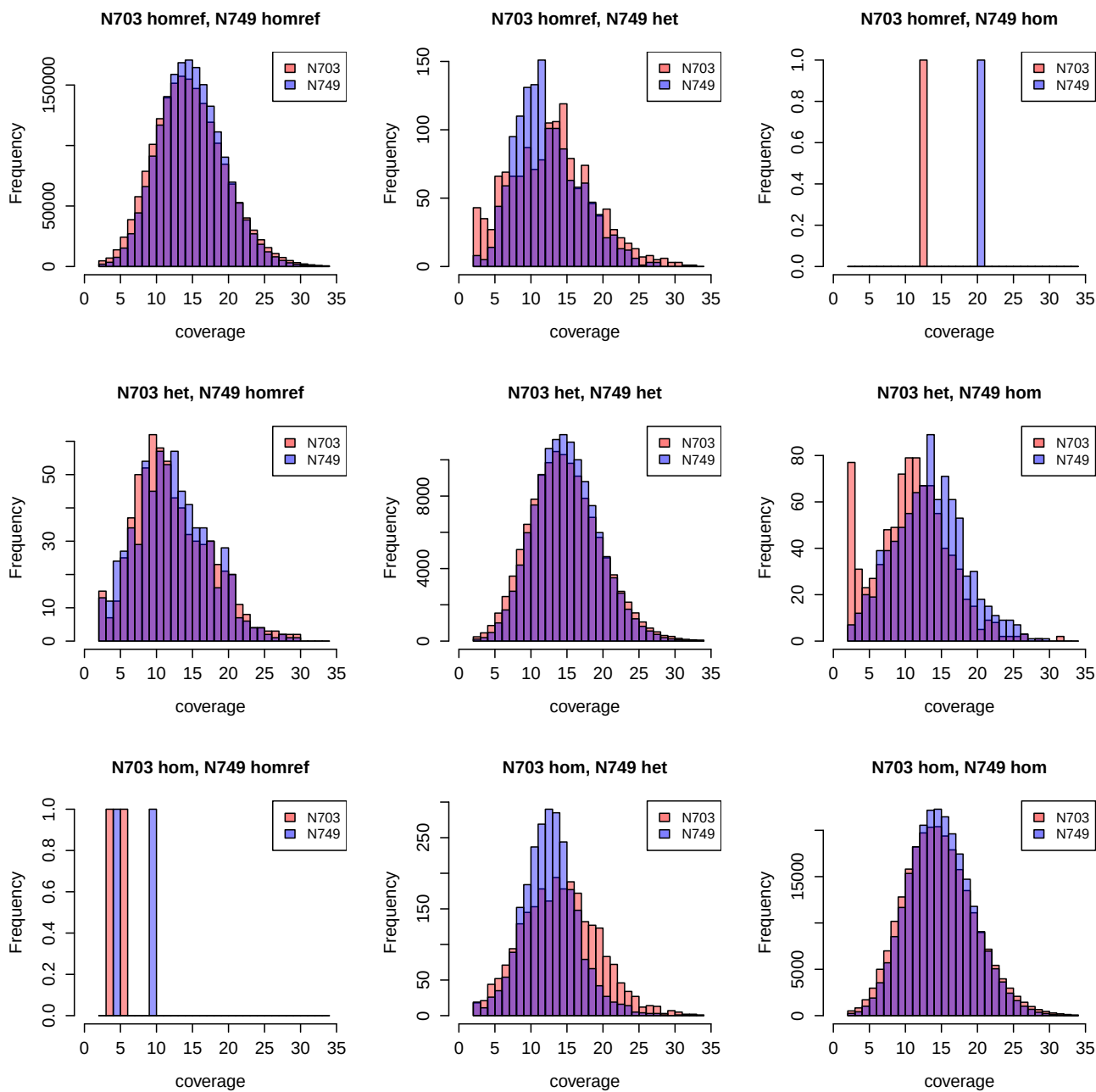


Figure 4.A.11: Distribution of coverage at Platypus/CORTEX intersection SNPs in the duplicate samples. See legend to Figure 4.A.8.

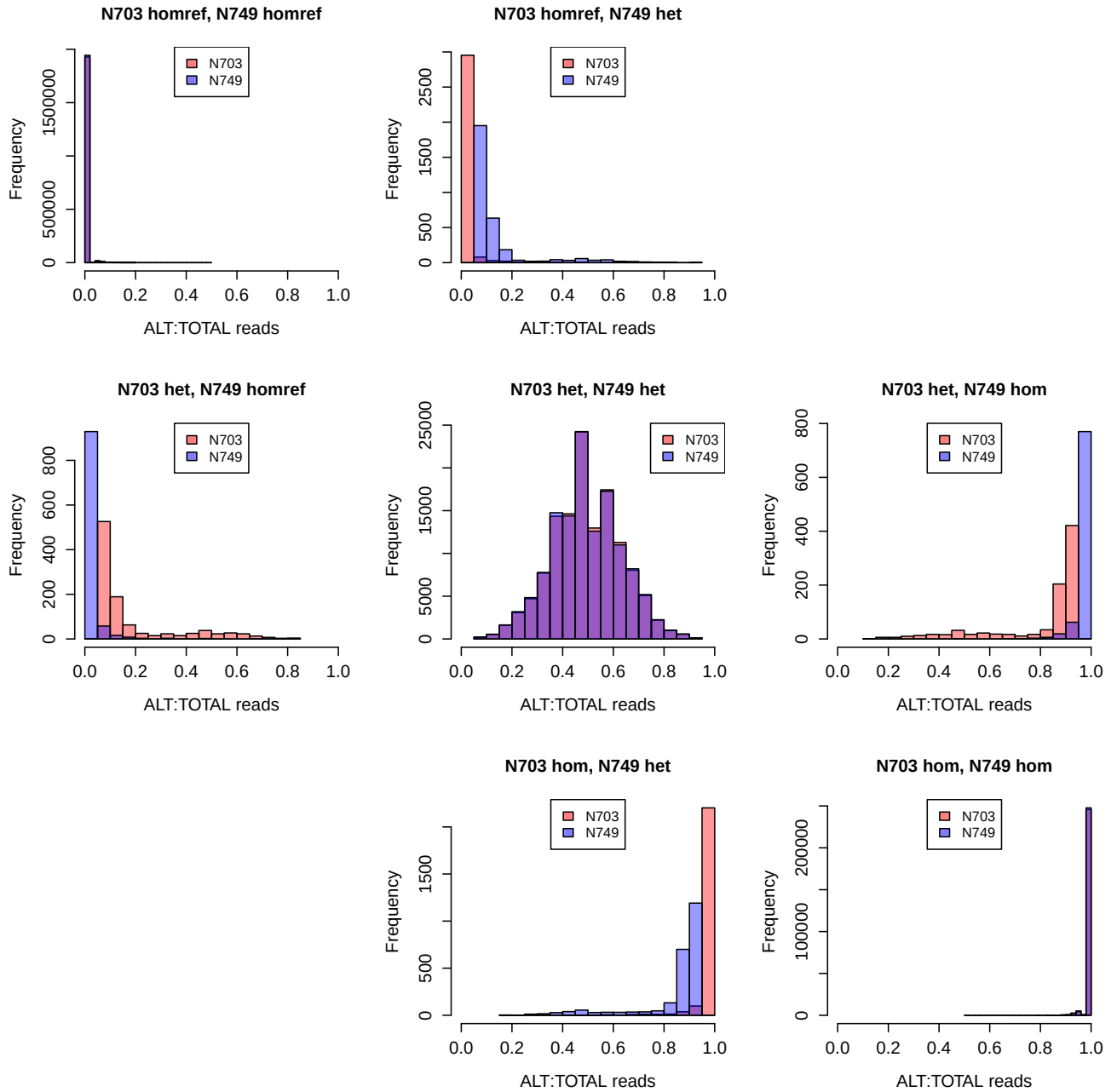


Figure 4.A.12: **Distribution of the fraction of reads carrying the alternate allele at GATK/CORTEX intersection SNPs in the duplicate samples.** See legend to Figure 4.A.9. Note that blank spaces have been left if there were no SNPs in a particular genotype category, to facilitate comparison with the other figures.

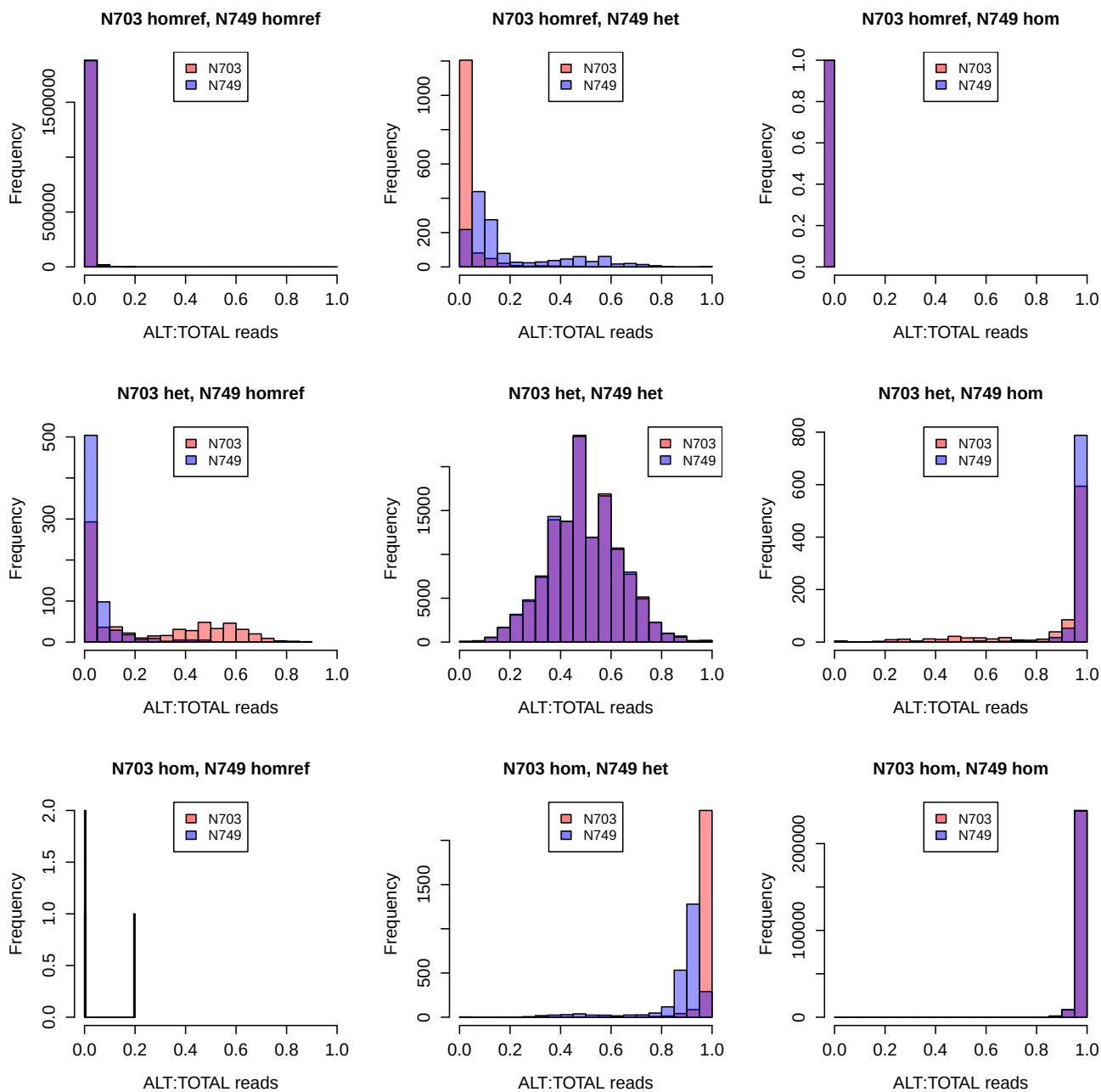


Figure 4.A.13: Distribution of the fraction of reads carrying the alternate allele at Platypus/CORTEX intersection SNPs in the duplicate samples. See legend to Figure 4.A.9.

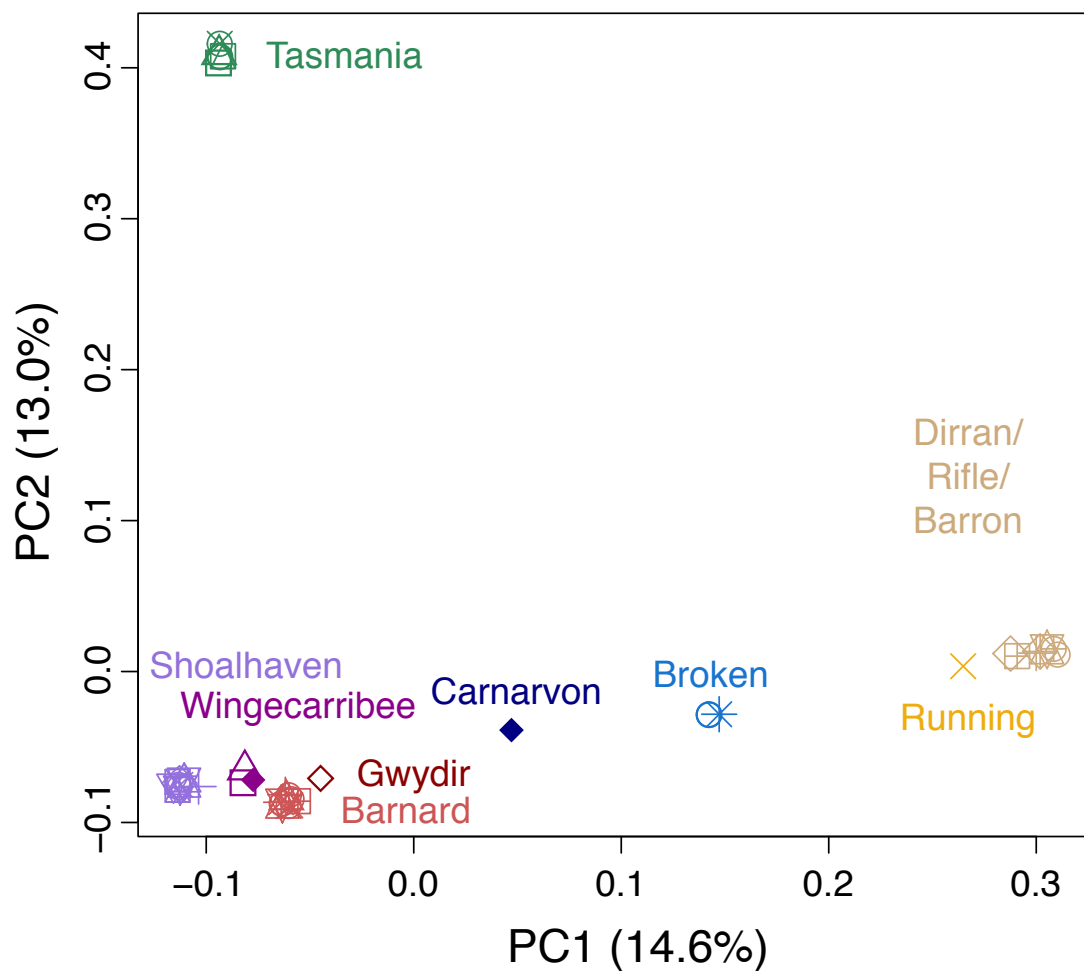


Figure 4.A.14: **Principal component analysis using LD-pruned SNPs.** As for Figure 4.18 but with the SNPs pruned based on LD, as described in Section 4.2.6.

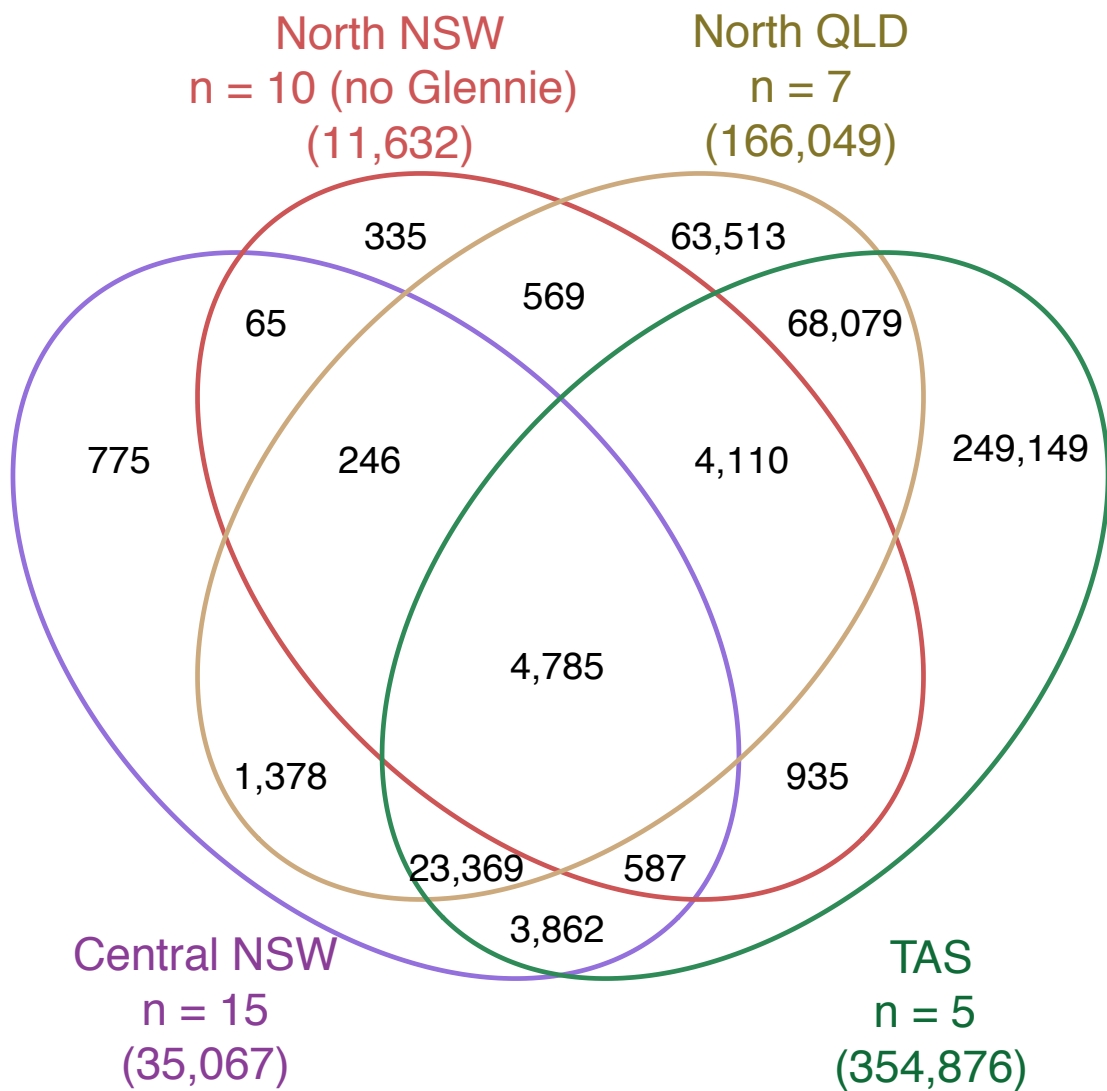


Figure 4.A.15: Venn diagram showing number of fixed differences from the reference in different populations. The numbers in parentheses are the totals for that population, out of the ~2.3 million SNPs segregating in the 58 samples.

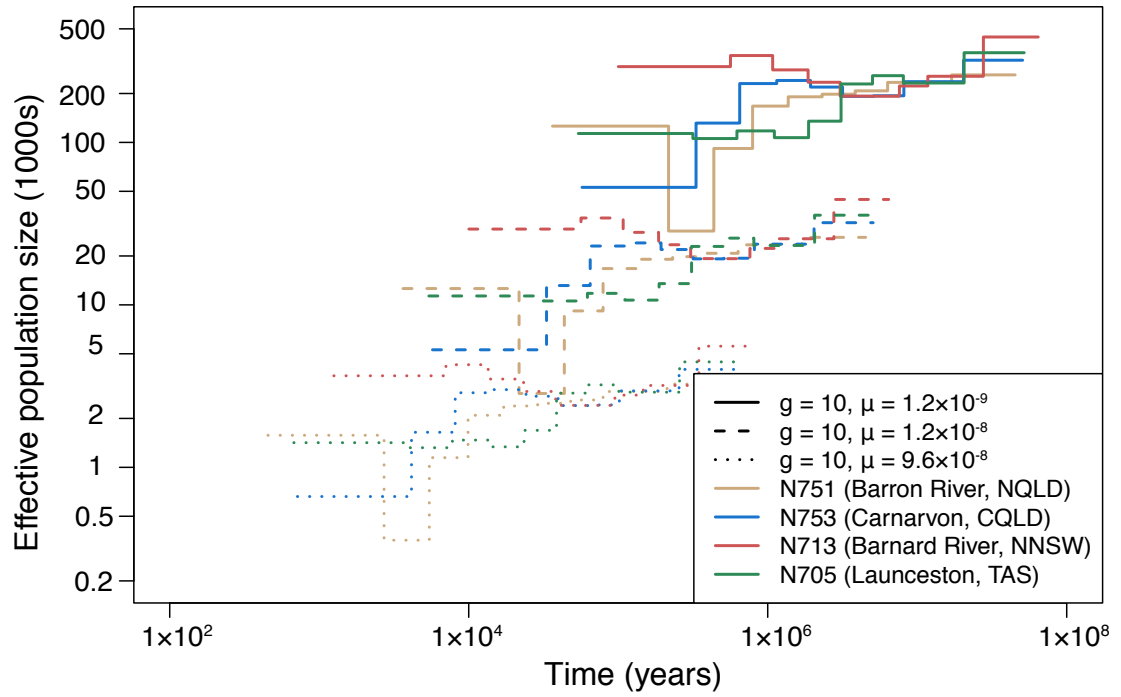


Figure 4.A.16: **Effect of mutation rate on PSMC results.** This plot illustrates how the choice of μ shifts the estimates of N_e along the X and Y axes. For other plots and for the dashed lines here, we used $\mu = 1.2 \times 10^{-8}$ /bp/generation, a conservative estimate of the *de novo* mutation rate from the platypus quartet that also happens to be the same as the estimate given by Scally and Durbin (2012) for humans. The lower value of 1.2×10^{-9} /bp/generation (solid lines) was chosen arbitrarily for comparison, and the higher value of 9.6×10^{-8} /bp/generation (dotted lines) comes from the less conservative GATK PASS calls with GQ filter in the platypus quartet (see Section 4.3.2).

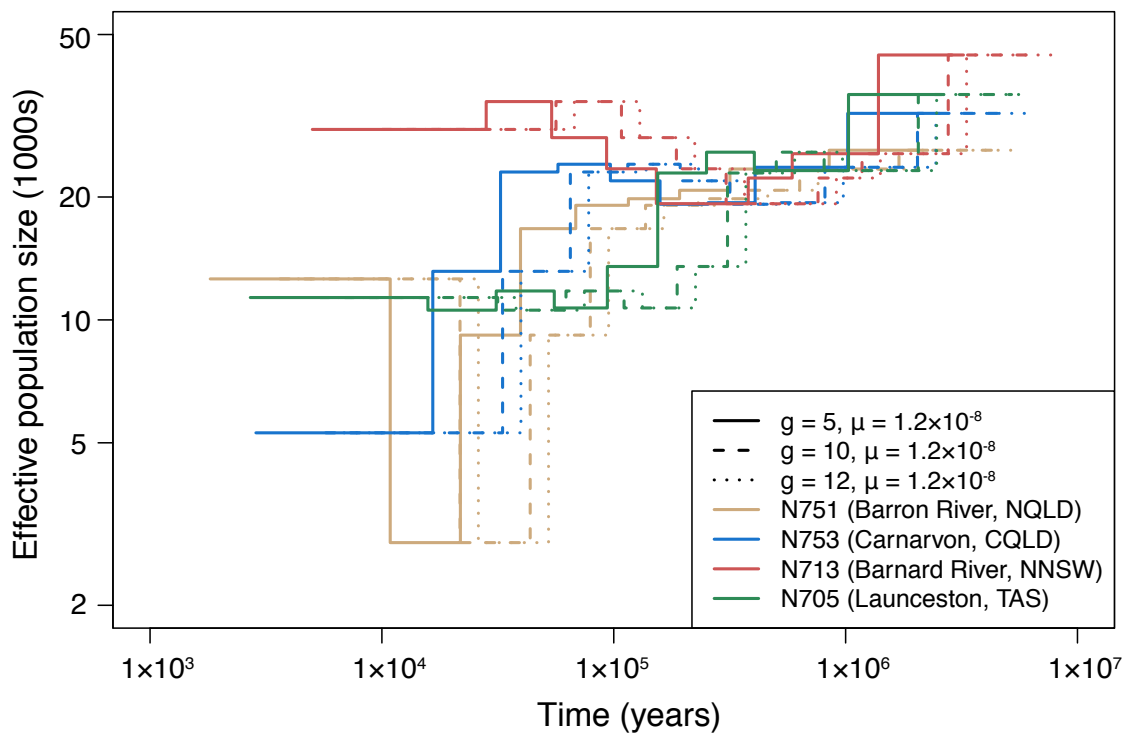


Figure 4.A.17: **Effect of generation time on PSMC results.** This plot illustrates how the choice of g shifts the estimates of N_e along the X axis. Furlan *et al.* (2012) used $g = 10$, and we also show $g = 5$ and $g = 12$ here for comparison.

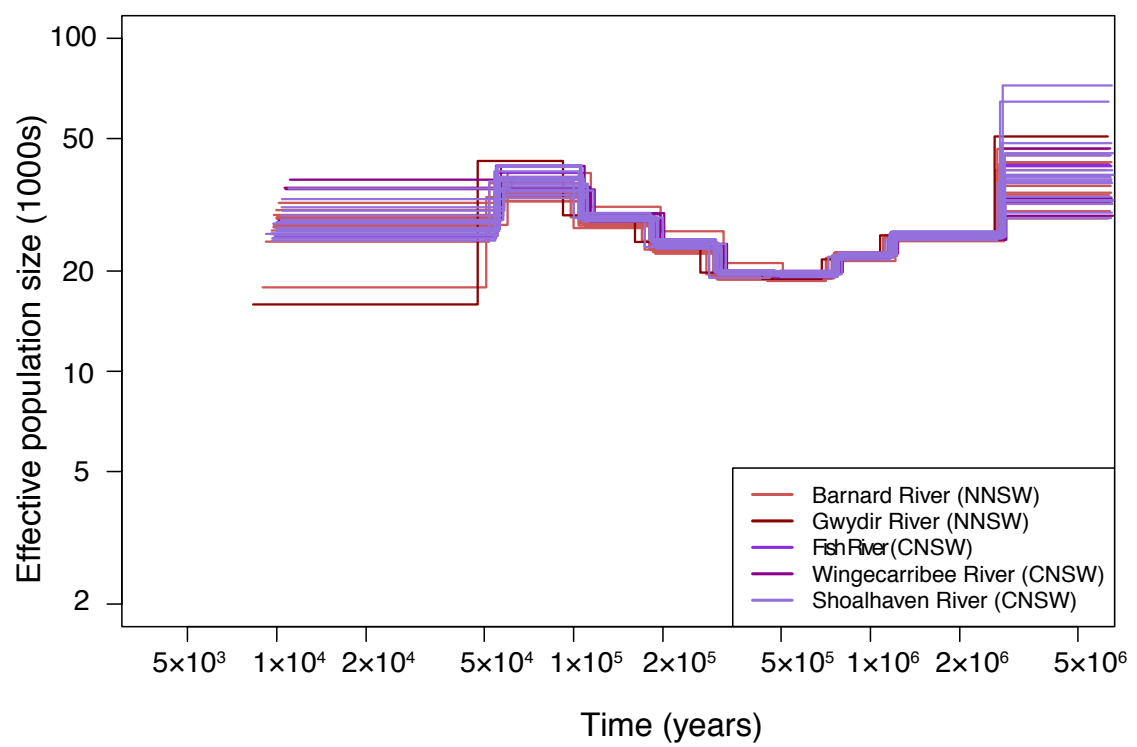


Figure 4.A.18: **Historical effective population sizes inferred from New South Wales samples using PSMC.** These were scaled using $g = 10$ and $\mu = 1.2 \times 10^{-8}$.

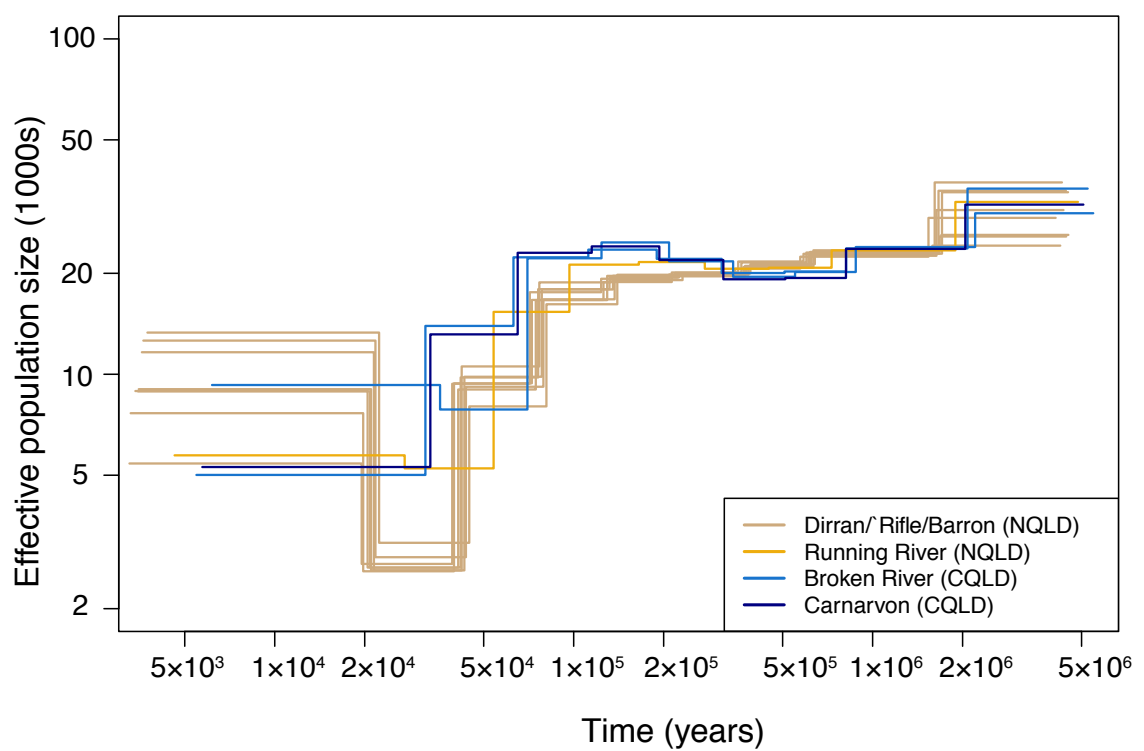


Figure 4.A.19: **Historical effective population sizes inferred from Queensland samples using PSMC.** These were scaled using $g = 10$ and $\mu = 1.2 \times 10^{-8}$.

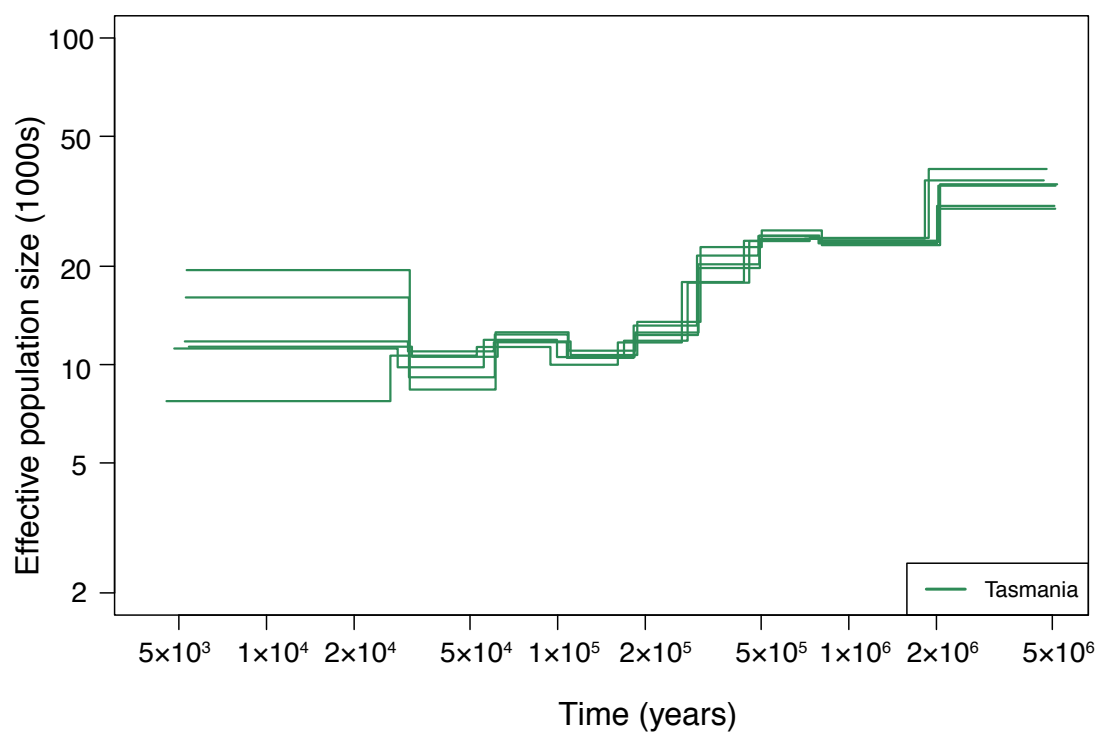


Figure 4.A.20: **Historical effective population sizes inferred from Tasmanian samples using PSMC.** These were scaled using $g = 10$ and $\mu = 1.2 \times 10^{-8}$.

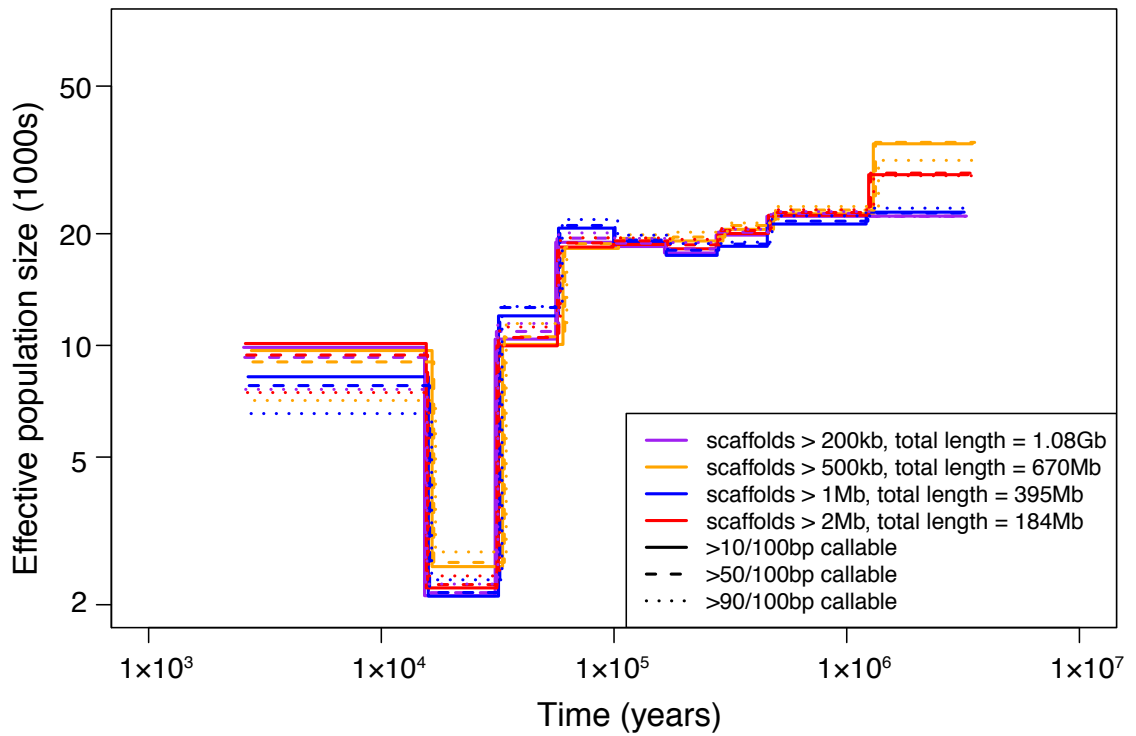


Figure 4.A.21: **Historical effective population sizes inferred from sample N702 using PSMC, with different subsets of the data.** This illustrates the effect of setting different minimum length thresholds for the scaffolds, and different thresholds for the number of callable bases required in a 100bp bin in order to consider it. The N_e estimates were scaled using $g = 10$ and $\mu = 1.2 \times 10^{-8}$.

4.A.2 Supplementary Tables

Table 4.A.1: **Comparison of repeat content between platypus and other species.** These statistics are taken from RepeatMasker tracks. Platypus have the highest number of individual repeats per kb, although other species have overall higher repeat content.

Species	Genome size (Gb)	Repeat density (per kb)	Percentage of genome covered by repeats
Zebrafinch	1.23	0.50	8.37%
Chimpanzee	3.31	1.65	45.36%
Human	3.21	1.75	49.95%
Mouse	2.79	1.90	42.87%
Platypus	2.00	2.56	44.32%

Table 4.A.2: **Sequencing coverage and read span by sample.** The coverage is calculated using all mapped reads with MAPQ > 0. The read span is the total distance between the leftmost and rightmost mapping position of the reads in a pair (i.e. twice the read length, plus the distance between the two reads.)

Sample	Tranche	Mean coverage	Mean read span (bp)
N701	1	17.9	224.2
N702	1	20.5	257.1
N703	1	17.8	295.6
N704	1	21.6	250.6
N705	1	18.4	272.3
N707	1	19.8	285.4
N710	2	20.3	361.5
N711	2	20.4	300.8
N712	2	18.8	347.7
N713	3	14.0	367.0
N714	3	13.5	358.8
N715	3	14.8	368.5
N716	3	13.3	371.7
N717	3	15.2	376.9
N718	3	15.9	379.4
N719	3	16.5	375.4
N720	3	17.7	360.8
N721	3	17.8	367.3
N722	3	15.3	367.0
N723	3	18.1	359.5
N724	3	16.1	363.7
N725	3	17.4	368.8
N726	4	14.2	339.5

Continued on next page

Table 4.A.2 Continued

Sample	Tranche	Mean coverage	Mean read span (bp)
N727	4	16.0	353.9
N728	4	15.7	352.8
N729	4	15.8	375.8
N730	4	16.3	354.3
N731	4	15.1	363.9
N732	4	13.8	373.8
N733	4	14.7	363.4
N734	4	13.3	394.5
N735	4	15.0	370.2
N736	4	14.3	369.1
N737	4	15.2	351.7
N738	4	16.2	350.1
N739	4	14.1	359.7
N740	4	14.5	349.2
N741	4	13.1	381.5
N742	4	16.8	324.5
N743	4	20.3	376.2
N744	4	20.4	359.0
N745	4	17.5	363.8
N746	4	17.4	343.3
N747	4	17.3	376.7
N748	4	17.2	349.5
N749	4	16.5	355.7
N750	4	16.8	358.9
N751	4	13.5	363.4
N752	4	12.9	382.8
N753	4	16.3	337.6
N754	4	15.5	353.3
N755	4	15.2	352.3
N756	4	14.7	356.1
N757	4	19.7	355.3
N758	4	15.1	352.1
N759	4	14.9	346.9
N760	4	14.6	361.2
N761	4	16.0	366.9

Table 4.A.3: **Basic statistics of mate pair libraries.** The usable pairs in the rightmost column are those remaining after deduplication of pairs in which both reads were longer than 50bp after adaptor trimming.

Library preparation method	Library	Mean (bp)	Standard deviation (bp)	Number of read pairs	Proportion of pairs with duplicate	Proportion of pairs usable
gel-free	WTCHG_95726	2,537	905	147,138,520	2.70%	42.67%
gel-plus	WTCHG_122248_05	5,889	375	57,274,900	92.34%	2.69%
	WTCHG_122248_06	8,094	663	43,886,629	55.14%	2.18%
	WTCHG_122248_15	10,594	1,787	22,163,827	27.97%	1.67%

Table 4.A.4: **Filters used in GATK VariantFiltration.** *The depth filters are applied to the total depth across all 58 samples, and adjusted depending on whether the scaffold was inferred to be autosomal/pseudoautosomal (A/PA) or X-linked (XL). The minimum total coverage was set to $290 = N_{samples} \times 5$ for A/PA, and $212 \approx N_{females} \times 5 + N_{males} \times \frac{5}{2}$ for XL. The maximum total coverage was set to $1740 = N_{samples} \times 30$ for A/PA, and $1275 = N_{females} \times 30 + N_{males} \times \frac{30}{2}$ for XL. Alt: reads carrying the alternate allele; Ref: reads carrying the reference allele.

GATK filter expression	Metric description	Filter description
AF >= 1.0	allele frequency	variant fixed
DP < 290 (A/PA), < 212 (XL)*	read depth	coverage too low
DP > 1740(A/PA), > 1275 (XL)*	read depth	coverage too high
AN < 80	number of called alleles out of $2 \times 58 = 116$	too many missing genotypes
Dels > 0.4	fraction of reads containing spanning deletions	too many reads containing deletions
HaplotypeScore > 13.0	consistency of the site with ≤ 2 segregating haplotypes	evidence for > 2 segregating haplotypes
FS > 60.0	phred-scaled p-value using Fisher's exact test to detect strand bias	strand bias
ReadPosRankSum < -8.0	Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias	read position bias
MQ < 40.0	root mean squared mapping quality	mapping quality too low
MQRankSum < -12.5	Z-score from Wilcoxon rank sum test of Alt vs. Ref read mapping qualities	mapping quality bias
MQ0 > 75	total mapping quality 0 reads	too many reads with mapping quality 0

Table 4.A.5: **Proportion of sequence in the most common repeat classes for ornAna1 compared to filled gaps in ornAna1.2.** The gaps were filled using Molecule reads and *de novo* contigs, as described in Section 4.2.2.4. RepeatMasker (Smit, 2015) was then run on the original assembly (ornAna1) and on the gap-filling sequences longer than 200bp. Note that the total for the whole genome is slightly lower than the percentage given in Table 4.A.1 because it includes only the most common repeat classes.

repeat type	whole genome (2Gb)	filled intra-scaffold gaps (total 8.8Mb)	filled inter-scaffold gaps (total 5.8Mb)
Small RNA	0.08%	0.31%	0.21%
LTR elements	0.29%	0.41%	0.30%
Low complexity	0.55%	0.14%	0.17%
Simple repeats	0.67%	0.62%	0.66%
DNA elements	0.72%	1.05%	0.52%
Satellites	0.97%	0.23%	1.17%
LINEs	19.73%	31.36%	50.63%
SINEs	20.72%	19.88%	17.28%
Total	43.72%	54.00%	70.94%

Table 4.A.6: **Density of switch errors in the family quartet on ornAna1 versus ornAna1.2.** We identified places where the offspring changed from carrying the same haplotype from a particular parent to carrying a different haplotype, or vice versa. These switches are either due to misassemblies or to true crossovers. Given the large number of switches, we assume they are dominated by misassemblies. We removed switches that involved only a single SNP because these were likely due to genotyping errors. Note that we did not run GATK and CORTEX (Iqbal *et al.*, 2012) on the mapping to ornAna1 because of computational constraints (CORTEX takes about a week to run on 48 cores of a high-memory machine), but instead simply ran Platypus (Rimmer *et al.*, 2014) to produce a set of calls quickly. The switch error rate in ornAna1.2 was much lower when using the more stringent GATK/CORTEX intersection callset with the GQ filter than the Platypus PASS set (see Section 4.3.2).

		Platypus PASS on ornAna1	Platypus PASS on ornAna1.2	GATK/CORTEX intersection on ornAna1.2
Informative SNPs	maternal	636,713	675,491	170,232
	paternal	697,357	748,049	186,149
Switches	maternal	1,971	1,573	252
	paternal	2,680	2,196	188
Mean distance between switches	maternal	561,326bp	872,056bp	5,055,597bp
	paternal	413,090bp	634,566bp	6,960,950bp
Mean informative SNPs per switch	maternal	323	429	676
	paternal	260	341	990

Table 4.A.7: Related samples identified with KING.

ID 1	ID 2	Kinship	Sampling location	Relationship
N715	N721	0.1551	Barnard	2nd degree
N713	N714	0.0797	Barnard	3rd degree
N717	N720	0.0681	Barnard	3rd degree
N713	N722	0.0509	Barnard	3rd degree
N721	N725	0.0504	Barnard	3rd degree
N719	N722	0.0444	Barnard	3rd degree
N703	N749	0.4982	Dirran	duplicate
N702	N749	0.0839	Dirran	3rd degree
N702	N703	0.0728	Dirran	3rd degree
N749	N750	0.0587	Dirran	3rd degree
N703	N750	0.0442	Dirran	3rd degree
N712	N742	0.2446	Shoalhaven	1st degree - parent/child
N742	N752	0.257	Shoalhaven	1st degree - parent/child
N712	N757	0.2444	Shoalhaven	1st degree - parent/child
N752	N757	0.2616	Shoalhaven	1st degree - parent/child
N712	N727	0.2731	Shoalhaven	1st degree - siblings
N742	N757	0.2561	Shoalhaven	1st degree - siblings
N712	N734	0.1415	Shoalhaven	2nd degree
N727	N734	0.1473	Shoalhaven	2nd degree
N738	N741	0.1009	Shoalhaven	2nd degree
N727	N742	0.1381	Shoalhaven	2nd degree
N727	N757	0.1445	Shoalhaven	2nd degree
N729	N760	0.1015	Shoalhaven	2nd degree
N731	N741	0.0496	Shoalhaven	3rd degree
N734	N742	0.0838	Shoalhaven	3rd degree
N734	N743	0.0517	Shoalhaven	3rd degree
N734	N757	0.0782	Shoalhaven	3rd degree
N741	N757	0.045	Shoalhaven	3rd degree
N732	N733	0.0832	Wingecarribee	3rd degree
N732	N744	0.0705	Wingecarribee	3rd degree
N733	N744	0.0689	Wingecarribee	3rd degree
N732	N759	0.0369	Wingecarribee	3rd degree
N744	N759	0.0342	Wingecarribee	3rd degree
N735	N736	0.2513	Fish (Taronga)	1st degree - parent/child
N755	N756	0.0405	Brumby's Creek/Lake	3rd degree

Table 4.A.8: **Description of scaffolds with Y-homologous regions (YHRs).** The boundaries between the pseudoautosomal and X-specific regions (PAR boundaries) correspond to start or end points of contigs, or are otherwise approximate since they are difficult to determine accurately, due either to noisy coverage or to heterozygous variants seen in some but not all males that could be SNPs within the PAR or the YHR. Regions with an asterisk contain at least four discrete YHRs of length less than 1kb within the indicated region.

Scaffold	Scaffold length (bp)	Classification	Autosome or PAR	X-specific region	Y-homologous region(s)
Scaffold 0	6,392,720	PAR boundary	start-4,463,266	4,463,266-end	4,463,266-end
Scaffold 40	2,548,299	PAR/autosome			1,470,595-1,485,684
Scaffold 101 ²⁴	1,028,842	PAR boundary/ misassembly	start-968,988	968,988-end	968,988-1,003,811 1,013,316-1,027,832
Scaffold 179 ²⁵	1,350,745	X		whole scaffold	575,756-601,276 1,284,616-1,328,595 1,336,872-1,350,588
Scaffold 368	681,844	PAR boundary	651,656-end	start-651,656	643,648-651,656
Scaffold 481	701,419	X		whole scaffold	255,942-400,211* 481,054-701,419*
Scaffold 621	391,819	PAR/autosome	whole scaffold		207,703-208,075
Scaffold 627 ²⁶	556,693	X/ misassembly	554,362-556,693	whole scaffold	509,994-531,357
Scaffold 1480	455,301	PAR boundary	start-132,344	132,344-end	132,344-226,214 ²⁷
Scaffold 1937	529,027	X		whole scaffold	284,445-304,923* 318,005-322,954*
Scaffold 2552	231,902	X		whole scaffold	25,320-105,643 216,967-229,872*
Scaffold 2660 ²⁸	199,188	PAR boundary/ misassembly	start-23,428 148,916-199,188	23,428-75,687	28,270-75,682
Scaffold 2833 ²⁹	60,168	X?		whole scaffold	whole scaffold

Continued on next page

²⁴The junction between the PAR/autosome and XSR on Scaffold 101 is only weakly supported by MPs and BACs, so this is suspected to be a misassembly.

²⁵There was very high coverage in males on Scaffold 179:1,284,693-1,288,845, suggesting multiple copies on the Y.

²⁶The final contig in Scaffold 627, Contig13608.7 at 554362-556693, has equal coverage in males and females, and is suspected to have been joined spuriously to the scaffold due to MPs landing partially in repeats.

²⁷Within this YHR on Scaffold 1480, the males had ~4X the coverage of the females (including reads apparently deriving from the Y), suggesting multiple copies on the Y.

²⁸The junction between PAR/autosome and XSR at ~23kb on Scaffold 2660 is only supported by two fosmids, which land within repeats, so it is suspected to be a misassembly. The last PAR/autosomal piece at Scaffold 2660:148,916-199,188 may also have been erroneously added, since it is only connected by two BACs, which map overlapping repeats.

²⁹In the region from the start of Scaffold 2833 to ~42kb, the males had ~1.5X the coverage of the

Table 4.A.8 Continued

Scaffold	Scaffold length	Classification	Autosome or PAR	X-specific region	Y-homologous region(s)
Scaffold 2901	235,768	PAR/autosome	whole scaffold		47,037-50,899
Scaffold 2839	205,163	X		whole scaffold	235-913 13,583-23,679 120,045-124,972 126,673-131,640 150,168-205,163
Scaffold 3642	252,285	X		whole scaffold	38,278-55,248 56,364-59,960 64,031-71,497
Scaffold 4647 ³⁰	53,889	X?		whole scaffold	1-1,103 1,420-3,435 8,450-16,628 36,127-51,973
Scaffold 4712	88,971	X		whole scaffold	5,064-6,104 35,740-38,426
Scaffold 4926	65,686	X		whole scaffold	whole scaffold
Scaffold 5034	189,314	X		whole scaffold	172,033-185,100*
Scaffold 5038	86,290	X		whole scaffold	57,483-59,504 60,414-60,807
Scaffold 5201	489,444	X		whole scaffold	146,801-146,874 150,026-150,379 153,233-153,343
Scaffold 5237	50,518	X		whole scaffold	whole scaffold
Scaffold 5515	94,817	PAR/autosome	whole scaffold		50,358-55,953
Scaffold 5838	73,598	X		whole scaffold	52,964-57,559
Scaffold 8035	52,757	X		whole scaffold	176-2,115 7,153-8,732 44,527-45,247
Scaffold 8700	87,854	X		whole scaffold	whole scaffold*
Scaffold 10588	60,022	PAR/autosome	whole scaffold		26,416-34,839*
Scaffold 11479	55,316	X		whole scaffold	whole scaffold
Scaffold 15497	52,637	X		whole scaffold	whole scaffold
Scaffold 17598	52,590	X		whole scaffold	6,773-50,530*

females (including reads apparently deriving from the Y), suggesting a duplication on the Y.

³⁰Scaffold 4647 has two large blocks of contigs separated by a long gap. The first block (from the start to ~16.7kb) has a male-to-female coverage ratio ~0.5, except within YHRs, whereas in the block from ~35kb to the end this ratio is ~1.5 (including presumably Y reads), suggesting either that this is a PAR/autosome with paralagous copy on the Y, or that this is part of an X chromosome, and the gametologous Y copy has been duplicated. These two blocks are connected just by two fosmids, which map partially in repeats, so this may be a misassembly.

Table 4.A.9: **Density of X-Y divergences in Y-homologous regions (YHRs).**

If a scaffold had many discrete YHRs separated by gaps, we chose a representative subset of these to estimate the level of X-Y divergence. The number and total length of these subregions are shown in columns 2 and 3. Columns 4 and 5 show the number of X-Y divergences within these subregions, and the average distance between them. Note that some scaffolds have been classified as pseudo(autosomal) in Table 4.A.8, with a region transposed onto a Y, in which case an “X-Y divergence” simply refers to a heterozygous SNP called within this region only in males which fulfilled the criteria described in Section 4.2.8.

Scaffold	Number of subregions used	Total length used (bp)	Number of X-Y divergences	Mean distance between X-Y divergences (bp)
Scaffold 0:4,578,013-4,820,164 ³¹	1	242151	2015	120.2
Scaffold 0:5,606,235-5,848,386	1	242151	4739	51.1
Scaffold 40	1	15089	365	41.3
Scaffold 101	2	49339	2057	24.0
Scaffold 179:580,646-585,492 ³²	4	3481	296	11.8
Scaffold 179:1,290,481-1,328,283	5	20827	1447	14.4
Scaffold 179:1,336,872-1,350,580	1	13708	94	145.8
Scaffold 368	1	8008	427	18.8
Scaffold 481	5	16508	1277	12.9
Scaffold 621	1	372	31	12.0
Scaffold 627	1	21363	355	60.2
Scaffold 1480	1	93870	1690	55.5
Scaffold 1937	13	6339	541	11.7
Scaffold 2552	3	2930	184	15.9
Scaffold 2660	1	7474	586	12.8
Scaffold 2833	2	11916	770	15.5
Scaffold 2839	5	75663	3881	19.5
Scaffold 2901	1	3862	398	9.7
Scaffold 3642	1	16954	1172	14.5
Scaffold 4505	3	14431	251	57.5

Continued on next page

³¹We show two different subregions for Scaffold 0, to differentiate the YHR which is missing in N702 (first subregion) from that which is seen in all males (second subregion).

³²We show three different subregions for Scaffold 179, taken from the YHRs at 575,756-601,276, 1,284,616-1,328,595, and 1,336,872-1,350,588 respectively, excluding the suspected duplicated region on at 284,693-1,288,845.

Table 4.A.9 Continued

Scaffold	Number of subregions used	Total length used (bp)	Number of X-Y divergences	Mean distance between X-Y divergences (bp)
Scaffold 4647:1-1,103, 1,420-3,435, 8,450-16,628 ³³	3	11295	485	23.3
Scaffold 4647:36,201- 43,381	1	7180	168	42.7
Scaffold 4647:44,435-51,944	1	7509	72	104.3
Scaffold 4712	2	3726	184	20.3
Scaffold 4926	1	65686	2746	23.9
Scaffold 5034	4	978	95	10.3
Scaffold 5038	3	2443	209	11.7
Scaffold 5201	2	426	46	9.3
Scaffold 5237	1	50518	2466	20.5
Scaffold 5515	1	5595	338	16.6
Scaffold 5838	2	1219	114	10.7
Scaffold 8035	3	4238	257	16.5
Scaffold 8700	2	12468	824	15.1
Scaffold 10588	1	1970	138	14.3
Scaffold 11479	1	55316	2000	27.7
Scaffold 15497	1	52637	2707	19.4
Scaffold 17598	1	1091	44	24.8

³³We show three different subregions for Scaffold 4647. The first is comprised of three chunks from the 16.7kb region at the beginning of the scaffold, which seems to have haploid coverage on the X and Y copies in males. The second two come from the block from ~35kb to the end, which is suspected have been added on spuriously, as mentioned in footnote 28 below Table 4.A.8. This has been split into two blocks since the X-Y divergences were noticeably denser from ~35-43kb than from ~43kb to the end.

Table 4.A.10: **Synteny between platypus scaffolds with Y-homologous regions and chicken and human chromosomes.** This was determined using orthologous genes, as explained in Section 4.2.9. If the block contains only one platypus gene, the coordinates for chicken and human indicate the position of its predicted ortholog. (This should not be called “synteny”, which requires at least two genes.) Scaffolds 5237 and 8035 contain no predicted genes and thus are excluded.

Platypus				Chicken			Human		
Scaffold	Start	End	N. genes	Chr.	Start	End	Chr.	Start	End
0	385	4,128,039	8	2	74,276,764	69,235,105	5	19,472,951	31,555,058
0	4,132,640	6,366,781	13	2	67,702,996	66,252,176	6 ³⁴	2,786,952	3,268,962
40	189,856	2,463,629	14	2	6,187,456	8,053,409	7	151,341,699	155,781,485
101	123,197	921,756	12	4	36,714,905	57,469,617	4	109,433,772	110,565,285
101	974,022	978,165	1	13	14,934,288	14,940,878	5	135,027,735	135,034,813
179	31,647	51,948	1	Z ³⁵	34,022,935	34,119,106	9	20,658,309	20,995,955
179	195,650	287,229	1	Z	33,889,453	34,000,379	9	20,341,665	20,622,543
179	553,606	602,494	3	2	88,935,921	85,697,949	5	353,856	196,871
179	612,629	666,688	2	2	88,410,899	88,578,292	18	36,180,556	36,108,532
179	693,006	743,769	1	2	88,499,507	88,560,120	9	109,013,529	108,934,181
179	761,363	790,224	1	2	88,645,530	88,745,491	9 ³⁶	100,442,271	100,577,636
179	791,992	795,386	1	2	88,746,284	88,761,395	9	100,578,079	100,587,906
179	800,891	918,267	1	2	88,761,816	88,910,236	NA	NA	NA
179	924,079	935,020	1	2	55,897,544	55,900,827	5	191,511	195,353
179	940,241	1,016,440	1	20	10,700,368	10,707,694	5	140,258	189,970
179	1,040,794	1,049,649	1	2	67,753,130	67,763,712	18	63,871,692	63,903,890
179	1,074,815	1,094,192	3 ³⁷	19	353,752	359,048	4	188,139,419	188,147,743
179	1,099,735	1,100,566	1	2	564,030	564,628	4	188,091,273	188,109,603
179	1,106,131	1,123,308	1	2	88,911,152	88,929,189	5	443,158	471,937
179	1,275,724	1,298,610	1	2	64,685,429	64,700,170	6	7,389,496	7,418,037
368	341,906	617,190	3	12 ³⁸	85,233	2,944,369	3	52,048,919	52,214,327
368	622,097	624,811	1	NA	NA	NA	3	52,221,080	52,226,163
368	643,533	652,966	1	13	12,782,159	12,786,548	5	179,806,398	179,838,078
481	1,215	542,062	4	2	65,203,624	65,647,950	6	6,007,605	4,994,732
481	574,659	641,549	1	2	65,668,286	65,773,998	6 ³⁹	4,706,159	4,955,551
621	3,457	271,614	3	15	11,172,365	7,795,744	22	29,259,597	28,883,592

Continued on next page

³⁴Gene order has been changed in humans.

³⁵Chicken also has two putative orthologs of this gene, focadhesin, on unassigned scaffolds.

³⁶The human ortholog is duplicated in tandem (ENSG00000251349 and ENSG00000241697).

³⁷Platypus has three copies of this TRIM gene (tripartite motif family-like) arranged in tandem, but there is only one predicted ortholog in human and chicken.

³⁸These genes (*DUSP7*, *POC1A*, *ALAS1*) are in a different order in chicken.

³⁹Human also has four putative orthologs of this gene on chrY, which encode chromodomain proteins (*CDY1*, *CDY1B*, *CDY2A*, *CDY2B*).

Table 4.A.10 Continued

Platypus				Chicken			Human		
Scaffold	Start	End	N. genes	Chr.	Start	End	Chr.	Start	End
627	50,065	331,280	3	NA	NA	NA	19	18,418,864	18,671,714
627	402,304	545,737	5	28	3,502,158	3,613,712	19	18,683,677	18,896,727
1480	3,480	4,618	1	2 ⁴⁰	97,952,712	97,956,997	NA	NA	NA
1480	322,786	420,331	1	Z	52,082,066	52,346,243	4	784,957	826,198
1937	134,873	171,916	1	2	64,976,838	65,001,770	6	6,588,108	6,654,983
2552	219,234	229,751	1	2	64,710,800	64,720,473	6	7,268,306	7,347,446
2660	3,729	20,350	1	8	28,011,855	28,023,641	1	71,063,291	71,081,297
2660	37,289	54,988	1	13	14,427,581	14,435,845	5	136,132,845	136,188,747
2833	1,454	42,172	2	17	554,237	565,821	9	136,850,943	136,866,286
2839	25,723	26,244	1	NA	NA	NA	5	135,444,218	135,452,399
2839	174,906	202,402	1	13	14,736,403	14,782,876	5	135,333,900	135,399,914
2901	9,827	53,999	1	19	353,752	359,048	4	188,139,419	188,147,743
2901	21,595	24,102	1	NA	NA	NA	NA	NA	NA
3642	3,481	205,031	3	2	65,994,143	66,151,793	6	4,115,689	3,752,026
4505	16,903	21,014	1	13	13,695,529	13,725,044	NA	NA	NA
4505	30,158	56,314	1	13	13,728,686	13,739,566	5	137,887,968	137,942,747
4505	58,480	81,161	1	NA	NA	NA	5	137,867,791	137,887,851
4647	41,992	44,677	1	NA	NA	NA	NA	NA	NA
4647	45,693	50,794	1	19	831,531	838,160	9	136,796,350	136,807,741
4712	569	4,119	1	2 ⁴¹	97,952,712	97,956,997	NA	NA	NA
4712	41,938	86,356	4	NA	NA	NA	19 ⁴²	53,258,292	57,457,142
5034	10,453	184,928	4	28	2,771,422	2,743,935	19	925,781	797,075
5038	27,956	56,749	1	13	14,295,426	14,509,385	5	136,028,895	136,063,818
5201	146,394	153,657	1	12	16,380,861	16,415,135	3	72,371,825	72,446,918
5515	71,650	89,731	1	Z ⁴³	44,067,508	44,176,409	9	91,213,815	91,361,913
5838	45,019	46,770	1	NA	NA	NA	17 ⁴⁴	8,753,106	8,758,559
8700	5,417	13,073	1	2	64,735,015	64,854,227	6	7,107,597	7,251,980
10588	5,193	5,426	1 ⁴⁵	7	23,128,082	23,144,659	2	127,257,290	127,294,176
11479	2,816	41,639	1	13	14,971,863	14,987,932	5	134,904,906	135,007,959
17598	7,055	33,182	1	28	3,071,247	3,095,645	19	1,407,569	1,435,687

⁴⁰The uncharacterised platypus gene ENSOANG00000006512 has several chicken orthologs: ENSGALG000000022875 on chr2, plus seven on chr16 and one on an unassigned scaffold.

⁴¹This platypus gene, ENSOANG000000013368, has the same chicken orthologs as the gene ENSOANG000000006512 on Scaffold 1480 mentioned in the previous footnote.

⁴²Human has three copies of this block of four genes (ENSOANG000000006847, ENSOANG000000009744, ENSOANG00000000776, ENSOANG000000007763).

⁴³Chicken has a second putative ortholog at chr8:23429040-23434465, and human at chr1:52895910-52927212.

⁴⁴Human also has seven predicted orthologs of this gene (*SPDYE*) on chr17.

⁴⁵Only one of the two exons of platypus gene ENSOANG000000030279 is found on Scaffold 10588.

Chapter 5

Multi-cohort analysis of the maternal age effect on recombination

5.1 Introduction

Failure of recombination is known to contribute to human aneuploidies by producing nondisjunction of homologous chromosomes during meiosis I (reviewed in Hassold and Hunt, 2001; Hou *et al.*, 2013). Aneuploidies are the leading genetic cause of miscarriage and developmental disabilities in humans, and the vast majority are of maternal origin (Nagaoka *et al.*, 2012). The frequency of aneuploidies increases dramatically in older mothers (Chiang *et al.*, 2012): only about 2% of clinically-recognised pregnancies in women under 25 are trisomic, compared to 35% in women over 40 (Hassold and Hunt, 2001).

Henderson and Edwards (1968) proposed that changes in recombination rate were responsible for the maternal age effect on aneuploidies, in what they termed the “production line” model. This stemmed from two important observations. Firstly, they saw an age-dependent decrease in chiasmata frequency in mouse oocytes. Secondly, it was known that female meiosis is a very long process that begins in the foetus and does not complete until fertilisation, potentially decades later (Figure 5.1). The crossovers begin forming during prophase I, during the first trimester of pregnancy, and then the oocytes go into meiotic arrest until ovulation, years or decades later, at which point the crossovers resolve (reviewed in Nagaoka *et al.*, 2012). The production line model involves two postulates: (1) that oocytes are ovulated in the same order they enter meiosis, and (2) that meiotic recombination rates decrease with foetal age. Thus, the

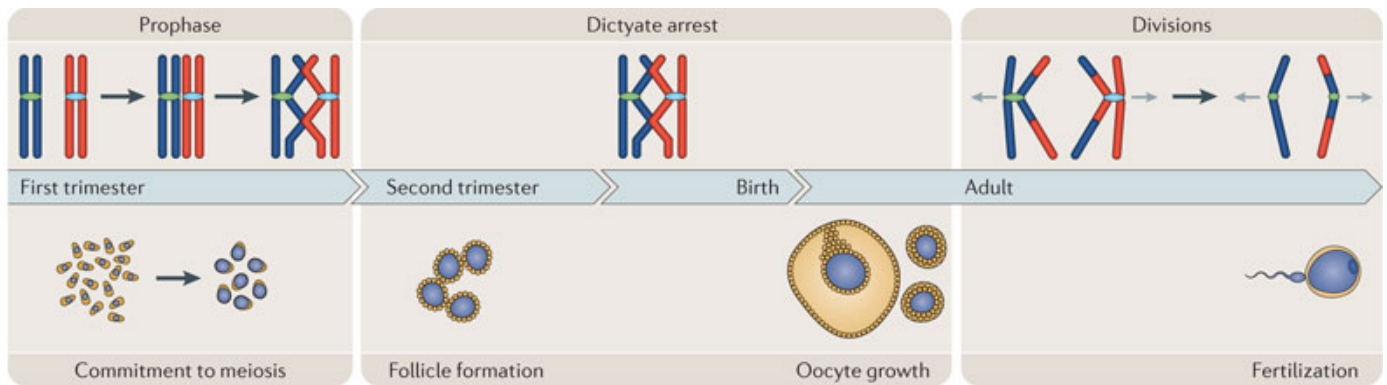


Figure 5.1: **The timeline of female meiosis and oogenesis in humans. Meiosis initiates at 8-10 weeks gestation.** After DNA replication, the homologous chromosomes pair, synapse and recombine, and then enter dictyate arrest, in which state they remain until the female reaches maturity. The follicles form during the second trimester of pregnancy, and the oocyte grows in the mature female, typically taking 85 days and culminating in the ovulation of a single egg. Ovulation completes the first meiotic division, in which the homologous chromosomes come apart, resolving the crossovers. The ovulated egg then arrests again, and meiosis II completes only if the egg is fertilised, causing the separation of sister chromatids. This figure was taken directly from Nagaoka *et al.* (2012). See also Figure 5.2 for an illustration of normal and abnormal chromosome segregation during meiosis.

hypothesis was that oocytes ovulated later in a woman's life have fewer crossovers, increasing the risk of nondisjunction. Polani and Crolla (1991) demonstrated that the first postulate held, at least in mice, and there is evidence supporting the second postulate in both mice (Henderson and Edwards, 1968) and hamsters (Sugawara and Mikamo, 1983).

There is evidence from cytogenetic studies of human gametes that the number and placement of crossovers affects the risk of developing trisomies. However, these patterns differ between chromosomes (Hassold *et al.*, 2000). For example, achiasmate bivalents are responsible for 20-40% of cases of trisomy 15, 18 and 21 of maternal origin, but do not seem to contribute to risk of trisomy 16. Brown *et al.* (2000) found that there was a genome-wide reduction in recombination in eggs with a nondisjoined chromosome 21 that had no detectable crossovers, and argued that, once the number of recombination events in a cell falls below some threshold, certain chromosomes become more likely to undergo nondisjunction.

Robinson *et al.* (1998) found that the number of crossovers on chromosome 15

that occurred in meioses with a non-disjunction event leading to trisomy for that chromosome increased with maternal age, whereas Hassold *et al.* (2000) found no effect of age on recombination levels for trisomies 16 and 21. Thus, it seems there is heterogeneity across chromosomes in the causes of aneuploidies and in the extent to which this is influenced by maternal age. Nonetheless, there has still been considerable interest in testing the effect of maternal age on recombination in humans. To test it directly, one would need to examine recombination levels in oocytes taken from the same women at multiple timepoints. Obtaining such samples is difficult, so several groups have instead used pedigree studies to examine the effect of maternal age on the number of crossovers in liveborn offspring (Kong *et al.*, 2004; Coop *et al.*, 2008; Hussin *et al.*, 2011; Bleazard *et al.*, 2013), with conflicting results. The focus has largely been on genome-wide recombination levels, and although this may not be relevant to changes in the aneuploidy rates for specific chromosomes, it is nonetheless of biological interest because recombination is such an important process for evolutionary and population genetic reasons.

Kong *et al.* (2004) analysed about 1000 microsatellite markers in 14,140 maternal meioses from Iceland, and used data on maternal age at birth rounded to the nearest 5 years. They fitted a linear mixed model that accounted for differences in baseline rates between mothers to estimate the expected change in the number of crossovers per year of maternal age, β_{age} . They reported that $\beta_{age} = 0.082$ (standard error = 0.012, $p < 10^{-8}$) but noted that the relationship between recombination rate and maternal age was not linear in their data, and that there was a large jump from age 30 to 35. They hypothesised that the apparent increase in the number of crossovers with age was due to a selection effect, whereby eggs with more crossovers were more likely to give rise to a live birth. Thus, they predicted that women with higher recombination rates would have more children, and indeed, they found that the number of crossovers was positively associated with family size (+0.0109 children per crossover, standard error 0.0041, $p = 0.0076$). It is for this reason that it is important to account for the baseline recombination rate of the mother when analysing the maternal age effect, since the women who have children later might simply have higher rates throughout their lives.

In the first study of recombination using genome-wide SNP data, Coop *et al.* (2008) analysed Affymetrix 500K chip data from 725 Hutterite individuals, part of a large 13-generation pedigree. They found an increase of 0.19 crossovers/year, using

crossover counts and ages adjusted by the mean for the parent (standard error 0.093, $p=0.035$ by permutation). Like Kong *et al.* (2004), they found a more pronounced increase in older women: mothers over 35 had an average of 3.1 more crossovers than those under 35 ($p=0.028$ using a within-family permutation test). Coop *et al.* also found that the mean maternal recombination rate was significantly correlated with family size (Spearman $\rho = 0.24, p = 0.027$).

Two more recent pedigree studies (Hussin *et al.*, 2011; Bleazard *et al.*, 2013) found that the recombination rate *decreases* with maternal age in humans, as it does in mice (Polani and Jagiello, 1976). Hussin *et al.* (2011) analysed 195 maternal meioses from a French Canadian cohort. Using a linear mixed model with the mother as a random effect, they estimated that $\beta_{age} = -0.49$ crossovers per year ($p=0.0017$), and they obtained similar results using adjusted counts within families ($\beta_{age} = -0.42, p = 0.0047$). Bleazard *et al.* (2013) analysed 338 meioses from Mongolian and Korean families. They regressed maternal crossover count on maternal age, including mother as a factor, and found that $\beta_{age} = -0.29$ crossovers/year ($p = 0.03$).

While this project was underway, Campbell *et al.* (2015) reported an analysis of recombination in about 4200 families from the 23andMe dataset, in which they found an increase of 0.067 maternal crossovers per year. Since they fitted a simple linear regression that did not account for the baseline rate of the mother, this analysis could be confounded for the reason described above, as they noted. When Campbell *et al.* restricted the analysis to 2184 “phase-known” meioses (from families with more than two children) and ran the regression on the adjusted counts, they did not find a significant association with age ($p=0.11$).

Thus, there is a lack of consensus about the size and direction of effect of maternal age on recombination rate.

There are likely several factors contributing to the conflicting results between different cohorts in the previous studies. There was considerable heterogeneity in the number of genetic markers used, in genotype data filtering and quality control, in the accuracy of the age data available, and in the methods used to call crossovers and to test the age effect. Here, we revisited this question in nine different cohorts (Table 5.1; Table 5.A.1). These included the French Canadians in which Hussin *et al.* had previously investigated the maternal age effect, six cohorts that O’Connell *et al.*

(2014) had used in a study on phasing in populations with high levels of relatedness, and two cohorts from twin registries not previously studied for this purpose. We tried to eliminate some of the possible confounders by applying the same methodology, as far as possible, to all cohorts. The aim was to resolve the direction of effect and to ascertain whether applying a standardised approach would make the results for each cohort more consistent.

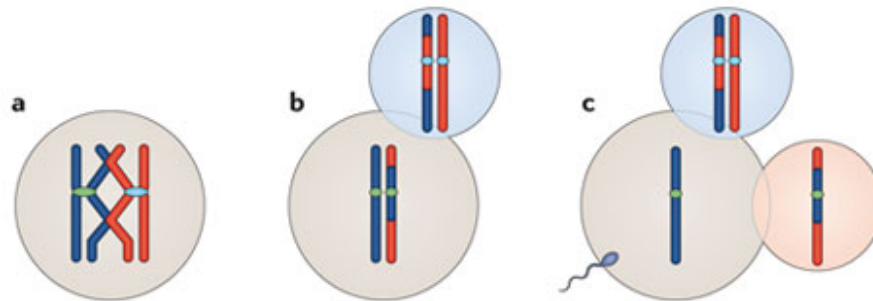
This chapter is primarily taken from Martin *et al.* (2015). As mentioned in the Acknowledgements section of the paper, Ryan Christ substantially assisted with the development of the hierarchical models, Jared O’Connell and Julie Hussin advised on the use of duoHMM and NFTOOLS respectively, and Scott Gordon, Hamdi Mbarek, Kerrie MacAloney and Jouke-Jan Hottenga helped with the data preparation and cleaning. The data were contributed by many different people, detailed in the Acknowledgements, with the largest dataset (Queensland Twin Registry) from Nick Martin.

Figure 5.2 (*following page*): **Normal and abnormal chromosome segregation.**

A) In the meiosis I oocyte (a), the homologous chromosomes form chiasmata which tether them at the crossover sites. They remain like this throughout the dictyate arrest (see Figure 5.1) and then separate at anaphase I, with one remaining in the egg and the other going to the first polar body (b). The ovulated egg is then arrested again at metaphase II, until fertilisation triggers the second meiotic division, causing the sister chromatids to separate, one into the egg, and the other to the second polar body (c). Various errors at meiosis I (panel B) or meiosis II (panel C) can lead to aneuploidies. The expected outcomes assume random segregation of the unassociated homologues or sister chromatids. Regardless of whether the errors involve whole chromosomes or sister chromatids, the outcome will be a monosomic or trisomic conception. The error of relevance in this chapter is recombination failure (panel B, part a), which results in two unpaired univalent chromosomes at meiosis I. Rowsey *et al.* (2014) suggest that age-dependent loss of cohesin, leading to premature sister chromatid separation (panel B, part d), may be responsible for the maternal age effect on aneuploidies. This figure was taken directly from Nagaoka *et al.* (2012).

A

Normal chromosome segregation



B

Meiosis I errors

		Expected outcome in egg			
		Balanced	Extra or missing chromosome	Extra or missing chromatid	
a	Recombination failure		50%	50%	*
b	Premature homologue separation		50%	50%	*
c	True nondisjunction		100%		
d	Premature sister chromatid separation		50%	‡	50%

C

Meiosis II errors

		Expected outcome in egg		
		Balanced	Extra or missing chromosome	Extra or missing chromatid
a	Nondisjunction			100%
b	Premature sister chromatid separation		50%	50%

5.2 Methods

5.2.1 Cohorts

The cohorts used are detailed in Table 5.1, along with the chips on which they were genotyped. The Queensland Twin Registry (QTR) and Netherlands Twin Registry (NTR) cohorts were collected for a variety of studies on behavioural and medical traits. In both cohorts, all families include one or more pairs of monozygotic (MZ) and/or dizygotic (DZ) twins, as well as one or both parents and, in some cases, additional siblings. For the analyses in this paper, one child of each MZ pair was removed. Permission was received to use these data for studying recombination from the Central University Research Ethics Committee of the University of Oxford, from the Ethics Committee of the Queensland Institute of Medical Research, and from the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam.

The French Canadian cohort consists of 478 individuals from 89 overlapping nuclear families. This is the same cohort studied by Hussin *et al.* (2011). The remaining cohorts were studied by O’Connell *et al.* (2014), and are composed of both unrelated individuals and pedigrees of various sizes. These are from isolated populations: the Orkney Islands (ORCADES (McQuillan *et al.*, 2008)), the Dalmatian islands of Vis and Korcula off Croatia (Vitart *et al.*, 2010) (treated separately), and three Italian populations, Val Borbera (VB) (Colonna *et al.*, 2013), Friuli Venezia Giulia (FVG) (Esko *et al.*, 2013) and Carlantino (CARL) (Esko *et al.*, 2013).

5.2.2 Auxiliary data on alternative reproductive technologies and oral contraception

Some data on the use of alternative reproductive technologies (ART) were available for a subset of families in the QTR. Specifically, some mothers had been surveyed during clinic visits or online surveys about whether they had ever had hormone treatment, *in vitro* fertilisation (IVF) or alternative or natural fertility treatments. However, the questions either did not ask which children were conceived by ART, or only asked about the twins. Additionally, it was often not possible to distinguish responses that were negative *versus* absent. Thus, we designed a more comprehensive survey which we conducted on several hundred QTR mothers.

Table 5.1: **Cohorts included in the study.** The three rightmost columns show the number of maternal meioses for which we had age data, followed by the number of families in parentheses. *The total number analysed excludes the cohorts (or, for the rightmost column, family configurations within a cohort) with fewer than 20 meioses. The SNP chips used, number of SNPs, and sample sizes for paternal meioses are given in Table 5.A.1.

Cohort	Number of samples	Duos from informative nuclear families	Informative duos	Duos from families with > 1 child
Carlantino, Italy (CARL)	630	6 (2)	29 (11)	53 (23)
French Canadians (FC)	477	106 (29)	158 (50)	218 (80)
Friuli Venezia Giulia, Italy (FVG)	1236	13 (4)	72 (26)	160 (70)
Korcula, Croatia	897	0 (0)	7 (3)	34 (17)
Netherlands Twin Registry (NTR)	2729	298 (96)	398 (130)	889 (379)
ORCADES, Orkney, Scotland	2215	48 (15)	118 (40)	252 (109)
Queensland Twin Registry (QTR370)	3754	890 (239)	1310 (323)	1441 (389)
Queensland Twin Registry (QTR610)	7364	1283 (403)	1337 (420)	2898 (1202)
Queensland Twin Registry (QTRCoreExome)	4444	234 (75)	329 (106)	781 (332)
Val Borbera, Italy (VB)	1664	30 (9)	72 (26)	266 (123)
Vis, Croatia	960	0 (0)	11 (4)	37 (17)
Total		3036 (910)	4253 (1268)	7688 (3003)
Total used*		2889 (866)	3823 (1132)	6011 (2305)

Data on the use of ART were also available for a subset of families in the NTR and ORCADES cohorts, as well as data on oral contraception. For the NTR, the mothers had been asked whether each child was conceived spontaneously or not (e.g. using hormones or IVF), whether they had ever taken oral contraception, and whether they had taken it before the birth of the twins. For the ORCADES cohort, we had data on whether each mother had ever taken oral contraception.

For some analyses, we stratified QTR, NTR and/or ORCADES mothers by ART or oral contraceptive use before the birth of their children.

5.2.3 Preparation of genotype data

We first ran KING (Manichaikul *et al.*, 2010) on each cohort to detect any pedigree errors and find any cryptic first-degree relationships. In addition to the quality control performed on individual cohorts prior to this project, we also removed SNPs with a minor allele frequency in the founders less than 5%, a p-value from a test of Hardy-Weinberg equilibrium in the founders less than 0.01, or more than 1% Mendelian errors. We set all remaining Mendelian errors to missing, as well as any genotypes flagged as unlikely by Merlin’s error detection algorithm (Abecasis *et al.*, 2002). SNPs with more than 1% missingness were then removed. For the QTR, we split the sample into three sets of families to maximise the intersection of the SNPs genotyped on all family members: QTR610 - families in which all individuals had been genotyped on the Illumina Human610 or Human660W chip; QTR370 - families in which at least one individual had been genotyped on the Illumina CNV370 chip; QTRCoreExome - families in which all individuals had been genotyped on the HumanCoreExome chip.

For ORCADES, individuals were genotyped on either the Illumina HumanHap300 or the Illumina Omni Express chip. Since most families had individuals genotyped on different chips, for all families we used the intersection of the chips across the whole cohort. The final number of SNPs used for all cohorts is shown in Table 5.A.1.

5.2.4 Calling crossovers

There are several different methods for calling crossovers using genotype data from pedigrees (Kong *et al.*, 2004; Coop *et al.*, 2008; Hussin *et al.*, 2011; Bleazard *et al.*, 2013). Essentially, these rely on inferring the transmission patterns of polymorphic

markers from parents to their offspring. A key requirement is that they take into account the possibility of genotype errors which can look like recombination events over short intervals. We used two different methods for calling crossovers. The first, NFTOOLS (Hussin *et al.*, 2011), based on the approach of Coop *et al.* (2008), is a heuristic method that phases the parental chromosomes using patterns of transmission to the children, and then identifies positions at which a child’s chromosome switches from copying one parental haplotype to the other. Following Hussin *et al.* (2011), we removed double crossovers within 5 informative markers of each other, since these were likely due to genotyping errors.

The second method, duoHMM (O’Connell *et al.*, 2014) involves first phasing the samples with SHAPEIT2 (Delaneau *et al.*, 2013) ignoring family information, and then running a Hidden Markov Model (HMM) on every parent-child duo. This HMM corrects phasing errors inconsistent with the pedigree structure, detects genotyping errors, and calls crossovers, assigning a posterior probability p to each. Following O’Connell *et al.*, we kept only the crossovers with $p > 0.5$. We found the number of crossovers called at this threshold was somewhat higher than expected in some cohorts (relative to established genetic maps). Hence we applied some further filtering, excluding double crossovers within X SNPs of each other, where X was the average number of SNPs per megabase for that cohort, given in Table 5.A.1. We also fitted the hierarchical models described below on the raw counts, to verify that the results were robust to this filtering (compare Table 5.1 to Table 5.A.2).

NFTOOLS requires families with both parents and more than two siblings genotyped in order to assign crossovers to particular children; we refer to these as “informative nuclear families”. In contrast, duoHMM can, in theory, call crossovers in any parent-child duo, because it uses all individuals in the cohort to phase the parents rather than relying on having three offspring. However, it has reduced power if there are insufficient first-degree relatives to phase the parents exactly. We refer to duos/meioses from families with both parents and at least three children, or from three-generation pedigrees, as “informative”, and all other duos/meioses as “partially informative” (these were termed “uninformative” in O’Connell *et al.* (2014)).

We applied the hierarchical models presented below to either fully informative meioses (Models 1 and 2), or to fully and partially informative meioses (Model 2*), in both cases from families with at least two children. For Model 2*, we use the following

six family configurations: families with at least three children and one or both parents (configurations 1 and 2 respectively), families with two children and one or both parents (configurations 3 and 4), and families with two children, the grandparents (on the side of the relevant parent) and one or both parents (configurations 5 and 6). Configurations 2 and 6 are fully informative, and the rest partially informative.

5.2.5 Linear mixed models and meta-analysis

To test the effect of parental age on the number of crossovers, while accounting for differences in baseline recombination rates between individuals, we began by following the approach of Hussin *et al.* (2011) and fitted a linear mixed model on the maternal or paternal crossover counts:

$$Y_{i,j} = \mu + \beta_{age} \times age + \alpha_j + \epsilon_i \quad (5.1)$$

where $Y_{i,j}$ is the number of crossovers for child i of parent j , μ is an intercept, β_{age} is the effect of the age of parent j at the birth of child i (in crossovers per year), α_j is the normally-distributed random effect of parent j , and ϵ_i is a normally-distributed error term. Hussin *et al.* (2011) included the number of children as a covariate, but since, for most cohorts, we only knew the number of genotyped children, not the total, this was not possible here. This model was fitted to the meioses from informative nuclear families for every cohort separately using the nlme package in R (Pinheiro *et al.*, 2014). We then carried out a fixed-effect meta-analysis on the β_{age} estimates using the metafor package (Viechtbauer, 2010), weighting them by their inverse variances (Cooper *et al.*, 2009).

5.2.6 Bayesian hierarchical models

We next fitted a variety of hierarchical models to test the maternal age effect. These accounted for differences between families and between cohorts. All models were fitted on the paternal crossovers too, using the priors detailed at the end of this section. The first model was a simple normal model on crossover counts:

$$\left. \begin{aligned}
Y_{i,j,c} &\sim N(\beta_{age} \times \text{age}_{i,j,c} + \alpha_{j,c}, \tau^2) \\
\alpha_{j,c} &\sim N(\mu_c, \sigma_c^2) \\
\mu_c &\sim N(41, 100) \\
\sigma_c^2 &\sim \text{Inv-Gamma}(2, 40) \\
\beta_{age} &\sim N(0, 1) \\
\tau^2 &\sim \text{Inv-Gamma}(2, 70)
\end{aligned} \right\} \text{Model 1} \quad (5.2)$$

where $Y_{i,j,c}$ is the number of maternal crossovers for child i of mother j from cohort c , $\alpha_{j,c}$ is the effect of the mother, μ_c and σ_c^2 are the mean and variance respectively of the mother effects for cohort c , and τ^2 is the residual variance.

We chose conjugate distributions for the priors and hyperpriors, and used an empirical Bayes approach to choose some of the parameters for the prior on the residual variance τ^2 and for the hyperpriors for μ_c and σ_c^2 . We specified the mean value of the prior for μ_c to match approximately the mean number of crossovers for females (41) reported in Hussin *et al.* (2011), and we set the variance to be large (100) since we found that a tight prior on μ_c influenced the estimates of β_{age} . The σ_c^2 prior was chosen so that its 95% credible interval encompassed the empirical variances in the average number of crossovers for a parent for each cohort. Given that the previous estimates for β_{age} had ranged between -0.49 and +0.19, we chose a $N(0, 1)$ prior for this parameter. We chose the prior for τ^2 so that its mean would be approximately equal to the variance in the number of maternal crossovers across all fully informative meioses.

We also fitted a more general model which allowed for heterogeneity in the age effect between cohorts:

$$\left. \begin{aligned}
Y_{i,j,c} &\sim N(\beta_{age,c} \times \text{age}_{i,j,c} + \alpha_{j,c}, \tau^2) \\
\beta_{age,c} &\sim N(\beta_{age,global}, \sigma_{age,global}^2) \\
\beta_{age,global} &\sim N(0, 1) \\
\sigma_{age,global}^2 &\sim \text{Inv-Gamma}(3, 0.5)
\end{aligned} \right\} \text{Model 1.2} \quad (5.3)$$

where $\beta_{age,c}$ is the age effect for cohort c , and $\beta_{age,global}$ and $\sigma_{age,global}^2$ are the mean and variance respectively for the age effects. We used the same hierarchy for $\alpha_{j,c}$ and

the same priors for τ^2 , μ_c and σ_c^2 as in Model 1.

Although the normal model has an easy interpretation and seemed to fit the data quite well in posterior predictive checks, it is not strictly appropriate for count data. Rather than use a Poisson model, which constrains the variance to be equal to the mean, we considered a negative binomial model, using the parameterisation:

$$X \sim \text{NegBinom}(a, b) \text{ when } Pr(X = x) = \binom{a+x-1}{a-1} \left(\frac{b}{1+b}\right)^a \left(\frac{1}{1+b}\right)^x \quad (5.4)$$

for $a \in \mathbb{R}^+$, $b \in \mathbb{R}^+$ and $x \in \mathbb{N}$. This gives:

$$E(X) = \frac{a}{b} \text{ and } \text{Var}(X) = \frac{(1+b)a}{b^2} = \frac{1+b}{b} E(X). \quad (5.5)$$

We fitted the following model:

$$\left. \begin{array}{ll} Y_{i,j,c} \sim \text{NegBinom} & \left(\frac{\exp(\beta_{age} \times \text{age}_{i,j,c} + \alpha_{j,c})}{\omega - 1}, \frac{1}{\omega - 1} \right) \\ \alpha_{j,c} & \sim N(\mu_c, \sigma_c^2) \\ \mu_c & \sim N(3.7, 0.2) \\ \sigma_c^2 & \sim \text{Inv-Gamma}(3, 0.05) \\ \beta_{age} & \sim N(0, 0.01) \\ \frac{1}{\omega} & \sim U(0, 1) \end{array} \right\} \text{Model 2} \quad (5.6)$$

This gives $E(Y_{i,j}) = \exp(\beta_{age} \times \text{age}_{i,j,c} + \alpha_{j,c})$ and $\text{Var}(Y_{i,j}) = \omega E(Y_{i,j})$, where ω is the over-dispersion parameter. We put a uniform $U(0, 1)$ prior on ω^{-1} so that ω would be constrained to the range $(1, \infty)$. We used the exponential link function since the shape parameter of the negative binomial distribution, a , needs to be strictly positive. We adjusted the priors to account for the transformation of the parameters, so that they would be similar to the priors for Model 1 e.g. $E(\mu_c) = \log(41) \approx 3.7$.

The models described so far were fitted to data from informative meioses from families with at least two children (with only one child, it is not possible to estimate $\alpha_{j,c}$), excluding cohorts with fewer than 20 such meioses. Since including partially informative meioses substantially increased the available sample size, but, as O'Connell *et al.* (2014) showed, there was decreased power to call crossovers for these meioses, we adapted Model 2 to account for this by simply drawing the mother effects, α , from a distribution with a mean and variance that were specific to the family configuration f from cohort c :

$$\left. \begin{aligned}
Y_{i,j,c} &\sim \text{NegBinom} \left(\frac{\exp(\beta_{age} \times \text{age}_{i,j,c} + \alpha_{j,c,f})}{\omega - 1}, \frac{1}{\omega - 1} \right) \\
\alpha_{j,c,f} &\sim N(\mu_{c,f}, \sigma_{c,f}^2) \\
\mu_{c,f} &\sim N(3.7, 0.2) \\
\sigma_{c,f}^2 &\sim \text{Inv-Gamma}(3, 0.05) \\
\beta_{age} &\sim N(0, 0.01) \\
\frac{1}{\omega} &\sim U(0, 1) .
\end{aligned} \right\} \text{Model 2}^* \quad (5.7)$$

We fitted Model 2* to all meioses from families with at least two genotyped children; we excluded a particular family configuration for a cohort if it included fewer than 20 meioses. The family configurations used are explained in Section 5.2.4.

As for Model 1, we also fitted a more general version of Model 2 with cohort-specific age effects. The following model was fitted to the informative meioses:

$$\left. \begin{aligned}
Y_{i,j,c} &\sim \text{NegBinom} \left(\frac{\exp(\beta_{age,c} \times \text{age}_{i,j,c} + \alpha_{j,c})}{\omega - 1}, \frac{1}{\omega - 1} \right) \\
\beta_{age,c} &\sim N(\beta_{global}, \sigma_{global}^2) \\
\beta_{age,global} &\sim N(0, 0.01) \\
\sigma_{age,global}^2 &\sim \text{Inv-Gamma}(3, 0.02) .
\end{aligned} \right\} \text{Model 2.2} \quad (5.8)$$

with the hyperpriors on $\alpha_{j,c}$ and the other priors the same as for Model 2. For the informative and partially informative meioses combined, we kept the μ and σ^2 parameters cohort- and family configuration-specific, but had a common $\beta_{age,c}$ for all families in cohort c :

$$\left. \begin{aligned}
Y_{i,j,c} &\sim \text{NegBinom} \left(\frac{\exp(\beta_{age,c} \times \text{age}_{i,j,c} + \alpha_{j,c,f})}{\omega - 1}, \frac{1}{\omega - 1} \right) \\
\beta_{age,c} &\sim N(\beta_{global}, \sigma_{global}^2) \\
\beta_{age,global} &\sim N(0, 0.01) \\
\sigma_{age,global}^2 &\sim \text{Inv-Gamma}(3, 0.02)
\end{aligned} \right\} \text{Model 2.2}^* \quad (5.9)$$

with the hyperpriors on $\alpha_{j,c,f}$ and the other priors the same as for Model 2*.

To check the results were robust to the choice of priors, we set each parameter in turn to have an uninformative prior, and fitted Model 2* again. Specifically, we used

these priors:

$$\mu_{c,f} \sim \text{U}(-\infty, \infty) \quad (5.10)$$

$$\sigma_{c,f} \sim \text{U}(-\infty, \infty) \quad (5.11)$$

$$\omega \sim \text{U}(1, \infty) \quad (5.12)$$

$$\beta_{age} \sim \text{U}(-\infty, \infty). \quad (5.13)$$

We chose a uniform prior on $\sigma_{c,f}$ following the recommendation of Gelman *et al.* (2006).

We fitted these models on the paternal crossovers too, using the same priors as above, with the following exceptions:

$$\left. \begin{aligned} \mu_c &\sim \text{N}(27, 64) \\ \sigma_c^2 &\sim \text{Inv-Gamma}(2, 15) \end{aligned} \right\} \text{for Model 1} \quad (5.14)$$

$$\mu_c \sim \text{N}(3.2, 0.2) \text{ for Model 2.} \quad (5.15)$$

We fitted all models using a Markov Chain Monte Carlo (MCMC) method, specifically the No U-Turn Sampler extension to Hamiltonian Monte Carlo (HMC), implemented in RSTAN (Stan Development Team, 2014). HMC takes a series of steps informed by the first-order gradient of the log-posterior in order to avoid the random walk behaviour inherent in other MCMC methods such as Metropolis and Gibbs sampling (Hoffman and Gelman, 2011).

We used 4 chains with 10,000 draws each, and discarded the first half as burn-in. Convergence was assessed using the $\hat{R}$ statistic, which measures the between- versus within-chain variance (Gelman *et al.*, 2004). This should be close to 1 if convergence has been reached. If $\hat{R}$ exceeded 1.05 for any parameter, the chains were run for a further 10,000 draws or until $\hat{R}$ was less than 1.05. For some models, we also tried running the MCMC for 100,000 draws, discarded the first half, and thinned the remainder by taking every 100th draw. The 95% credible intervals were almost identical, and so we report results from 10,000 draws here.

5.2.7 Posterior predictive checks

To assess the fit of these hierarchical models, we carried out posterior predictive checks to compare empirical properties of the observed data, Y , to data simulated under a model with the estimated parameters, which we denote Y^{rep} . For all parameters in all models, the draws were concentrated in a single mode, and we took the draw with the highest posterior density as the Bayesian estimator. We denote the parameter estimates from this draw as $\hat{\beta}_{age}$, $\hat{\mu}_c$ etc. Using these estimated parameters, we simulated 1000 datasets of the same size as the actual dataset, starting from the highest level of the hierarchy in the model. As an example, for Model 1, we simulated $\alpha_{j,c}$ for each mother from $N(\hat{\mu}_c, \hat{\sigma}_c^2)$, then simulated $Y_{i,j,c}^{rep}$ for each child from $N(\hat{\beta}_{age} \times \text{age}_{i,j,c} + \alpha_{j,c}, \hat{\tau}^2)$. Then we compared these simulated datasets to the observed data by examining the empirical cumulative distribution functions and distributions of summary statistics that are ancillary to the inference. We also tried simulating datasets using 1000 different sets of parameters drawn from the joint posterior distribution and comparing these to the observed data. Results were very similar to those from the simulations using the draw with the highest posterior density, so we show only results for the latter.

5.2.8 Assessing the frequentist properties of the Bayesian method

We wanted to assess the probability that, if β_{age} had some true non-zero value $\beta_{age,true}$, our Bayesian procedure would have estimated it well. We therefore estimated the coverage and power of the credible intervals under the Bayesian model. Formally, the 95% credible interval has a $100\gamma\%$ coverage for a parameter θ if

$$\Pr(F^{-1}(0.025) < \theta < F^{-1}(0.975)) = \gamma, \quad (5.16)$$

and $100\pi\%$ power if

$$\Pr(F^{-1}(0.025) < 0 < F^{-1}(0.975)) = 1 - \pi, \quad (5.17)$$

where $F^{-1}(n)$ is the $100n^{th}$ quantile of the cumulative distribution function of the posterior of θ .

We selected 10 bins of β_{age} such that there were at least 1000 draws in each bin, and then sampled, with replacement, 100 sets of parameters from among those draws.

Note that we did this rather than specifying a particular value of β_{age} to test, since all the parameters in the model needed to be consistent. For each of those sets of parameters, we simulated a dataset and fitted the model on it in RSTAN, using 4 chains with 10,000 draws. We then asked in what proportion of datasets the credible interval, at different levels, overlapped 0 (1-power) or overlapped the value of β_{age} used for simulation (coverage).

5.3 Results

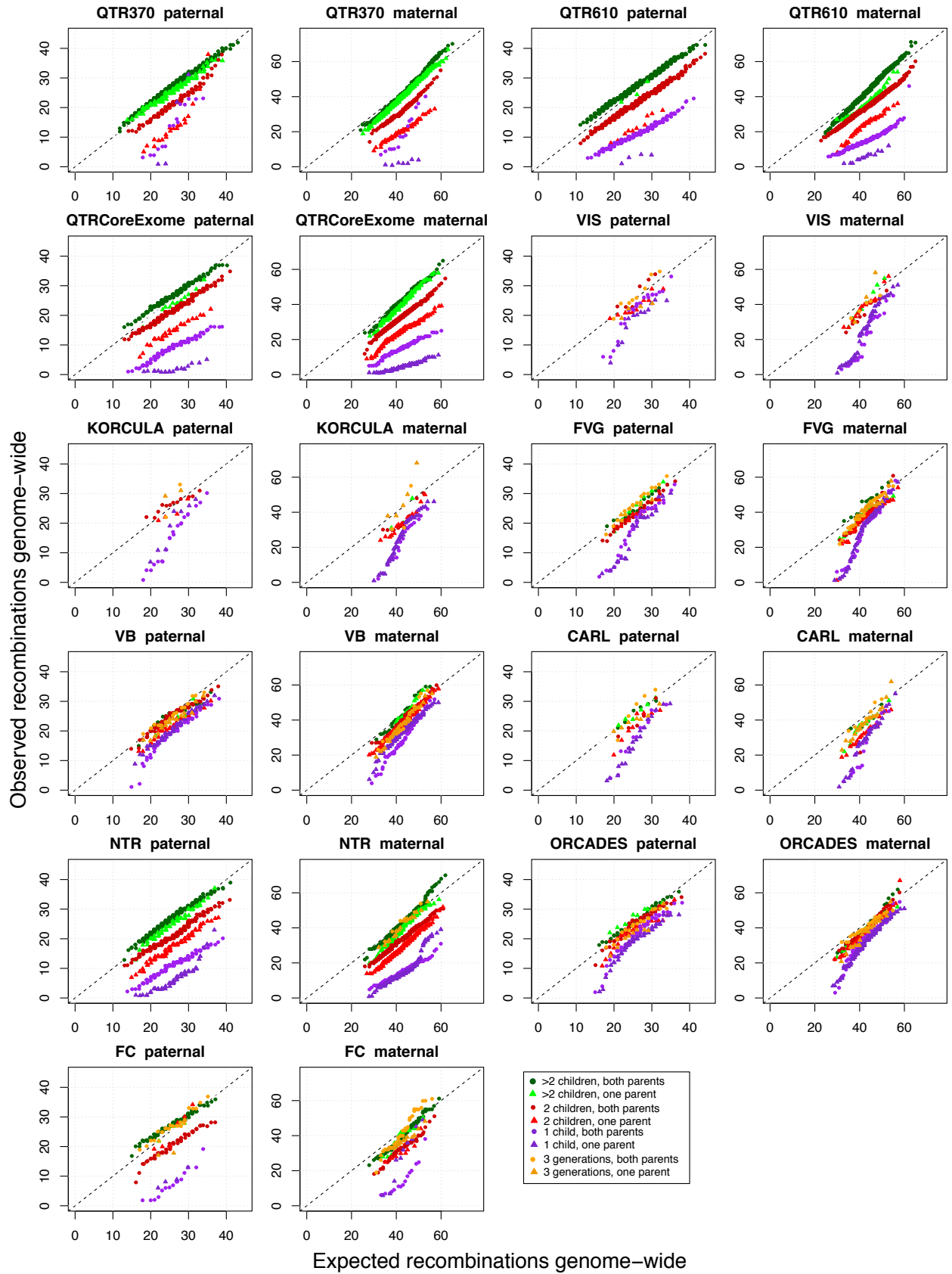
5.3.1 Comparison of duoHMM to NFTOOLS

We applied two methods to call crossovers, NFTOOLS (Hussin *et al.*, 2011) and duoHMM (O’Connell *et al.*, 2014) (Section 5.2.4). The number of crossovers called by NFTOOLS agreed well with expectation (Figure 5.A.1), but duoHMM seemed to be over-calling for several of the cohorts, notably those from more outbred populations (QTR, NTR and FC) (Figure 5.A.2; Tables 5.A.3, 5.A.4 and 5.A.5). It seemed likely this was due to phasing errors, since one would expect the phasing to be better in cohorts with more low-level relatedness. We found that removing double crossovers within short intervals, which were probably due to switch or genotype errors, brought the number in line with expectation for all cohorts (Figure 5.A.2). All duoHMM results described below were based on these filtered calls, but we also obtained very similar results using the raw calls (compare Table 5.1 and 5.A.2). There was good agreement between the final NFTOOLS and duoHMM callsets on the same meioses (i.e. from informative nuclear families) (Figure 5.A.3), with 96% of crossovers overlapping. We estimated that 70-76% of maternal crossovers occurred in HapMap II recombination hotspots, after correcting for the number expected by chance (Tables 5.A.6 and 5.A.7). This is similar to previous reports in non-African cohorts (Hussin *et al.*, 2011; Bleazard *et al.*, 2013; Fledel-Alon *et al.*, 2009).

5.3.2 Linear mixed models and meta-analysis

Exploratory analyses of fully informative meioses suggested a small increase in the number of crossovers with maternal age (Figure 5.2), with some suggestion of a larger increase in mothers over 35, as described by Coop *et al.* (2008) and Kong *et al.* (2004), although not in mothers over 39, as described by Campbell *et al.* (2015). However,

Figure 5.1 (*following page*): **Distribution of number of crossovers called by duoHMM.** QQ-plots for the observed versus expected number of crossovers called by duoHMM, after filtering. The expected counts were generated from a Poisson distribution using the genetic lengths from the deCODE 2002 map (Kong *et al.*, 2002). The number of crossovers called is in line with expectation for fully informative meioses (green and yellow squares), but somewhat lower than expected for families in which only one parent or fewer than three children were genotyped. The extent to which this happens depends on the cohort, reflecting differences in sample size and in the amount of low-level relatedness, since these factors affect the phasing accuracy.



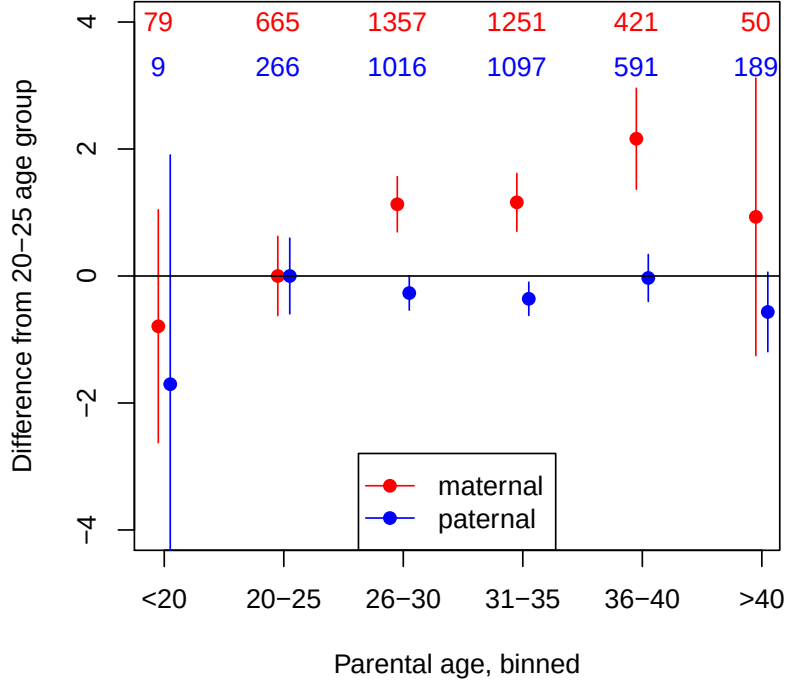


Figure 5.2: **Plot of the number of crossovers as a function of binned parental age.** Ages are grouped into 5-year bins, relative to parents of between 20 and 25 years old. The number of meioses in each age bin is indicated along the top, in red for maternal and blue for paternal. Note that age data were not always available for both parents, so the total sample size differs for maternal and paternal meioses. Points show means and error bars show 95% confidence intervals. This plot is based on data from fully informative duos analysed with duoHMM.

this simple analysis does not account for differences in baseline recombination rates between mothers. Thus, for each cohort separately, we fitted a linear mixed model to the duoHMM counts from fully informative meioses, treating mothers as a random effect (Table 5.A.8; see Table 5.A.9 for paternal crossovers), following Kong *et al.* (2004) and Hussin *et al.* (2011). Estimates of the effect of maternal age on the number of crossovers, β_{age} , varied between -0.26 for the FC cohort, to 0.34 for VB (Figure 5.3; Table 5.A.8), but only the three QTR cohorts were significantly different from 0 at the 0.05 level (QTR370: $\beta_{age} = 0.09$, $p = 0.03$, QTR610: $\beta_{age} = 0.15$, $p = 0.01$, QTRCoreExome: $\beta_{age} = 0.28$, $p = 0.01$).

Combining the estimates from all cohorts in a fixed-effects meta-analysis, we found that the effect of maternal age was small but significant ($\beta_{age} = 0.09$, $p = 0.002$; Z-test) (Table 5.A.10). Similar results were obtained using only informative nuclear

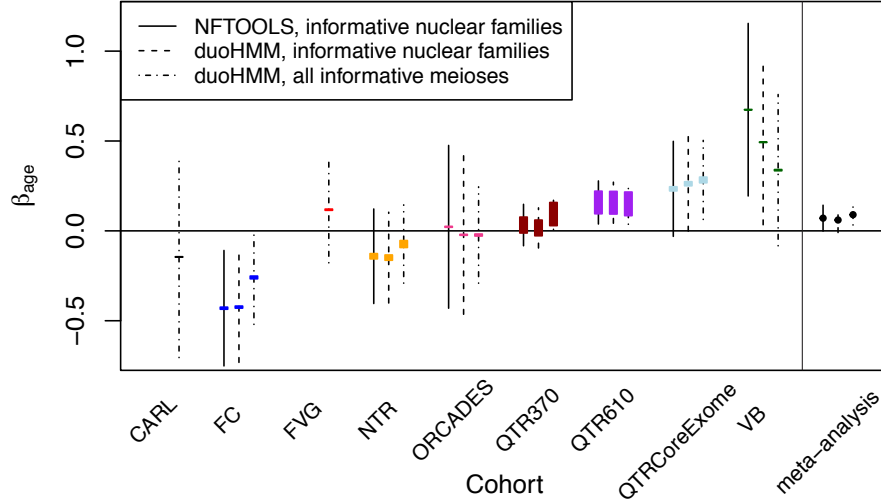


Figure 5.3: **Estimates of the maternal age effect from linear mixed models.** The length of the coloured boxes is proportional to sample size. The black lines show 95% confidence intervals, and the line type indicates which dataset was used. On the far right are estimates from the meta-analysis, indicated as points with their confidence intervals.

families with both NFTOOLS and duoHMM data (Figure 5.3, Tables 5.A.8 and 5.A.10), but the effects for FC and VB were more extreme ($\beta_{age} = 0.49, p = 0.049$ for VB; $\beta_{age} = -0.42, p = 0.008$ for FC, using duoHMM counts). It seems plausible that these estimates were more affected by outliers, given the smaller sample size imposed by restricting analysis to informative nuclear families (Figure 5.4), but we discuss this further below. In the meta-analysis, there was significant heterogeneity in the estimated maternal age effects between cohorts, as assessed by Cochran’s Q test (Table 5.A.10).

5.3.3 Bayesian analysis of the maternal age effect

Hierarchical models that share information between cohorts, while allowing for differences between them, may add power to discern an effect, and have a more natural interpretation. We fitted a variety of such models in a Bayesian framework (see Methods). We found very similar results for NFTOOLS and duoHMM counts on the same set of meioses from informative nuclear families (Table 5.A.11; Figure 5.5 *versus* 5.A.4). Thus, we focus on the duoHMM results below, because this method also allows us to include meioses from all families, not just informative nuclear families,

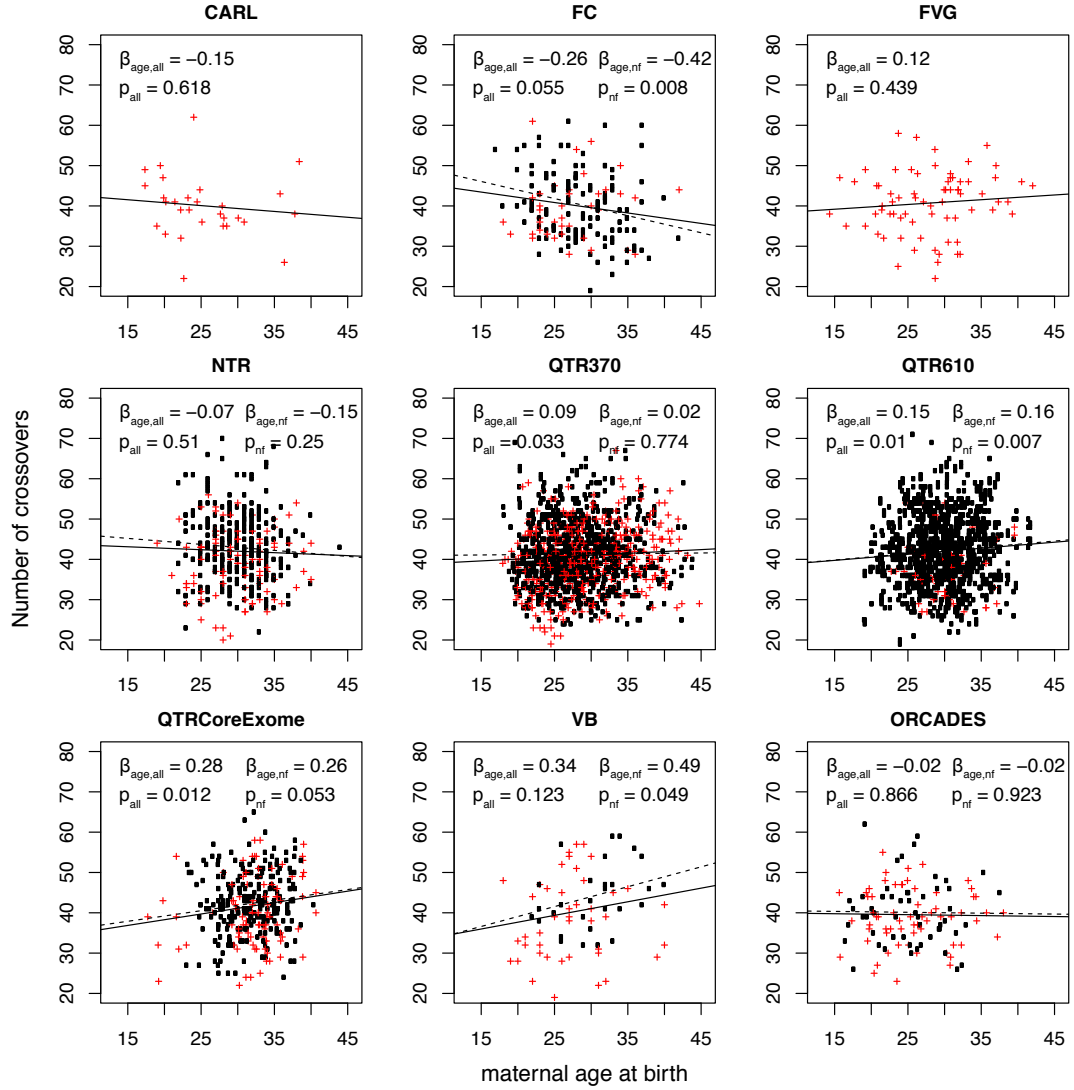


Figure 5.4: **Scatter plots of the number of crossovers versus maternal age for each cohort.** Counts are from duoHMM. Meioses from informative nuclear families (as could be analysed by NFTOOLS) are shown as black dots, and those from families with only two children but a third generation (fully informative for duoHMM) as red crosses. The lines are from a linear mixed model on the number of crossovers versus maternal age, fitted to either all fully informative meiosis (solid line) or only to meiosis from informative nuclear families (dotted line). The corresponding slopes and their p-values are indicated on the left for all fully informative meiosis (p_{all}) and on the right for meiosis from informative nuclear families (p_{nf}).

and hence to substantially increase the number studied.

We began by analysing only the fully informative meioses (n=3823 maternal meioses from 1132 families; 3558 paternal meioses from 1073 families). Figure 5.5 shows the posterior distribution of the age effect, β_{age} , under a normal model for the number of crossovers in which the mother (or father) effects for parent j in cohort c , $\alpha_{j,c}$, are drawn from a normal distribution with cohort-specific mean μ_c and variance σ_c^2 , with the cohort-specific means and variances being independently sampled from a common distribution across cohorts. Posterior distributions for the other parameters are shown in Figure 5.A.5. We fitted either a common β_{age} for all cohorts (Model 1; Figure 5.5 top), or a different $\beta_{age,c}$ for each cohort c , drawn from the same normal distribution with mean $\beta_{age,global}$ and variance $\sigma_{age,global}^2$ (Model 1.2; Figure 5.5 bottom). The results for Model 1 fitted to maternal crossovers (Figure 5.5a) strongly support a small positive effect (Table 5.1): the 95% credible interval was [0.0267, 0.1451], consistent with the effect found by Kong *et al.* (2004) using microsatellite markers, and 99.8% of the posterior distribution for β_{age} was concentrated on positive values. The results are not consistent with the larger positive effect reported in the Hutterites by Coop *et al.* (2008) or the negative effects reported in the original analysis of the French Canadians by Hussin *et al.* (2011) or in the Asian cohort by Bleazard *et al.* (2013).

Table 5.1: Summary of posteriors for β_{age} for different Bayesian models. This table shows several quantiles of the posterior of β_{age} for Model 1 and of $e^{\beta_{age}}$ for Models 2 and 2*, as well as the posterior probability that β_{age} is greater than 0. Models 1 and 2 were fitted to fully informative meioses, and Model 2* to fully and partially informative meioses. Note that the interpretation of β_{age} is additive for Model 1 and multiplicative for Models 2 and 2*. These results are for duoHMM counts.

Parent	Model	2.5%	25%	50%	75%	97.5%	$\Pr(\beta_{age} > 0)$
Maternal	Model 1	0.02672	0.06608	0.08582	0.10706	0.14513	0.99803
Maternal	Model 2	1.00081	1.00171	1.00217	1.00266	1.00367	0.99963
Maternal	Model 2*	1.00098	1.00172	1.00213	1.00251	1.00333	0.99995
Paternal	Model 1	-0.04766	-0.02582	-0.01492	-0.00410	0.01769	0.17883
Paternal	Model 2	0.99820	0.99906	0.99951	0.99997	1.00089	0.23465
Paternal	Model 2*	0.99849	0.99925	0.99965	1.00006	1.00082	0.28315

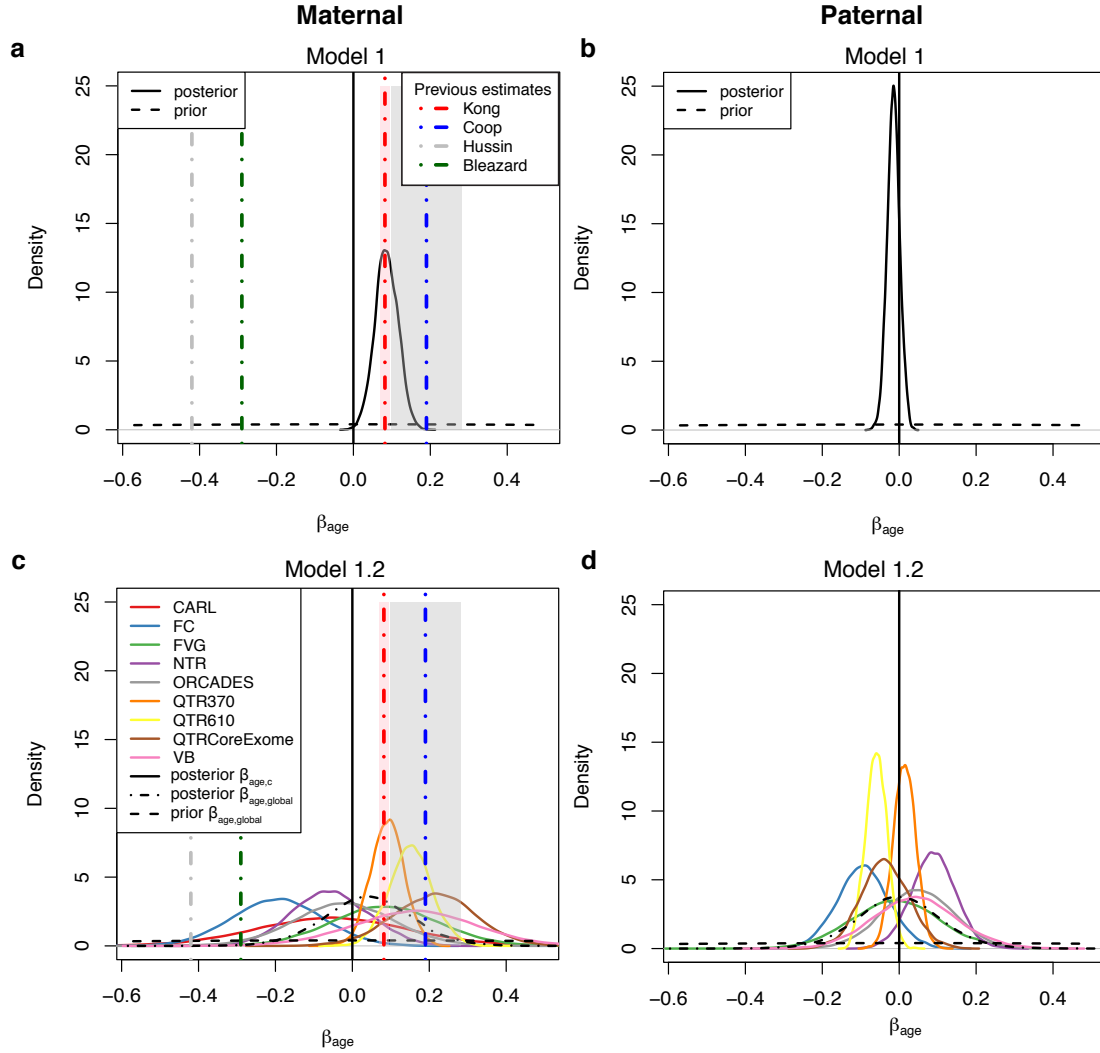


Figure 5.5: **Bayesian posteriors for the age effect from normal models fitted to fully informative meioses.** These plots show the priors (dashed lines) and posteriors (solid lines) for β_{age} from a normal model fitted to the number of maternal (left) or paternal (right) crossovers called in fully informative duos by duoHMM, with either the same (top plots) or different (bottom plots) age effects for each cohort (see Models 1 and 1.2 in Section 5.2.6). In the bottom plots, the posterior for $\beta_{age,global}$ is also indicated. The vertical lines show the estimates from previous studies, and the shaded boxes the corresponding standard errors, if reported. Under this model, the expected number of crossovers increases by β_{age} per year.

When we fitted a more general normal model with cohort-specific age effects (Model 1.2) to maternal meioses, there was some evidence for heterogeneity of effects, with FC showing the strongest evidence for a negative effect ($\Pr(\beta_{age,FC} < 0) = 0.95$), and QTR610 the strongest for a positive effect ($\Pr(\beta_{age,QTR610} > 0) = 0.98$).

(Figure 5.5c). Since it seemed more appropriate for count data, we also fitted a negative binomial model with a common age effect, in which the number of crossovers increased multiplicatively by a factor of $e^{\beta_{age}}$ per year (Model 2). The evidence in favour of a positive effect was very similar to that under the normal model (Figure 5.6; Figure 5.A.6): 99.96% of the posterior distribution for β_{age} was concentrated on positive values (Table 5.1).

The results presented thus far have all been based on data from 3823 fully informative maternal meioses from 1132 families. However, duoHMM can call crossovers in any duo, since it uses all individuals in the cohort for phasing. Its performance depends on the amount of low-level relatedness in the cohort, which affects phasing accuracy. Figure 5.1 shows that duoHMM is under-calling crossovers in meioses that are only partially informative, but to a different extent in each cohort, consistent with the results of O’Connell *et al.* (2014). We adapted the negative binomial model to allow for this effect by instead drawing the mother (or father) effects, $\alpha_{j,c,f}$, from a normal distribution with mean $\mu_{c,f}$ and variance $\sigma_{c,f}^2$ corresponding to the family configuration f in cohort c (see Section 5.2.6 for a description of Model 2* and the family configurations). We fitted this to the 6011 maternal meioses from families with at least two children, requiring multiple children so that the mother effect could be estimated. With this increased sample size, the results were broadly similar (Figure 5.7; Figure 5.A.7), though the posterior on β_{age} (or $\beta_{age,c}$) was tighter than when using only the fully informative meioses (compare Figure 5.6 to 5.7).

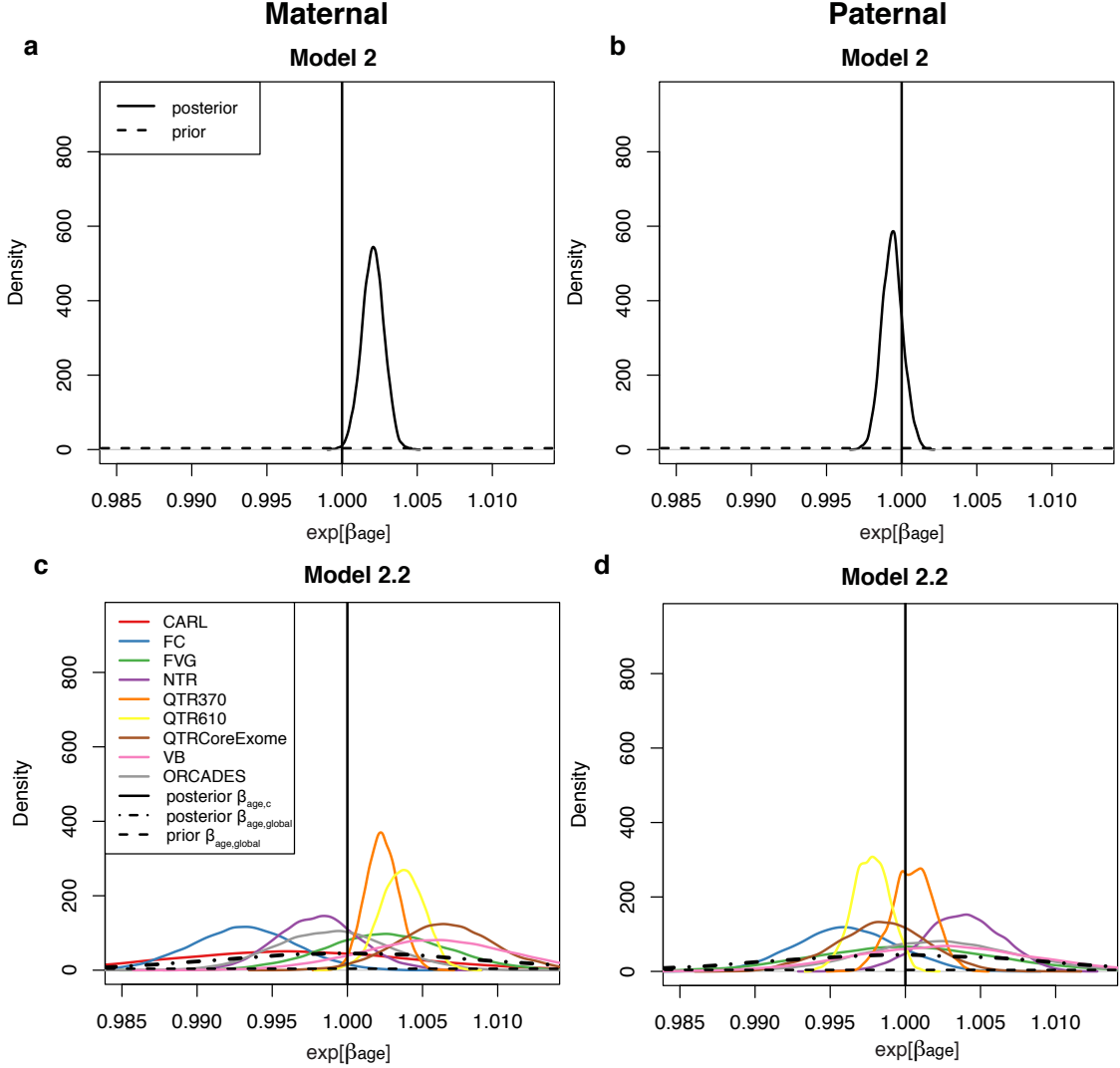


Figure 5.6: **Bayesian posteriors from a negative binomial model fitted to fully informative meioses.** These plots show the priors and posteriors for $e^{\beta_{age}}$ from a negative binomial model fitted to the number of crossovers called by duoHMM in informative duos, with either the same (Model 2) or different (Model 2.2) age effects for each cohort (see Section 5.2.6). We have plotted $e^{\beta_{age}}$ since, under the negative binomial model, the expected number of crossovers increases by this factor per year. The axes have been chosen to facilitate comparison with the normal model shown in Figure 5.5, assuming a baseline of 38. i.e. the bounds $[0.985, 1.013]$ on $e^{\beta_{age}}$ under the negative binomial model are approximately equivalent to the bounds $[38 \times \log(0.985), 38 \times \log(1.013)] = [-0.57, 0.49]$ on β_{age} under the normal model.

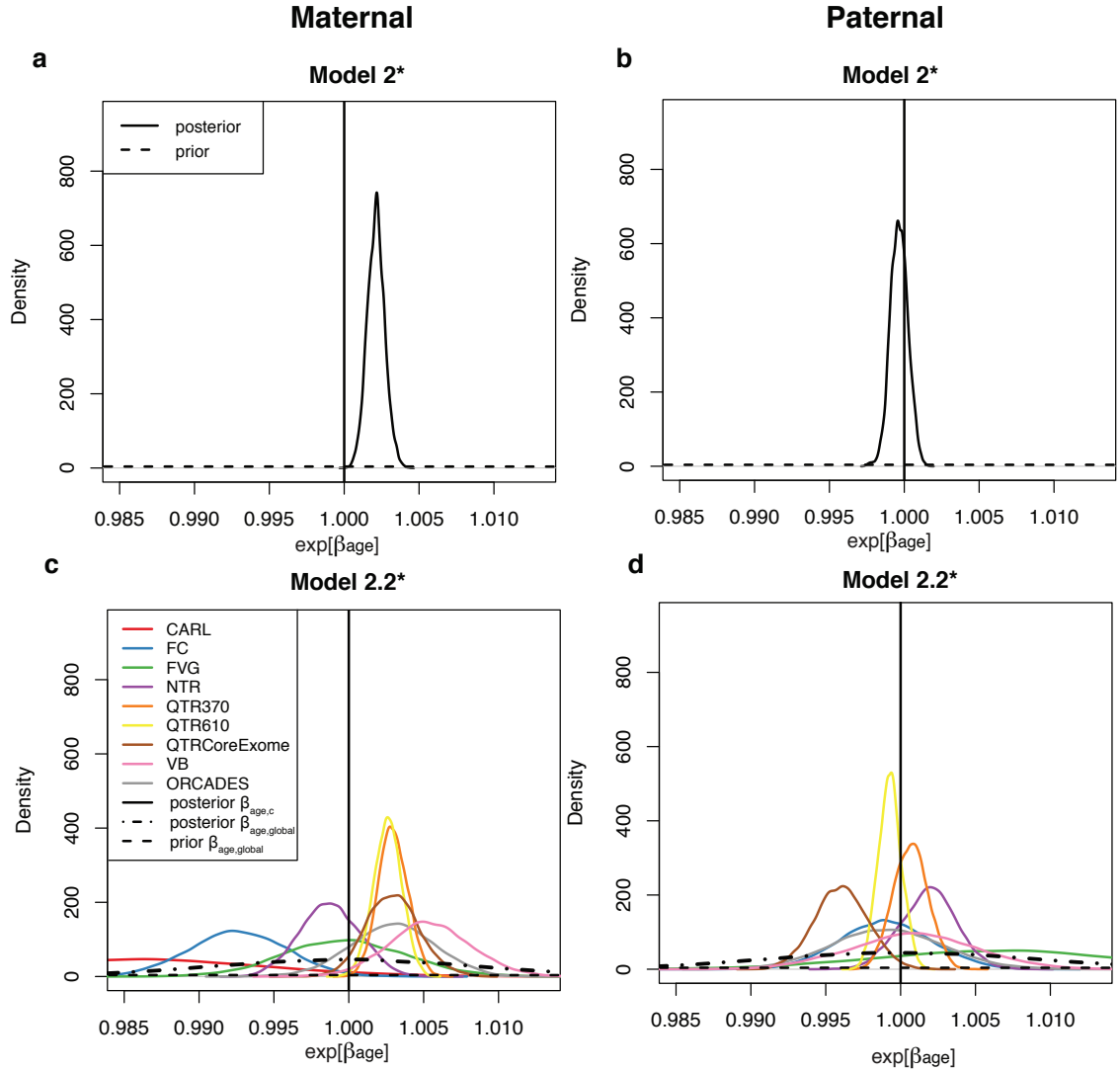


Figure 5.7: **Bayesian posteriors from a negative binomial model fitted to fully and partially informative meioses.** These plots show the priors and posteriors for $e^{\beta_{age}}$ from a negative binomial model fitted to the number of crossovers called by duoHMM in both informative and partially informative duos, with either the same (Model 2*) or different (Model 2.2*) age effects for each cohort (see Section 5.2.6 and legend to Figure 5.6).

It is important to emphasise that the interpretation of β_{age} is different for the negative binomial models *versus* the normal model. In the normal model (Model 1), the number of crossovers is expected to increase linearly from the baseline by β_{age} per year, whereas in the negative binomial models (Models 2 and 2*), it increases multiplicatively by a factor of $e^{\beta_{age}}$ per year. In practice, because β_{age} is so small, this makes very little difference. For our estimated β_{age} values, the multiplicative

$e^{\beta_{age}}$ -fold increase is close to a linear increase by $(\beta_{age} \times \text{baseline})$ per year, at least over a relatively small number of years.

To check that these results were robust to the choice of priors, we fitted Model 2* with an uninformative prior placed in turn on each parameter (see Methods). The inference about β_{age} was not substantially changed: in each model, at least 99% of the posterior density was concentrated on positive values (Figures 5.A.9, 5.A.10, 5.A.11 and 5.A.12).

All previous studies had found no significant effect of the father’s age at birth on the number of paternal crossovers (Kong *et al.*, 2004; Coop *et al.*, 2008; Hussin *et al.*, 2011; Bleazard *et al.*, 2013; Campbell *et al.*, 2015). Our results are consistent with this (Table 5.1; Figures 5.5, 5.6 and 5.7; Figures 5.A.4, 5.A.5, 5.A.6 and 5.A.8), but it is notable that there is some heterogeneity of estimated paternal age effects between cohorts. For example, for Model 1.2 (Figure 5.5d), we find $\Pr(\beta_{age} > 0) = 0.94$ for NTR, and 0.09 for FC. We discuss possible explanations for and implications of this heterogeneity later.

5.3.4 Posterior predictive checks

We conducted posterior predictive checks to assess how well these hierarchical models fitted the data, by comparing the distribution of the observed crossover counts to data simulated under a model with the estimated parameters (Figures 5.A.13, 5.A.14, 5.A.15, 5.A.16, 5.A.17). The median number of crossovers was similar between the observed and simulated data for all models. The minimum tended to be too low in data simulated under the normal model (Model 1), which is expected given that it has no lower bound. There was an asymmetry in the distribution of the observed data (Figure 5.A.13, red histograms), and, for the fully informative meioses, the negative binomial model (Model 2) captured this better than the normal model (Model 1) (Figure 5.A.17).

5.4 Discussion

We have investigated the maternal age effect on recombination in data from over 6000 meioses from nine cohorts. This is by far the largest SNP-based sample analysed for

this question. Kong *et al.* (2004) studied a much larger sample of over 14,000 meioses, using 1000 microsatellite markers, and had age data rounded to the nearest 5 years.

To estimate the age effect, we first applied the same statistical approach as Kong *et al.* (2004) and Hussin *et al.* (2011), namely a linear mixed model (Figure 5.3; Table 5.A.8). We found significant effects in some cohorts (positive in VB and the QTR cohorts, negative in FC), and a meta-analysis of all cohorts combined showed a small and significant positive effect (Table 5.A.10).

In addition to the approach of earlier authors, we also fitted various hierarchical models that accounted for differences between mothers and cohorts, fitting these in a Bayesian statistical framework. Model 2* allowed us to account for the uncalled crossovers in partially informative meioses, so we could increase the sample size by including these. These hierarchical models had good frequentist properties (see Section 5.2.8), including low type I error, and there was good power to detect a multiplicative effect greater than about 1.0015 crossovers per year with the informative meioses (Figure 5.A.18).

Our results on 3823 fully informative meioses (posterior mean for $\beta_{age} = 0.0862$ crossovers/year) were very similar to the small positive maternal age effect reported by Kong *et al.* (2004) (0.082 crossovers/year). They are also in line with the estimate reported recently by Campbell *et al.* (2015) ($\beta_{age} = 0.067$ crossovers/year), although, notably, this effect became nonsignificant when they restricted the analysis to the families in which crossovers could be assigned to children (2184 meioses). There was extremely strong support for a positive maternal age effect: 99.96% of the posterior density was concentrated on positive values (Model 2 fitted to informative meioses). When we included partially informative meioses (Model 2*), increasing the sample size to 6011 meioses, this gave a tighter posterior on the age effect, and we estimated, using the posterior mean of β_{age} , that the expected number of crossovers increases by 2% over 10 years (95% credible interval [1.1%,3.0%]). Translation between parameters in this multiplicative model and the additive model used here and by others for fully informative meioses is not immediate. Nonetheless, assuming a baseline number of crossovers of 38, as inferred from the data (Figure 5.A.7), shows the multiplicative parameter estimate to be in good agreement with the additive change of ~ 0.08 crossovers/year inferred here and by Kong *et al.* (2004).

Following previous studies (Coop *et al.*, 2008; Hussin *et al.*, 2011), we removed double crossovers over short intervals since these were likely to be due to genotyping or, for duoHMM, switch errors. This brought the distribution of the number of crossovers closer to expectation for the QTR, NTR and FC cohorts. If, however, some of these double crossover events were real, this would lead to underestimation of the number of crossovers, and if their frequency changed with age, as might be expected given the recent report by Campbell *et al.* (2015), this could be a confounding factor in the analysis presented here. We thus fitted the hierarchical models again using the raw duoHMM data. The 95% credible intervals for β_{age} were very similar to those obtained using the filtered data (compare Table 5.A.2 to Table 5.1), demonstrating that the filtering did not greatly affect the results.

Although we applied the same methods to all cohorts, there was still substantial heterogeneity between cohorts in the magnitude and direction of the maternal age effect. There were differences between cohorts in the number of SNPs (Table 5.A.1), but although this does affect μ_c , the mean of the mother effects for cohort c (Figures 5.A.5, 5.A.6, and 5.A.7), it is unlikely to affect the estimate of $\beta_{age,c}$, since, within a cohort, there is no association between maternal age at birth and the number of SNPs. There could also be inter-cohort differences in family size, which has previously been found to be associated with recombination rate (Kong *et al.*, 2004; Coop *et al.*, 2008). If women with higher baseline recombination rates have more children later in life, the number of recombinations will appear to be higher in older mothers. We could not explicitly include family size as a covariate in the analyses, since we did not have that information, but the model should, in theory, be capturing this effect with the mother effect, α . Since the α for each mother is drawn from the same distribution as the other mothers in the cohort, one might expect some shrinkage towards the mean mother effect. This could potentially, for example, reduce the estimate of α for high-recombining mothers and thus inflate $\beta_{age,c}$. However, it is not clear how such a phenomenon could cause the negative $\beta_{age,c}$ estimates in some cohorts, notably in the French Canadians. Notably, there was still heterogeneity when fitting the linear mixed model to each cohort separately (Figure 5.3), in which the mother effects are independent of one another, so it seems unlikely differences in family size can account for this.

There is information about the levels of between-cohort heterogeneity expected

under the null from analyses of paternal meioses. All previous studies (and the analyses presented here) have suggested that there is no paternal age effect on numbers of crossovers. Further, the explanations proposed for the maternal age effect on aneuploidies (i.e. the production line model (Henderson and Edwards, 1968), or loss of cohesin during the long meiotic arrest (Chiang *et al.*, 2010)) are not relevant to males. There is clearly still some heterogeneity between cohorts for paternal age effect estimates (Figures 5.5d, 5.6d and 5.7d), which must just reflect variation under the null hypothesis.

The number of crossovers was more over-dispersed in females than in males, relative to the mean: in Models 2 and 2*, the posteriors for ω , the over-dispersion parameter, put more weight on smaller values for males *versus* females (Figures 5.A.6, 5.A.7 and 5.A.8). One should thus expect more variation, and less precise estimation, in females than in males, and indeed this is the case in our data: the posteriors for β_{age} are tighter for fathers than for mothers (e.g. compare Figure 5.5c to 5.5d). Taken together, this suggests that the differences in β_{age} seen between cohorts may simply reflect chance effects. Notably, the large negative effect reported by Hussin *et al.* (2011) in informative nuclear families from the French Canadians was attenuated when the sample was expanded from 106 to 158 to include all fully informative meioses (i.e. adding families with only 2 children, but with a third generation; see Figure 5.4).

It is possible that the heterogeneity in effects between cohorts is due, at least in part, to biological differences or to confounders. If the relationship between maternal age and recombination rate were nonlinear, and if the cohorts had different maternal age distributions, this might explain why they show varying directions and magnitudes of effect. The age distributions do differ between cohorts (Figure 5.A.19), but there was no evidence for nonlinearity with age in the residuals (Figures 5.A.20, 5.A.21 and 5.A.22), so this not a likely explanation.

There are at least two possible confounders that may be contributing to the heterogeneity. The first is the use of assisted reproductive technologies (ART), which might interfere with normal recombination processes (e.g. *in vitro* fertilisation might select for eggs with more crossovers). These are not uncommon in mothers of twins, who make up a substantial fraction of our sample (the QTR and NTR cohorts). About 27% and 10% of the mothers of informative duos in the QTR and the NTR reported

they had had some form of ART. Stratifying QTR and NTR mothers by ART use in the hierarchical models did not reveal any clear differences in the age effect (Figure 5.A.23A), although we had limited power for these smaller sample sizes.

A second possible confounder is oral contraception. This suppresses ovulation, and so, if eggs are released in the order they are produced during foetal development, as happens in mice (Polani and Crolla, 1991), the age of eggs released at a certain point in a woman’s lifetime will differ depending on whether and for how long she has taken oral contraception. If there is an association between meiotic entry and recombination, as in the production line model suggested by Henderson and Edwards (1968), oral contraceptive use could confound the measured effect of maternal age on the number of crossovers. Importantly, a recent study argued that the production line model could not explain the maternal age effect on aneuploidies, because it found no evidence of a difference in recombination levels between oocytes entering meiosis early *versus* late in foetal life (Rowsey *et al.*, 2014). However, there is evidence that mothers over 40 who have taken oral contraception for longer have a significantly lower risk of having a foetus with a common trisomy (Nagy *et al.*, 2013), suggesting that suppressing ovulation could somehow affect the recombination levels of the eggs subsequently released. It is notable that the largest positive estimate of β_{age} previously reported was in the Hutterites (Coop *et al.*, 2008), most of whom probably do not use any birth control. There were some data on oral contraceptive use for the NTR and ORCADES cohorts. There were no clear difference in β_{age} between those who had taken oral contraception before the birth of their children and those who had not (Figure 5.A.23B). Nonetheless, there was reduced power for these smaller samples, and thus, the effects of oral contraception on recombination levels and aneuploidy risk merit further investigations in much larger samples for which more detailed data, for example on the length of oral contraceptive use, are available.

In conclusion, we analysed over 6000 meioses from across nine cohorts, and found that the maternal age effect on the number of crossovers was small and positive. These results rule out the larger positive and negative effects previously reported, and do not replicate the larger increase in number of crossovers for mothers over 39 described by Campbell *et al.* (2015) (Figure 1). Neither the results presented here nor those of other studies apply directly to either maternal recombination rates nor to the maternal age effect on aneuploidies because all studies count crossovers in live births. Since the chance of foetal survival may increase with the amount of

recombination, observed effects on numbers of crossovers may not reflect changes in the recombination rate in oocytes. Nonetheless, these results are still of inherent biological interest, and further insights into the recombination process could shed light on factors affecting aneuploidy rates. The positive effect observed can probably be explained by the selection hypothesis proposed by Kong *et al.* (2004): because the factors that cause aneuploidies increase with maternal age, eggs with more crossovers are more likely to be able to overcome these and give rise to a live birth in older mothers. There was some evidence for heterogeneity between cohorts, and we can exclude differences in crossover-calling methodology or in statistical approaches as explanations for this. It may instead be attributable to hidden confounders, possibly including, but not limited to, assisted reproductive technologies or oral contraception, or simply to chance effects in smaller samples. Thus, these findings also emphasise that care should be taken in interpreting results from any specific cohort.

Appendix

5.A Appendices

5.A.1 Supplementary Figures

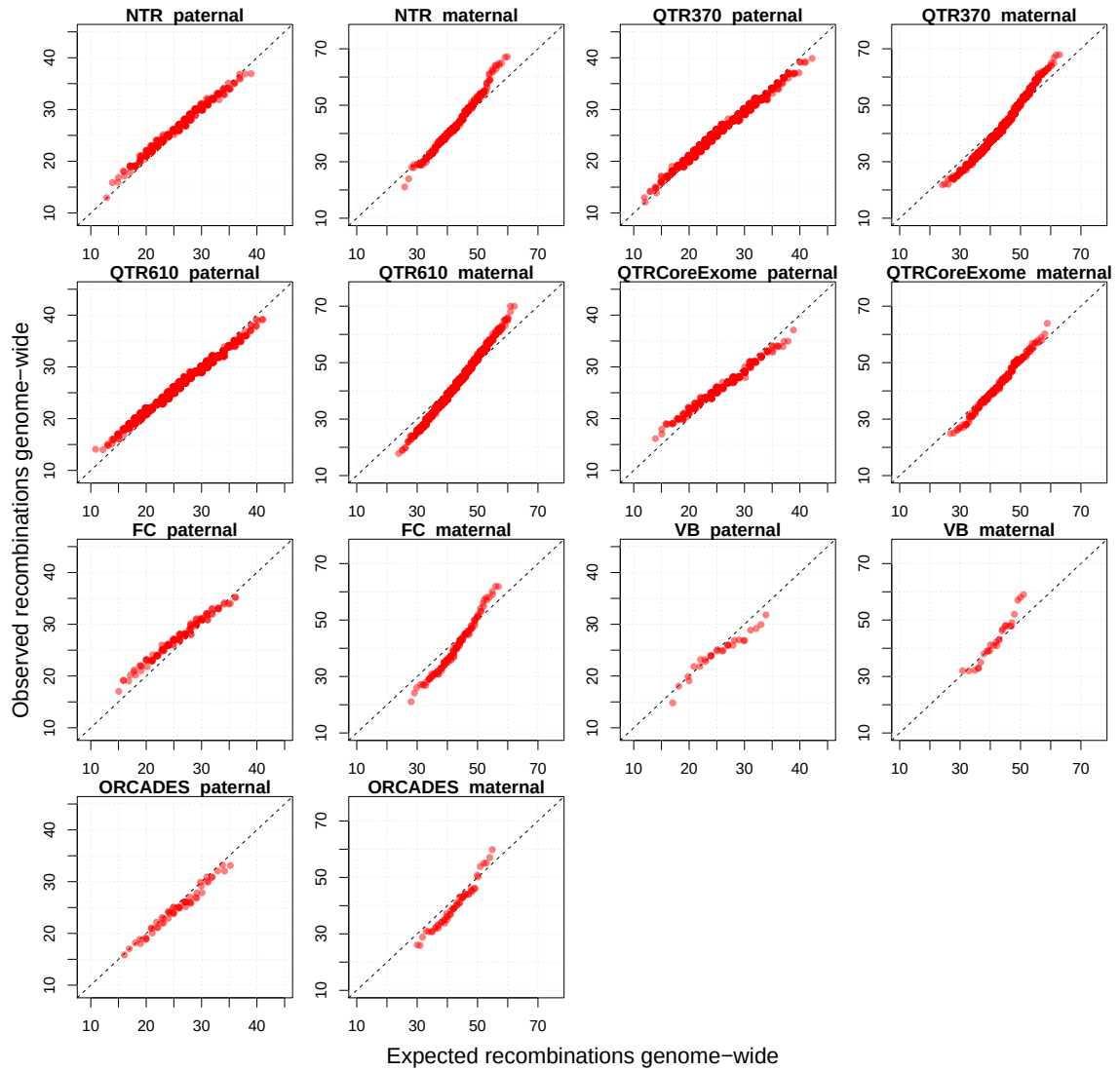


Figure 5.A.1: **Distribution of number of crossovers called by NFTOOLS.** QQ-plots for the observed versus expected number of crossovers called by NFTOOLS for paternal and maternal meioses in each cohort. The expected counts were generated from a Poisson distribution using the genetic lengths from the deCODE 2002 map (Kong *et al.*, 2002). Note that the points have been jittered to aid visualisation, so the counts do not always appear to be strictly increasing.

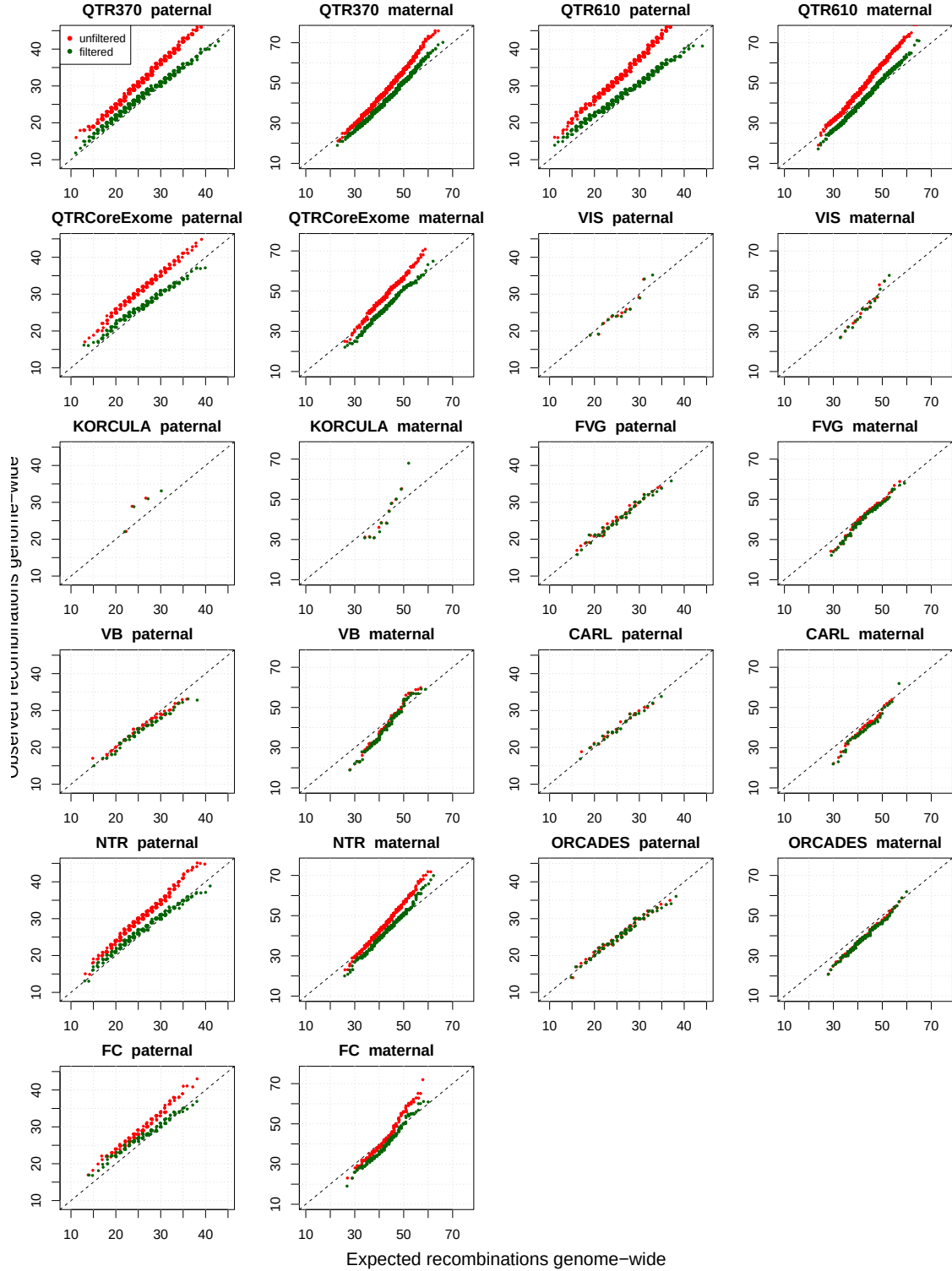


Figure 5.A.2: **Distribution of number of crossovers called by duoHMM, before and after filtering.** The qq-plots were constructed as for Figure 5.A.1, but using data from informative meioses called with duoHMM. Red points show the raw counts before filtering, and green points the filtered counts (see Section 5.2.4).

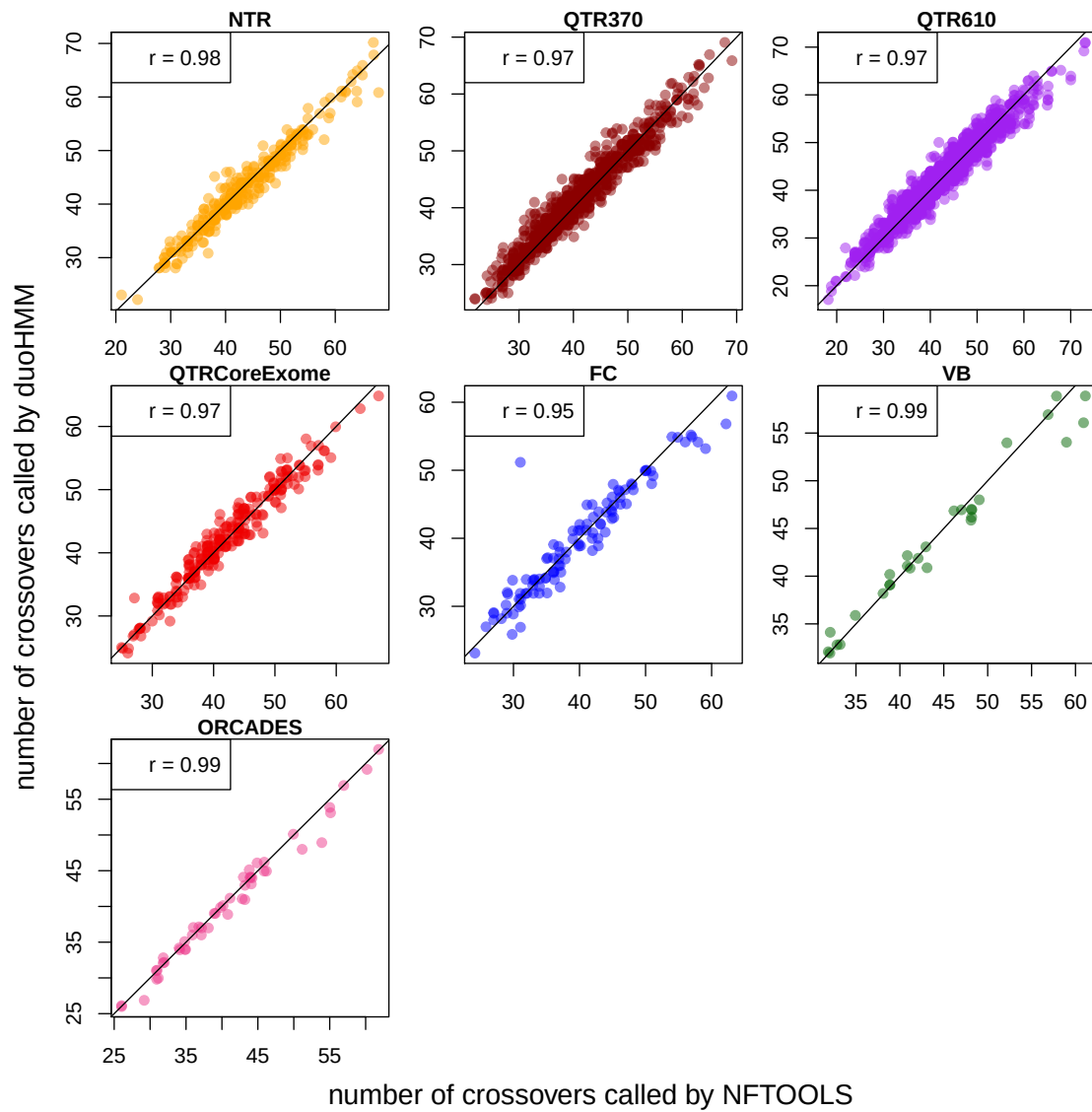


Figure 5.A.3: Comparison of number of maternal crossovers called by duoHMM versus NFTOOLS in informative nuclear families. The Pearson correlation is shown in the top left.

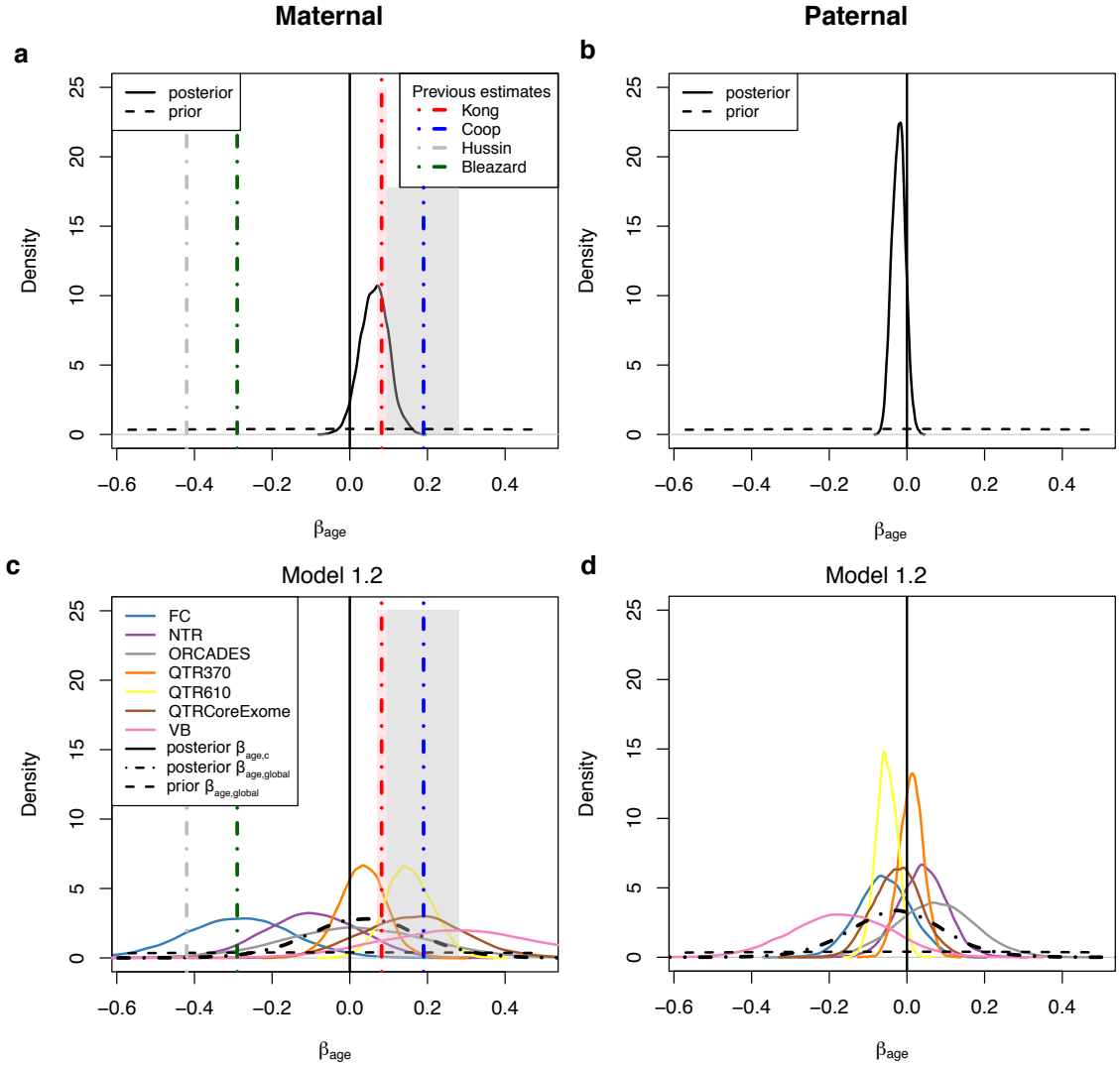


Figure 5.A.4: **Bayesian posteriors for the age effect, with NFTOOLS data.** These plots show the priors and posteriors for β_{age} from Models 1 and 1.2 fitted to the number of crossovers called in informative nuclear families by NFTOOLS. For Model 1.2, the posterior for $\beta_{age,global}$ is also indicated. The vertical lines show the estimates from previous studies, and the shaded boxes the corresponding standard errors, if reported. Note that these results are based on meioses from informative nuclear families, whereas Figure 5.5 is based on all informative meioses.

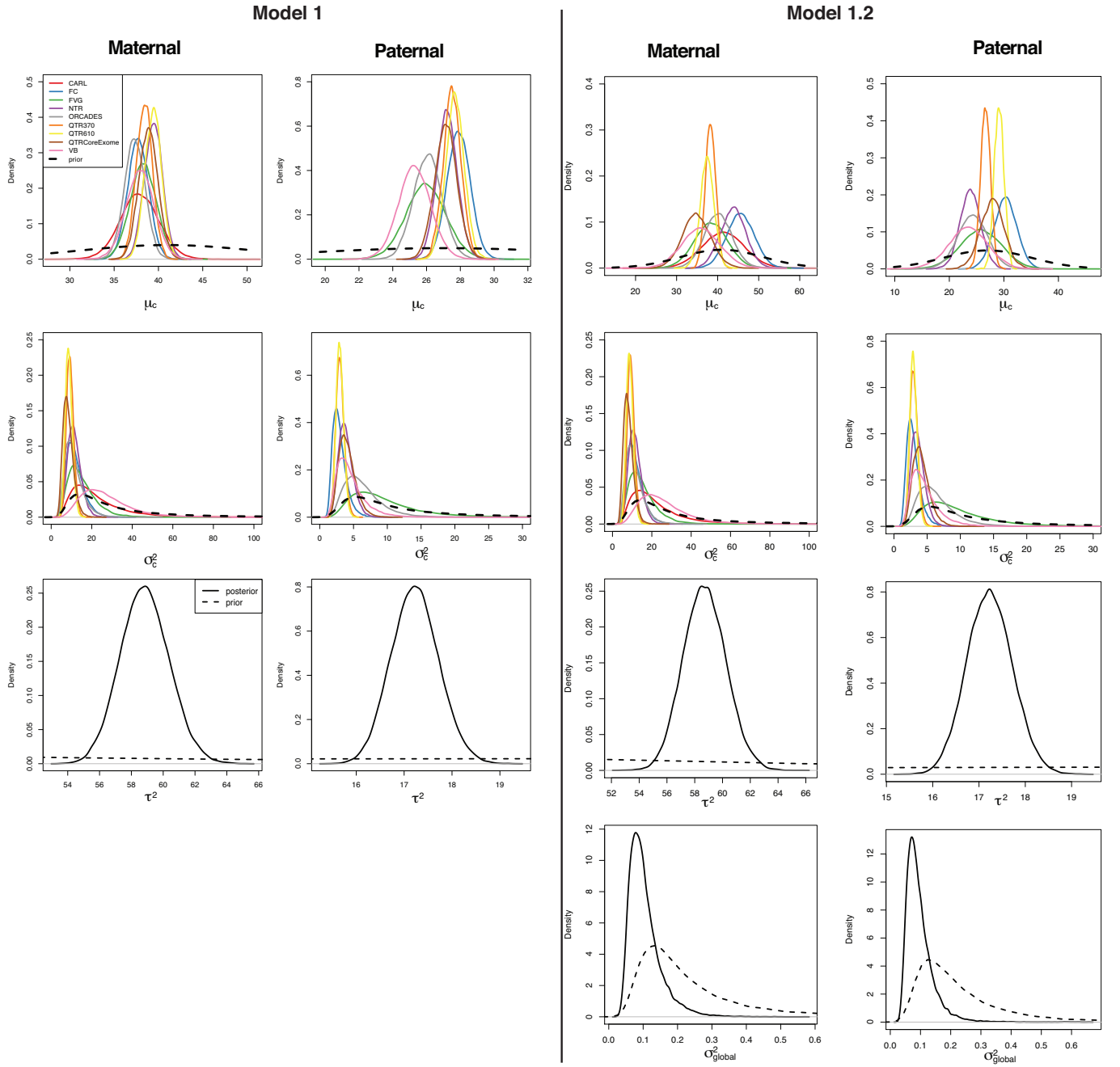


Figure 5.A.5: **Bayesian posteriors for parameters from Models 1 and 1.2, for duoHMM data.** We fitted a normal model on the number of crossovers called in fully informative meioses, with either the same or different age effects for all cohorts (Model 1 or 1.2; see Methods). The posteriors for the age effects are shown in Figure 5.5. This figure shows the posteriors for μ_c and σ_c^2 , the cohort-specific mean and variance of the parent effects respectively, τ^2 , the residual variance, and $\sigma_{age,global}^2$, the variance of the cohort-specific age effects.

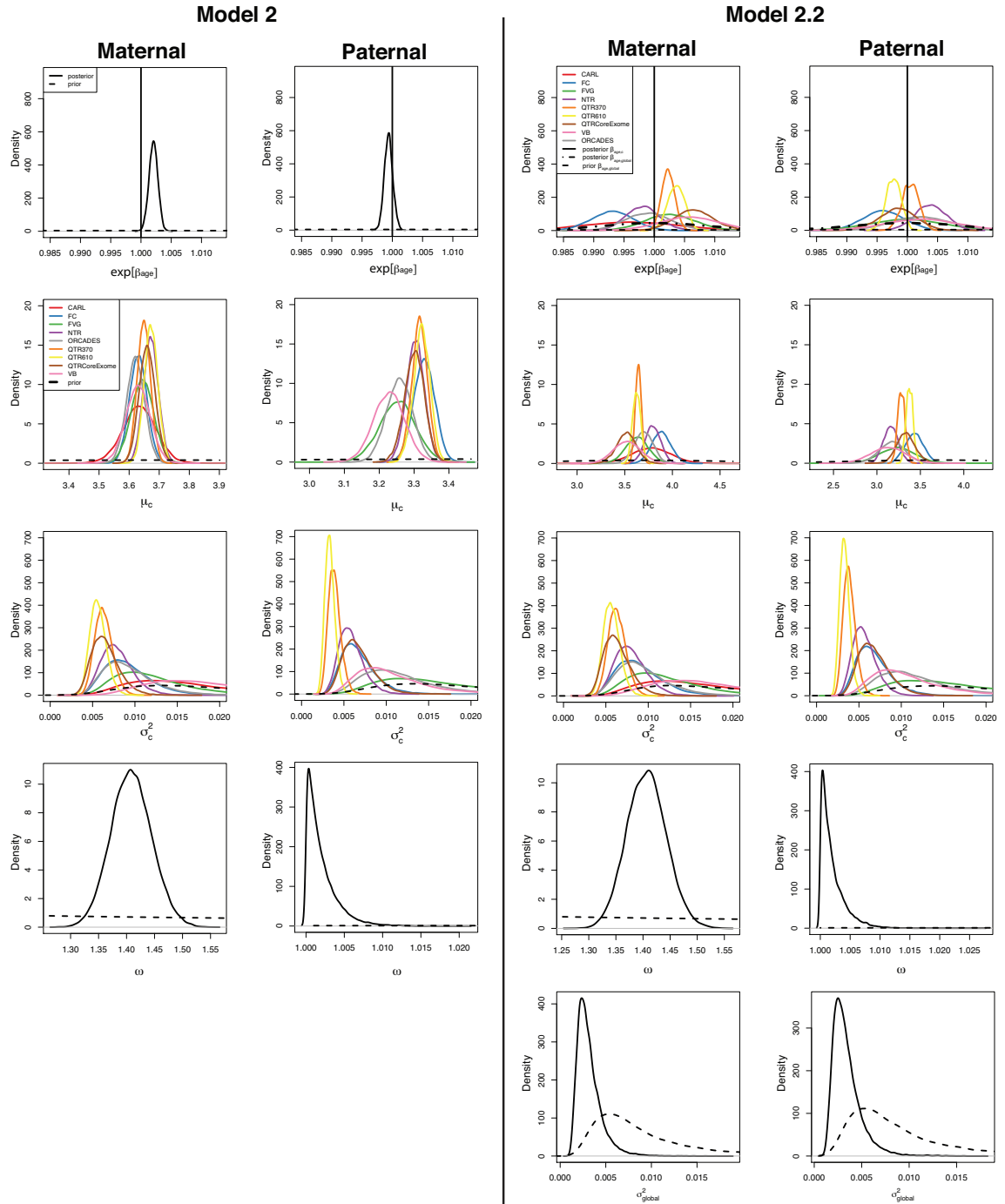


Figure 5.A.6: **Bayesian posteriors from a negative binomial model fitted to fully informative meioses.** These plots show results from Model 2, described in the Methods, fitted to duoHMM counts from informative duos. Parameters are as described for Figure 5.7 and Figure 5.A.5, except that ω is the over-dispersion parameter, replacing τ^2 from the normal model. Note that the expected number of crossovers increases by $e^{\beta_{age}}$ -fold per year.

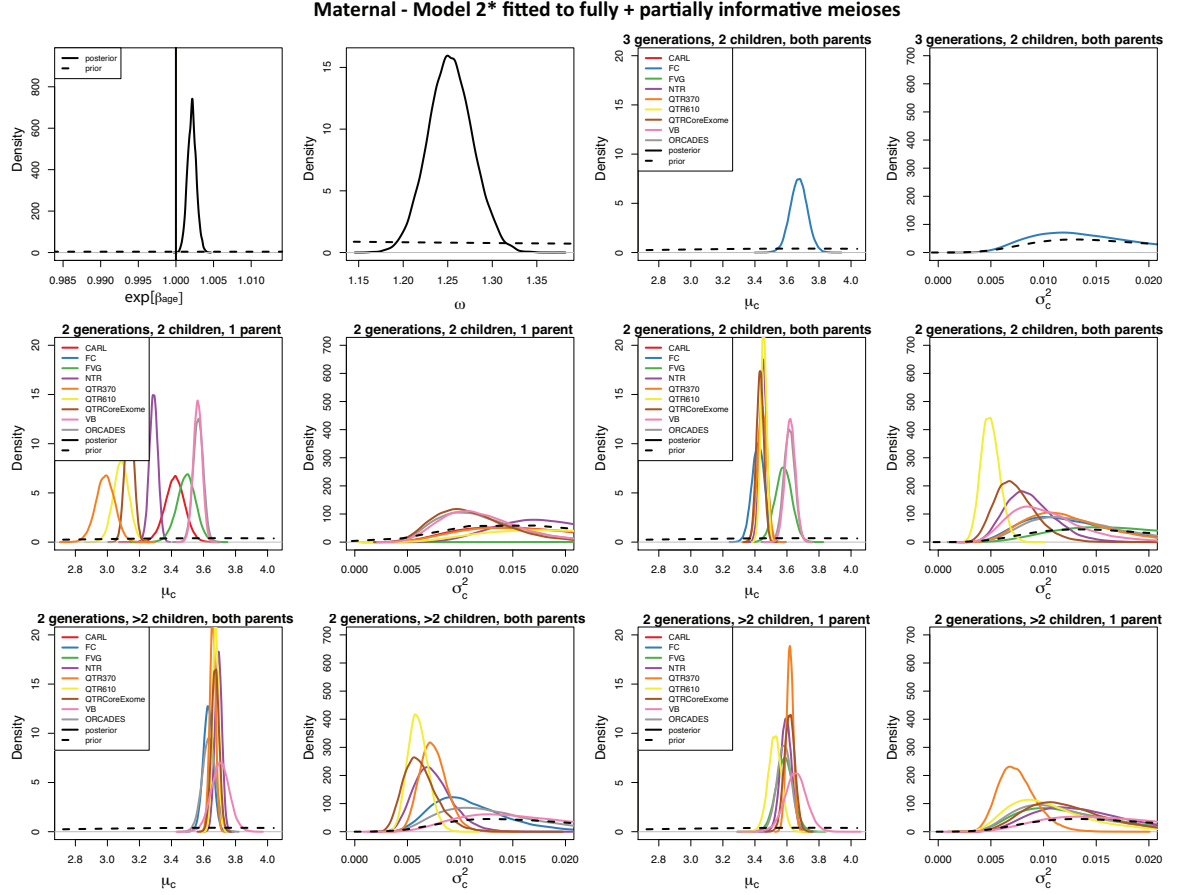


Figure 5.A.7: **Bayesian posteriors from a negative binomial model fitted to fully and partially informative maternal meioses.** This is from Model 2* (see Methods) fitted to maternal crossovers from all families with more than one child. Parameters are as described for Model 2 in Figure 5.A.6, except that there are different μ and σ for each family type-cohort combination. As for Model 2, note that the expected number of crossovers increases by $e^{\beta_{age}}$ -fold per year.

Paternal - Model 2* fitted to fully + partially informative meioses

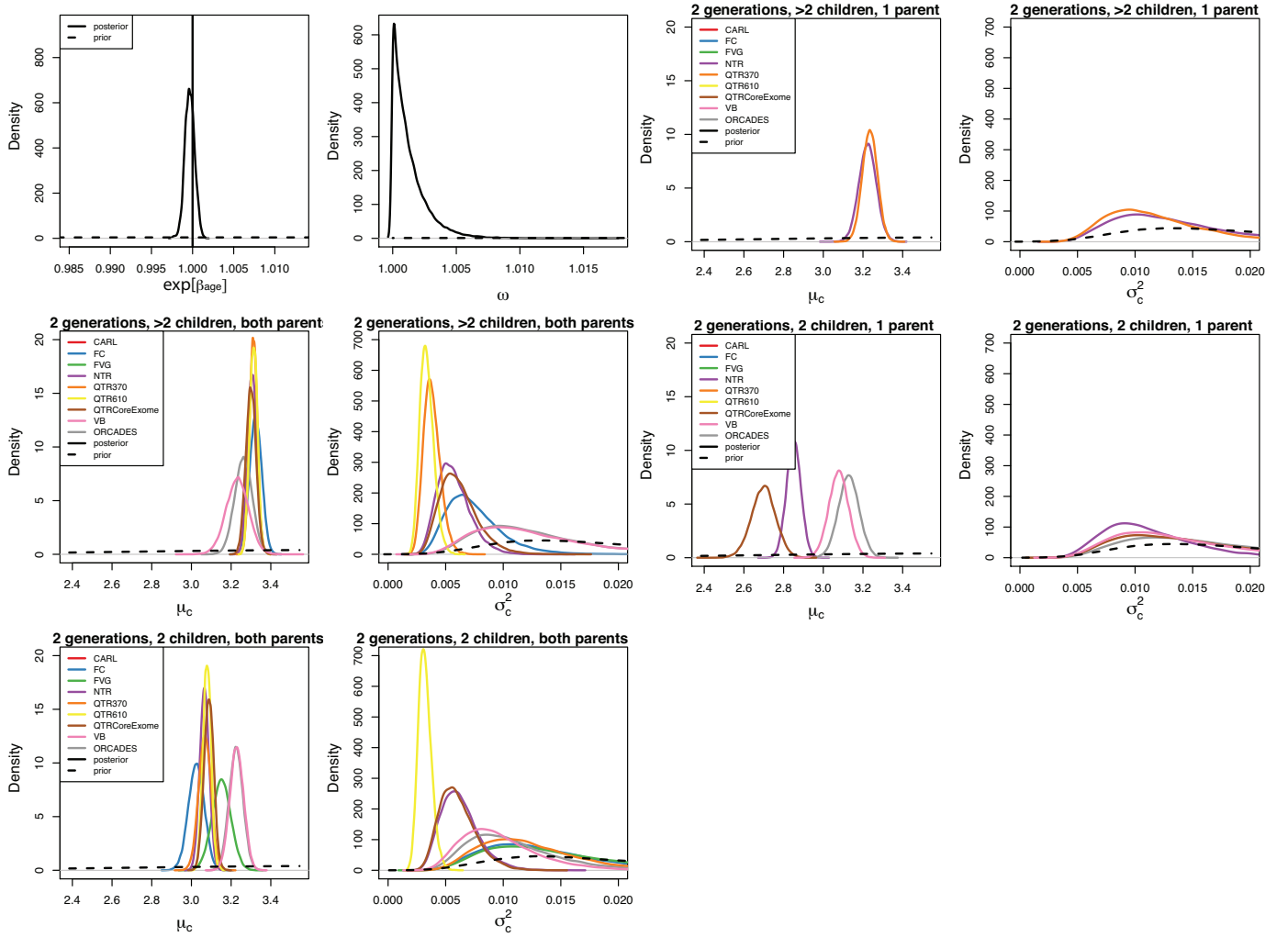


Figure 5.A.8: Bayesian posteriors from a negative binomial model fitted to fully and partially informative paternal meioses. As for Figure 5.A.7 but for paternal meioses.

Maternal - Model 2* fitted to fully + partially informative meioses
Uninformative prior on β_{age}

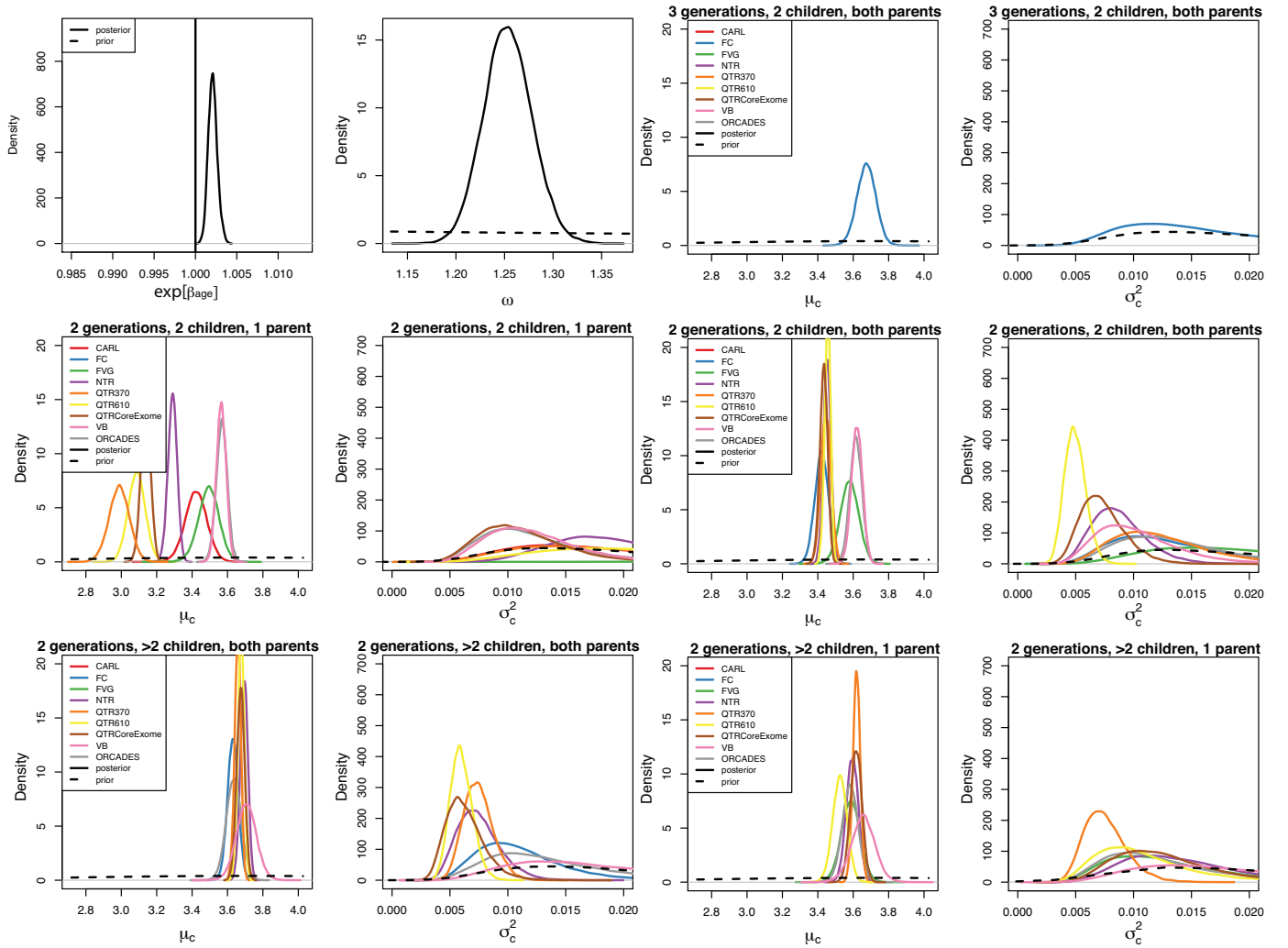


Figure 5.A.9: Bayesian posteriors from a negative binomial model fitted to fully and partially informative meioses (Model 2*), with uninformative prior on β_{age} .

Maternal - Model 2* fitted to fully + partially informative meioses
Uninformative prior on μ_c

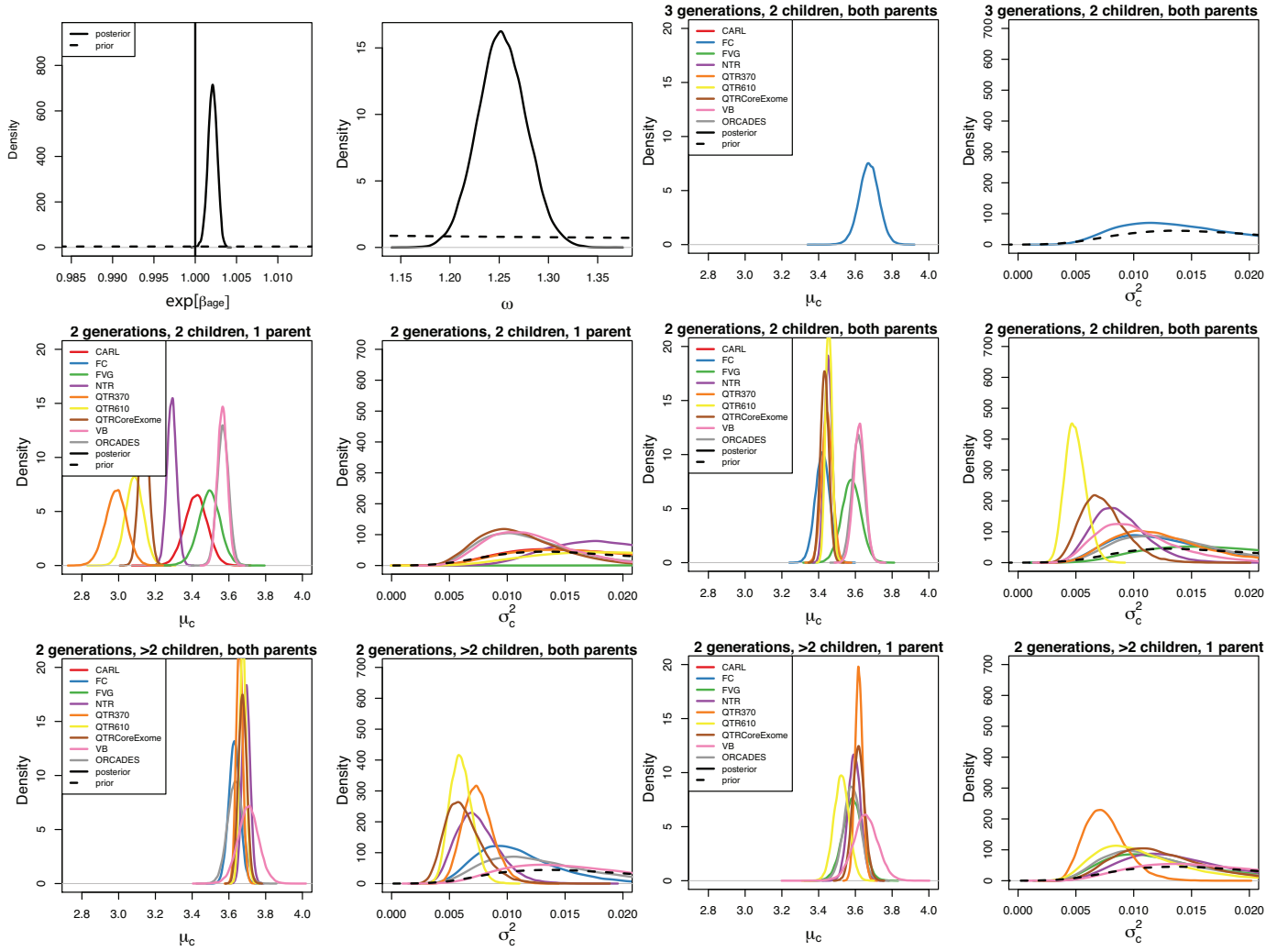


Figure 5.A.10: Bayesian posteriors from a negative binomial model fitted to fully and partially informative meioses (Model 2*), with uninformative prior on μ_c .

Maternal - Model 2* fitted to fully + partially informative meioses
Uninformative prior on σ_c

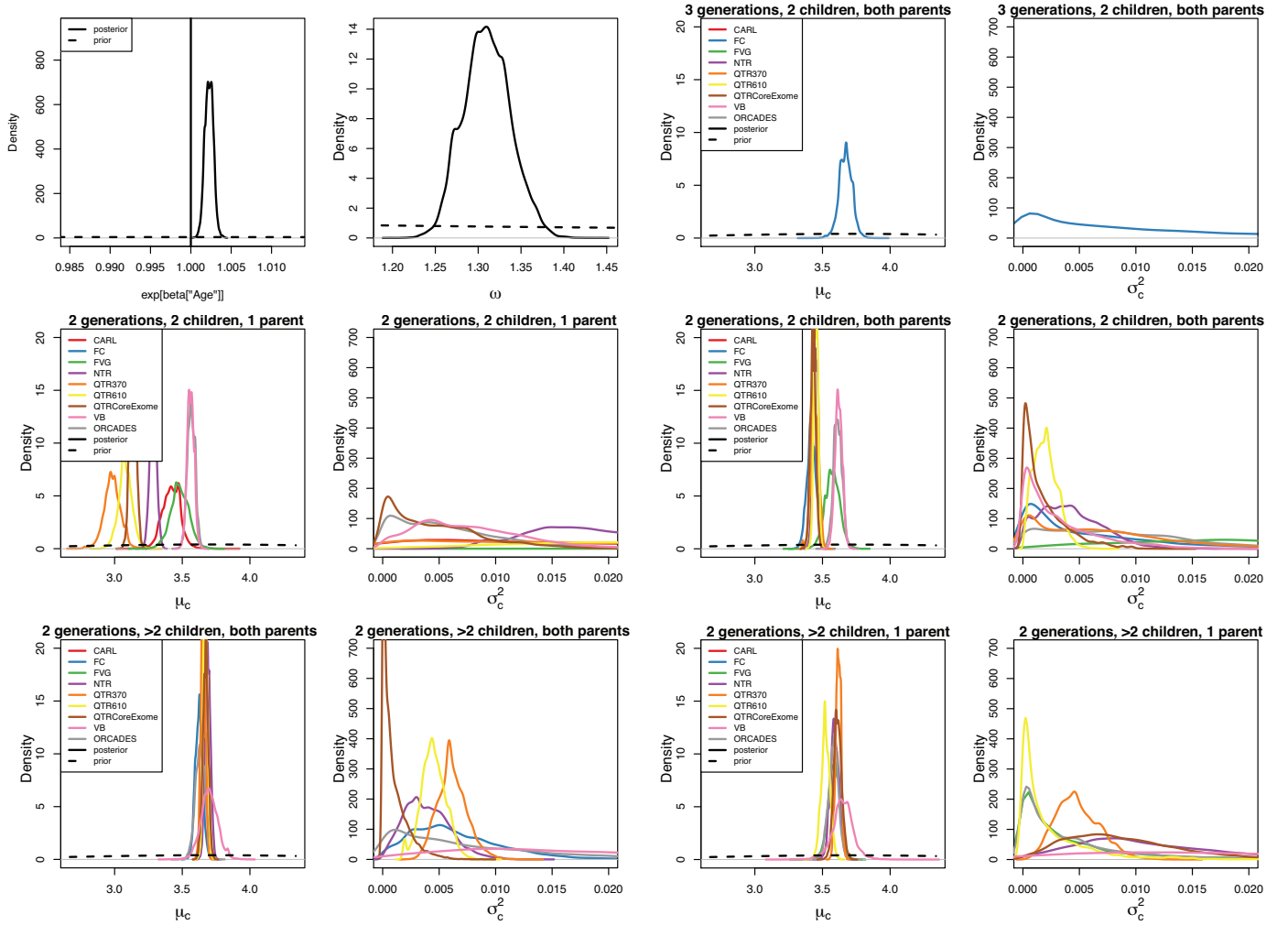


Figure 5.A.11: Bayesian posteriors from a negative binomial model fitted to fully and partially informative meioses (Model 2*), with uninformative prior on σ_c .

Maternal - Model 2* fitted to fully + partially informative meioses
Uninformative prior on ω

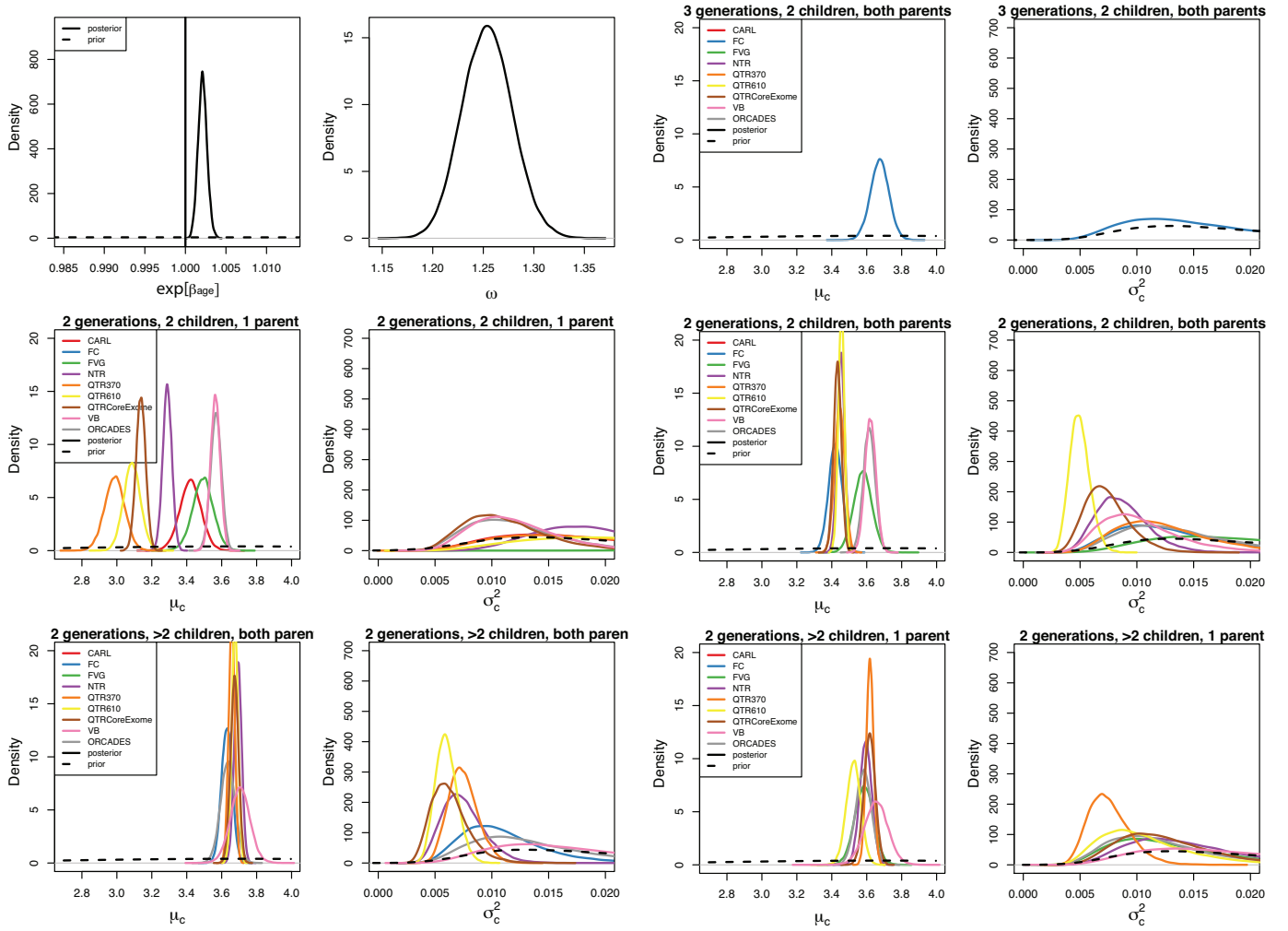


Figure 5.A.12: Bayesian posteriors from a negative binomial model fitted to fully and partially informative meioses (Model 2*), with uninformative prior on ω .

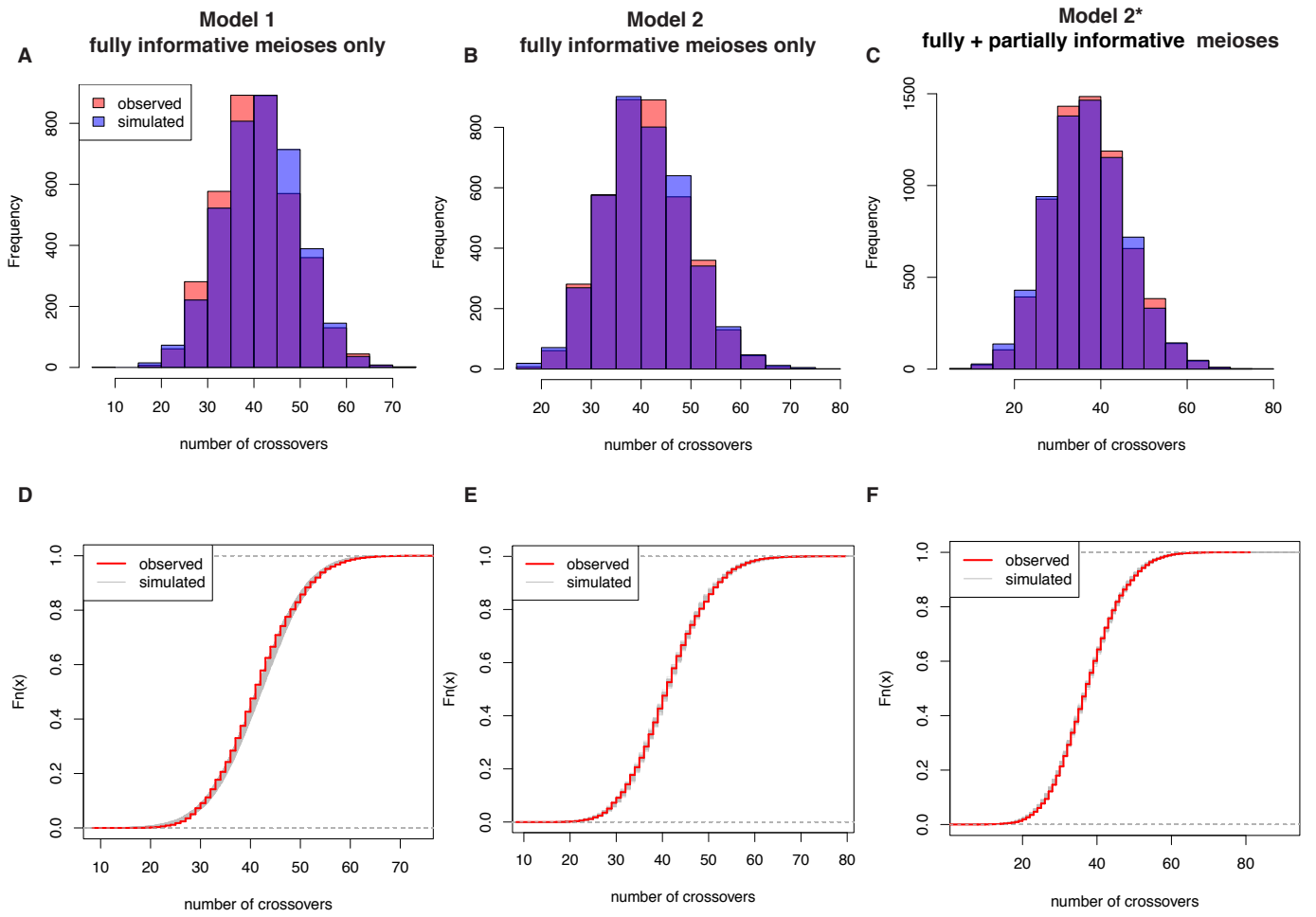


Figure 5.A.13: **Replications of the maternal crossover counts from the posterior predictive distributions, compared to observed data.** We took the parameter estimates from the draw with the highest log probability, and simulated, from the highest level of the hierarchy, 1000 datasets of the same size as the actual dataset under the model. Panels A-C show histograms of the observed data (duoHMM counts for fully informative meioses) and a randomly selected simulated dataset. Panels D-F show empirical cumulative distribution functions for 100 randomly selected datasets (grey lines), compared to the observed data (red line). The simulated datasets clearly match the observed data well.

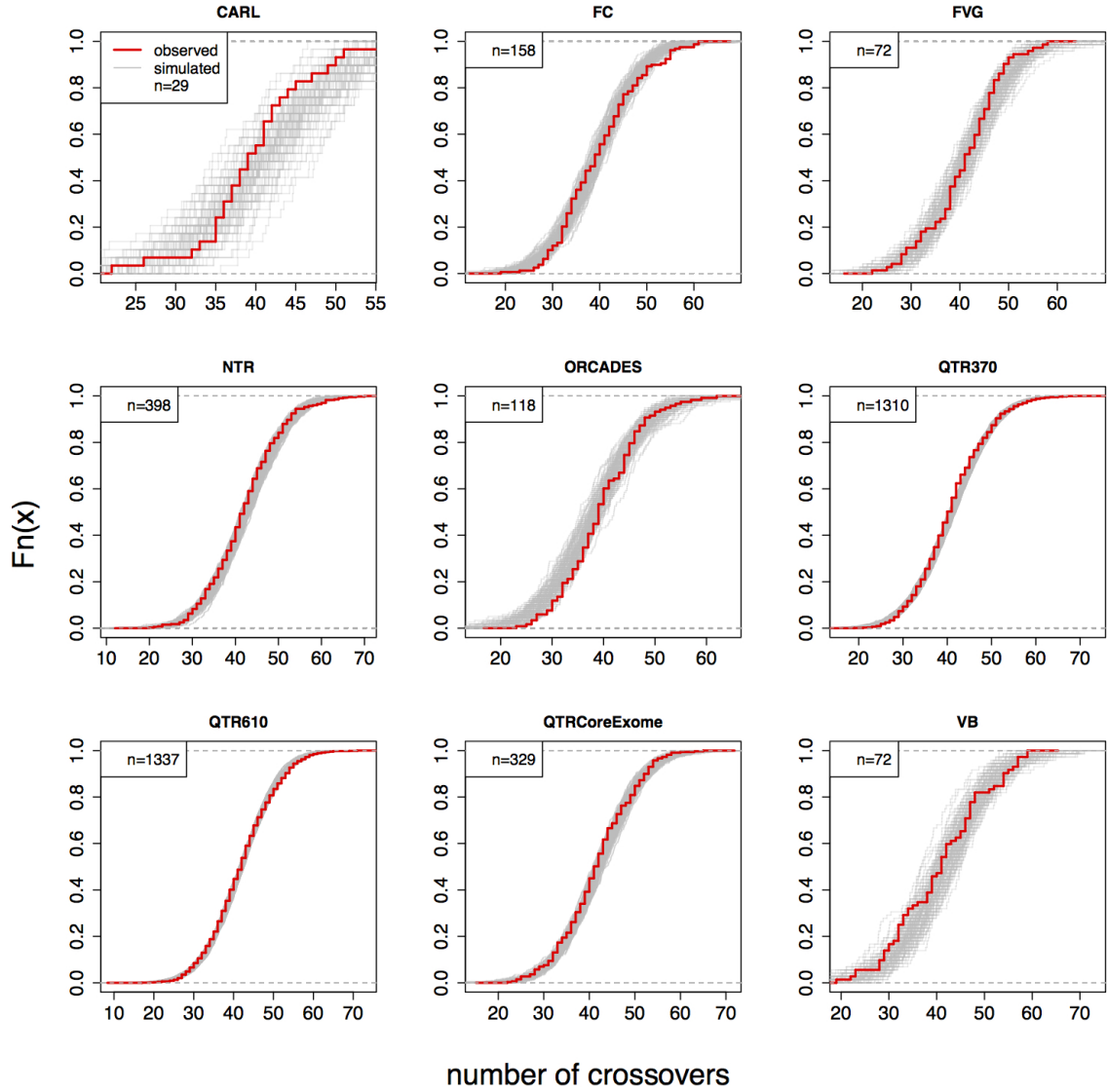


Figure 5.A.14: **Empirical cumulative distribution functions for 100 posterior predictive datasets simulated from Model 1, by cohort.** As for Figure 5.A.13D, but with the samples stratified by cohort.

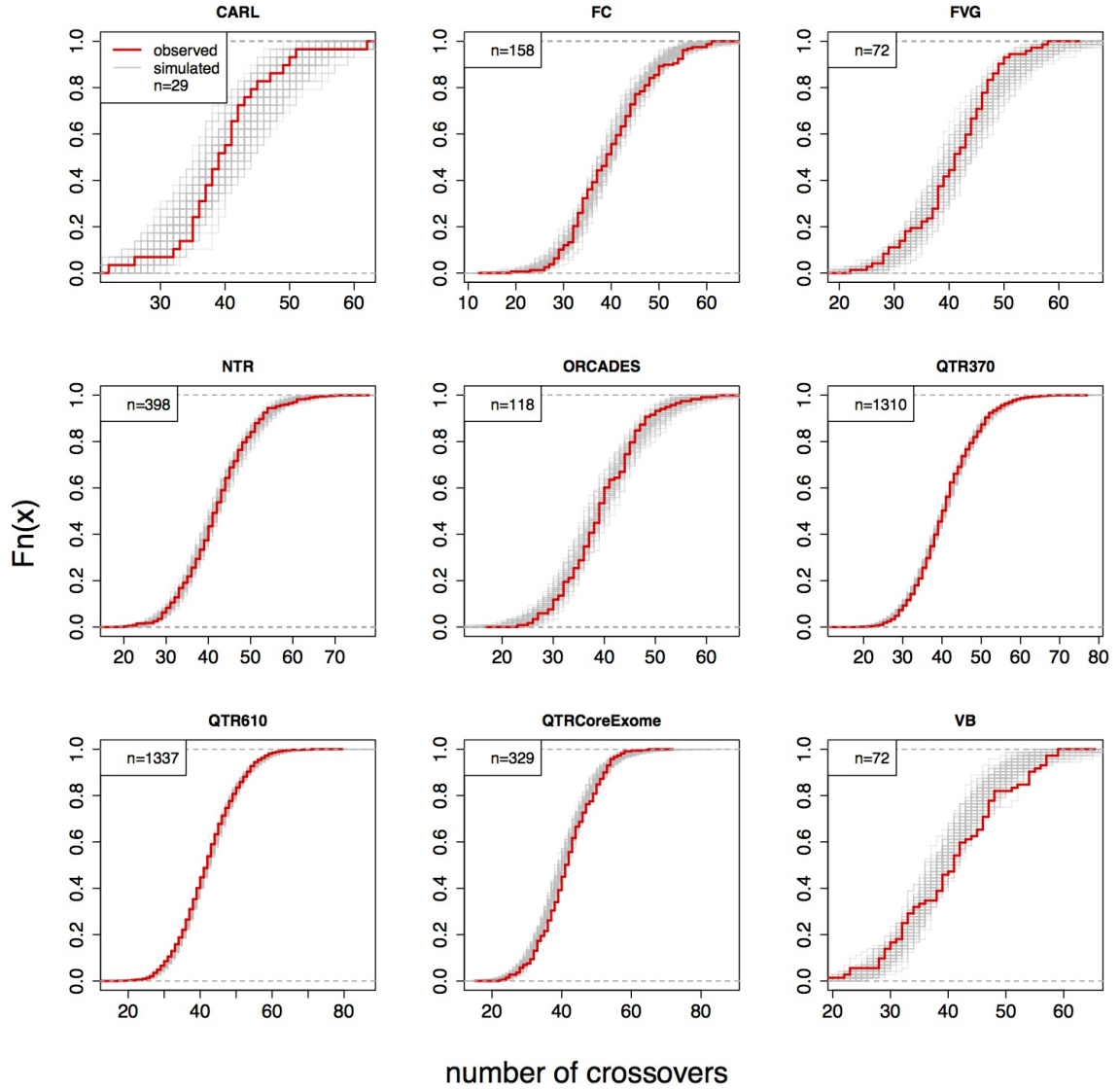


Figure 5.A.15: **Empirical cumulative distribution functions for 100 posterior predictive datasets simulated from Model 2, by cohort.** As for Figure 5.A.13E, but with the samples stratified by cohort.

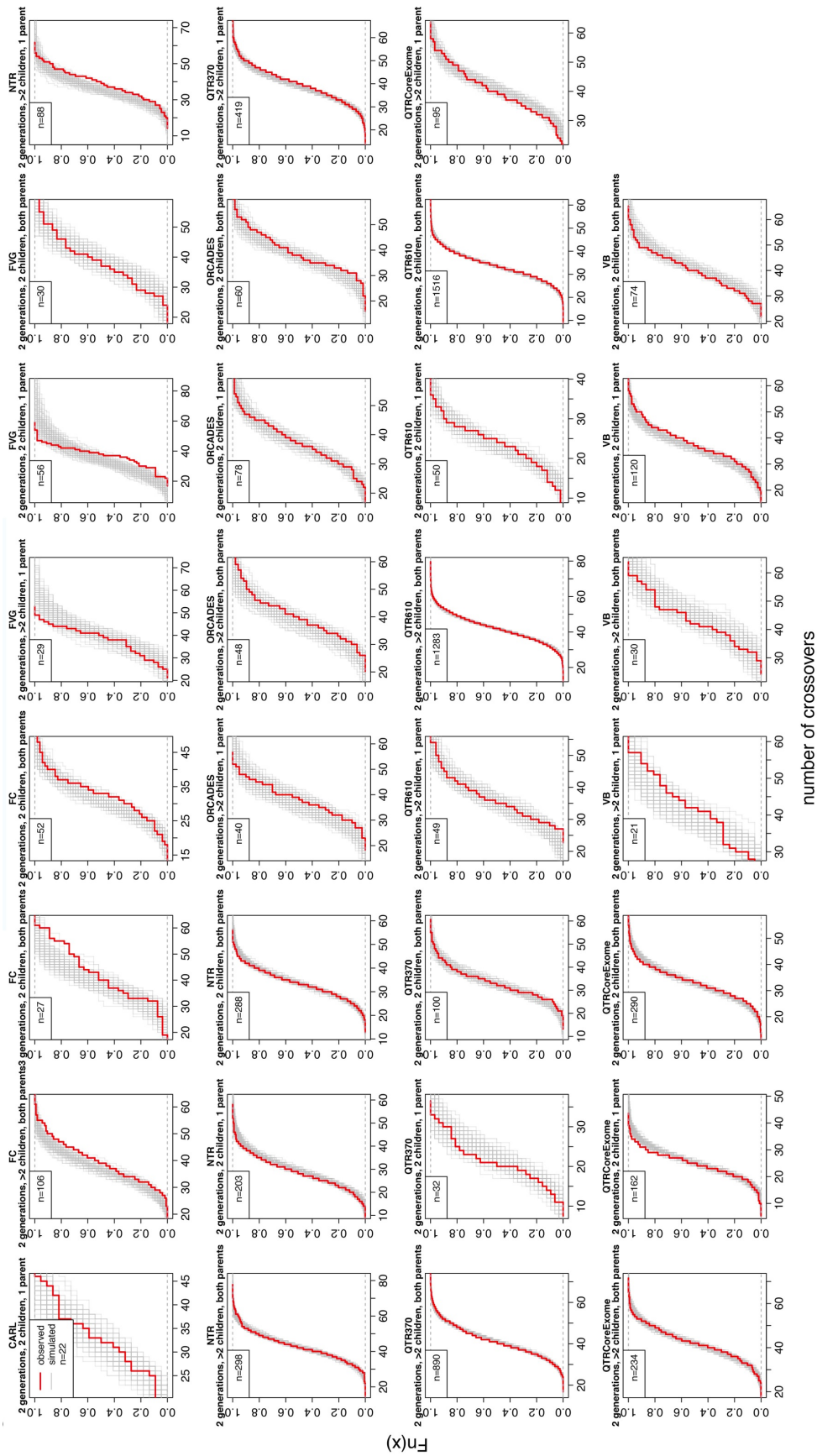


Figure 5.A.16: Empirical cumulative distribution functions for 100 posterior predictive datasets simulated from Model 2*, by cohort. As for Figure 5.A.13F, but with the samples stratified by cohort.

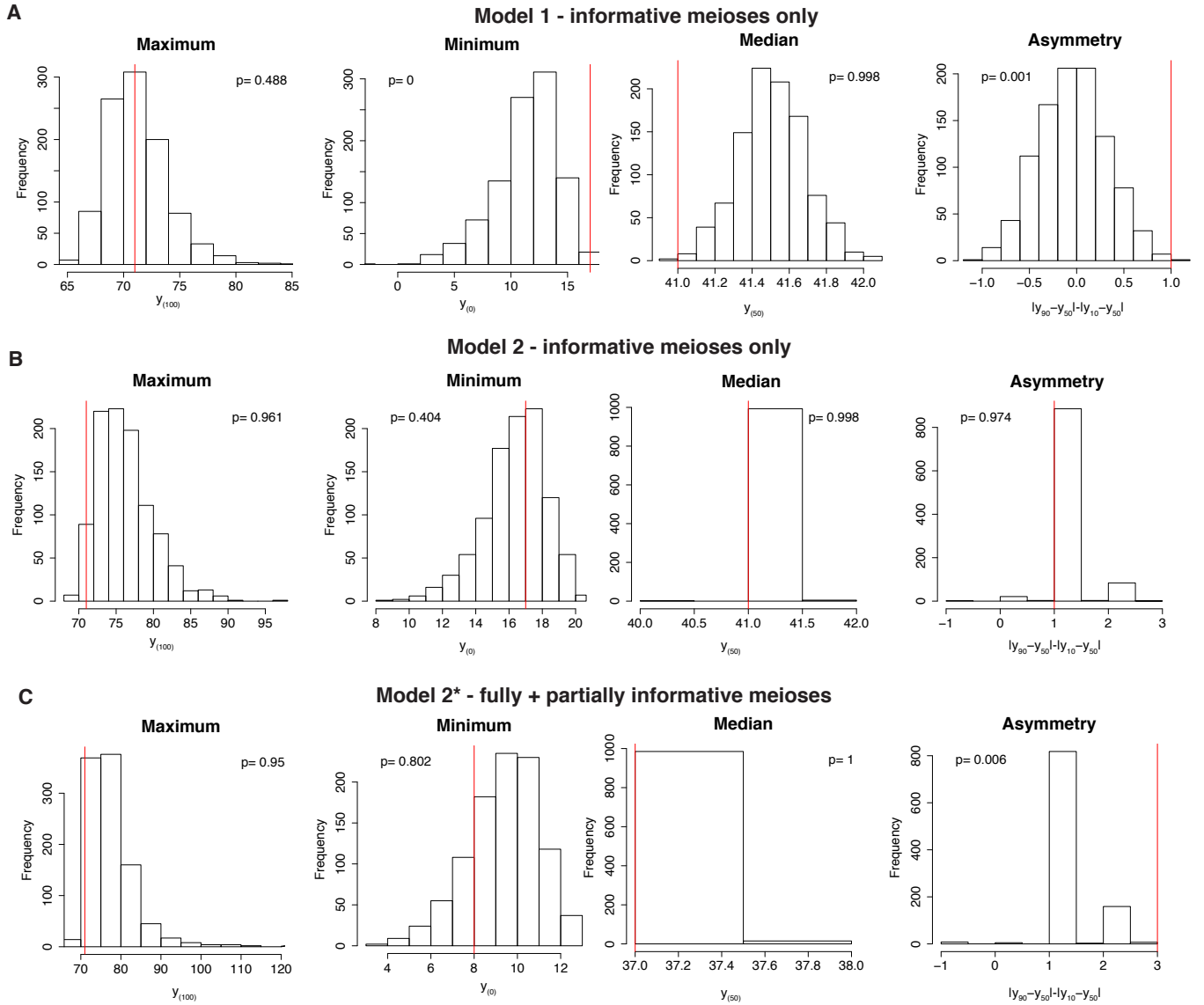


Figure 5.A.17: **Distributions of summary statistics for datasets simulated from posterior predictive distributions.** Histograms of an asymmetry statistic and the maximum, minimum and median crossover counts for the simulated datasets shown in Figure 5.A.13D-F. The asymmetry is defined as $|y_{90} - y_{50}| - |y_{10} - y_{50}|$, where y_i is the i th percentile of the distribution of the number of crossovers. Thus, it should be 0 for a symmetric distribution. The red line indicates the value from the observed data, and p is the proportion of simulated datasets with a statistic greater than or equal to this.

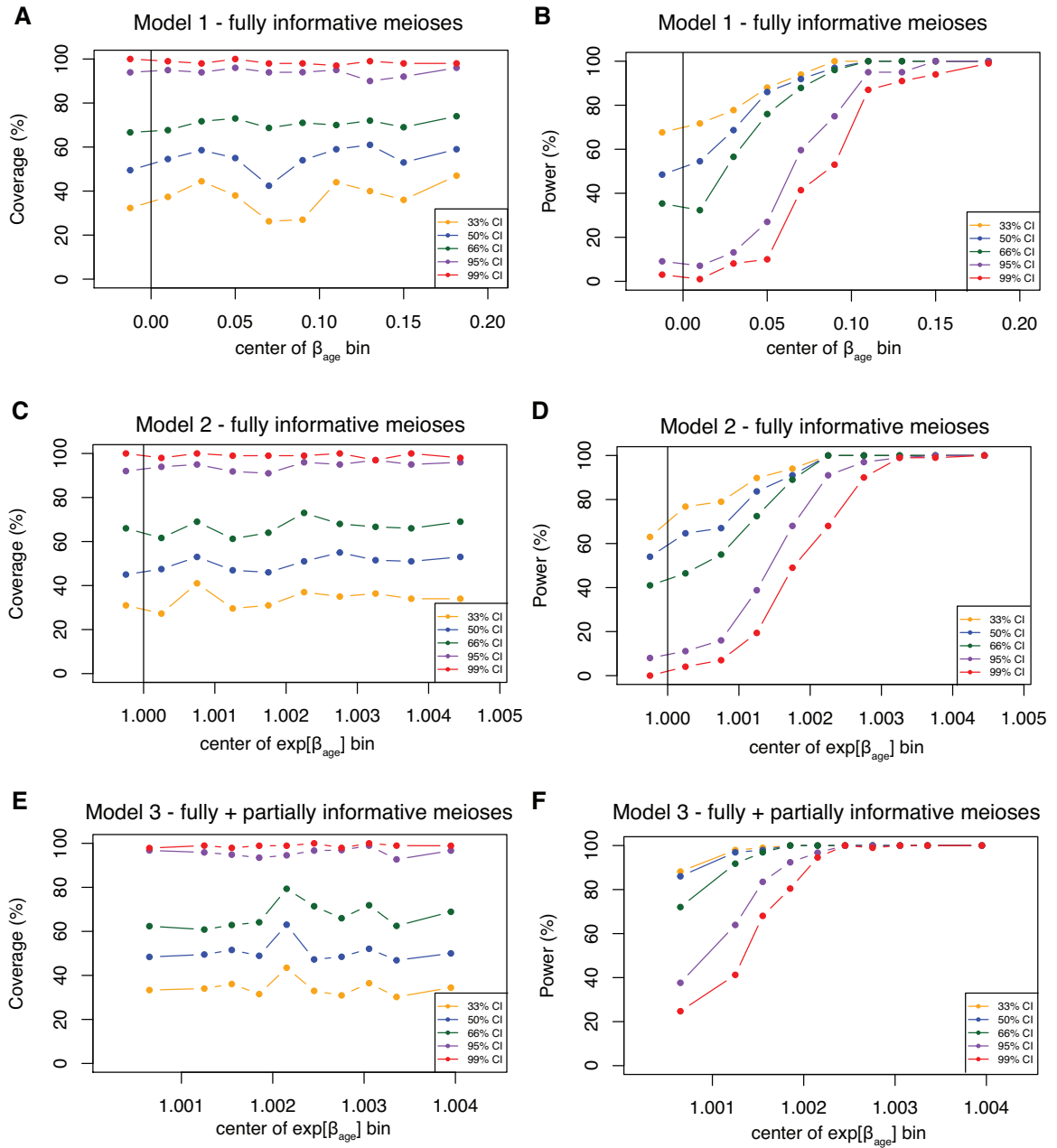


Figure 5.A.18: **Frequentist properties of the Bayesian procedure.** We divided β_{age} into 10 bins based on the posterior, sampled 100 draws from each bin, and used those parameters to simulate datasets on which we then refitted the Bayesian model. The left-hand plots show the coverage, or percentage of simulations in which the indicated credible interval overlapped the value of β_{age} used for simulation. The coverage for the bin centred on 0 is equivalent to the Type I error. The right-hand plots show the power, or percentage of simulations in which the interval did not overlap 0.

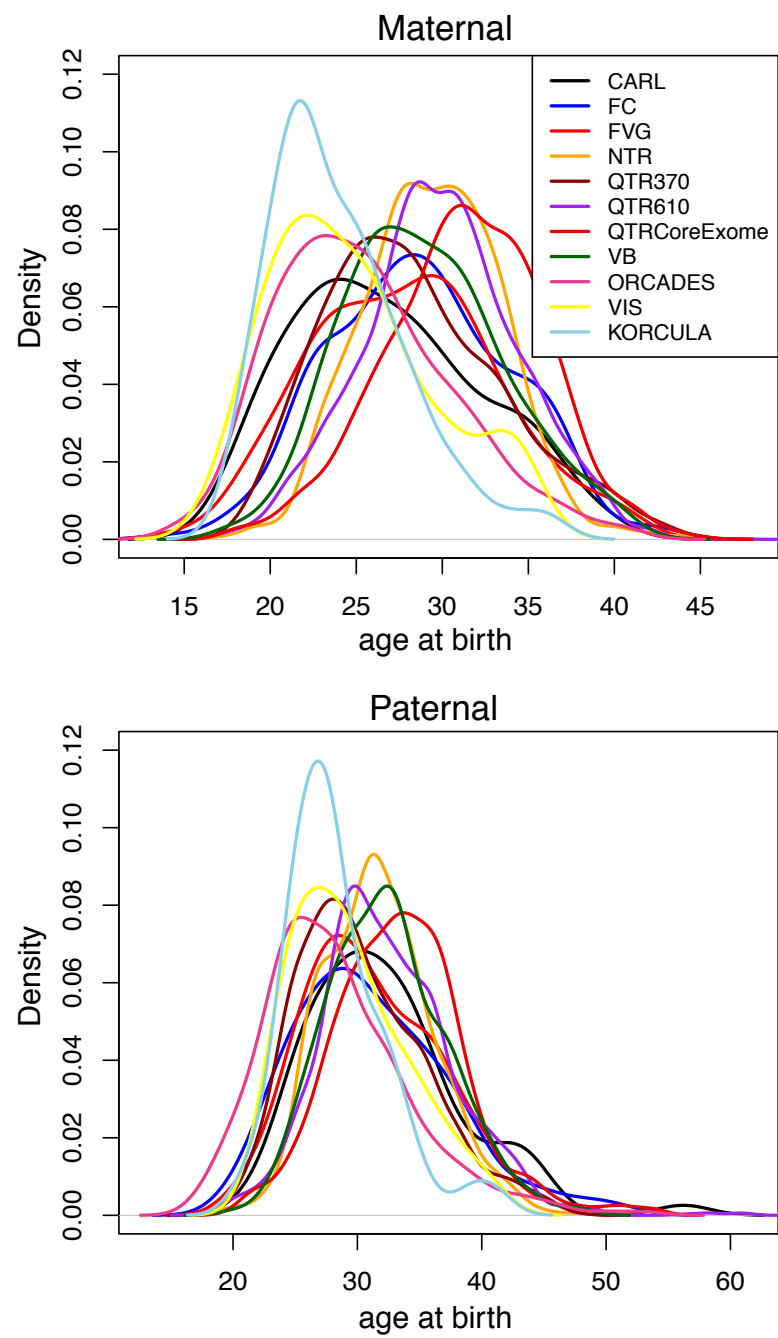


Figure 5.A.19: Distributions of parental age at birth by cohort.

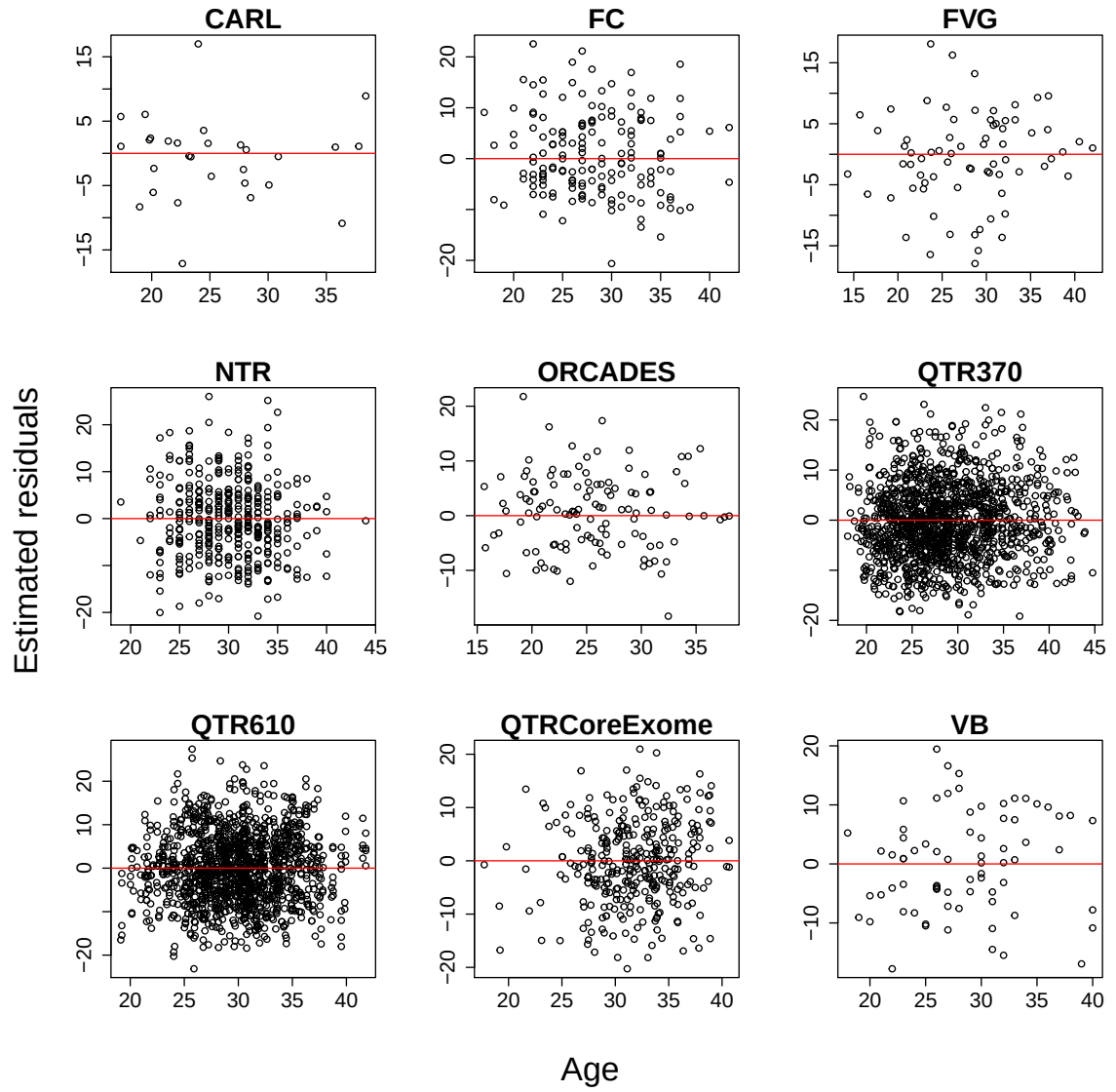


Figure 5.A.20: **Residuals from Model 1 vs. maternal age.** The residuals were calculated by subtracting the number of crossovers expected (given the parameters for the draw with the highest likelihood) from the observed number of crossovers.

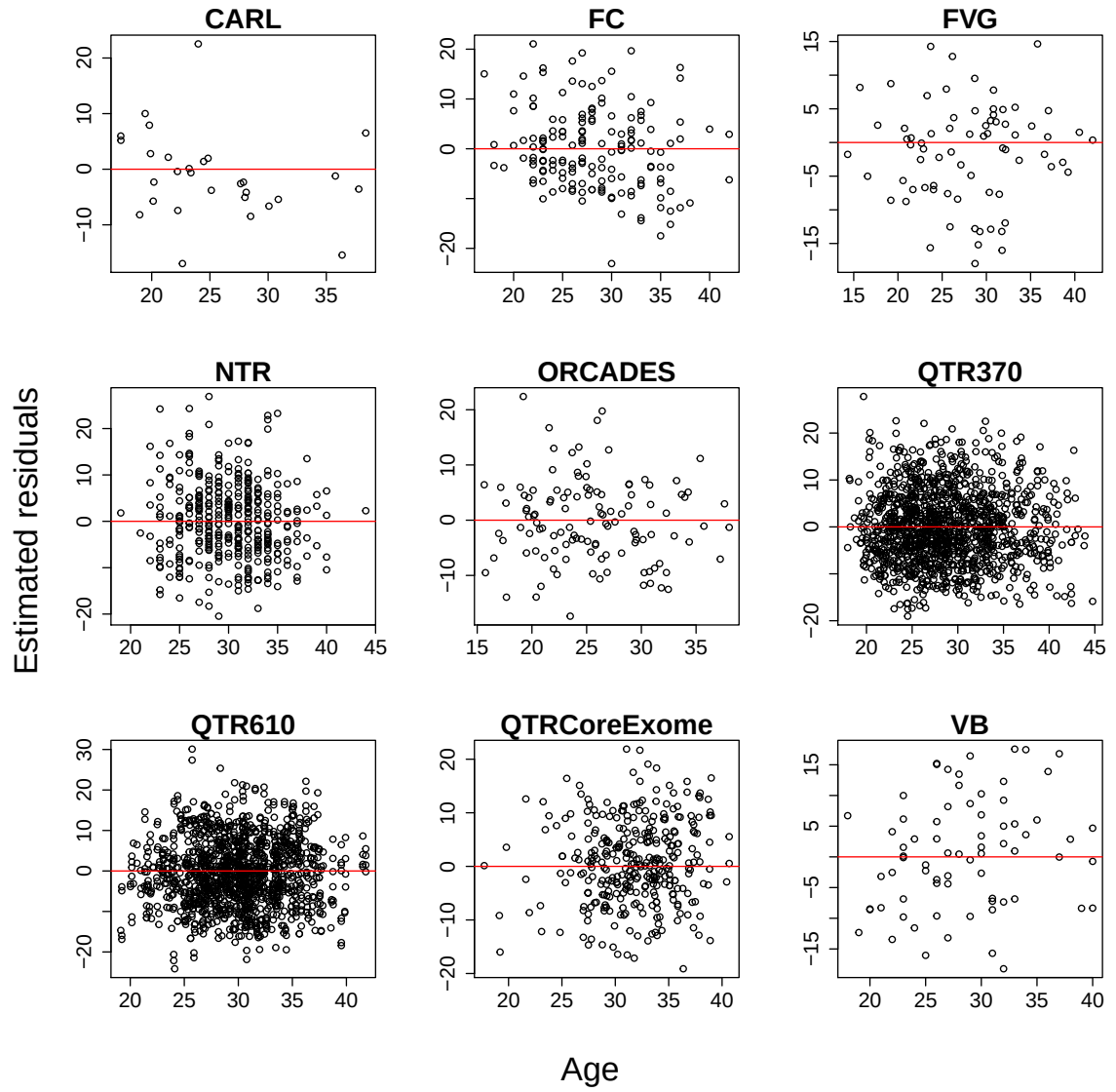


Figure 5.A.21: Residuals from Model 2 vs. maternal age. See legend of Figure 5.A.20.

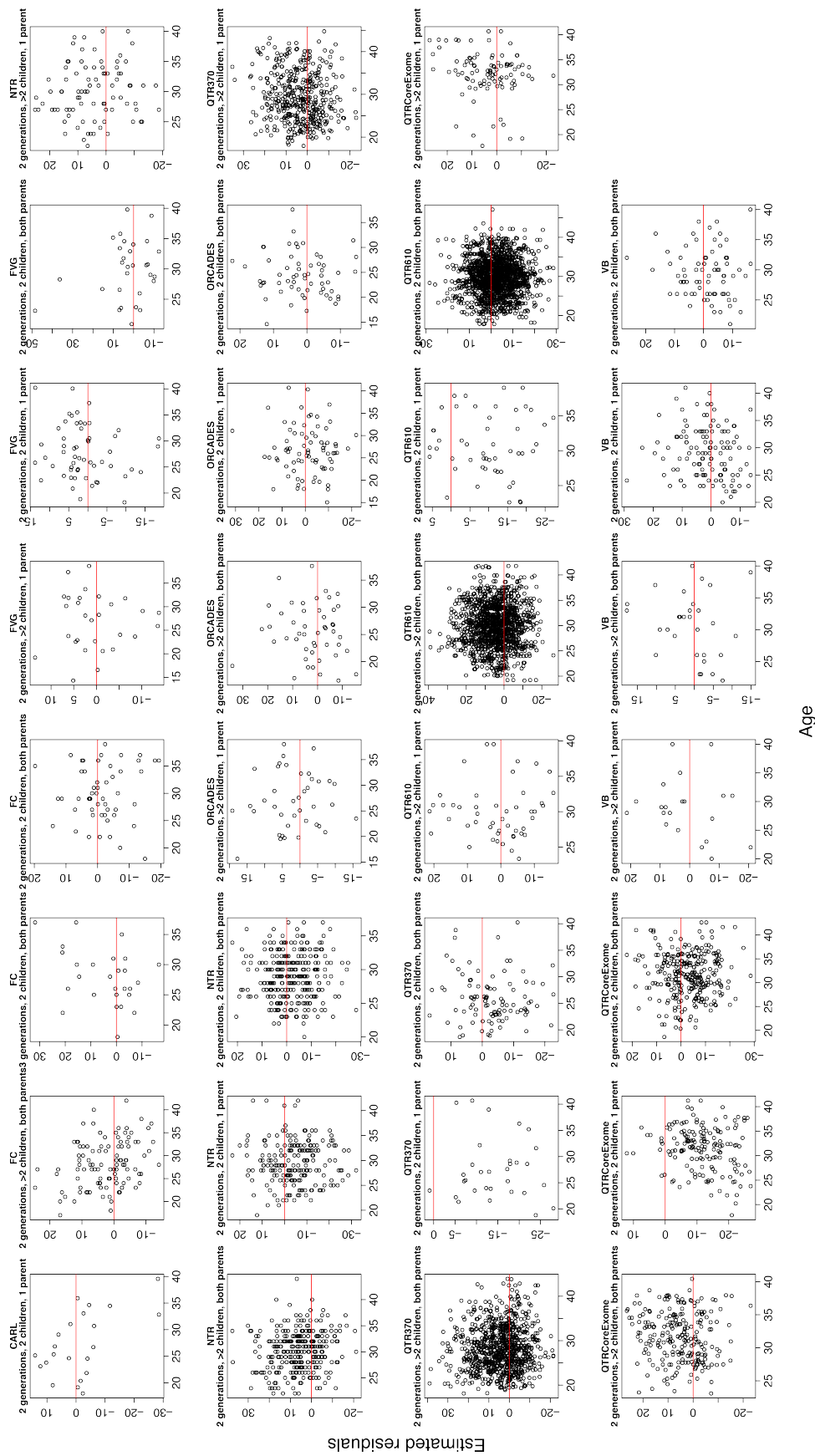


Figure 5.A.22: Residuals from Model 2* vs. maternal age. See legend of Figure 5.A.20

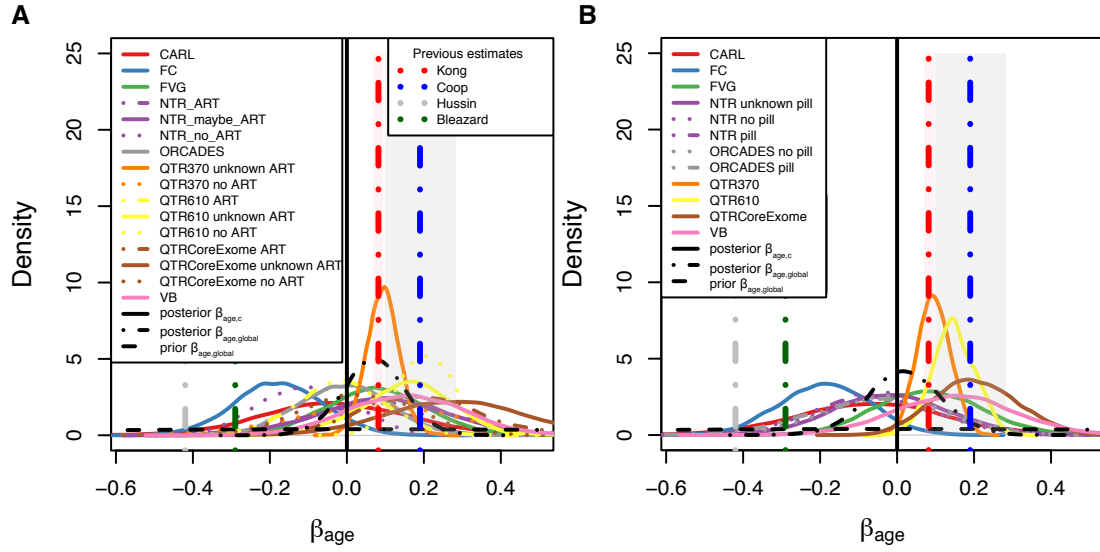


Figure 5.A.23: Bayesian posteriors for β_{age} from Model 1.2, with mothers stratified by usage of alternative reproductive technologies (ART) or by oral contraceptive (OC) usage. A) QTR and NTR mothers are stratified into those who used ART, those who did not, and those with missing data, or who had given conflicting answers at different times. B) NTR and ORCADES mothers are stratified into those who had taken oral contraception before the birth of their children (OC), those who had not (no OC) and those for whom this information was not available (unknown OC). Parameters are as defined in Figure 5.5 and Figures 5.A.5 and 5.A.6. We fitted the model to the duoHMM data.

5.A.2 Supplementary Tables

Table 5.A.1: **Number of paternal meioses, by cohort.** These are slightly lower than the number of maternal meioses given in Table 5.1, because age data were not always available for the fathers for whom we had genotype data. Also shown are the genotyping chip and number of SNPs used for analysis.

Cohort	Chip	Number of SNPs	Duos from informative nuclear families	Informative duos	Duos from families with > 1 child
Carlantino, Italy (CARL)	Illumina HumanHap300	289671	6 (2)	16 (6)	29 (13)
French Canadians (FC)	Affymetrix 6.0 (900K)	478093	107 (29)	130 (38)	197 (72)
Friuli Venezia Giulia, Italy (FVG)	Illumina HumanHap300	302661	18 (5)	39 (14)	83 (37)
Korcula, Croatia	Human370CNV	294148	0 (0)	2 (1)	16 (8)
Netherlands Twin Registry (NTR)	Affymetrix 6.0 (900K)	413268	303 (98)	356 (115)	735 (305)
ORCADES, Orkney, Scotland	Illumina HumanHap300, Omni	152398	49 (15)	74 (25)	171 (75)
Queensland Twin Registry (QTR370)	Illumina HumanCNV370v1_C, HumanCNV370-Quadv3_C, Human610-Quadv1_B	273926	897 (241)	986 (260)	1100 (317)
Queensland Twin Registry (QTR610)	Illumina Human660W-Quad_v1, Human610-Quadv1	451905	1289 (405)	1292 (406)	2820 (1170)
Queensland Twin Registry (QTRCoreExome)	Illumina HumanCoreExome-12v1-0_C	245153	234 (75)	243 (78)	557 (235)
Val Borbera, Italy (VB)	Illumina HumanHap300	310788	30 (9)	48 (17)	156 (71)
Vis, Croatia	Illumina HumanHap300	252205	0 (0)	4 (2)	20 (10)
Total			3061 (924)	3558 (1073)	6311 (2484)
Total used			2909 (872)	3168 (953)	5117 (1986)

Table 5.A.2: **Summary of posteriors for β_{age} from different Bayesian hierarchical models fitted to raw duoHMM data.** As for Table 5.1 in the main text, but using the raw duoHMM counts, before the removal of double crossovers over short intervals.

Parent	Model	Interpretation	Dataset	2.50%	25%	50%	75%	97.50%	$\Pr(\beta_{\text{age}} > 0)$
Maternal	Model 1	additive	raw data, fully informative meioses	0.0433	0.0816	0.1022	0.1232	0.1659	0.9995
	Model 2	multiplicative	raw data, fully informative meioses	1.001	1.0018	1.0023	1.0028	1.0037	0.9999
	Model 2*	multiplicative	raw data, fully + partially informative meioses	1.001	1.0017	1.0021	1.0025	1.0033	1
Paternal	Model 1	additive	raw data, fully informative meioses	-0.0375	-0.0104	0.0051	0.0196	0.048	0.5912
	Model 2	multiplicative	raw data, fully informative meioses	0.9987	0.9997	1.0001	1.0006	1.0016	0.5754
	Model 2*	multiplicative	raw data, fully + partially informative meioses	0.999	0.9997	1.0001	1.0004	1.001	0.5391

Table 5.A.3: **Summary of number of crossovers called duoHMM, before filtering.** The number of crossovers called by duoHMM was somewhat higher than expected in some cohorts, relative to the deCODE 2002 genetic map (Kong *et al.*, 2002) and Hussin *et al.* (2011). We therefore filtered the calls by removing double crossovers within X SNPs of each other, where X was the average number of SNPs per megabase for that cohort. The filtered counts are shown in Table 5.A.4. SD: standard deviation; perc.: percentile.

duoHMM, fully informative meioses, raw counts											
Parental meiosis	Cohort	Mean N.		N. duos	minimum	25th perc.	median	mean	75th perc.	maximum	SD
		SNPs within 1Mb									
Maternal	CARL	104	47	22	35	41	40.17	43.5	62	7.81	
	FC	172	166	23	36.25	42	43.71	49	98	10.44	
	FVG	109	92	24	36.75	41	41.20	46.25	60	7.82	
	KORCULA	106	11	31	33.5	38	42.73	49	68	11.72	
	NTR	148	426	23	40	46	46.25	52	80	9.13	
	ORCADES	55	143	21	34	39	39.55	45	62	7.46	
	QTR370	98	1766	21	40	46	46.22	52	82	8.99	
	QTR610	162	1399	19	43	48	48.90	56	89	9.38	
	QTRCoreExome	88	364	25	42	47	47.77	53	76	8.71	
	VB	111	104	19	33	41	40.80	47	61	9.52	
VIS	91	17	27	35	41	41.29	46	58	8.72		
Paternal	CARL	104	24	19	23	26	26.00	29	34	4.10	
	FC	172	137	17	26	29	29.73	33	68	6.03	
	FVG	109	50	17	23	26	26.12	29.75	36	4.57	
	KORCULA	106	4	22	27.25	30	28.75	31.5	33	4.79	
	NTR	148	361	15	27	30	30.44	34	48	5.67	
	ORCADES	55	86	14	23	25	25.64	30	36	4.56	
	QTR370	98	1306	16	27	31	31.42	35	52	5.85	
	QTR610	162	1348	16	29	32	32.79	36.25	53	5.91	
	QTRCoreExome	88	268	17	28	31	31.38	35	46	5.15	
	VB	111	69	17	22	25	25.35	28	34	4.12	
VIS	91	12	19	22.75	24.5	25.50	26.75	35	5.07		

Table 5.A.4: **Summary of number of crossovers called by duoHMM, after filtering.** In the raw calls (Table 5.A.3), the number of crossovers called by duoHMM was somewhat higher than expected in some cohorts, relative to the deCODE 2002 genetic map (Kong *et al.*, 2002) and Hussin *et al.* (2011). When we excluded double crossovers within X SNPs of each other, where X was the average number of SNPs per megabase for that cohort, the numbers were in line with expectation, and so we used these filtered calls for all analyses except those presented in Table 5.A.2. SD: standard deviation; perc.: percentile.

duoHMM, fully informative meioses, filtered counts									
Parental meiosis	Cohort	Mean N.		N. duos	minimum	25th perc.	median	mean	75th perc.
		SNPs within 1Mb							
Maternal	CARL	104	47	20	34	38	38.23	42	62
	FC	172	166	19	33	39	39.83	45	61
	FVG	109	92	22	36	41	40.26	45.25	58
	KORCULA	106	11	29	32.5	38	41.82	48.5	67
	NTR	148	426	20	36	41	41.25	46.75	70
	ORCADES	55	143	20	33	38	38.54	43.5	59
	QTR370	98	1766	19	34	39	39.70	45	66
	QTR610	162	1399	17	33	38	38.72	44	66
	QTRCoreExome	88	364	22	36	41	40.91	46.25	64
	VB	111	104	19	31	38	38.23	43.25	59
	VIS	91	17	26	35	41	40.35	45	58
Paternal	CARL	104	24	17	23	25	25.54	29	33
	FC	172	137	17	25	27	27.15	29	46
	FVG	109	50	16	23	26	25.78	29	36
	KORCULA	106	4	22	26.5	29.5	28.50	31.5	33
	NTR	148	361	13	23	26	25.97	29	37
	ORCADES	55	86	13	22.25	25	25.23	28.75	36
	QTR370	98	1306	10	22	26	25.40	29	40
	QTR610	162	1348	11	21	24	24.02	27	40
	QTRCoreExome	88	268	16	23.75	26	26.07	29	37
	VB	111	69	13	21	23	23.04	26	33
	VIS	91	12	17	22.75	24	24.58	25	35

Table 5.A.5: **Summary of number of crossovers called NFTOOLS.** Note that this is for informative nuclear families only.
SD: standard deviation; perc.: percentile.

NFTOOLS, from informative nuclear families								
Parental meiosis	Cohort	N. duos	minimum	25th perc.	median	mean	75th perc.	maximum SD
Maternal	FC	122	21	33.25	40	40.49	46	63 9.13
	NTR	308	21	38	42	43.32	49	68 8.42
	ORCADES	56	26	34	39.5	40.38	44	62 8.20
	QTR370	1159	22	35	40	41.15	46	69 8.40
	QTR610	1343	18	36	42	42.03	48	73 8.75
	QTRCoreExome	240	25	37	41	42.21	48	67 7.83
	VB	29	32	39	43	44.31	48	61 8.92
Paternal	FC	122	17	24.25	27	27.33	30	36 4.05
	NTR	308	13	24	26	26.65	30	38 4.48
	ORCADES	56	16	22	25	24.98	27.25	35 4.44
	QTR370	1159	12	23	26	26.13	29	41 4.48
	QTR610	1345	14	23	26	26.27	29	41 4.28
	QTRCoreExome	240	16	24	26	26.11	28	39 3.97
	VB	29	15	23	25	25.28	27	44 5.10

Table 5.A.6: **Proportion of maternal crossovers overlapping hotspots, by cohort.** We determined the percentage of crossovers resolved to within 30kb that overlapped HapMap II hotspots, for the duoHMM calls on all meioses. Using the method of Coop *et al.* (2008), we then estimated the proportion of crossovers overlapping hotspots by chance, and the proportion actually occurring in hotspots after correcting for this (α).

Maternal									
Cohort	Number of crossovers	% crossovers < 30kb	% crossovers overlapping hotspots	% crossovers overlapping hotspots by chance	α	α 95% lower bound	α 95% upper bound		
CARL	1333	33.8%	82.2%	32.3%	73.7%	70.6%	76.7%		
FC	3618	39.6%	80.6%	30.4%	72.2%	70.3%	74.0%		
FVG	3574	35.4%	82.0%	31.5%	73.8%	71.9%	75.6%		
KORCULA	1065	34.1%	82.7%	32.4%	74.4%	70.9%	77.7%		
NTR	12302	32.1%	82.0%	31.3%	73.8%	72.8%	74.8%		
ORCADES	3879	21.9%	84.5%	34.6%	76.4%	74.6%	78.0%		
QTR370	27574	34.9%	81.6%	32.3%	72.8%	72.1%	73.4%		
QTR610	56585	47.2%	78.7%	28.2%	70.3%	69.8%	70.7%		
QTRCoreExome	9094	27.8%	82.5%	34.3%	73.4%	72.2%	74.5%		
VB	6509	37.8%	82.4%	31.2%	74.4%	73.0%	75.7%		
VIS	1184	30.7%	82.2%	31.8%	73.9%	70.6%	77.0%		

Table 5.A.7: Proportion of paternal crossovers overlapping hotspots, by cohort. See legend of Table 5.A.6.

Cohort	Number of crossovers	Paternal						α 95% lower bound	α 95% upper bound
		% crossovers < 30kb	% crossovers overlapping hotspots	% crossovers overlapping hotspots by chance	α				
CARL	531	34.7%	85.1%	34.3%	77.3%	72.4%	81.7%		
FC	2058	38.5%	83.3%	31.6%	75.6%	73.2%	77.9%		
FVG	1210	34.4%	86.6%	32.8%	80.1%	77.1%	82.8%		
KORCULA	336	33.7%	84.8%	32.8%	77.4%	71.2%	82.8%		
NTR	6037	31.1%	85.4%	32.8%	78.3%	77.0%	79.6%		
ORCADES	1522	21.5%	87.1%	36.0%	79.9%	77.1%	82.4%		
QTR370	13363	34.3%	84.4%	33.3%	76.7%	75.8%	77.6%		
QTR610	34857	46.2%	83.2%	29.8%	76.1%	75.6%	76.7%		
QTRCoreExome	4439	28.4%	84.9%	35.2%	76.6%	75.0%	78.2%		
VB	2495	37.3%	84.8%	32.3%	77.5%	75.4%	79.5%		
VIS	480	31.8%	86.9%	33.9%	80.2%	75.2%	84.5%		

Table 5.A.8: **Results from linear mixed models of the effect of maternal age on the number of crossovers.** This table gives the estimated effect size, β_{age} , its standard error, and the associated two-sided p-value. Note that we did not analyse the meioses from informative nuclear families in CARL and FVG because there were fewer than 20 per cohort. SE: standard error.

Maternal									
Cohort	duoHMM data, all fully informative duos			duoHMM data, duos from informative nuclear families			NFTOOLS data, duos from informative nuclear families		
	β_{age}	SE	p-value	β_{age}	SE	p-value	β_{age}	SE	p-value
CARL	-0.145	0.286	0.618	NA	NA	NA	NA	NA	NA
FC	-0.259	0.133	0.055	-0.424	0.157	0.008	-0.430	0.164	0.010
FVG	0.118	0.151	0.439	NA	NA	NA	NA	NA	NA
NTR	-0.073	0.111	0.510	-0.149	0.129	0.250	-0.141	0.134	0.294
ORCADES	-0.023	0.137	0.866	-0.022	0.226	0.923	0.022	0.231	0.923
QTR370	0.093	0.044	0.033	0.016	0.057	0.774	0.032	0.059	0.582
QTR610	0.150	0.058	0.010	0.157	0.058	0.007	0.158	0.061	0.010
QTRCoreExome	0.283	0.112	0.012	0.262	0.134	0.053	0.234	0.135	0.085
VB	0.338	0.215	0.123	0.493	0.234	0.049	0.674	0.245	0.013

Table 5.A.9: Results from linear mixed models of the effect of paternal age on the number of crossovers. See legend of Table 5.A.8. SE: standard error.

Paternal									
Cohort	duoHMM data, all fully informative duos			duoHMM data, duos from informative nuclear families			NFTOOLS data, duos from informative nuclear families		
	β_{age}	SE	p-value	β_{age}	SE	p-value	β_{age}	SE	p-value
CARL	NA	NA	NA	NA	NA	NA	NA	NA	NA
FC	-0.094	0.064	0.147	-0.095	0.068	0.169	-0.074	0.064	0.250
FVG	-0.044	0.122	0.719	NA	NA	NA	NA	NA	NA
NTR	0.105	0.057	0.068	0.099	0.062	0.113	0.060	0.063	0.344
ORCADES	0.073	0.098	0.459	0.107	0.114	0.357	0.118	0.109	0.283
QTR370	0.015	0.029	0.616	0.027	0.031	0.382	0.009	0.029	0.750
QTR610	-0.061	0.028	0.028	-0.060	0.028	0.031	-0.055	0.027	0.045
QTRCoreExome	-0.039	0.063	0.530	-0.045	0.060	0.451	-0.027	0.054	0.619
VB	0.045	0.104	0.665	-0.157	0.142	0.282	-0.322	0.179	0.088

Table 5.A.10: **Results from a meta-analysis of the effect of parental age on the number of crossovers.** We analysed the effect sizes shown in Tables 5.A.8 and 5.A.9 using a fixed-effects meta-analysis, weighting them by their inverse variances (Cooper *et al.*, 2009). This table gives the estimated effect size, β_{age} , its standard error, and the associated two-sided p-value, as well as the test statistic and p-value from Cochran's Q test for heterogeneity.

Dataset	Method	estimated β_{age}	standard error	p-value	test statistic from Cochran's Q test for heterogeneity	p-value from Cochran's Q test for heterogeneity
Maternal	duoHMM data, all informative duos	0.090	0.029	0.002	15.796	0.045
	duoHMM data, informative nuclear families	0.060	0.035	0.089	21.278	0.002
	NFTOOLS data, informative nuclear families	0.071	0.037	0.055	21.862	0.001
Paternal	duoHMM data, all informative duos	-0.015	0.017	0.364	11.050	0.136
	duoHMM data, informative nuclear families	-0.017	0.018	0.327	11.584	0.072
	NFTOOLS data, informative nuclear families	-0.022	0.017	0.203	9.433	0.151

Table 5.A.11: **Comparison of results from Model 1 fitted to NFTOOLS *versus* duoHMM data.** As for Table 5.1, but for meioses from informative nuclear families.

Parent	Method	2.50%	25%	50%	75%	97.50%	$\Pr(\beta_{age} > 0)$
Maternal	NFTOOLS	-0.0081	0.0389	0.0641	0.0884	0.1363	0.9597
	duoHMM	-0.0283	0.0243	0.0513	0.0774	0.1228	0.8990
Paternal	NFTOOLS	-0.0543	-0.0333	-0.0210	-0.0096	0.0121	0.1079
	duoHMM	-0.0540	-0.0295	-0.0172	-0.0049	0.0187	0.1729

Chapter 6

Conclusion

In this thesis, we have presented results from three different applications of genomic technologies. The first was the use of WGS in the study of rare disease, as part of the WGS500 project. In Chapter 2, we described findings from case studies of several Mendelian neurological diseases. Sequencing of six trios with early-onset epilepsy led to the discovery of two new genes for Ohtahara syndrome, *KCNT1* and *PIGQ*, and two candidate genes for nonsyndromic epilepsy, *CBL* and *CSNK1G1*. This case study emphasises the efficiency of WGS for detecting causal mutations in genetically heterogeneous Mendelian disorders. Our discovery of the *GRIA3* mutation in the brothers with severe sleep disturbance suggests a role for this gene in circadian rhythm, which we hope will be better understood through ongoing studies in a knock-in mouse strain. Our experience also illustrates the difficulty that can be encountered in finding a causal mutation in a consanguineous family, even when there is a well-defined linkage region (as in the family with the novel learning disability/spasticity syndrome), and if the inheritance mode is not clear (as in the sisters with mental retardation and epileptic encephalopathy). Our successes also support the notion that WES is more cost-effective than WGS for families with severe Mendelian disorders. On the other hand, having WGS data on more patients would give us more power to start identifying causal noncoding variants, if these do account for some of the unsolved Mendelian phenotypes.

The analyses in Chapter 3 demonstrate some important points about optimal design of clinical sequencing studies. We showed that sequencing trios greatly reduces the number of candidates to consider for patients with a sporadic disease, as has been reported in other large studies (Wright *et al.*, 2015). Additionally, we showed that a large number of artefacts of the mapping and calling pipeline can be ruled out by screening against locally sequenced controls, but would otherwise be considered as

candidates if only public databases are consulted. Our results also suggest that the strategy of focusing on candidate genes generated through high-throughput database searches is prone to both false positives and false negatives, and that filtering based on family information first is more effective. In addition to pinpointing pathogenic non-coding mutations, homing in on causal genes quickly through sensitive, specific and automated literature searches remains the biggest challenge for future large clinical sequencing projects such as Genomics England, in which laborious and time-consuming manual review of multiple candidates will not be feasible.

In Chapter 4, we described the first population sequencing study of the platypus. Much of our effort in this project was devoted to trying to improve the highly fragmented reference genome, and this is ongoing. We managed to fix some of the errors, but many remain, and the assembly still contains a large number of small scaffolds. Despite this, we demonstrated that our SNP callset had a low false positive rate and accurate genotypes, as assessed by the Mendelian error rate and discordant genotype rate between two duplicates. We therefore considered this a reasonable substrate for our population genetic analyses. These confirmed the strong population structure that had previously been noted in microsatellite and mitochondrial DNA studies (Kolomyjec *et al.*, 2009; Gongora *et al.*, 2012; Furlan *et al.*, 2013), and also revealed unexpected similarities between platypus from different river systems several hundred kilometres apart, which may be explained by overland travel or historical connections between the basins. Analysis of historical effective population sizes showed the distinct histories of the different populations, and suggested a different tree from that previously reported based on mitochondrial sequences. Our results suggest that the North Queensland population should be a particular priority for conservation, since it appears quite distinct from populations further south and has low nucleotide diversity resulting from a bottleneck. We also discovered some apparently recently-formed strata on the platypus sex chromosomes, which were not reported in an earlier study (Cortez *et al.*, 2014). These appear to be at PAR boundaries, and in one case, there seems to have been a crossover between the X and the Y that has led to the extension of the PAR by 600kb in one individual. Our observations support the notion that the platypus sex chromosomes are still undergoing the transformation from autosomes to fully-fledged sex chromosomes. We hope that further reference genome improvements will allow us to assemble the full sequences of the sex chromosomes, which will help to understand how (and possibly even why) they formed.

Finally, in Chapter 5, we investigated patterns of meiotic recombination in human pedigrees using genome-wide SNP data. Our analysis of nine different cohorts using Bayesian hierarchical models revealed a small positive effect of maternal age on the number of crossovers (an increase of 2% over ten years). However, despite standardising the data cleaning and crossover calling as far as possible between cohorts, we still saw substantial inter-cohort heterogeneity. We think this is most likely to be sampling noise, but there may also be confounding factors at play. Our results may not be relevant to the maternal age effect on aneuploidies since our approach only considers live births, but future studies using single-cell genotyping or sequencing of meiotic products (Handyside *et al.*, 2010; Hou *et al.*, 2013; Ottolini *et al.*, 2015) may further our understanding of this.

Over the next few years, falling costs will undoubtedly lead to the sequencing of millions of genomes of clinical cases and healthy people. This will give us much greater power to uncover the genetic causes of many more rare and common diseases, leading to an increased understanding of aetiology and ultimately, it is hoped, better treatments. For rare diseases, this will be greatly facilitated if genotype and phenotype information can be securely shared between studies. For common diseases and clinically relevant quantitative traits, international collaborations and biobanks will allow the collection of larger and larger samples, increasing our power to find variants of smaller effect sizes. The integration of data from projects such as ENCODE and the Epigenomics Roadmap may help identify pathogenic noncoding variants, but (ideally high-throughput) experimental approaches for determining the function of such variants in the relevant tissue type will also be required. Furthermore, lower costs will spearhead the sequencing of population samples from many other species, giving rise to a better understanding of their biology, as well as improving agricultural practices (e.g. Daetwyler *et al.*, 2014; Brozynska *et al.*, 2015) and identifying conservation priorities (e.g. Li *et al.*, 2014; Perdiguero *et al.*, 2015). “Third generation” long-read sequencing technologies such as PacBio and Oxford Nanopore will greatly facilitate *de novo* assembly of new genomes, and if the error rates decrease dramatically, they will also be useful for phasing, which has applications in the study of disease, of recombination, and of demographic history.

Bibliography

- 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.
- 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., *et al.* (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.
- Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, 30(1):97–101.
- Adachi, M., Abe, Y., Aoki, Y., and Matsubara, Y. (2012). Epilepsy in RAS/MAPK syndrome: two cases of cardio-facio-cutaneous syndrome with epileptic encephalopathy and a literature review. *Seizure*, 21(1):55–60.
- Adzhubei, I. A., Schmidt, S., Peshkin, L., *et al.* (2010). A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248–9.
- Afonso, J. M., Volz, A., Hernandez, M., *et al.* (1990). Mitochondrial DNA variation and genetic structure in Old-World populations of *Drosophila subobscura*. *Mol Biol Evol*, 7(2):123–142.
- Ai, H., Fang, X., Yang, B., *et al.* (2015). Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat Genet*, 47(3):217–225.
- Aird, D., I. B. H. H., Mehigan, J. A., and Roberts, J. A. F. (1954). The blood groups in relation to peptic ulceration and carcinoma of colon, rectum, breast, and bronchus; an association between the ABO groups and peptic ulceration. *Br Med J*, 2(4883):315–321.

- Al Turki, S., Manickaraj, A. K., Mercer, C. L., *et al.* (2014). Rare variants in NR2F2 cause congenital heart defects in humans. *Am J Hum Genet*, 94(4):574–585.
- Alazami, A. M., Patel, N., Shamseldin, H. E., *et al.* (2015). Accelerating Novel Candidate Gene Discovery in Neurogenetic Disorders via Whole-Exome Sequencing of Prescreened Multiplex Consanguineous Families. *Cell Rep*, 10(2):148–161.
- Aldahmesh, M. A., Khan, A. O., Mohamed, J. Y., *et al.* (2012). Genomic analysis of pediatric cataract in Saudi Arabia reveals novel candidate disease genes. *Genet Med*, 14(12):955–62.
- Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 19(9):1655–1664.
- Allen, A. S., Berkovic, S. F., Cossette, P., *et al.* (2013). De novo mutations in epileptic encephalopathies. *Nature*.
- Almeida, A. M., Murakami, Y., Layton, D. M., *et al.* (2006). Hypomorphic promoter mutation in PIGM causes inherited glycosylphosphatidylinositol deficiency. *Nat Med*, 12(7):846–51.
- Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *science*, 322(5903):881–888.
- Arboleda, V. A., Lee, H., Parnaik, R., *et al.* (2012). Mutations in the PCNA-binding domain of CDKN1C cause IMAGE syndrome. *Nat Genet*, 44(7):788–92.
- Arnheim, N., Calabrese, P., and Nordborg, M. (2003). Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. *Am J Hum Genet*, 73(1):5–16.
- Asan, Xu, Y., Jiang, H., *et al.* (2011). Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol*, 12(9):R95.
- Ashburner, M., Ball, C. A., Blake, J. A., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9.
- Atzmon, G., Hao, L., Pe’er, I., *et al.* (2010). Abraham’s children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry. *Am J Hum Genet*, 86(6):850–859.

- Au, P. Y. B., You, J., Caluseriu, O., *et al.* (2015). GeneMatcher Aids in the Identification of a New Malformation Syndrome with Intellectual Disability, Unique Facial Dysmorphisms, and Skeletal and Connective Tissue Abnormalities Caused by De Novo Variants in HNRNPK. *Hum Mutat.*
- Auton, A., Fledel-Alon, A., Pfeifer, S., *et al.* (2012). A fine-scale chimpanzee genetic map from population sequencing. *Science*, 336(6078):193–8.
- Auton, A., Li, Y. R., Kidd, J., *et al.* (2013). Genetic recombination is targeted towards gene promoter regions in dogs.
- Auton, A. and McVean, G. (2012). Estimating recombination rates from genetic variation in humans. *Methods Mol Biol*, 856:217–237.
- Bachmann-Gagescu, R., Ishak, G. E., Dempsey, J. C., *et al.* (2012). Genotype-phenotype correlation in CC2D2A-related Joubert syndrome reveals an association with ventriculomegaly and seizures. *J Med Genet*, 49(2):126–37.
- Bahlo, M. and Bromhead, C. J. (2009). Generating linkage mapping files from Affymetrix SNP chip data. *Bioinformatics*, 25(15):1961–2.
- Bainbridge, M. N., Wiszniewski, W., Murdock, D. R., *et al.* (2011). Whole-genome sequencing for optimized patient management. *Sci Transl Med*, 3(87):87re3.
- Bamshad, M. J., Ng, S. B., Bigham, A. W., *et al.* (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet*, 12(11):745–55.
- Barcia, G., Fleming, M. R., Deligniere, A., *et al.* (2012). De novo gain-of-function KCNT1 channel mutations cause malignant migrating partial seizures of infancy. *Nat Genet*, 44(11):1255–9.
- Barge-Schaapveld, D. Q. C. M., Brooks, A. S., Lequin, M. H., *et al.* (2011). Beare-Stevenson syndrome: two Dutch patients with cerebral abnormalities. *Pediatr Neurol*, 44(4):303–307.
- Baudat, F., Buard, J., Grey, C., *et al.* (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, 327(5967):836–40.
- Beaudet, A., Bowcock, A., Buchwald, M., *et al.* (1986). Linkage of cystic fibrosis to two tightly linked DNA markers: joint report from a collaborative study. *American journal of human genetics*, 39(6):681.

- Beaulieu, C. L., Majewski, J., Schwartzentruber, J., *et al.* (2014). FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *The American Journal of Human Genetics*, 94(6):809–817.
- Begun, D. J. and Aquadro, C. F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, 356(6369):519–520.
- Bell, C. J., Dinwiddie, D. L., Miller, N. A., *et al.* (2011). Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med*, 3(65):65ra4.
- Berg, I. L., Neumann, R., Lam, K. W., *et al.* (2010). PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat Genet*, 42(10):859–63.
- Berg, J. S., Evans, J. P., Leigh, M. W., *et al.* (2011). Next generation massively parallel sequencing of targeted exomes to identify genetic mutations in primary ciliary dyskinesia: implications for application to clinical testing. *Genet Med*, 13(3):218–229.
- Berkovic, S. F., Kennerson, M. L., Howell, R. A., *et al.* (1994). Phenotypic expression of benign familial neonatal convulsions linked to chromosome 20. *Arch Neurol*, 51(11):1125–8.
- Bhatia, G., Patterson, N., Sankararaman, S., and Price, A. L. (2013). Estimating and interpreting F_{ST} : the impact of rare variants. *Genome Res*, 23(9):1514–1521.
- Bhattacharjee, A., Gan, L., and Kaczmarek, L. K. (2002). Localization of the Slack potassium channel in the rat central nervous system. *J Comp Neurol*, 454(3):241–54.
- Bick, Y. and Sharman, G. (1975). The chromosomes of the platypus (*Ornithorhynchus*: Monotremata). *Cytobios*, 14:1728.
- Biesecker, L. G. and Green, R. C. (2014). Diagnostic clinical genome and exome sequencing. *N Engl J Med*, 371(12):1170.
- Bilgüvar, K., Oztürk, A. K., Louvi, A., *et al.* (2010). Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature*, 467(7312):207–210.

- Bininda-Emonds, O. R. P., Cardillo, M., Jones, K. E., *et al.* (2007). The delayed rise of present-day mammals. *Nature*, 446(7135):507–512.
- Bleazard, T., Ju, Y. S., Sung, J., and Seo, J. S. (2013). Fine-scale mapping of meiotic recombination in Asians. *BMC Genet*, 14:19.
- Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 27(4):578–579.
- Bokil, N. J., Baisden, J. M., Radford, D. J., and Summers, K. M. (2010). Molecular genetics of long QT syndrome. *Mol Genet Metab*, 101(1):1–8.
- Borrie, S. C., Baeumer, B. E., and Bandtlow, C. E. (2012). The Nogo-66 receptor family in the intact and diseased CNS. *Cell Tissue Res*, 349(1):105–17.
- Bosch, A. M., Stroek, K., Abeling, N. G., *et al.* (2012). The Brown-Vialetto-Van Laere and Fazio Londe syndrome revisited: natural history, genetics, treatment and future perspectives. *Orphanet J Rare Dis*, 7:83.
- Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, 32(3):314.
- Boycott, K. M., Vanstone, M. R., Bulman, D. E., and MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet*, 14(10):681–691.
- Boyd, L. J., Livingston, J. S., Brown, M. G., *et al.* (2005). Meiotic exchange event within the stalk region of an inverted chromosome 22 results in a recombinant chromosome with duplication of the distal long arm. *Am J Med Genet A*, 138(4):355–360.
- Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L., and Weber, J. L. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet*, 63(3):861–9.
- Broman, K. W. and Weber, J. L. (2000). Characterization of human crossover interference. *Am J Hum Genet*, 66(6):1911–1926.
- Brown, A. S., Feingold, E., Broman, K. W., and Sherman, S. L. (2000). Genome-wide variation in recombination in female meiosis: a risk factor for non-disjunction of chromosome 21. *Hum Mol Genet*, 9(4):515–523.

- Brozynska, M., Furtado, A., and Henry, R. J. (2015). Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnol J*.
- Brunetti-Pierri, N., Berg, J. S., Scaglia, F., *et al.* (2008). Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat Genet*, 40(12):1466–1471.
- Buske, O. J., Girdea, M., Dumitriu, S., *et al.* (2015). PhenomeCentral: A Portal for Phenotypic and Genotypic Matchmaking of Patients with Rare Genetic Diseases. *Hum Mutat*.
- Cabral, R. M., Kurban, M., Wajid, M., *et al.* (2012). Whole-exome sequencing in a single proband reveals a mutation in the CHST8 gene in autosomal recessive peeling skin syndrome. *Genomics*, 99(4):202–208.
- Campbell, C. L., Furlotte, N. A., Eriksson, N., Hinds, D., and Auton, A. (2015). Escape from crossover interference increases with maternal age. *Nat Commun*, 6:6260.
- Campbell, I. M., Yuan, B., Robberecht, C., *et al.* (2014). Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. *Am J Hum Genet*, 95(2):173–182.
- Cann, R. L., Stoneking, M., and Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, 325(6099):31–36.
- Carvill, G. L., Heavin, S. B., Yendle, S. C., *et al.* (2013). Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. *Nat Genet*, 45(7):825–830.
- Castro, J. A., Picornell, A., and Ramon, M. (1998). Mitochondrial DNA: a tool for populational genetics studies. *Int Microbiol*, 1(4):327–332.
- Cavalli-Sforza, L. and Bodmer, W. (1999). *The genetics of human populations*. Dover Publications.
- Cavalli-Sforza, L.L., M. P. . P. A. (1994). *The History and Geography of Human Genes*. Princeton University Press, Princeton, New Jersey.
- Chaisson, M. J. and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, 13:238.

- Chambille, I. (1999). Circadian rhythm of AMPA receptor GluR2/3 subunit-immunoreactivity in the suprachiasmatic nuclei of Syrian hamster and effect of a light-dark cycle. *Brain Res*, 833(1):27–38.
- Charlesworth, D., Charlesworth, B., and Marais, G. (2005). Steps in the evolution of heteromorphic sex chromosomes. *Heredity (Edinb)*, 95(2):118–128.
- Chergui, K., Svenningsson, P., and Greengard, P. (2005). Physiological role for casein kinase 1 in glutamatergic synaptic transmission. *J Neurosci*, 25(28):6601–9.
- Cheung, V. G., Burdick, J. T., Hirschmann, D., and Morley, M. (2007). Polymorphic variation in human meiotic recombination. *Am J Hum Genet*, 80(3):526–530.
- Chiang, T., Duncan, F. E., Schindler, K., Schultz, R. M., and Lampson, M. A. (2010). Evidence that weakened centromere cohesion is a leading cause of age-related aneuploidy in oocytes. *Curr Biol*, 20(17):1522–1528.
- Chiang, T., Schultz, R. M., and Lampson, M. A. (2012). Meiotic origins of maternal age-related aneuploidy. *Biol Reprod*, 86(1):1–7.
- Chiyonobu, T., Hayashi, S., Kobayashi, K., *et al.* (2007). Partial tandem duplication of GRIA3 in a male with mental retardation. *Am J Med Genet A*, 143A(13):1448–55.
- Choi, G., Kopplin, L. J., Tester, D. J., *et al.* (2004). Spectrum and frequency of cardiac channel defects in swimming-triggered arrhythmia syndromes. *Circulation*, 110(15):2119–24.
- Choi, J., Song, S.-H., Bak, C. W., *et al.* (2012). Impaired spermatogenesis and gr/gr deletions related to Y chromosome haplogroups in Korean men. *PLoS One*, 7(8):e43550.
- Choi, M., Scholl, U. I., Ji, W., *et al.* (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A*, 106(45):19096–101.
- Chong, J. X., Buckingham, K. J., Jhangiani, S. N., *et al.* (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet*, 97(2):199–215.

- Cleland, R. (1972). *Oenothera; Cytogenetics and Evolution*. Academic Press, New York.
- Colavito, V., Tesoriero, C., Wirtu, A. T., Grassi-Zuconi, G., and Bentivoglio, M. (2015). Limbic thalamus and state-dependent behavior: The paraventricular nucleus of the thalamic midline as a node in circadian timing and sleep/wake-regulatory networks. *Neurosci Biobehav Rev*, 54:3–17.
- Collins, F. S. (1995). Positional cloning moves from perditional to traditional. *Nature genetics*, 9(4):347–350.
- Colonna, V., Pistis, G., Bomba, L., *et al.* (2013). Small effective population size and genetic homogeneity in the Val Borbera isolate. *Eur J Hum Genet*, 21(1):89–94.
- Colver, A. F. and Sethumadhavan, T. (2003). The term diplegia should be abandoned. *Arch Dis Child*, 88(4):286–290.
- Connolly, J., Obendorf, D., Whittington, R., and Muir, D. (1998). Causes of morbidity and mortality in platypus (*Ornithorhynchus anatinus*) from Tasmania, with particular reference to Mucor amphibiorum infection. *Australian Mammalogy*, 20:177–187.
- Coop, G. and Przeworski, M. (2007). An evolutionary view of human recombination. *Nat Rev Genet*, 8(1):23–34.
- Coop, G., Wen, X., Ober, C., Pritchard, J. K., and Przeworski, M. (2008). High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science*, 319(5868):1395–8.
- Cooper, H., Hedges, L. V., and Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.
- Cortez, D., Marin, R., Toledo-Flores, D., *et al.* (2014). Origins and functional evolution of Y chromosomes across mammals. *Nature*, 508(7497):488–493.
- Couto, J. A., Vivero, M. P., Kozakewich, H. P. W., *et al.* (2015). A somatic MAP3K3 mutation is associated with verrucous venous malformation. *Am J Hum Genet*, 96(3):480–486.
- Craig, D. W. and Stephan, D. A. (2005). Applications of whole-genome high-density SNP genotyping. *Expert Rev Mol Diagn*, 5(2):159–170.

- Crisci, J. L., Poh, Y.-P., Bean, A., Simkin, A., and Jensen, J. D. (2012). Recent progress in polymorphism-based population genetic inference. *J Hered*, 103(2):287–296.
- Daetwyler, H. D., Capitan, A., Pausch, H., *et al.* (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet*, 46(8):858–865.
- D’Alessandro, L. C., Al Turki, S., Manickaraj, A. K., *et al.* (2015). Exome sequencing identifies rare variants in multiple genes in atrioventricular septal defect. *Genetics in Medicine*.
- Davidson, G., Wu, W., Shen, J., *et al.* (2005). Casein kinase 1 gamma couples Wnt receptor activation to cytoplasmic signal transduction. *Nature*, 438(7069):867–72.
- Davydov, E. V., Goode, D. L., Sirota, M., *et al.* (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*, 6(12):e1001025.
- de Knijff, P. (2000). Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. *Am J Hum Genet*, 67(5):1055–1061.
- Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, 519(7542):223–228.
- Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods*, 10(1):5–6.
- Della Mina, E., Ciccone, R., Brustia, F., *et al.* (2015). Improving molecular diagnosis in epilepsy by a dedicated high-throughput sequencing platform. *Eur J Hum Genet*, 23(3):354–362.
- Denayer, E. and Legius, E. (2007). What’s new in the neuro-cardio-facial-cutaneous syndromes? *Eur J Pediatr*, 166(11):1091–8.
- Deprez, L., Weckhuysen, S., Holmgren, P., *et al.* (2010). Clinical spectrum of early-onset epileptic encephalopathies associated with STXBP1 mutations. *Neurology*, 75(13):1159–65.

- DePristo, M. A., Banks, E., Poplin, R., *et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43(5):491–498.
- Dib, C., Fauré, S., Fizames, C., *et al.* (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, 380(6570):152–154.
- Djukic, A., Lado, F. A., Shinnar, S., and Moshe, S. L. (2006). Are early myoclonic encephalopathy (EME) and the Ohtahara syndrome (EIEE) independent of each other? *Epilepsy Res*, 70 Suppl 1:S68–76.
- Dohm, J. C., Tsend-Ayush, E., Reinhardt, R., Grutzner, F., and Himmelbauer, H. (2007). Disruption and pseudoautosomal localization of the major histocompatibility complex in monotremes. *Genome Biol*, 8(8):R175.
- Dong, C., Wei, P., Jian, X., *et al.* (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet*, 24(8):2125–2137.
- Donis-Keller, H., Green, P., Helms, C., *et al.* (1987). A genetic linkage map of the human genome. *Cell*, 51(2):319–337.
- Doyle, A. J., Doyle, J. J., Bessling, S. L., *et al.* (2012). Mutations in the TGF-beta repressor SKI cause Shprintzen-Goldberg syndrome with aortic aneurysm. *Nat Genet*, 44(11):1249–54.
- Duncan, E., Brown, M., and Shore, E. M. (2014). The revolution in human monogenic disease mapping. *Genes (Basel)*, 5(3):792–803.
- Emery, L. S., Magnaye, K. M., Bigham, A. W., Akey, J. M., and Bamshad, M. J. (2015). Estimates of continental ancestry vary widely among individuals with the same mtDNA haplogroup. *Am J Hum Genet*, 96(2):183–193.
- English, A. C., Richards, S., Han, Y., *et al.* (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One*, 7(11):e47768.
- Esko, T., Mezzavilla, M., Nelis, M., *et al.* (2013). Genetic characterization of north-eastern Italian population isolates in the context of broader European genetic diversity. *Eur J Hum Genet*, 21(6):659–665.

- Evans, D. M. and Cardon, L. R. (2004). Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. *Am J Hum Genet*, 75(4):687–692.
- Ewart, A. K., Morris, C. A., Atkinson, D., *et al.* (1993). Hemizyosity at the elastin locus in a developmental disorder, Williams syndrome. *Nature genetics*, 5(1):11–16.
- Falush, D., Stephens, M., and Pritchard, J. K. (2003a). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.
- Falush, D., Wirth, T., Linz, B., *et al.* (2003b). Traces of human migrations in *Helicobacter pylori* populations. *Science*, 299(5612):1582–1585.
- Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, 78(2):737–756.
- Fisher, R. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.
- Fledel-Alon, A., Leffler, E. M., Guan, Y., *et al.* (2011). Variation in human recombination rates and its genetic determinants. *PLoS One*, 6(6):e20321.
- Fledel-Alon, A., Wilson, D. J., Broman, K., *et al.* (2009). Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genet*, 5(9):e1000658.
- Fokkema, I. F. A. C., Taschner, P. E. M., Schaafsma, G. C. P., *et al.* (2011). LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat*, 32(5):557–563.
- Foote, A. D., Liu, Y., Thomas, G. W. C., *et al.* (2015). Convergent evolution of the genomes of marine mammals. *Nat Genet*, 47(3):272–275.
- Fossat, N., Jones, V., Garcia-Garcia, M. J., and Tam, P. P. (2012). Modulation of WNT signaling activity is key to the formation of the embryonic head. *Cell Cycle*, 11(1):26–32.
- Fu, W., O’Connor, T. D., Jun, G., *et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–20.
- Fujita, M. and Kinoshita, T. (2012). GPI-anchor remodeling: potential functions of GPI-anchors in intracellular trafficking and membrane dynamics. *Biochim Biophys Acta*, 1821(8):1050–8.

- Furlan, E., Griffiths, J., Gust, N., *et al.* (2013). Dispersal patterns and population structuring among platypuses, *Ornithorhynchus anatinus*, throughout south-eastern Australia. *Conservation Genetics*, 14(4):837–853.
- Furlan, E., Stoklosa, J., Griffiths, J., *et al.* (2012). Small population size and extremely low levels of genetic diversity in island populations of the platypus, *Ornithorhynchus anatinus*. *Ecol Evol*, 2(4):844–857.
- Gasparini, P., Rabionet, R., Barbujani, G., *et al.* (2000). High carrier frequency of the 35delG deafness mutation in European populations. Genetic Analysis Consortium of GJB2 35delG. *Eur J Hum Genet*, 8(1):19–23.
- Gattepaille, L. M., Jakobsson, M., and Blum, M. G. B. (2013). Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. *Heredity (Edinb)*, 110(5):409–419.
- Gecz, J., Barnett, S., Liu, J., *et al.* (1999). Characterization of the human glutamate receptor subunit 3 gene (GRIA3), a candidate for bipolar disorder and nonspecific X-linked mental retardation. *Genomics*, 62(3):356–68.
- Gelman, A. *et al.* (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3):515–534.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.
- Ghorbel, M., Baklouti-Gargouri, S., Keskes, R., *et al.* (2014). Deletion of CDY1b copy of Y chromosome CDY1 gene is a risk factor of male infertility in Tunisian men. *Gene*, 548(2):251–255.
- Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., *et al.* (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511(7509):344–347.
- Goldman, A. M., Glasscock, E., Yoo, J., *et al.* (2009). Arrhythmia in heart and brain: KCNQ1 mutations link epilepsy and sudden unexplained death. *Sci Transl Med*, 1(2):2ra6.
- Gongora, J., Swan, A. B., Chong, A. Y., *et al.* (2012). Genetic structure and phylogeography of platypuses revealed by mitochondrial DNA. *Journal of Zoology*, 286(2):110–119.

- Gossmann, T. I., Santure, A. W., Sheldon, B. C., Slate, J., and Zeng, K. (2014). Highly variable recombinational landscape modulates efficacy of natural selection in birds. *Genome Biol Evol.*
- Grant, T. (2007). *Platypus*. Australian Natural History Series. CSIRO Publishing, 4th edition.
- Grant, T. and Temple-Smith, P. (1983). Size, seasonal weight change and growth in platypuses, *Ornithorhynchus anatinus* (Monotremata: *Ornithorhynchidae*), from rivers and lakes of New South Wales. *Australian Mammalogy*, 6(1-2):51–60.
- Graves, J. A. M. (2006). Sex chromosome specialization and degeneration in mammals. *Cell*, 124(5):901–914.
- Groth, K. A., Gaustadnes, M., Thorsen, K., *et al.* (2015). Difficulties in diagnosing Marfan syndrome using current FBN1 databases. *Genet Med.*
- Grutzner, F., Rens, W., Tsend-Ayush, E., *et al.* (2004). In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird Z and mammal X chromosomes. *Nature*, 432(7019):913–7.
- Guergueltcheva, V., Azmanov, D. N., Angelicheva, D., *et al.* (2012). Autosomal-recessive congenital cerebellar ataxia is caused by mutations in metabotropic glutamate receptor 1. *Am J Hum Genet*, 91(3):553–64.
- Gusella, J. F., Wexler, N. S., Conneally, P. M., *et al.* (1983). A polymorphic DNA marker genetically linked to Huntington’s disease. *Nature*, 306(5940):234–238.
- Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*, 5(10):e1000695.
- Hale, L. R. and Singh, R. S. (1987). Mitochondrial DNA variation and genetic structure in populations of *Drosophila melanogaster*. *Mol Biol Evol*, 4(6):622–637.
- Handsaker, R. E., Korn, J. M., Nemesh, J., and McCarroll, S. A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet*, 43(3):269–76.
- Handyside, A. H., Harton, G. L., Mariani, B., *et al.* (2010). Karyomapping: a universal method for genome wide analysis of genetic disease based on mapping crossovers between parental haplotypes. *J Med Genet*, 47(10):651–658.

- Harter, A. V., Gardner, K. A., Falush, D., *et al.* (2004). Origin of extant domesticated sunflowers in eastern North America. *Nature*, 430(6996):201–205.
- Hassold, T. and Hunt, P. (2001). To err (meiotically) is human: the genesis of human aneuploidy. *Nat Rev Genet*, 2(4):280–91.
- Hassold, T., Sherman, S., and Hunt, P. (2000). Counting cross-overs: characterizing meiotic recombination in mammals. *Hum Mol Genet*, 9(16):2409–2419.
- Hayman, D. and Martin, P. (1974). *Mammalia I: Monotremata and Marsupialia.*, volume Volume 4 of *Animal cytogenetics*. Gebruder Borntraege.
- Henderson, S. A. and Edwards, R. G. (1968). Chiasma frequency and maternal age in mammals. *Nature*, 218(5136):22–28.
- Heron, S. E., Crossland, K. M., Andermann, E., *et al.* (2002). Sodium-channel defects in benign familial neonatal-infantile seizures. *Lancet*, 360(9336):851–2.
- Heron, S. E., Smith, K. R., Bahlo, M., *et al.* (2012). Missense mutations in the sodium-gated potassium channel gene KCNT1 cause severe autosomal dominant nocturnal frontal lobe epilepsy. *Nat Genet*, 44(11):1188–90.
- Hill, W. G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genet Res*, 8(3):269–294.
- Hinch, A. G., Tandon, A., Patterson, N., *et al.* (2011). The landscape of recombination in African Americans. *Nature*, 476(7359):170–5.
- Hino-Fukuyo, N., Haginoya, K., Iinuma, K., Uematsu, M., and Tsuchiya, S. (2009). Neuroepidemiology of West syndrome and early infantile epileptic encephalopathy in Miyagi Prefecture, Japan. *Epilepsy Res*, 87(2-3):299–301.
- Hirabayashi, Y., Itoh, Y., Tabata, H., *et al.* (2004). The Wnt/beta-catenin pathway directs neuronal differentiation of cortical neural precursor cells. *Development*, 131(12):2791–801.
- Ho, S. Y. W. and Shapiro, B. (2011). Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol Ecol Resour*, 11(3):423–434.
- Hoffman, M. D. and Gelman, A. (2011). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *arXiv preprint arXiv:1111.4246*.

- Hoischen, A., Gilissen, C., Arts, P., *et al.* (2010a). Massively parallel sequencing of ataxia genes after array-based enrichment. *Hum Mutat*, 31(4):494–499.
- Hoischen, A., van Bon, B. W., Gilissen, C., *et al.* (2010b). De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat Genet*, 42(6):483–5.
- Hong, Y., Ohishi, K., Watanabe, R., *et al.* (1999). GPI1 stabilizes an enzyme essential in the first step of glycosylphosphatidylinositol biosynthesis. *J Biol Chem*, 274(26):18582–8.
- Hou, Y., Fan, W., Yan, L., *et al.* (2013). Genome analyses of single human oocytes. *Cell*, 155(7):1492–1506.
- Huang, X., Yang, S.-P., Chinwalla, A. T., *et al.* (2006). Application of a superword array in genome assembly. *Nucleic acids research*, 34(1):201–205.
- Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour*, 9(5):1322–1332.
- Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*, 159(4):1805–1817.
- Hughes, J. F., Skaletsky, H., Brown, L. G., *et al.* (2012). Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature*, 483(7387):82–86.
- Hughes, J. F., Skaletsky, H., Pyntikova, T., *et al.* (2010). Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature*, 463(7280):536–539.
- Hurles, M. E., Dermitzakis, E. T., and Tyler-Smith, C. (2008). The functional impact of structural variation in humans. *Trends Genet*, 24(5):238–245.
- Hussin, J., Roy-Gagnon, M. H., Gendron, R., Andelfinger, G., and Awadalla, P. (2011). Age-dependent recombination rates in human pedigrees. *PLoS Genet*, 7(9):e1002251.
- Ingram, V. M. (1956). A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. *Nature*, 178(4537):792–794.

- International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.
- Iossifov, I., Ronemus, M., Levy, D., *et al.* (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron*, 74(2):285–99.
- Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet*, 44(2):226–232.
- James, C. and Moritz, C. (2000). Intraspecific phylogeography in the sedge frog *Litoria fallax* (Hylidae) indicates pre-Pleistocene vicariance of an open forest species from eastern Australia. *Molecular Ecology*, 9(3):349–358.
- Jamuar, S. S., Lam, A.-T. N., Kircher, M., *et al.* (2014). Somatic mutations in cerebral cortical malformations. *N Engl J Med*, 371(8):733–743.
- Jarvis, E. D., Mirarab, S., Aberer, A. J., *et al.* (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331.
- Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet*, 29(2):217–222.
- Johnston, J. J., Gropman, A. L., Sapp, J. C., *et al.* (2012). The phenotype of a germline mutation in PIGA: the gene somatically mutated in paroxysmal nocturnal hemoglobinuria. *Am J Hum Genet*, 90(2):295–300.
- Jones, M. A., Bhide, S., Chin, E., *et al.* (2011). Targeted polymerase chain reaction-based enrichment and next generation sequencing for diagnostic testing of congenital disorders of glycosylation. *Genet Med*, 13(11):921–932.
- Joseph D’Ercole, A. and Ye, P. (2008). Expanding the mind: insulin-like growth factor I and brain development. *Endocrinology*, 149(12):5958–62.
- Just, W., Baumstark, A., Süss, A., *et al.* (2007). *Ellobius lutescens*: sex determination and sex chromosome. *Sexual Development*, 1(4):211–221.
- Kaiser, J. (2010). Affordable ‘exomes’ fill gaps in a catalog of rare diseases. *Science*, 330(6006):903–903.

- Kales, S. C., Ryan, P. E., Nau, M. M., and Lipkowitz, S. (2010). Cbl and human myeloid neoplasms: the Cbl oncogene comes of age. *Cancer Res*, 70(12):4789–94.
- Kan, Y. W. and Dozy, A. M. (1978). Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. *Proceedings of the National Academy of Sciences*, 75(11):5631–5635.
- Kato, M., Saitoh, S., Kamei, A., *et al.* (2007). A longer polyalanine expansion mutation in the ARX gene causes early infantile epileptic encephalopathy with suppression-burst pattern (Ohtahara syndrome). *Am J Hum Genet*, 81(2):361–6.
- Keane, T. M., Goodstadt, L., Danecek, P., *et al.* (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294.
- Keightley, P. D. and Otto, S. P. (2006). Interference among deleterious mutations favours sex and recombination in finite populations. *Nature*, 443(7107):89–92.
- Kemp, S., Berger, J., and Aubourg, P. (2012). X-linked adrenoleukodystrophy: Clinical, metabolic, genetic and pathophysiological aspects. *Biochim Biophys Acta*.
- Kirkpatrick, M. and Guerrero, R. F. (2014). Signatures of sex-antagonistic selection on recombining sex chromosomes. *Genetics*, 197(2):531–541.
- Knowlton, R. G., Cohen-Haguenauer, O., Van Cong, N., *et al.* (1985). A polymorphic DNA marker linked to cystic fibrosis is located on chromosome 7.
- Kodera, H., Kato, M., Nord, A. S., *et al.* (2013). Targeted capture and sequencing for detection of mutations causing early onset epileptic encephalopathy. *Epilepsia*, 54(7):1262–9.
- Köhler, S., Doelken, S. C., Mungall, C. J., *et al.* (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res*, 42(Database issue):D966–D974.
- Kolomyjec, S. H., Chong, J. Y., Blair, D., *et al.* (2009). Population genetics of the platypus (*Ornithorhynchus anatinus*): a fine-scale look at adjacent river systems. *Australian Journal of Zoology*, 57(4):225–234.
- Kong, A., Barnard, J., Gudbjartsson, D. F., *et al.* (2004). Recombination rate and reproductive success in humans. *Nat Genet*, 36(11):1203–6.

- Kong, A. and Cox, N. J. (1997). Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet*, 61(5):1179–88.
- Kong, A., Gudbjartsson, D. F., Sainz, J., *et al.* (2002). A high-resolution recombination map of the human genome. *Nat Genet*, 31(3):241–7.
- Kong, A., Thorleifsson, G., Frigge, M. L., *et al.* (2014). Common and low-frequency variants associated with genome-wide recombination rate. *Nat Genet*, 46(1):11–16.
- Kong, A., Thorleifsson, G., Gudbjartsson, D. F., *et al.* (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–103.
- Kong, A., Thorleifsson, G., Stefansson, H., *et al.* (2008). Sequence variants in the RNF212 gene associate with genome-wide recombination rate. *Science*, 319(5868):1398–1401.
- Krawitz, P. M., Murakami, Y., Hecht, J., *et al.* (2012). Mutations in PIGO, a Member of the GPI-Anchor-Synthesis Pathway, Cause Hyperphosphatasia with Mental Retardation. *Am J Hum Genet*, 91(1):146–51.
- Krawitz, P. M., Schweiger, M. R., Rodelsperger, C., *et al.* (2010). Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet*, 42(10):827–9.
- Kripke, D. F., Klimecki, W. T., Nievergelt, C. M., *et al.* (2014). Circadian polymorphisms in night owls, in bipolars, and in non-24-hour sleep cycles. *Psychiatry Investig*, 11(4):345–362.
- Ku, C. S., Cooper, D. N., Polychronakos, C., *et al.* (2012). Exome sequencing: dual role as a discovery and diagnostic tool. *Ann Neurol*, 71(1):5–14.
- Kumar, R. A., KaraMohamed, S., Sudi, J., *et al.* (2008). Recurrent 16p11. 2 microdeletions in autism. *Human molecular genetics*, 17(4):628–638.
- Kurian, M. A., Meyer, E., Vassallo, G., *et al.* (2010). Phospholipase C beta 1 deficiency is associated with early-onset epileptic encephalopathy. *Brain*, 133(10):2964–70.
- Labasque, M. and Faivre-Sarrailh, C. (2010). GPI-anchored proteins at the node of Ranvier. *FEBS Lett*, 584(9):1787–92.

- Lahn, B. T. and Page, D. C. (1999). Four evolutionary strata on the human X chromosome. *Science*, 286(5441):964–967.
- Lander, E. S., Linton, L. M., Birren, B., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.
- Lanzone, C., Rodriguez, D., Cuello, P., *et al.* (2011). XY1Y2 chromosome system in *Salinomys delicatus* (Rodentia, Cricetidae). *Genetica*, 139(9):1143–1147.
- Lappin, F. M., Medert, C. M., Hawkins, K. K., *et al.* (2015). A polymorphic pseudoautosomal boundary in the *Carica papaya* sex chromosomes. *Mol Genet Genomics*, 290(4):1511–1522.
- Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G., and Maniatis, T. (1978). The isolation and characterization of linked delta- and beta-globin genes from a cloned library of human DNA. *Cell*, 15(4):1157–1174.
- Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genet*, 8(1):e1002453.
- Lee, W.-H., Bookstein, R., Hong, F., *et al.* (1987). Human retinoblastoma susceptibility gene: cloning, identification, and sequence. *Science*, 235(4794):1394–1399.
- Leffler, E. M., Gao, Z., Pfeifer, S., *et al.* (2013). Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science*, 339(6127):1578–1582.
- Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D., and Caccamo, M. (2014). NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics*, 30(4):566–568.
- Lenzi, M. L., Smith, J., Snowden, T., *et al.* (2005). Extreme heterogeneity in the molecular events leading to the establishment of chiasmata during meiosis i in human oocytes. *Am J Hum Genet*, 76(1):112–127.
- Leslie, S., Winney, B., Hellenthal, G., *et al.* (2015). The fine-scale genetic structure of the British population. *Nature*, 519(7543):309–314.
- Levine, M. A. (2012). An update on the clinical and molecular characteristics of pseudohypoparathyroidism. *Curr Opin Endocrinol Diabetes Obes*, 19(6):443–451.

- Levinsohn, J. L., McNiff, J. M., Antaya, R. J., and Choate, K. A. (2015). A Somatic p.G45E GJB2 Mutation Causing Porokeratotic Eccrine Ostial and Dermal Duct Nevus. *JAMA Dermatol*, 151(6):638–641.
- Li, G., Davis, B. W., Raudsepp, T., *et al.* (2013a). Comparative analysis of mammalian Y chromosomes illuminates ancestral structure and lineage-specific evolution. *Genome Res*, 23(9):1486–1495.
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.
- Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496.
- Li, M., Yang, L., Li, C., *et al.* (2010). Mutational spectrum of the ADAR1 gene in dyschromatosis symmetrica hereditaria. *Arch Dermatol Res*, 302(6):469–76.
- Li, M. X., Kwan, J. S., Bao, S. Y., *et al.* (2013b). Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet*, 9(1):e1003143.
- Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233.
- Li, S., Li, B., Cheng, C., *et al.* (2014). Genomic signatures of near-extinction and rebirth of the crested ibis and other endangered bird species. *Genome Biol*, 15(12):557.
- Lindhurst, M. J., Parker, V. E. R., Payne, F., *et al.* (2012). Mosaic overgrowth with fibroadipose hyperplasia is caused by somatic activating mutations in PIK3CA. *Nat Genet*, 44(8):928–933.
- Lindhurst, M. J., Sapp, J. C., Teer, J. K., *et al.* (2011). A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *N Engl J Med*, 365(7):611–619.
- Litt, M. and Luty, J. A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet*, 44(3):397–401.

- Lo Giudice, T., Lombardi, F., Santorelli, F. M., Kawarai, T., and Orlacchio, A. (2014). Hereditary spastic paraplegia: clinical-genetic characteristics and evolving molecular mechanisms. *Exp Neurol*, 261:518–539.
- Lunter, G. and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res*, 21(6):936–9.
- Lupski, J. R., de Oca-Luna, R. M., Slaugenhaupt, S., *et al.* (1991). DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell*, 66(2):219–232.
- Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., *et al.* (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med*, 362(13):1181–91.
- MacArthur, D. G., Manolio, T. A., Dimmock, D. P., *et al.* (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469–476.
- MacDonald, M. E., Ambrose, C. M., Duyao, M. P., *et al.* (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington’s disease chromosomes. *Cell*, 72(6):971–983.
- Manase, D., DAlessandro, L. C., Manickaraj, A. K., *et al.* (2014). High throughput exome coverage of clinically relevant cardiac genes. *BMC medical genomics*, 7(1):67.
- Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., *et al.* (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873.
- Marshall, C. R., Noor, A., Vincent, J. B., *et al.* (2008). Structural variation of chromosomes in autism spectrum disorder. *The American Journal of Human Genetics*, 82(2):477–488.
- Martin, H. C., Christ, R., Hussin, J. G., *et al.* (2015). Multicohort analysis of the maternal age effect on recombination. *Nat Commun*, 6:7846.
- Martin, H. C., Kim, G. E., Pagnamenta, A. T., *et al.* (2014). Clinical whole-genome sequencing in severe early-onset epilepsy reveals new genes and improves molecular diagnosis. *Hum Mol Genet*, 23(12):3200–3211.
- Martinelli, S., De Luca, A., Stellacci, E., *et al.* (2010). Heterozygous germline mutations in the CBL tumor-suppressor gene cause a Noonan syndrome-like phenotype. *Am J Hum Genet*, 87(2):250–7.

- Martins, V. G. and Mesa, A. (1995). Two permanent linear chains of sex chromosomes in *Neotermes fulvescens* and karyotypes of two other neotropical *Kalotermitidae* species (Insecta, Isoptera). *Genome*, 38(5):958–967.
- Mathieson, I. and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nat Genet*, 44(3):243–246.
- Maydan, G., Noyman, I., Har-Zahav, A., *et al.* (2011). Multiple congenital anomalies-hypotonia-seizures syndrome is caused by a mutation in PIGN. *J Med Genet*, 48(6):383–9.
- McCarthy, D. J., Humburg, P., Kanapin, A., *et al.* (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Med*, 6(3):26.
- McEvoy, B. P., Powell, J. E., Goddard, M. E., and Visscher, P. M. (2011). Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res*, 21(6):821–829.
- McGaugh, S. E., Heil, C. S., Manzano-Winkler, B., *et al.* (2012). Recombination modulates how selection affects linked sites in *Drosophila*.
- McKean, D. M. and Niswander, L. (2012). Defects in GPI biosynthesis perturb Cripto signaling during forebrain development in two new mouse models of holoprosencephaly. *Biol Open*, 1(9):874–83.
- McMillan, D., Miethke, P., Alsop, A. E., *et al.* (2007). Characterizing the chromosomes of the platypus (*Ornithorhynchus anatinus*). *Chromosome Res*, 15(8):961–974.
- McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., *et al.* (2008). Runs of homozygosity in European populations. *Am J Hum Genet*, 83(3):359–372.
- McTague, A., Appleton, R., Avula, S., *et al.* (2013). Migrating partial seizures of infancy: expansion of the electroclinical, radiological and pathological disease spectrum. *Brain*, 136(Pt 5):1578–1591.
- McVean, G., Awadalla, P., and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160(3):1231–1241.

- McVean, G. A., Myers, S. R., Hunt, S., *et al.* (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–4.
- Melani, F., Mei, D., Pisano, T., *et al.* (2011). CDKL5 gene-related epileptic encephalopathy: electroclinical findings in the first year of life. *Dev Med Child Neurol*, 53(4):354–60.
- Mensah, M. A., Hestand, M. S., Larmuseau, M. H. D., *et al.* (2014). Pseudoautosomal region 1 length polymorphism in the human population. *PLoS Genet*, 10(11):e1004578.
- Middleton, F. A., Pato, M. T., Gentile, K. L., *et al.* (2004). Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. *Am J Hum Genet*, 74(5):886–897.
- Mills, P. B., Surtees, R. A., Champion, M. P., *et al.* (2005). Neonatal epileptic encephalopathy caused by mutations in the PNPO gene encoding pyridox(am)ine 5'-phosphate oxidase. *Hum Mol Genet*, 14(8):1077–86.
- Molinari, F., Raas-Rothschild, A., Rio, M., *et al.* (2005). Impaired mitochondrial glutamate transport in autosomal recessive neonatal myoclonic epilepsy. *Am J Hum Genet*, 76(2):334–9.
- Møller, R. S., Heron, S. E., Larsen, L. H. G., *et al.* (2015). Mutations in KCNT1 cause a spectrum of focal epilepsies. *Epilepsia*, 56(9):e114–e120.
- Monaco, A. P., Neve, R. L., Colletti-Feener, C., *et al.* (1986). Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. *Nature*, 323(6089):646–650.
- Montenegro, G., Powell, E., Huang, J., *et al.* (2011). Exome sequencing allows for rapid gene identification in a Charcot-Marie-Tooth family. *Ann Neurol*, 69(3):464–470.
- Muller, H. J. (1932). Some genetic aspects of sex. *American Naturalist*, pages 118–138.
- Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutat Res*, 106:2–9.

- Mundia, M. M., Desai, V., Magwood, A. C., and Baker, M. D. (2014). Nascent DNA synthesis during homologous recombination is synergistically promoted by the rad51 recombinase and DNA homology. *Genetics*, 197(1):107–119.
- Murtagh, C. (1977). A Unique Cytogenetic System in Monotremes. *Chromosoma*, 65:37– 57.
- Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324.
- Myers, S., Bowden, R., Tumian, A., *et al.* (2010). Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, 327(5967):876–9.
- Nadachowska-Brzyska, K., Burri, R., Olason, P. I., *et al.* (2013). Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. *PLoS Genet*, 9(11):e1003942.
- Nagaoka, S. I., Hassold, T. J., and Hunt, P. A. (2012). Human aneuploidy: mechanisms and new insights into an age-old problem. *Nat Rev Genet*, 13(7):493–504.
- Nagy, G. R., Gyrfly, B., Nagy, B., and Rigó, Jr, J. (2013). Lower risk for Down syndrome associated with longer oral contraceptive use: a case-control study of women of advanced maternal age presenting for prenatal diagnosis. *Contraception*, 87(4):455–458.
- Najmabadi, H., Hu, H., Garshasbi, M., *et al.* (2011). Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature*, 478(7367):57–63.
- Nakamura, K., Kato, M., Osaka, H., *et al.* (2013). Clinical spectrum of SCN2A mutations expanding to Ohtahara syndrome. *Neurology*, 81(11):992–998.
- Neale, B. M., Kou, Y., Liu, L., *et al.* (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397):242–5.
- Need, A. C., Shashi, V., Hitomi, Y., *et al.* (2012). Clinical application of exome sequencing in undiagnosed genetic conditions. *J Med Genet*, 49(6):353–61.
- Nei, M. (1969). Linkage modification and sex difference in recombination. *Genetics*, 63(3):681.

- Nei, M. and Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76(10):5269–5273.
- Ng, P. C. and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31(13):3812–4.
- Ng, S. B., Buckingham, K. J., Lee, C., *et al.* (2010a). Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*, 42(1):30–5.
- Ng, S. B., Nickerson, D. A., Bamshad, M. J., and Shendure, J. (2010b). Massively parallel sequencing and rare disease. *Human molecular genetics*, 19(R2):R119–R124.
- Ng, S. B., Turner, E. H., Robertson, P. D., *et al.* (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–6.
- Niemeyer, C. M., Kang, M. W., Shin, D. H., *et al.* (2010). Germline CBL mutations cause developmental abnormalities and predispose to juvenile myelomonocytic leukemia. *Nat Genet*, 42(9):794–800.
- Nordli, D. R., J. (2012). Epileptic encephalopathies in infants and children. *J Clin Neurophysiol*, 29(5):420–4.
- Norton, N., Robertson, P. D., Rieder, M. J., *et al.* (2012). Evaluating pathogenicity of rare variants from dilated cardiomyopathy in the exome era. *Circ Cardiovasc Genet*, 5(2):167–174.
- Novembre, J., Johnson, T., Bryc, K., *et al.* (2008). Genes mirror geography within Europe. *Nature*, 456(7218):98–101.
- Novembre, J. and Ramachandran, S. (2011). Perspectives on human population structure at the cusp of the sequencing era. *Annu Rev Genomics Hum Genet*, 12:245–274.
- Novembre, J. and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nat Genet*, 40(5):646–649.
- O’Connell, J., Gurdasani, D., Delaneau, O., *et al.* (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet*, 10(4):e1004234.

- Ogiwara, I., Ito, K., Sawaishi, Y., *et al.* (2009). De novo mutations of voltage-gated sodium channel alphaII gene SCN2A in intractable epilepsies. *Neurology*, 73(13):1046–53.
- Ohba, C., Kato, M., Takahashi, N., *et al.* (2015). De novo KCNT1 mutations in early-onset epileptic encephalopathy. *Epilepsia*.
- Ohtahara, S. and Yamatogi, Y. (2003). Epileptic encephalopathies in early infancy with suppression-burst. *J Clin Neurophysiol*, 20(6):398–407.
- O’Roak, B. J., Vives, L., Fu, W., *et al.* (2012a). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science*, 338(6114):1619–22.
- O’Roak, B. J., Vives, L., Girirajan, S., *et al.* (2012b). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397):246–50.
- Otley, H. M., Munks, S. A., and Hindell, M. A. (2000). Activity patterns, movements and burrows of platypuses (*Ornithorhynchus anatinus*) in a sub-alpine Tasmanian lake. *Australian Journal of Zoology*, 48(6):701–713.
- Otto, S. P. and Lenormand, T. (2002). Resolving the paradox of sex and recombination. *Nat Rev Genet*, 3(4):252–261.
- Otto, S. P., Pannell, J. R., Peichel, C. L., *et al.* (2011). About PAR: the distinct evolutionary dynamics of the pseudoautosomal region. *Trends Genet*, 27(9):358–67.
- Ottolini, C. S., Newnham, L. J., Capalbo, A., *et al.* (2015). Genome-wide maps of recombination and chromosome segregation in human oocytes and embryos show selection for maternal recombination rates. *Nat Genet*, 47(7):727–735.
- Oza, A. U., Lovett, K. E., Williams, S. E., and Moritz, C. (2012). Recent speciation and limited phylogeographic structure in Mixophyes frogs from the Australian Wet Tropics. *Mol Phylogenet Evol*, 62(1):407–413.
- Pagel, P., Kovac, S., Oesterheld, M., *et al.* (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–4.

- Pagnamenta, A. T., Lise, S., Harrison, V., *et al.* (2012a). Exome sequencing can detect pathogenic mosaic mutations present at low allele frequencies. *J Hum Genet*, 57(1):70–2.
- Pagnamenta, A. T., Lise, S., Harrison, V., *et al.* (2012b). Exome sequencing can detect pathogenic mosaic mutations present at low allele frequencies. *J Hum Genet*, 57(1):70–72.
- Paigen, K., Szatkiewicz, J. P., Sawyer, K., *et al.* (2008). The recombinational anatomy of a mouse chromosome. *PLoS Genet*, 4(7):e1000119.
- Pardo-Manuel de Villena, F. and Sapienza, C. (2001). Recombination is proportional to the number of chromosome arms in mammals. *Mamm Genome*, 12(4):318–322.
- Park, G., Gim, J., Kim, A. R., *et al.* (2013). Multiphasic analysis of whole exome sequencing data identifies a novel mutation of ACTG1 in a nonsyndromic hearing loss family. *BMC Genomics*, 14:191.
- Parvanov, E. D., Petkov, P. M., and Paigen, K. (2010). Prdm9 controls activation of mammalian recombination hotspots. *Science*, 327(5967):835.
- Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190.
- Pavone, P., Spalice, A., Polizzi, A., Parisi, P., and Ruggieri, M. (2012). Ohtahara syndrome with emphasis on recent genetic discovery. *Brain Dev*, 34(6):459–68.
- Peltonen, L. and McKusick, V. A. (2001). Dissecting human disease in the postgenomic era. *Science*, 291(5507):1224–1229.
- Perdiguerro, P., Venturas, M., Cervera, M. T., Gil, L., and Collada, C. (2015). Massive sequencing of *Ulmus minor*’s transcriptome provides new molecular tools for a genus under the constant threat of Dutch elm disease. *Front Plant Sci*, 6:541.
- Perez, B., Mechinaud, F., Galambrun, C., *et al.* (2010). Germline mutations of the CBL gene define a new genetic syndrome with predisposition to juvenile myelomonocytic leukaemia. *J Med Genet*, 47(10):686–91.
- Phillips, M. J., Bennett, T. H., and Lee, M. S. Y. (2009). Molecules, morphology, and ecology indicate a recent, amphibious ancestry for echidnas. *Proc Natl Acad Sci U S A*, 106(40):17089–17094.

- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2014). *nlme: Linear and Nonlinear Mixed Effects Models*.
- Pinkel, D., Segreaves, R., Sudar, D., *et al.* (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature genetics*, 20(2):207–211.
- Piry, S., Alapetite, A., Cornuet, J.-M., *et al.* (2004). GENECLASS2: a software for genetic assignment and first-generation migrant detection. *Journal of heredity*, 95(6):536–539.
- Plagnol, V., Curtis, J., Epstein, M., *et al.* (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, 28(21):2747–2754.
- Polani, P. E. and Crolla, J. A. (1991). A test of the production line hypothesis of mammalian oogenesis. *Hum Genet*, 88(1):64–70.
- Polani, P. E. and Jagiello, G. M. (1976). Chiasmata, meiotic univalents, and age in relation to aneuploid imbalance in mice. *Cytogenet Cell Genet*, 16(6):505–529.
- Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, 20(1):110–21.
- Pool, J. E., Hellmann, I., Jensen, J. D., and Nielsen, R. (2010). Population genetic inference from genomic sequence variation. *Genome Res*, 20(3):291–300.
- Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., *et al.* (2013). Great ape genetic diversity and population history. *Nature*, 499(7459):471–475.
- Price, A. L., Tandon, A., Patterson, N., *et al.* (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*, 5(6):e1000519.
- Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.
- Puliti, A., Caridi, G., Ravazzolo, R., and Ghiggeri, G. M. (2007). Teaching molecular genetics: chapter 4-positional cloning of genetic disorders. *Pediatr Nephrol*, 22(12):2023–2029.

- Purcell, S., Neale, B., Todd-Brown, K., *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–575.
- Putnam, N. H., O’Connell, B., Stites, J. C., *et al.* (2015). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *arXiv preprint arXiv:1502.05331*.
- Qiu, S., Bergero, R., Guirao-Rico, S., *et al.* (2015). RAD mapping reveals an evolving, polymorphic and fuzzy boundary of a plant pseudoautosomal region. *Mol Ecol*.
- Raile, K., Klammt, J., Schneider, A., *et al.* (2006). Clinical and functional characteristics of the human Arg59Ter insulin-like growth factor i receptor (IGF1R) mutation: implications for a gene dosage effect of the human IGF1R. *J Clin Endocrinol Metab*, 91(6):2264–71.
- Raja Rayan, D. L. and Hanna, M. G. (2010). Skeletal muscle channelopathies: nondystrophic myotonias and periodic paralysis. *Curr Opin Neurol*, 23(5):466–76.
- Reich, D., Price, A. L., and Patterson, N. (2008). Principal component analysis of genetic data. *Nat Genet*, 40(5):491–492.
- Reis, L. M., Tyler, R. C., Muheisen, S., *et al.* (2013). Whole exome sequencing in dominant cataract identifies a new causative factor, CRYBA2, and a variety of novel alleles in known genes. *Hum Genet*.
- Rens, W., O’Brien, P., Grutzner, F., *et al.* (2007). The multiple sex chromosomes of platypus and echidna are not completely identical and several share homology with the avian Z. *Genome Biol*, 8(11):R243.
- Riazuddin, S., Belyantseva, I. A., Giese, A. P., *et al.* (2012). Alterations of the CIB2 calcium- and integrin-binding protein cause Usher syndrome type 1J and nonsyndromic deafness DFNB48. *Nat Genet*, 44(11):1265–71.
- Rilstone, J. J., Alkhater, R. A., and Minassian, B. A. (2013). Brain dopamine-serotonin vesicular transport disease and its treatment. *N Engl J Med*, 368(6):543–550.
- Rimmer, A., Phan, H., Mathieson, I., *et al.* (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet*, 46(8):912–918.

- Riordan, J. R., Rommens, J. M., Kerem, B.-s., *et al.* (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*, 245(4922):1066–1073.
- Ritchie, G. R. S., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. *Nat Methods*, 11(3):294–296.
- Riviere, J. B., van Bon, B. W., Hoischen, A., *et al.* (2012). De novo mutations in the actin genes ACTB and ACTG1 cause Baraitser-Winter syndrome. *Nat Genet*, 44(4):440–4, S1–2.
- Roach, J. C., Glusman, G., Smit, A. F., *et al.* (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978):636–9.
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., *et al.* (2011). Integrative genomics viewer. *Nat Biotechnol*, 29(1):24–26.
- Robinson, W. P., Kuchinka, B. D., Bernasconi, F., *et al.* (1998). Maternal meiosis I non-disjunction of chromosome 15: dependence of the maternal age effect on level of recombination. *Hum Mol Genet*, 7(6):1011–1019.
- Ronemus, M., Iossifov, I., Levy, D., and Wigler, M. (2014). The role of de novo mutations in the genetics of autism spectrum disorders. *Nat Rev Genet*, 15(2):133–141.
- Ropers, H. H. (2010). Genetics of early onset cognitive impairment. *Annu Rev Genomics Hum Genet*, 11:161–187.
- Rosenberg, N. A., Burke, T., Elo, K., *et al.* (2001). Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics*, 159(2):699–713.
- Rosenberg, N. A. and Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet*, 3(5):380–390.
- Rosenberg, N. A., Pritchard, J. K., Weber, J. L., *et al.* (2002). Genetic structure of human populations. *Science*, 298(5602):2381–2385.
- Rowsey, R., Gruhn, J., Broman, K. W., Hunt, P. A., and Hassold, T. (2014). Examining variation in recombination levels in the human female: a test of the production-line hypothesis. *Am J Hum Genet*, 95(1):108–112.

- Royer-Pokora, B., Kunkel, L. M., Monaco, A. P., *et al.* (1986). Cloning the gene for an inherited human disorder—chronic granulomatous disease—on the basis of its chromosomal location. *Nature*, 322(6074):32–38.
- Saitsu, H., Kato, M., Koide, A., *et al.* (2012). Whole exome sequencing identifies KCNQ2 mutations in Ohtahara syndrome. *Ann Neurol*, 72(2):298–300.
- Saitsu, H., Kato, M., Mizuguchi, T., *et al.* (2008). De novo mutations in the gene encoding STXBP1 (MUNC18-1) cause early infantile epileptic encephalopathy. *Nat Genet*, 40(6):782–8.
- Saitsu, H., Tohyama, J., Kumada, T., *et al.* (2010). Dominant-negative mutations in alpha-II spectrin cause West syndrome with severe cerebral hypomyelination, spastic quadriplegia, and developmental delay. *Am J Hum Genet*, 86(6):881–91.
- Samocha, K. E., Robinson, E. B., Sanders, S. J., *et al.* (2014). A framework for the interpretation of de novo mutation in human disease. *Nat Genet*, 46(9):944–950.
- Sanders, S. J., Murtha, M. T., Gupta, A. R., *et al.* (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485(7397):237–41.
- Sandstedt, S. A. and Tucker, P. K. (2004). Evolutionary strata on the mouse X chromosome correspond to strata on the human X chromosome. *Genome Res*, 14(2):267–272.
- Sawcer, S. J., Maranian, M., Singlehurst, S., *et al.* (2004). Enhancing linkage analysis of complex disorders: an evaluation of high-density genotyping. *Hum Mol Genet*, 13(17):1943–1949.
- Scally, A. and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet*, 13(10):745–753.
- Schaid, D. J., Guenther, J. C., Christensen, G. B., *et al.* (2004). Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer–susceptibility loci. *The American Journal of Human Genetics*, 75(6):948–965.
- Schäuble, C. and Moritz, C. (2001). Comparative phylogeography of two open forest frogs from eastern Australia. *Biological Journal of the Linnean Society*, 74(2):157–170.

- Scheinfeldt, L. B. and Tishkoff, S. A. (2013). Recent human adaptation: genomic approaches, interpretation and insights. *Nature Reviews Genetics*, 14(10):692–702.
- Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nat Genet*, 46(8):919–925.
- Schossig, A., Wolf, N. I., Fischer, C., *et al.* (2012). Mutations in ROGDI Cause Kohlschutter-Tonz Syndrome. *Am J Hum Genet*, 90(4):701–7.
- Schrader, K. A., Heravi-Moussavi, A., Waters, P. J., *et al.* (2011). Using next-generation sequencing for the diagnosis of rare disorders: a family with retinitis pigmentosa and skeletal abnormalities. *J Pathol*, 225(1):12–18.
- Schwarz, J. M., Cooper, D. N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*, 11(4):361–362.
- Sehat, B., Andersson, S., Girnita, L., and Larsson, O. (2008). Identification of c-Cbl as a new ligase for insulin-like growth factor-I receptor with distinct roles from Mdm2 in receptor ubiquitination and endocytosis. *Cancer Res*, 68(14):5669–77.
- Seielstad, M., Bekele, E., Ibrahim, M., Touré, A., and Traoré, M. (1999). A view of modern human origins from Y chromosome microsatellite variation. *Genome Res*, 9(6):558–567.
- Sella, G., Petrov, D. A., Przeworski, M., and Andolfatto, P. (2009). Pervasive natural selection in the *Drosophila* genome? *PLoS Genet*, 5(6):e1000495.
- Sen, S., Ambulkar, P., Hinduja, I., *et al.* (2015). Susceptibility of gr/gr rearrangements to azoospermia or oligozoospermia is dependent on DAZ and CDY1 gene copy deletions. *J Assist Reprod Genet*.
- Sharman, G. and Barber, H. (1952). Multiple sex chromosomes in the marsupial Potorous. *Heredity*, 6:345–355.
- Shashi, V., McConkie-Rosell, A., Rosell, B., *et al.* (2013). The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. *Genetics in Medicine*, 16(2):176–182.
- Shaw-Smith, C. J., Lewis, S. J., and Reid, E. (2004). X-linked adrenoleukodystrophy presenting as autosomal dominant pure hereditary spastic paraparesis. *J Neurol Neurosurg Psychiatry*, 75(5):686–8.

- Sheehan, S., Harris, K., and Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. *Genetics*, 194(3):647–662.
- Shen, J., Gilmore, E. C., Marshall, C. A., *et al.* (2010). Mutations in PNKP cause microcephaly, seizures and defects in DNA repair. *Nat Genet*, 42(3):245–9.
- Sherry, S. T., Ward, M. H., Kholodov, M., *et al.* (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29(1):308–311.
- Shi, X., Yasumoto, S., Kurahashi, H., *et al.* (2011). Clinical spectrum of SCN2A mutations. *Brain Dev*.
- Siepel, A., Bejerano, G., Pedersen, J. S., *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–50.
- Simpson, J. T., Wong, K., Jackman, S. D., *et al.* (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res*, 19(6):1117–1123.
- Singh, N. A., Charlier, C., Stauffer, D., *et al.* (1998). A novel potassium channel gene, KCNQ2, is mutated in an inherited epilepsy of newborns. *Nat Genet*, 18(1):25–9.
- Singh, N. A., Westenskow, P., Charlier, C., *et al.* (2003). KCNQ2 and KCNQ3 potassium channel genes in benign familial neonatal convulsions: expansion of the functional and mutation spectrum. *Brain*, 126(Pt 12):2726–37.
- Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., *et al.* (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423(6942):825–837.
- Smit, AFA, H. R. . G. P. (2013-2015). RepeatMasker Open-4.0.
- Smith, K. N. and Nicolas, A. (1998). Recombination at work for meiosis. *Curr Opin Genet Dev*, 8(2):200–11.
- Smith, K. R., Bromhead, C. J., Hildebrand, M. S., *et al.* (2011). Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol*, 12(9):R85.

- Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene. *Hum Mutat.*
- Soh, Y. Q. S., Alföldi, J., Pyntikova, T., *et al.* (2014). Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell*, 159(4):800–813.
- Spencer, C. C. A., Deloukas, P., Hunt, S., *et al.* (2006). The influence of recombination on human genetic diversity. *PLoS Genet*, 2(9):e148.
- Srour, M., Schwartzentruber, J., Hamdan, F. F., *et al.* (2012). Mutations in C5ORF42 Cause Joubert Syndrome in the French Canadian Population. *Am J Hum Genet*, 90(4):693–700.
- St. Hilaire, C., Ziegler, S. G., Markello, T. C., *et al.* (2011). NT5E mutations and arterial calcifications. *New England Journal of Medicine*, 364(5):432–442.
- Stan Development Team (2014). RStan: the R interface to Stan, Version 2.5.0.
- Stankiewicz, P. and Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annu Rev Med*, 61:437–455.
- Steenland, H. W., Kim, S. S., and Zhuo, M. (2008). GluR3 subunit regulates sleep, breathing and seizure generation. *Eur J Neurosci*, 27(5):1166–73.
- Stenson, P. D., Mort, M., Ball, E. V., *et al.* (2009). The Human Gene Mutation Database: 2008 update. *Genome Med*, 1(1):13.
- Stevison, L. S., Woerner, A. E., Kidd, J. M., *et al.* (2015). The Time-Scale of Recombination Rate Evolution in Great Apes. *bioRxiv*, page 013755.
- Stumpf, M. P. and Goldstein, D. B. (2001). Genealogical and evolutionary inference with the human Y chromosome. *Science*, 291(5509):1738–1742.
- Sugawara, S. and Mikamo, K. (1983). Absence of correlation between univalent formation and meiotic nondisjunction in aged female Chinese hamsters. *Cytogenet Cell Genet*, 35(1):34–40.
- Suzuki, Y., Kobayashi, M., Miyashita, H., *et al.* (2010). Isolation of a small vasohibin-binding protein (SVBP) and its role in vasohibin secretion. *J Cell Sci*, 123(Pt 18):3094–101.

- Swaminathan, G. and Tsygankov, A. Y. (2006). The Cbl family proteins: ring leaders in regulation of cell signaling. *J Cell Physiol*, 209(1):21–43.
- Swaminathan, G. J., Bragin, E., Chatzimichali, E. A., *et al.* (2012). DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. *Hum Mol Genet*, 21(R1):R37–R44.
- Syren, R. M. and Luykx, P. (1977). Permanent segmental interchange complex in the termite *Incisitermes schwarzi*.
- Tarpey, P. S., Smith, R., Pleasance, E., *et al.* (2009). A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nat Genet*, 41(5):535–43.
- Tavyev Asher, Y. J. and Scaglia, F. (2012). Molecular bases and clinical spectrum of early infantile epileptic encephalopathies. *Eur J Med Genet*, 55(5):299–306.
- Taylor, J. C., Martin, H. C., Lise, S., *et al.* (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat Genet*, 47(7):717–726.
- Tease, C. and Hulten, M. A. (2004). Inter-sex variation in synaptonemal complex lengths largely determine the different recombination rates in male and female germ cells. *Cytogenet Genome Res*, 107(3-4):208–215.
- Tennessen, J. A., Bigham, A. W., O’Connor, T. D., *et al.* (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64–69.
- Thevenon, J., Callier, P., Andrieux, J., *et al.* (2013). 12p13.33 microdeletion including ELKS/ERC1, a new locus associated with childhood apraxia of speech. *Eur J Hum Genet*, 21(1):82–88.
- Thomas, C. G., Wang, W., Jovelin, R., *et al.* (2015). Full-genome evolutionary histories of selfing, splitting, and selection in *Caenorhabditis*. *Genome Res*, 25(5):667–678.
- Thompson, M. D., Roscioli, T., Marcelis, C., *et al.* (2012). Phenotypic variability in hyperphosphatasia with seizures and neurologic deficit (Mabry syndrome). *Am J Med Genet A*, 158A(3):553–8.

- Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2012). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.*
- Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat*, 32(4):358–68.
- Tian, C., Gregersen, P. K., and Seldin, M. F. (2008). Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet*, 17(R2):R143–R150.
- Timmerman, V., Strickland, A. V., and Züchner, S. (2014). Genetics of Charcot-Marie-Tooth (CMT) Disease within the Frame of the Human Genome Project Success. *Genes (Basel)*, 5(1):13–32.
- Torgasheva, A. A., Rubtsov, N. B., and Borodin, P. M. (2013). Recombination and synaptic adjustment in oocytes of mice heterozygous for a large paracentric inversion. *Chromosome Res*, 21(1):37–48.
- Touma, M., Joshi, M., Connolly, M. C., *et al.* (2013). Whole genome sequencing identifies SCN2A mutation in monozygotic twins with Ohtahara syndrome and unique neuropathologic findings. *Epilepsia*, 54(5):e81–e85.
- Trombetta, B., Sellitto, D., Scozzari, R., and Cruciani, F. (2014). Inter- and intraspecies phylogenetic analyses reveal extensive X-Y gene conversion in the evolution of gametologous sequences of human sex chromosomes. *Mol Biol Evol*, 31(8):2108–2123.
- Turner, D. J., Miretti, M., Rajan, D., *et al.* (2008). Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet*, 40(1):90–95.
- Underhill, P. A., Shen, P., Lin, A. A., *et al.* (2000). Y chromosome sequence variation and the history of human populations. *Nat Genet*, 26(3):358–361.
- Utami, K. H., Hillmer, A. M., Aksoy, I., *et al.* (2014). Detection of chromosomal breakpoints in patients with developmental delay and speech disorders. *PLoS One*, 9(6):e90852.

- Utge, S., Kronholm, E., Partonen, T., *et al.* (2011). Shared genetic background for regulation of mood and sleep: association of GRIA3 with sleep duration in healthy Finnish women. *Sleep*, 34(10):1309–16.
- Veeramah, K. R. and Hammer, M. F. (2014). The impact of whole-genome sequencing on the reconstruction of human population history. *Nat Rev Genet*, 15(3):149–162.
- Veeramah, K. R., Johnstone, L., Karafet, T. M., *et al.* (2013). Exome sequencing reveals new causal mutations in children with epileptic encephalopathies. *Epilepsia*.
- Veltman, J. A. and Brunner, H. G. (2012). De novo mutations in human genetic disease. *Nat Rev Genet*, 13(8):565–75.
- Venkateswaran, S., Myers, K. A., Smith, A. C., *et al.* (2014). Whole-exome sequencing in an individual with severe global developmental delay and intractable epilepsy identifies a novel, de novo GRIN2A mutation. *Epilepsia*, 55(7):e75–e79.
- Veyrunes, F., Waters, P. D., Miethke, P., *et al.* (2008). Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome research*, 18(6):965–973.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48.
- Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24.
- Visscher, P. M. and Gibson, G. (2013). What if we had whole-genome sequence data for millions of individuals? *Genome Medicine*, 5(80).
- Vitart, V., Bencić, G., Hayward, C., *et al.* (2010). Heritabilities of ocular biometrical traits in two croatian isolates with extended pedigrees. *Invest Ophthalmol Vis Sci*, 51(2):737–743.
- Vogel, M. W., Caston, J., Yuzaki, M., and Mariani, J. (2007). The Lurcher mouse: fresh insights from an old mutant. *Brain Res*, 1140:4–18.
- Walenkamp, M. J., van der Kamp, H. J., Pereira, A. M., *et al.* (2006). A variable degree of intrauterine and postnatal growth retardation in a family with a missense mutation in the insulin-like growth factor I receptor. *J Clin Endocrinol Metab*, 91(8):3062–70.

- Wang, D. G., Fan, J. B., Siao, C. J., *et al.* (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–1082.
- Wang, H., Liu, H., and Zhang, Z.-w. (2011a). Elimination of redundant synaptic inputs in the absence of synaptic strengthening. *J Neurosci*, 31(46):16675–16684.
- Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 38(16):e164.
- Wang, X., Wang, H., Cao, M., *et al.* (2011b). Whole-exome sequencing identifies ALMS1, IQCB1, CNGA3, and MYO7A mutations in patients with Leber congenital amaurosis. *Hum Mutat*, 32(12):1450–9.
- Wang, Y., Lu, Y., Zhang, Y., *et al.* (2015). The draft genome of the grass carp (*Ctenopharyngodon idellus*) provides insights into its evolution and vegetarian adaptation. *Nat Genet*, 47(6):625–631.
- Ward, L. D. and Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*, 337(6102):1675–8.
- Warren, W. C., Hillier, L. W., Marshall Graves, J. A., *et al.* (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453(7192):175–83.
- Watanabe, R., Inoue, N., Westfall, B., *et al.* (1998). The first step of glycosylphosphatidylinositol biosynthesis is mediated by a complex of PIG-A, PIG-H, PIG-C and GPI1. *EMBO J*, 17(4):877–85.
- Watson, J. M., Spencer, J. A., Riggs, A. D., and Graves, J. (1990). The X chromosome of monotremes shares a highly conserved region with the eutherian and marsupial X chromosomes despite the absence of X chromosome inactivation. *Proceedings of the National Academy of Sciences*, 87(18):7125–7129.
- Weber, J. L. and May, P. E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American journal of human genetics*, 44(3):388.
- Weckhuysen, S., Mandelstam, S., Suls, A., *et al.* (2012). KCNQ2 encephalopathy: emerging phenotype of a neonatal epileptic encephalopathy. *Ann Neurol*, 71(1):15–25.

- Weiss, L. A., Shen, Y., Korn, J. M., *et al.* (2008). Association between microdeletion and microduplication at 16p11. 2 and autism. *New England Journal of Medicine*, 358(7):667–675.
- Weissenbach, J. (1993). Microsatellite polymorphisms and the genetic linkage map of the human genome. *Curr Opin Genet Dev*, 3(3):414–417.
- Wetterstrand, K. (2015). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).
- White, R., Woodward, S., Leppert, M., *et al.* (1985). A closely linked genetic marker for cystic fibrosis. *Nature*, 318:382–384.
- Williams, A. L., Genovese, G., Dyer, T., *et al.* (2015). Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife*, 4:e04637.
- Worthey, E. A., Mayer, A. N., Syverson, G. D., *et al.* (2011). Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med*, 13(3):255–62.
- Wright, C. F., Fitzgerald, T. W., Jones, W. D., *et al.* (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet*, 385(9975):1305–1314.
- Wu, Y., Arai, A. C., Rumbaugh, G., *et al.* (2007). Mutations in ionotropic AMPA receptor 3 alter channel properties and are associated with moderate cognitive impairment in humans. *Proc Natl Acad Sci U S A*, 104(46):18163–8.
- Xia, Q., Guo, Y., Zhang, Z., *et al.* (2009). Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science*, 326(5951):433–436.
- Xu, H., Guan, J., Yi, H., and Yin, S. (2014). A systematic review and meta-analysis of the association between serotonergic gene polymorphisms and obstructive sleep apnea syndrome. *PLoS One*, 9(1):e86460.
- Xue, Y., Chen, Y., Ayub, Q., *et al.* (2012). Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am J Hum Genet*, 91(6):1022–1032.

- Xue, Y., Prado-Martinez, J., Sudmant, P. H., *et al.* (2015). Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, 348(6231):242–245.
- Yang, Y., Muzny, D. M., Xia, F., *et al.* (2014). Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA*, 312(18):1870–1879.
- Yau, C. (2013). OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics*, 29(19):2482–2484.
- Yinan, M., Yu, Q., Zhiyue, C., *et al.* (2004). Polymorphisms of casein kinase I gamma 2 gene associated with simple febrile seizures in Chinese Han population. *Neurosci Lett*, 368(1):2–6.
- Zhang, X., Guo, B. R., Cai, L. Q., *et al.* (2012). Exome sequencing identified a missense mutation of EPS8L3 in Marie Unna hereditary hypotrichosis. *J Med Genet*, 49(12):727–30.
- Zhou, X., Ma, A., Liu, X., *et al.* (2006). Infantile seizures and other epileptic phenotypes in a Chinese family with a missense mutation of KCNQ2. *Eur J Pediatr*, 165(10):691–5.
- Zuo, J., De Jager, P. L., Takahashi, K. A., *et al.* (1997). Neurodegeneration in Lurcher mice caused by mutation in delta2 glutamate receptor gene. *Nature*, 388(6644):769–73.