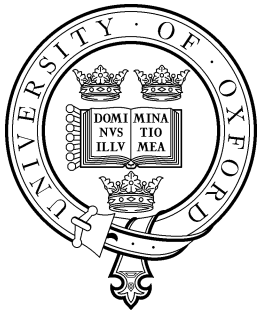
Inferring the fine-scale structure and evolution of recombination from high-throughput genome sequencing



Oliver Claude Venn

The Queen's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2012

Acknowledgements

It has been an great opportunity to study on the Medical Genomics and Statistics Programme. I would like to thank the course directors Jonathan Flint, Chris Holmes, Gil McVean and Richard Mott, and the Wellcome Trust for generously funding my studies and research.

I would to thank my supervisor Gil McVean. He has inspired, and forged, my scientific development through his intellect, rigour and trust. Additionally Peter Donnelly, Jonathan Flint, Richard Mott, Simon Myers and Molly Przeworski have been great mentors and I would like to thank them for allowing me to contribute to their research projects. I have thoroughly enjoyed working with them and their talented group members in Oxford and Chicago. I would also like to thank Rob Klose, Hanna Long, and David Sims (Non-methylated regions in humans); Anjali Gupta and Simon Myers (African-specific genetic map), Zam Iqbal and Isaac Turner (chimpanzee variants called through *de novo* assembly) Simon Myers (probabilities of HapMap hotspots being driven by the 13 bp recombination motif) for sharing unpublished data with me.

It has been fantastic to do research alongside talented students in the McVean, Myers and Marchini groups. I enjoyed the inspiring discussions, companionship, and joint troubleshooting sessions. I would like to thank Alex Dilthey, Zam Iqbal, and Hillary Martin for reading thesis chapters.

Finally, I would like to thank my family for their unquestioning support and love, and Emily who has shared every moment of this DPhil experience with me.

Inferring the fine-scale structure and evolution of recombination from high-throughput genome sequencing

Oliver Claude Venn, The Queen's College
(Wellcome Trust Centre for Human Genetics)

Thesis submitted for the degree of Doctor of Philosophy
University of Oxford, Michaelmas Term 2012.

Abstract

In eukaryotes, recombination plays a critical role in both the production of viable gametes and as a population genetic process. Here, we are interested in studying recombination as it provides insight into a process that has shaped variation. To this end, we study the evolution of cross-over rates in chimpanzees and humans through two experiments.

Components of the recombination machinery are well described in yeast and *C. elegans*, but less so in other species. In humans, cross-over rates vary across physical scales and occur predominantly in narrow ∼2 kb regions called *hotspots*, where hotspot usage differs considerably between individuals. Differential hotspot usage is associated with specific DNA motifs, and DNA-contacting zinc finger array variants in the trans-acting PRDM9 H3K4 trimethyltransferase. The precise relationship between DNA motifs, PRDM9 and hotspot activity is not completely understood.

Experiment 1. To investigate the importance of PRDM9 motif recognition, which is predicted be different between humans and chimpanzees, and the effect of PRDM9 on the evolution of fine-scale cross-over rates, we sequenced 10 unrelated *Pan troglodytes verus* (Western chimpanzee) genomes to moderate coverage (∼10×). I validate the approach by demonstrating that fine-scale maps estimated from 10 human genomes of each African and European ancestry recapitulate independently estimated maps. Then I characterise the error modes in sequencing data arising from errors in chemistry, alignment, variant calling, and genotyping. I identify several cryptic error modes missed by state-of-the-art filters and develop methods to counteract them.

To guard against genotype error arising from stochastic variation in low to moderate coverage sequencing, I develop methods to incorporate the underlying statistical uncertainty into recombination analyses, evaluate the approaches through simulation (estimated 11% improvement) and empirical assessment (estimated 4% improvement), and discover that the reported genotype uncertainty is poorly calibrated, which limits the approaches. Consequently, a filtering approach was applied to the hard-called chimpanzee genotypes.

I estimate recombination rates in chimpanzees through an existing LD-based method. In contrast to humans, there is no increased cross-over localisation around chimpanzee PRDM9 binding predictions, nor motifs consistently associated with activity. Hotspots do not overlap between the two species, indicating that rates evolved rapidly and consistent with PRDM9 localising all hotspots. In contrast, gene promoters and CpG islands are common attractors of recombination (2.7-fold increase

in rate in chimpanzee, 1.5-fold increase in human), suggesting chromatin state influences hotspot placement but to varying degree in the species. I discuss the potential implications for PRDM9 mechanism.

Experiment 2. To enable a more representative characterisation of the spectrum of genome changes occurring in chimpanzee genomes, I analyse data from an extended three generation Western chimpanzee pedigree sequenced at high coverage ($\sim$30$\times$). I use Mendel transmission to filter variants, infer haplotypes, and identify recombination events through a Hidden Markov Model approach. We detect 375 recombination events, of which 3 are double cross-over events. Sex-specific recombination rate estimates in chimpanzees mirror sex differences in humans ($N_{♂}/N_{♀} = 0.58$) and have similar levels of total recombination. We resolve recombination events typically at $\sim$ 856 base-pair resolution. Additionally, analyses of Mendel inconsistencies suggest that extended pedigree sequencing opens the door on studying complex genome changes.

These experiments demonstrate the power of comparative analyses, the utility of high throughput sequencing in enabling the study of recombination in almost any species of interest, the challenges in sifting signal from noise in these data, and the need for experimental and algorithmic methods to guard against error.

Contents

1 Introduction **1**
1.1 The biology of recombination . . . 3
1.1.1 The recombination process in eukaryotes . . . 4
1.1.2 The distribution of meiotic recombination in yeast, mouse, and human . . . 6
1.1.2.1 The structure of recombination rate variation at broad scales . . . 6
1.1.2.2 The structure of recombination rate variation at fine-scales . . . 9
1.1.2.3 Predicting zinc finger binding sequences . . . 10
1.1.2.4 The position weight matrix . . . 11
1.1.2.5 The role of PRDM9 in recombination . . . 13
1.2 Methods to identify recombination events . . . 16
1.2.1 A genealogical view of recombination . . . 16
1.2.2 Pedigree studies . . . 18
1.2.2.1 The deCODE genetic map . . . 23
1.2.3 Linkage disequilibrium . . . 26
1.2.4 Counting recombination events . . . 28
1.2.5 Models of recombination . . . 29
1.2.5.1 The ancestral recombination graph . . . 29
1.2.5.2 Approximate likelihood approaches . . . 31
1.2.5.3 Approximate genealogy approaches . . . 33
1.2.5.4 The African American admixture-based recombination map . . . 34
1.2.6 Typing recombination events in single gamete . . . 35
1.3 Discussion . . . 36

2 Challenges in the analysis of genome sequence data **38**
2.1 Reference genomes . . . 39
2.2 CASE 1: Western Chimpanzee . . . 42
2.2.0.1 Properties of the chimpanzee reference . . . 43
2.2.1 Recombination map construction . . . 44
2.2.1.1 LDhat parameters . . . 45
2.2.2 Detection and removal of errors . . . 48

2.2.2.1 Errors arising from reference mis-assembly driven false-negatives . . . 48
2.2.2.2 Errors arising from reference mis-assembly driven false-positives . . . 52
2.2.2.3 The construction of a robust map of recombination . 58
2.2.2.4 Generation of genetic maps from 10 human genomes of African and European ancestry . . . 58
2.2.2.5 Assessing HTS-map quality by comparison to independent fine-scale maps . . . 66
2.2.2.6 Comparison of sampled genetic maps to reference maps at Winckler et al. (2005) regions . . . 73
2.3 CASE 2: Wild Mus musculus domesticus . . . 76
2.4 Discussion . . . 85
2.4.1 The anatomy of HTS-errors . . . 86
2.4.2 Approaches to guard against error modes . . . 87
2.4.3 Case study: estimating recombination rates from reference genome free variant calling . . . 89
2.5 Methods . . . 93
2.5.1 Chimpanzee background methods . . . 93
2.5.1.1 Sample preparation and sequencing . . . 93
2.5.1.2 Read mapping . . . 93
2.5.1.3 Variant discovery, genotyping and filtering . . . 93
2.5.1.4 Genotype refinement and phasing . . . 95
2.5.1.5 Human-chimpanzee syntenic regions . . . 95
2.5.1.6 False discovery rates and genotype accuracy . . . 98
2.5.1.7 SNP discovery power . . . 99
2.5.1.8 Assessment of genetic map filters by simulation . . . 99
2.5.1.9 Estimates of effective population size . . . 100
2.5.1.10 Recombination hotspot inference . . . 101
2.5.2 Wild mice sequencing . . . 102
2.5.2.1 Data generation . . . 102

3 Methods for estimating LD and testing for recombination in HTS 106
3.1 Inferring haplotype frequencies through Expectation Maximisation for LD estimation . . . 107
3.1.1 Deriving the likelihood of haplotype frequencies . . . 108
3.1.2 Estimating haplotype frequencies through Expectation Maximisation applied to diplotypes . . . 110
3.1.3 Incorporating genotype uncertainty . . . 111
3.1.3.1 Measuring genotype uncertainty using genotype likelihoods . . . 112
3.1.3.2 Likelihood calculation with genotype uncertainty . . 113
3.2 Inferring evidence for recombination activity through pair-wise likelihood ratio tests . . . 114
3.2.1 Pair-wise likelihood ratio test on hard calls . . . 114

3.2.2 Constructing compatible sets of genotypes from genotype likelihoods . . . 116
3.2.2.1 Simulation study 1: Assessment of the compatible set approach for different HTS paradigms . . . 118
3.2.2.2 Simulation study 2: Impact of order dependency in compatible set construction . . . 124
3.2.3 Incorporating compatible sets into the pair-wise likelihood ratio test . . . 126
3.3 Evaluation of uncertainty aware method performance . . . 128
3.3.1 Data . . . 128
3.3.2 Empirical study 1: LD . . . 129
3.3.3 Empirical study 2: pair-wise likelihood ratio test . . . 133
3.4 Challenges . . . 135
3.4.1 Assessment of genotype likelihood calibration . . . 135
3.4.2 Examples of artefactual LD patterns at SNP clusters in mapped HTS . . . 138
3.4.3 Challenges in predicting errors from assembly sequence statistics 141
3.5 Discussion . . . 142

4 Features of fine-scale recombination rates in chimpanzees 144
4.1 Methods to evaluate cross over activity at genome features . . . 144
4.1.1 Sampling to remove clustering . . . 145
4.1.1.1 Comparisons between human and chimpanzee . . . 148
4.1.2 Quantifying recombination-influencing activity . . . 148
4.2 Recombination hotspots . . . 150
4.3 The distribution of fine-scale recombination rates across the chimpanzee genome . . . 153
4.4 Fine-scale recombination variation structure at genome features . . . 157
4.4.1 Genes and CpG islands . . . 158
4.4.2 DNA repeat elements . . . 165
4.5 Detecting DNA sequence motifs associated with recombination . . . 167
4.5.1 Impact of DNA motifs on recombination in DNA repeats . . . 167
4.5.2 Refinement of predicted PRDM9 binding motifs . . . 178
4.6 Discussion . . . 187

5 Sequencing of an extended chimpanzee pedigree 189
5.1 Acknowledgements . . . 192
5.2 Methods . . . 192
5.2.1 Sample preparation and sequencing . . . 193
5.2.2 Alignment and variant calling . . . 193
5.2.2.1 Identifying a robust call set: intersection with variants called through Cortex . . . 195
5.2.3 Estimates of genotype-error in chimpanzee HTS . . . 196
5.2.4 Inferring inheritance vector scaffolds . . . 197
5.2.5 Calculating the likelihood of an inheritance vector at a site . . . 201

5.2.6 Generating a scaffold of inheritance vectors across genomic segments . . . 202
5.2.7 Model for recombination . . . 204
5.2.8 Hidden Markov Model to infer a broad-scale inheritance vector scaffold . . . 206
5.2.9 Inheritance vector path across segments . . . 208
5.2.10 Mapping recombination events at fine-scale through posterior decoding . . . 209
5.2.11 Genome-wide map of cross-over events . . . 210
5.3 Estimates of genetic map length in male and female chimpanzees . . . 214
5.3.1 Anatomy of a single and a double cross-over event . . . 219
5.4 Anatomy of Mendel inconsistent genotypes . . . 221
5.5 Discussion . . . 225

6 Discussion **226**

Bibliography **242**

List of Figures

1.1 Meiotic recombination pathways. 5
1.2 Chromatin and nucleosome organisation influence DSB formation in yeast. 8
1.3 Recombination occurs in 1-2 kb hotspots. 8
1.4 Model of Cys2His2 DNA recognition for three zinc fingers (Jamieson et al., 2003). 10
1.5 Schematic of the human PRDM9 protein and estimated zinc finger array binding-prediction. 12
1.6 The recombination motif associated with human hotspots matches the human PRDM9 B-allele DNA-binding prediction. 14
1.7 PRDM9 zinc finger (ZF) variant diversity in human populations and cross-over activity in sperm. 15
1.8 The genealogy of three chromosomes. 17
1.9 Example of allele transmission through a simple pedigree. 19
1.10 Lander-Green Hidden Markov Model for estimating recombination rates from genotype data. 24
1.11 Sex-specific recombination rates around genes from deCODE pedigree-based recombination maps. 25

2.1 Errors or incompleteness in reference assemblies affect mapping. . . . 40
2.2 The impact of population mutation rate parameter θ on LD-based recombination rates estimated from 10 HTS Western chimpanzee genomes on chromosome 21. 47
2.3 Artefacual increases in genetic maps estimated from HTS genomes. . 48
2.4 Mis-assembled contigs in chimpanzee reference. 51
2.5 Chimpanzee reference genome (panTro2) contigs tend to end in repetitive sequence. 52
2.6 Artefacts arising from mis-assembly around HERVH repeat sequences. 54
2.7 The majority of HERVH elements in the chimpanzee reference genome are fragments. 56
2.8 An example of reads mapped to a HERVH represented in the chimpanzee reference. 57
2.9 Mis-assembled contigs in human reference (B36). 63
2.10 Recombination rate estimates for SNPs lifted over from mis-assembled contigs in the human reference (B36, shown in Figure 2.9) to B37. . . 64

2.11 Recombination rate estimates from haplotypes inferred for the 10 YRI genomes, but now from the 1000 Genomes Phase I Integrated data set (Consortium et al., 2012) which was aligned to B37 (same region as displayed in Figures 2.9 and 2.10). 65
2.12 Comparison of estimates of human fine-scale recombination rates at the (Winckler et al., 2005) regions, estimated using different approaches. 74
2.13 Estimated heterozygosity across chromosome 7 for four of the 20 HTS mice genomes. 78
2.14 Erroneous heterozygous genotype calls at small homozygous alternative indels in the 20 HTS mice genomes. 81
2.15 Mis-mapping at a segmental duplication in the 20 HTS mouse genomes. 82
2.16 Genotyping error arising from the GATK genotype model assumptions, which are inappropriate for the 20 wild mice. 83
2.17 Considerable variation in the shotgun-library insert sizes for four of the 20 sequenced mice. 84
2.18 Pearson correlation coefficients at different scales between recombination rates estimated from the combined-population HapMap map, and the *filtered* alignment-based map in chimpanzees (black; regional filters zeroed 6 artefactual jumps in the estimated map), and the *un-filtered* map from the genotypes called by the GATK at sites called by Cortex in chimpanzees (blue; no regional filtering required). 91
2.19 Regions of broad-scale synteny between chimpanzee and human reference genomes. 97
2.20 Map of the capture sites for the 20 sequenced mice located near the town of Ambert, France. 104

3.1 Exemplar of method to incorporate uncertainty in to equivalent set construction. 117
3.2 Precision of the compatible set approach. 121
3.3 Computational cost of the compatible set approach. 123
3.4 Density estimates for the minimum 2-locus genotype probability for the genotype model with maximum likelihood in each sample (N = 100 simulations). 126
3.5 Differences in LD estimates at 7q21 across call sets. 131
3.6 LD structure estimated from SNP-typing genotypes (Winckler), hard called HTS genotypes (hGT), haplotypes from LD-refined genotype likelihoods (rHT), and through incorporating uncertainty into haplotype frequency estimation (uaGT). 132
3.7 Evidence for recombination events inferred by pair-wise pseudo-likelihood ratio test. 134
3.8 The proportion of correct HTS genotypes stratified by their estimated genotype probabilities. 136
3.9 Distribution of GATK genotype quality scores for HTS genotypes. . . 138
3.10 Fine-scale estimates of recombination rates and LD in chimpanzees at problematic HTS SNPs. 140

3.11 Fine-scale recombination rate estimates in 10 chimpanzees from SNP-typing genotypes, and from genotypes inferred from HTS. 142

4.1 Impact of feature clustering on average fine-scale recombination rate variation profile. 146
4.2 Maximum likelihood model to quantify fine-scale recombination-influencing activity at genome features. 149
4.3 Fine-scale recombination rate variation around inferred hotspots in chimpanzees and humans. 152
4.4 Recombination rates around human hotspots that have no clear match to the defined PRDM9 binding motif. 152
4.5 The concentration of recombination rate in fine-scale genetic maps estimated for chimpanzee and human populations. 154
4.6 Models of the impact of PRDM9 on recombination hotspots. 155
4.7 The fine-scale profile of recombination rate variation around genomic features in chimpanzees and humans. 160
4.8 The fine-scale profile of recombination variation at genes in human and chimpanzee stratified by Cytosine methylation proportion measured in sperm (Molaro et al., 2011) over the 1 kb surrounding the transcription start sites (TSS). 163
4.9 The fine-scale profile of recombination rate variation around hypomethylated regions (HMRs) in human testes (red, N = 13,731) and liver (black = 8,483) samples for **(a)** human recombination rates (HapMap combined-population map) and **(b)** chimpanzee recombination rates. 164
4.10 Recombination rates around DNA repeat elements in chimpanzees and humans. 166
4.11 Recombination rate estimates around selected DNA repeats that have signals of recombination activity in chimpanzees. 167
4.12 P-values for systematic length differences between hot and cold element backgrounds. 169
4.13 Recombination-activating motif candidates on AluSq, AluJo and THE1B repeat backgrounds. 170
4.14 Recombination-activating motif candidates on AluJo and L2 repeat backgrounds. 171
4.15 Recombination rate in AluY elements as a function of CpG count within the repeat sequence. Mean rate estimates are shown for chimpanzee (red), and HapMap (black), with error bars representing 95% CI. 171
4.16 Motif-based recombination-activity is systematically much stronger in humans than chimpanzees. 174
4.17 Logo plots (Schneider and Stephens, 1990) showing predicted binding sequences for PRDM9 Zinc Finger (ZF) variants predicted for Bonobo (A1), Eastern (A1, E1) and Western (W1-W11) chimpanzee, and human A- and C-alleles (HA, HC). 183

4.18 Information content in PRDM9 allele position weight matrix predictions for humans and chimpanzees. 184
4.19 Sequence and structural variation in chimpanzee PRDM9 and implications for hotspot motifs. 185
4.20 Predicted chimpanzee PRDM9 binding alleles show little, if any, evidence for activation in non-repeat DNA. 186

5.1 The sequenced Western chimpanzee pedigree. 189
5.2 Three example transmission classes. 192
5.3 Proportion of heterozygous genotype calls on chromosome X. 195
5.4 Relatedness between individuals in the sequenced pedigree. 196
5.5 Visual representation of an inheritance vector. 198
5.6 Labelling strategy for a simple pedigree. 199
5.7 Example of redundant inheritance vectors. 200
5.8 Example of the *onto* mapping from an observed genotype vector to the set of compatible inheritance vectors. 200
5.9 Difference in log-likelihood between the aggregated window maximum likelihood inheritance vector and the other inheritance vectors over chromosome 21. 204
5.10 Example of how recombination events are enumerated. 205
5.11 Inferred inheritance vector for segments across chromosome 14 for **(a)** the Mendel consistent proportions, and **(b)** the inheritance vector likelihoods. 208
5.12 Cross-over breakpoint resolution 210
5.13 The average distribution of recombination events across chromosomes. 211
5.14 Map of cross-over events across the genome. 212
5.15 The average distribution of recombination events across chromosomes. 213
5.16 Total number of cross-over events per individual. 215
5.17 Comparisons of total genetic distances between the sexes, and between species. 217
5.18 Posterior decoding of two recombination events. 220
5.19 False Mendel inconsistent (MIC) signals generated through genotyping error. 222
5.20 Genotype probabilities for sites in a segregating deletion on chromosome 21. 223
5.21 Alignments across a segregating deletion on chromosome 21. 224

6.1 Average profile of fine-scale recombination rate variation estimated from the 20 wild mice at anti-DMC1 ChIP-Seq foci measured in different mouse strains by Brick et al. (2012) 230
6.2 Average profile of fine-scale recombination rate variation estimated in 20 wild mice at anti-H3K4Me3 ChIP-Seq foci in different mouse strains. 232

6.3 Fine-scale recombination rate variation in humans around matches to the degenerate 13 bp motif (red), and matches to motifs one nucleotide change away from the degenerate motif (black). The average recombination rate profile was estimated from the combined-population HapMap recombination map over 500bp intervals after thinning to remove motif clustering. 233
6.4 Space of interactions determining recombination hotspot placement. . 234
6.5 Hypothetical model of differential distributions of DSB locations between sex-polarised crosses of 13R and 9R mice. 239

Chapter 1

Introduction

Recombination is a fundamental biological process occurring in most known organisms including eukaryotes, bacteria and viruses. The combined effect of recombination and mutation underlie all genetic variation, the raw material on which natural selection acts, and each are therefore important in the process of evolution. Mutation creates new sequence variants, whereas recombination generates new combinations within a population. Despite early measurement of the signals of recombination (Morgan, 1911), and its importance in genetics research (cf., Stumpf and McVean, 2003), many questions remain about aspects of the recombination process: what influences the placement of events and determines their outcomes, what is the extent of variability in meioses between and within individuals, and how do recombination rates evolve? A diversity of approaches have been used in the investigation of recombination over the past 100 years, from observing visible phenotypes in *Drosophila* and *Abraxas* (Morgan, 1911), cytogenetic analysis of chromosome structures in *Zea mays* (Creighton and McClintock, 1931), and recently marker-based genetic analyses in pedigrees (e.g., Broman et al., 1998), linkage disequilibrium (though the theoretical considerations of LD and recombination go back at least 30 years cf., Hill and Robertson, 1968a) and statistical inference (reviewed, Stumpf and McVean, 2003) in populations, sperm

typing in males (Jeffreys et al., 2001) and ChIP-Seq (Borde et al., 2009). A particularly a useful approach has been to study recombination from the perspective of its impact on the genealogy relating a sample of chromosomes from unrelated individuals (Hudson, 1983; Griffiths and Marjoram, 1996). These relationships are inferred from the patterns of genetic variation between the sampled sequences. However, the interpretation of the observed patterns can be problematic without an understanding of the generative process. Statistical modelling of the evolutionary process through which the variation was generated, described by the coalescent process, has been used to understand patterns of variation and has played an important role in the study of recombination in the last decade (Clark et al., 2010). A recent advance in the field is the development of high-throughput methods for measuring polymorphism. High-throughput sequencing (HTS) has transformed genetics research, enabling rapid measurement of whole genome sequence data from any sample or species of interest. However, the problems of detection and interpretation of genetic variation from HTS data are far from solved. Rather, false positive and false negative errors can obscure biological signal. In this thesis I present methods to address the challenges of inferring historical recombination rates from population HTS of chimpanzees, mice, and humans.

For the remainder of this chapter, I will introduce the process of recombination, review the current state of knowledge of its properties, and describe the various methods by which it can be detected from genetic data. In the subsequent chapters, I turn to the following topics.

Chapter 2: I describe sources of error in variant discovery and genotyping from HTS, and the characteristic symptoms of the different error modes. Importantly many of the errors I describe are not detected by best-practice filters and are only revealed by analysing haplotype structure.

Chapter 3: I develop a new method to incorporate the genotype uncertainty under-

lying variants called from HTS. I assess the methods through simulation and empirical assessment.

Chapter 4: I apply recombination rate estimation methods to genome-wide variation data from HTS of 10 Western chimpanzee genomes sequenced to moderate coverage (10×). I consider the relationship between human and chimpanzee fine-scale recombination rates around recombination hotspots, genes, CpG islands, cytosine-methylation, DNA repeat element families, and DNA sequence motifs.

Chapter 5: I analyse data from an extended Western chimpanzee pedigree sequenced to high coverage (30×). I describe and apply methods to detect cross-over events within the pedigree. In addition, I use genetic transmission to validate variants called from the sequencing of the 10 unrelated chimpanzees.

Chapter 6: I conclude with a discussion of how work presented in this thesis contributes to the field and suggest potential avenues for further research.

1.1 The biology of recombination

In eukaryotes, recombination plays a critical role in both the production of viable gametes and as a population genetic processes. Mechanistically, in many eukaryotes, chiasmata (the physical manifestations of cross-overs) generate the physical connection between homologous chromosomes required for proper reductional segregation at the first meiotic division in prophase I (c.f. Keeney et al., 1997). In addition to ensuring correct homolog segregation, recombination has fundamental importance in the process of evolution because it generates novel haplotypes on which natural selection acts (e.g., Marais and Charlesworth, 2003). The production of new haplotypes within a population results in both the creation and separation of allelic combinations. The

balance between these opposing outcomes shapes the adaptive value of recombination, a topic of active research (Brandvain and Coop, 2012; Eshel and Feldman, 1970; Feldman et al., 1996; Keightley and Otto, 2006; Otto and Lenormand, 2002). Illegitimate recombination between homologous, but non-allelic, stretches of DNA (non allelic homologous recombination, NAHR) such as gene families, segmental duplications, and repeat elements are an important source of mutation (e.g., Lupski, 2007) and their outcomes (deletions, duplications and inversions) have been implicated in multiple genetic disorders (cf., Sasaki et al., 2010; Myers et al., 2008).

1.1.1 The recombination process in eukaryotes

The main pathways of cross-over formation were uncovered in yeast and are conserved in other eukaryotes (Lichten and de Massy, 2011). Recombination is initiated through the formation of localised DNA double-strand breaks (DSBs) catalysed by the evolutionarily conserved Spo11 protein, which is a homolog of the catalytic subunit of topoisomerase VI with little, or no, sequence specificity (Figure 1.1a, Bergerat et al., 1997; Keeney et al., 1997). After the initial cleavage, Spo11 remains covalently linked to 5' ends of DSBs. A nick in the 5'-end releases the Spo11-bound oligonucleotide complex. The DSB ends then undergo 5' to 3' resection (Figure 1.1b), generating 3'-overhangs that are bound by the stabilising recombinazes Rad51 and DMC1 (homologs of *E. coli* RecA) and which promote DNA strand exchange with homologous double-stranded DNA (Figure 1.1c, Mimitou and Symington, 2008). The alternate homolog is used as a template for the repair of exposed overhangs. Holliday junctions may be predominantly resolved as reciprocal exchanges referred to as cross-overs (CO, Figure 1.1e), while non-cross-overs (gene conversions, Figure 1.1e) are the result are the result of D-loop formation and synthesis dependent strand annealing (SDSA, Figure 1.1f).

Gene conversion is a nonreciprocal process affecting short lengths of DNA where

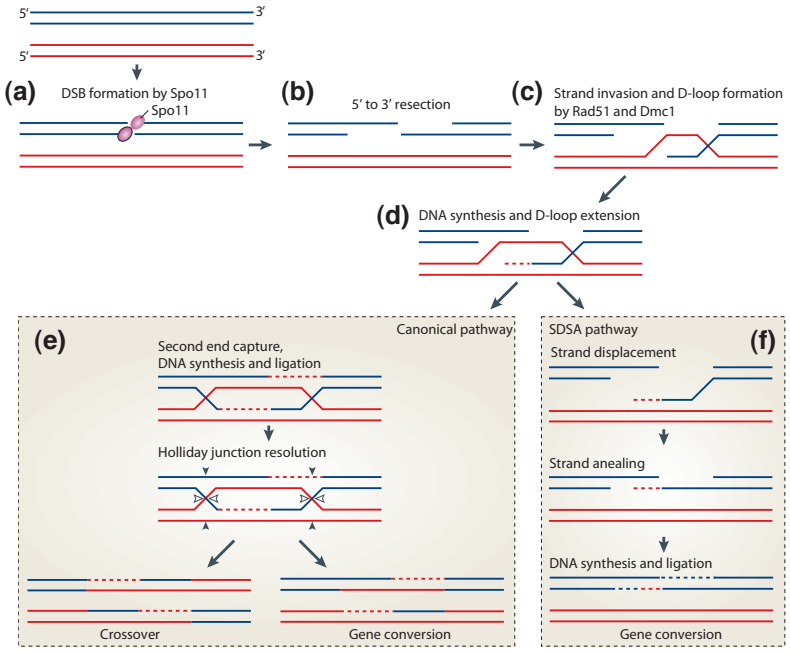


Figure 1.1: Meiotic recombination pathways. Blue and red chromosomes represent homologous paternal and maternal chromosomes in a gamete progenitor. **(a)** A double strand break is introduced by Spo11 and **(b)** resected to expose 3' single stranded DNA. **(c)** Strand exchange proteins Rad51 and DMC1 bind these exposed strands, search for homologous sequence, and facilitate strand invasion to form a D-loop DNA structure. **(d)** The D-loop structure is proposed to resolve via two pathways. For **(e)**, the canonical pathway, the second single stranded DNA tail is captured, extended and annealed to form a double Holliday junction. Holliday junctions are resolved through nicking endonucleases cutting the flanking strands (closed arrowheads) or crossed strands (open arrowheads), resulting in cross-over and gene-conversion (a minor pathway) products respectively. **(f)** An alternative pathway which is thought to account for the majority of gene-conversion events, synthesis-dependent strand annealing, leads primarily to gene-conversion events. Taken from (Sasaki et al., 2010).

genetic material from one parental chromosome is incorporated into the alternate homolog during meiotic exchange (Szostak et al., 1983). Hence it plays an important part in shaping genetic diversity in populations (Coop and Myers, 2007). But GCO detection is technically challenging. There is substantial uncertainty over many of its physical properties: the estimated tract length of GCO (thought to occur over 50 bp to 200 bp (cf., Borts and Haber, 1989; Hilliker et al., 1994; Jeffreys and May, 2004)),

the rate of events per meiosis, and the mechanism that specifies whether DSBs are resolved as GCO or CO. The focus of this thesis is on cross-over events; the reader should note that the term *recombination* should be understood to refer to *cross-over* unless otherwise stated.

1.1.2 The distribution of meiotic recombination in yeast, mouse, and human

Recombination rates are observed to vary at both broad, megabase scales and at fine, kilobase scales. The structure of recombination rate variation at each scale is informative about factors determining the placement and control of recombination events.

1.1.2.1 The structure of recombination rate variation at broad scales

I will follow the sequence of discoveries in recombination research, which, until recently, has been limited by the coarse resolution at which recombination events were detected. Hence I will first consider the properties of broad-scale rate variation, and then turn to fine-scales. In humans, the first genetic maps were based on only a few hundred minisatellites in pedigrees, and therefore resolution was limited to hundreds of megabases. Nevertheless, these maps provided the first insights into how recombination rates vary along chromosomes and between the sexes (e.g. the Marshfield map, Broman et al., 1998). For example, greater levels of recombination activity were observed in females than males, and the changes in rate were correlated with different chromosome components, especially with karyotypes representing regions with high G/C nucleotide content called R-bands (Baudat and Nicolas, 1997).

While the factors responsible for broad-scale recombination rate patterns remain partially understood [*RNF212*, *REC8* have been associated with changes in broad-scale rate in cattle (Sandor et al., 2012), and *RNF212* and the 17q21.31 inversion

in humans (Kong et al., 2008)], studies point towards chromosome organisation and structure. At the level of chromosomes, recombination rates are, in most eukaryotes, suppressed at centromeres and pericentric regions and increase towards the telomeres (Clark et al., 2010). At the time of DSB formation, chromosomes are organised into chromatin loops anchored to a linear protein structure, called the chromosome axis, which is enriched for cohesins and meiosis-specific proteins (Lichten and de Massy, 2011). In addition, these axes contain protein complexes of Hop1p, Rev7p and MAD2 (HORMA domains) that are thought to mediate checkpoints in chromosome synapsis and DNA repair (Aravind and Koonin, 1998). Several properties of chromosome axes have been observed. There is an inverse relation between loop size and axis length (Revenkova et al., 2004), DNA arranged into smaller loops has higher DSB potential (Kleckner et al., 2003), and autosomal cross-over frequency per chromosome in male mice correlates with axis length (Froenicke et al., 2002). In yeast, Panizza et al. (2011) demonstrated that a subset of the Spo11 accessory proteins Rec114, Mei4, and Mer2 are axis associated (Figure 1.2).

Most DSBs occur in DNA loops (Blat et al., 2002), suggesting that DSB formation involves recruitment of Spo11-bound sequence to the chromosome axes through interactions with accessory proteins. In mouse, human, and some other organisms, meiotic chromosome axes are longer and recombination frequencies are higher in females than males (Kleckner et al., 2003). The pseudo-autosomal region (PAR), the region(s) of homology between sex chromosomes obligate for correct sex chromosome segregation, has both exceptionally long axis segments (Kauppi et al., 2011) and high recombination activity (Smagulova et al., 2011). It remains to be determined whether DSB levels are influenced by regional differences in loop size, the frequency of axes within chromosome domains, or by axis protein composition.

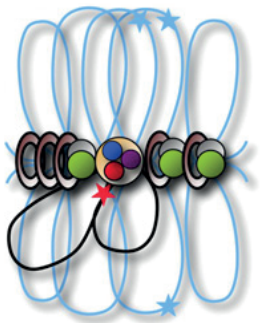
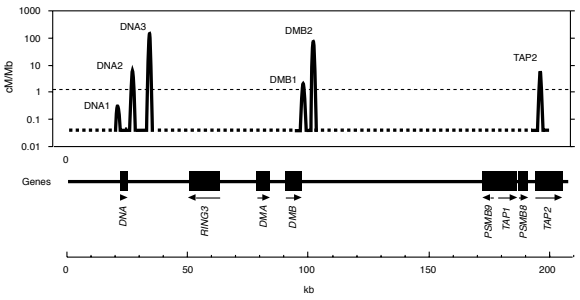


Figure 1.2: Chromatin and nucleosome organisation influence DSB formation in yeast. At meiosis, chromosomes condense into loops (light blue), separated by linear chromosome axes generated by bound cohesins (brown rings), axial element components Red1 and Hop1 (green and grey) and pre-DSB subunits Rec114, Mei4 and Mer2 (blue, red, purple). DSB occur in the loops at sites where Spo11 is recruited. However of the potential DSB sites (light blue stars), only one site per axis is cut (red star on black loop). Taken from (Panizza et al., 2011).



ID	95% width (kb)	Peak rate (cM/Mb)	Crossovers
DNA1	1.9	0.5	69
DNA2	1.3	3.7	237
DNA3	1.2	130	661
DMB1	1.8	3.1	36
DMB2	1.2	28	358
TAP2	1	5.8	141

Figure 1.3: Recombination occurs in 1-2 kb hotspots. Six sperm recombination hotspots in the human MHC class II region. The profile of fine-scale recombination rates measured by sperm-typing (Jeffreys et al., 2001) is depicted by the thick black line. Due to limitations in the method, the recombination rate was not measured between the typed hotspots and are therefore represented as broken lines. The average male recombination rate (1.55 cM/Mb, Kong et al., 2010) is represented by the thin dotted line. Almost 90% of the recombination in this region occurs in the six hotspots. The table contains information on: the width of each hotspot (defined as the region within which 95% of the cross-overs occur), the peak recombination rate, and the number of cross-overs detected. Adapted from (Jeffreys et al., 2001).

1.1.2.2 The structure of recombination rate variation at fine-scales

As the resolution of data increased, it became apparent that recombination events are not evenly distributed across the genome. Instead recombination occurs in *hotspots*, narrow ∼2 kb DNA regions where the cross-over rate is far higher than the flanking sequence (Figure 1.3, Chakravarti et al., 1984; Gerton et al., 2000; Steinmetz et al., 1982; Jeffreys et al., 2001; McVean et al., 2004). How hotspot locations and intensities are specified is an area of active research. An epigenetic mark associated with active transcription, tri-methylated histone 3 lysine 4 (H3K4Me3), was found to mark DSB sites in mouse (Buard et al., 2009; Smagulova et al., 2011) and yeast (Borde et al., 2009). In humans a 13 bp sequence motif CCNCCNTNNCCNC was found to be enriched in historical hotspots compared to coldspots (Myers et al., 2008), and shown to modulate cross-over activity (e.g. Myers et al., 2008; Jeffreys and Neumann, 2009). Comparisons of maps from populations with different ancestry and patterns of LD breakdown revealed startling variation in the placement and intensity of hotspots among humans (Kong et al., 2002; Coop et al., 2008; Berg et al., 2011; Hinch et al., 2011; Weitkamp, 1974; Jorgenson et al., 2005), mice (Paigen et al., 2008) and between humans and primates (Auton et al., 2012; Ptak et al., 2005, 2004; Wall et al., 2003; Winckler et al., 2005).

Through a confluence of experimental and statistical approaches, PRDM9 was associated with differential cross-over activity between mouse strains (Grey et al., 2009; Parvanov et al., 2010), and the 13 bp motif in humans. This potentially provides a connection between epigenetic marks and local sequence features:

- PRDM9 has a zinc finger (ZF) domain able to bind DNA (Figure 1.5c).
- PRDM9 has methyltransferase activity (Figure 1.5b) and is expressed only in mouse testes and ovaries at prophase I of meiosis.
- In *Prdm9*$^{-/-}$ mice, although both sexes are proficient in initiating DSBs, they

cannot repair the breaks and are sterile (Hayashi et al., 2005).

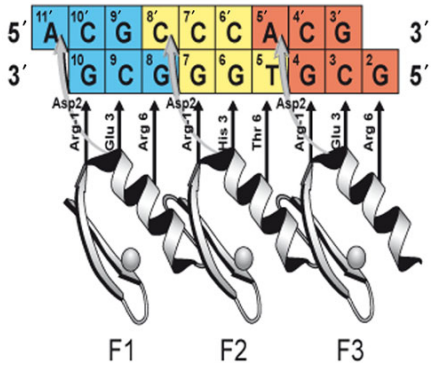


Figure 1.4: Model of Cys2His2 DNA recognition for three zinc fingers (Jamieson et al., 2003). Each finger interacts with a 3 bp subsite on one strand of the DNA, using amino acid residues in helical positions -1, 3 and 6. In addition, position 2 in the recognition helix binds the complementary strand. The fingers ideally bind 4 bp overlapping subsites, so that adjacent fingers are functionally synergistic though structurally independent. Taken from (Klug, 2010).

The gene also appeared in a different context as the only locus known to underlie hybrid sterility in mammals. The F1 offspring from specific crosses of *Mus mus musculus* × *Mus mus domesticus* (e.g. PWD × B6) arrest in prophase I of meiosis due to a Muller-Dobzhansky incompatibility involving the *Prdm9* locus together with chromosome X (Mihola et al., 2009). Questions of the potential role of PRDM9 in speciation, though extremely interesting, are outside the scope of this thesis and will not be considered.

1.1.2.3 Predicting zinc finger binding sequences

The PRDM9 DNA-binding functionality is provided by tandem arrays of Cys2His2 zinc fingers (ZFs) that cooperate to recognise specific nucleotide sequences (Baudat et al., 2010; Grey et al., 2011). Each finger has a $\beta\beta\alpha$ structure coordinated by a central zinc ion, of which three α-helix positions (commonly referred to as -1, 3,

and 6) bind directly to the primary DNA strand at nucleotides $n+3$,$n+2$, and $n+1$, while the amino acid at position 2 interacts with the complementary nucleotide at position $n'+4$, where n is an integer representing position in the sequence and n' represents position on the complementary strand (Choo and Klug, 1997) (Figure 1.4). Zinc fingers link in tandem in a linear, polar fashion to form ZF arrays that recognise DNA sequences of different lengths (Klug, 2010). This characteristic structure defines an amino acid ZF motif based on the spacing of the zinc ligands $X_2CysX_{2,4}CysX_{12}HisX_{3,4,5}His$, where X represents any amino acid residue. The precise spacing between the two cysteines (Cys) and the two histidines (His) is variable (Miller et al., 1985). This Cys2His2 sequence, and the modular properties of ZF cooperation in arrays, is used in bioinformatic approaches to predict the DNA sequence bound by Cys2His2 ZF arrays. The leading binding-prediction method developed by Persikov et al. (2009) applies a pre-trained support vector machine (SVM) to input Cys2His2 ZF-array amino acid sequence and returns a probabilistic prediction of the DNA sequence bound by the ZFs. This was applied successfully by Myers et al. (2010) to identify that the DNA sequence binding prediction for the human PRDM9 A allele ZFs matched the 13 bp motif associated with hotspots, validated by successful *in vitro* binding assays for human (Baudat et al., 2010) and mouse (Grey et al., 2011) PRDM9 alleles.

1.1.2.4 The position weight matrix

DNA sequence binding predictions are often represented in the form of a position weight matrix (PWM) - a $N \times M$ matrix in which element (n, m) gives the probability of binding nucleotide n in position m of a motif of length M. The variable N denotes the number of letters in the alphabet from which the motif was derived. In this thesis the alphabet is composed of nucleotides A,T,C, and G. An important summary measure of a given PWM is its information content profile. The information content

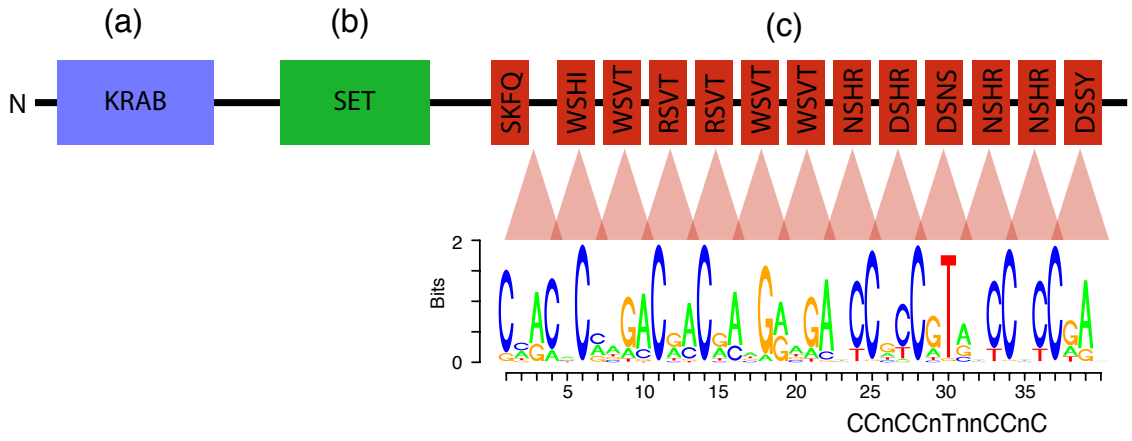


Figure 1.5: Schematic of the human PRDM9 protein and estimated zinc finger array binding-prediction. The N-terminal half contains **(a)** a KRAB domain, which is thought to be involved in transcriptional repression and **(b)** a SET domain that trimethylates histone H3 lysine 4 - an epigenetic mark associated with the initiation of meiotic recombination in yeast and mice (Buard et al., 2009; Borde et al., 2009; Smagulova et al., 2011). The C-terminal half **(c)** consists of an array of highly variable zinc fingers (human A-allele represented). Below the ZF array its binding-prediction is displayed (predicted through the bioinformatic method of (Persikov et al., 2009) under the polynomial support vector machine model). Letter heights represent binding energies. Below the binding-prediction, the 13 bp motif associated with human hotspots is displayed (Myers et al., 2008).

at a position m of the motif is given by

$$I(m) = \log_2(N) + \sum_{i=1}^{N} p_{im} \log_2(p_{im}).$$

The information content is measured in bits, and in the case of DNA sequences ranges from 0 to 2 bits. A position in the motif at which all nucleotides are bound with equal probability has an information content of 0 bits. The information content at a given position can therefore be thought of as giving a measure of the tolerance for alternative binding in that position. Positions that are bound strongly, and thus have low tolerance for alternative nucleotides, correspond to high information content, while positions with high tolerance for alternatives correspond to low information

content.

There are two PWM predictions for the human PRDM9 A variant. The first was estimated from over-enrichment of LD-hotspot associated motifs (Myers et al., 2008), the second was predicted through the SVM model of Persikov et al. (2009) (Figure 1.6). Predictions for the A-allele exactly match the observed degeneracy at positions 3, 6, 8, 9, and 12 within the 13 bp motif and lack of degeneracy at the other eight positions. PWMs provide a flexible representation of the uncertainty in motif binding predictions (though limited in representing only motifs of a particular size, i.e. one ZF cooperation confirmation) and will be used extensively in this thesis. Position weight matrices are often represented through a sequence logo plot (Schneider and Stephens, 1990). Figure 1.6 provides an example of a logo plot, where characters representing the potential bound bases are stacked on top of each other for each position in the prediction, and the height of each character is proportional to the information content in the binding of that base at that position.

1.1.2.5 The role of PRDM9 in recombination

Myers et al. (2010) made binding predictions for PRDM9 in several mammalian species and observed that (1) all the sequenced alleles were highly differentiated across species, likely due to tremendously rapid evolution at PRDM9 ZFs (Oliver et al., 2009); (2) that the 13 bp human recombination motif does not function in chimpanzees; and (3) that binding-motifs are being removed from the human genome by self-destructive drive. Baudat et al. (2010) sequenced several human PRDM9 variants and found that the variation among humans correlates with variability in hotspot usage. Additional studies catalogued tremendous variation in PRDM9 alleles and their distribution across human ethnic groups, and found that the different alleles correspond to differential hotspot activity in sperm (Berg et al., 2010).

Specifically, the A-allele (predicted to bind the 13 bp motif CCNCCNTNNC-

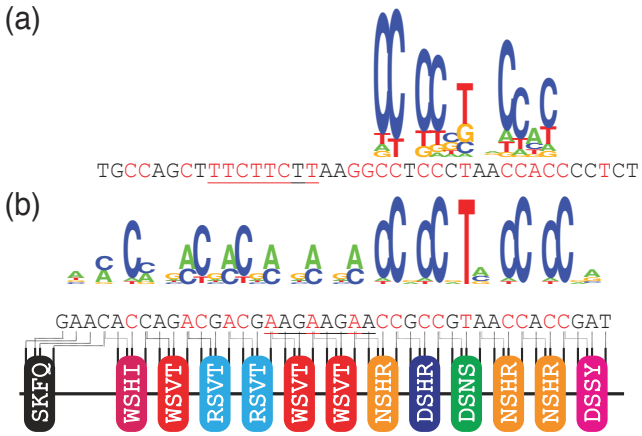


Figure 1.6: The recombination motif associated with human hotspots matches the human PRDM9 C-allele DNA-binding prediction. **(a)** Estimated degeneracy of the 13 bp hotspot motif (logo plot; relative letter height proportional to estimated probability of hotspot activity and total letter height determined by degree of base specificity, Myers et al., 2008) as well as an extended ∼39 bp motif (text below logo; influential positions ($p < 0.01$) shown in red). **(b)** In silico prediction of the binding consensus for PRDM9 A allele, aligned with the 13 bp oligomer, with more influential positions shown in red. Underlined in both (a) and (b) is an additional 8 bp of matching sequence. The logo shows predicted degeneracy within this consensus. Below the text is the sequence of four predicted DNA-contacting amino acids for the 13 successive human PRDM9 zinc fingers (one oval per finger, differing colours for differing fingers, and the separated finger is gapped N-terminal from others) and their predicted base contacts within the motif. Taken from (Myers et al., 2010).

CNC) was observed at 0.86 frequency in individuals of European ancestry and 0.50 frequency in individuals of African ancestry, whereas for the C-allele (predicted to bind CCgcngtnnnCgtnnCCg [1]) was observed at 0.01 frequency in individuals of European ancestry, and 0.13 frequency in individuals of African ancestry (Berg et al., 2010). Hinch et al. (2011) identified a 17 bp motif **CCCC**a**GT**GAG**CGT**tg**Cc** [2] enriched in ancestral African-specific hotspots in admixed African Americans. Indeed, differences in PRDM9 ZFs in humans explain approximately 80% of heritable variation in the proportion of cross-overs that occur in hotspots genome-wide (Hinch

[1] upper case letters represent the nucleotides most confidently predicted

[2] upper case letters represent the nucleotides with the strongest signal, bold represent exact matches to the PRDM9-binding prediction

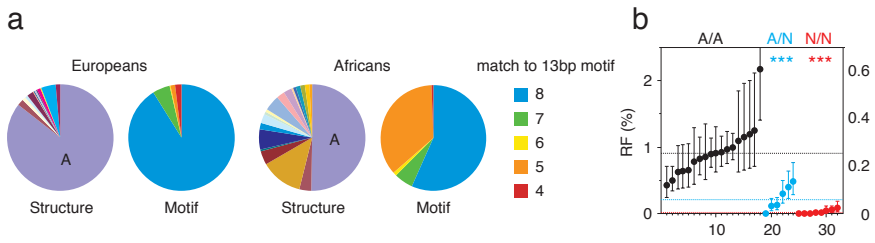


Figure 1.7: PRDM9 zinc finger (ZF) variant diversity in human populations and cross-over activity in sperm. **(a)** PRDM9 ZF array diversity in males of European or African ancestry. Alleles classified by their ZF structure or by their strength of predicted-binding to the 13 bp motif CCNCCNTNNCCNC. **(b)** Variation in recombination activity at a the Jeffreys F hotspot for men with different PRDM9 alleles. Recombination fractions (RF) are presented for three classes of males, those carrying two A alleles (black), one A allele (blue), and no A alleles (red). Taken from (Berg et al., 2010).

et al., 2011). Collectively, the findings imply that PRDM9 determines hotspot locations and is responsible for differences in hotspot location between species. However, there remain a number of unanswered questions:

- What is the exact relationship between PRDM9 variants, their predicted motifs, and the resulting recombination activity?
- How might PRDM9 influence hotspots without a clear match to the motif?
- Why is a gene so important in recombination evolving so rapidly, and why is such a costly genetic mechanism (of self-destructive drive on binding motifs) maintained over evolutionary time?

A potential approach to assess whether different classes of hotspot evolve in different ways, and to study recombination evolution over broader scales, is the construction of genome-wide fine-scale genetic maps in closely related species (Chapter 4).

1.2 Methods to identify recombination events

In most animals meiotic cells progress through meiosis asynchronously in the gonads and are therefore hard to isolate in large numbers. Classically, recombination rates have been quantified by analysing the products of meiosis (gametes or progeny) as recombinant or nonrecombinant. By scoring polymorphisms genome-wide, it has been possible to construct whole-genome genetic maps from pedigrees through maximum-likelihood or other recombination rate estimators (Thompson, 2000). However, a key limitation is that the resolution of pedigree-based maps is determined by the marker density, the number of informative meioses, and the number of cross-overs per genome per generation. Hence, large numbers of related individuals are required. This limitation motivated the development of methods that indirectly sample a much larger number of meiotic events. The two leading approaches have been (1) to study historical recombination events by inferring the genealogy underlying a sample of genetic variation from the population, and (2) to identify recombination events in large samples of parent-offspring pairs in population isolates that are phased by using identity by descent (IBD) sharing within the population (Gudbjartsson et al., 2000). Recently, exquisitely precise microfluidic techniques have enabled the isolation and genotyping of individual chromosomes enabling the construction of genome-wide fine-scale personal genetic maps (Wang et al., 2012), but these approaches require further refinement. I will review each of these methods in turn.

1.2.1 A genealogical view of recombination

A powerful approach to model the impact of evolutionary stochasticity on polymorphisms is from the perspective of the genealogical history underlying a sample of chromosomes (Stumpf and McVean, 2003). Figure 1.8 shows a genealogy of three chromosome segments. The segments are traced backwards within the pedigree of

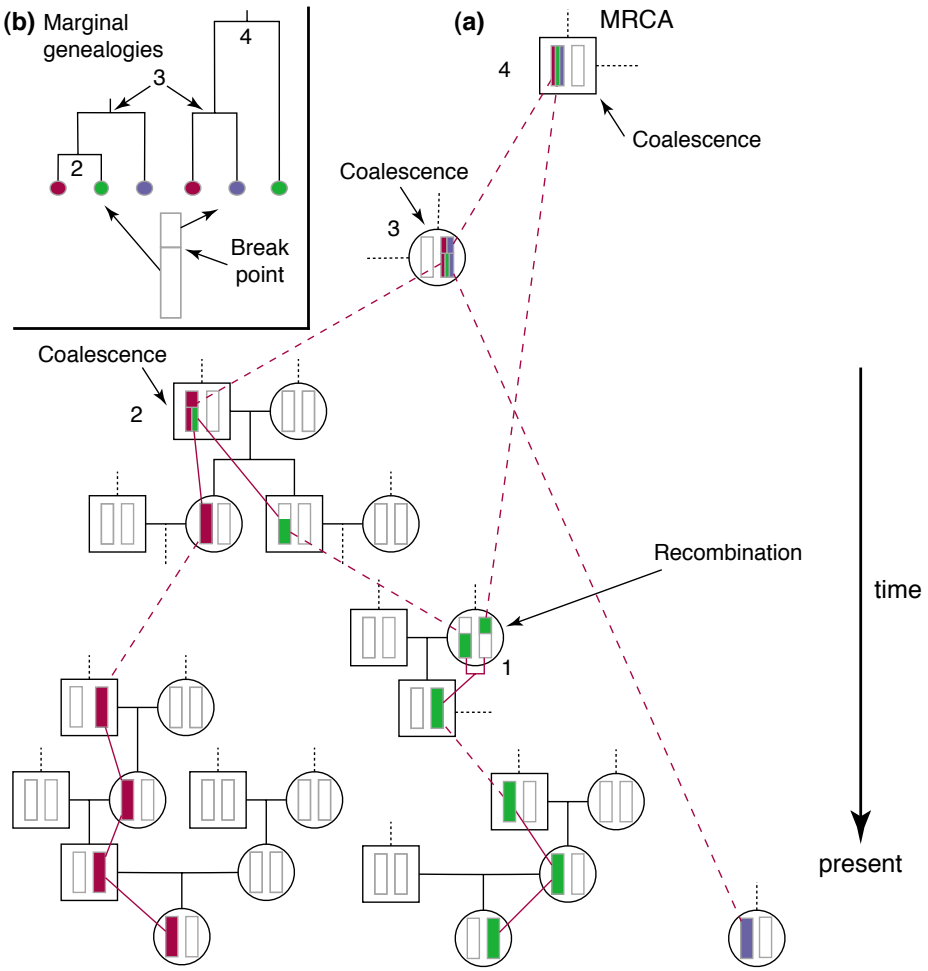


Figure 1.8: The genealogy of three chromosomes. Looking backwards in time, the three chromosomes (plum, green, and purple) are traced from offspring to parent (solid lines). The entire genealogy is only partially known; unobserved relationships are indicated by dashed lines. Lineages are traced back until they split (a recombination event, e.g. event 1), or until they merge together (a coalescent event, e.g. events 2, 3 and 4). Within the sampled chromosomes, coalescent events occur at a faster rate than recombination events, therefore guaranteeing that the chromosomes will meet in the most recent common ancestor (event 4, Griffiths and Marjoram, 1996). **(b)** Recombination events induce different genealogical trees either side of the event (Wiuf and Hein, 1999). Taken from (Nordborg and Tavaré, 2002).

the individuals. Recombination allows the segments to have different genealogical trees (Figure 1.8 b). As a result, the genealogy of a sample has a graph structure

rather than a tree. How does this affect patterns of polymorphism? New mutations arise on a single genetic background and are carried in complete association with all polymorphism on that chromosome. Over evolutionary time recombination serves to break down the association between sites. Those sites nestled very close to each other, which therefore rarely recombine, tend to remain associated. In contrast the association between more distantly spaced sites is gradually eroded. Hence, the extent of non-random patterns of association between polymorphisms along a chromosome reflects the recombination process. This association, whether observed within families or across distantly related individuals, forms the basis for detecting recombination events.

1.2.2 Pedigree studies

Cross-over results in the transmission of mosaic parental haplotypes to the next generation. Consequently, we can view sibling genomes as independent samples of the parental haplotypes and the action of recombination on those haplotypes. Hence, to detect cross-over in pedigrees the task is to infer the inheritance pattern underlying the genotypes observed in family members (c.f., Thompson, 2000). Note that artefact (genotyping errors) and biological process (*de novo* mutation, which are worthy of their own study) can mimic the action of recombination at local-scales. In this chapter we will ignore these potential sources, but we describe methods to distinguish these from cross-over in Chapter 5 - which is particularly effective if three or more generations are sequenced.

Consider the pedigree of six individuals in Figure 1.9a genotyped at two loci, the haplotypic phase of the genotypes are not known. First, focus on the quartet (C, D, E, F) (Figure 1.9b). The haplotypes in D are not informative about recombination events, $0-1$ is always transmitted. Nevertheless we can detect recombination events in C by enumerating all potential transmissions. We conclude that a recom-

bination event must have occurred in one of the transmitted chromosomes, but we do not know whether it is the chromosome inherited by E or F. Now if the entire pedigree is considered (Figure 1.9c), the phase of C's genotypes (her haplotypes) is fixed by the transmission patterns from A and B (B always transmits the haplotype $1-1$), so C must have the haplotypes $1-1$ and $0-1$. Since the haplotypic phase of genotypes in C is fixed by the grandparental genotypes, we conclude that E received grand-paternal haplotype $1-1$, and F received the recombined haplotype $1-0$.

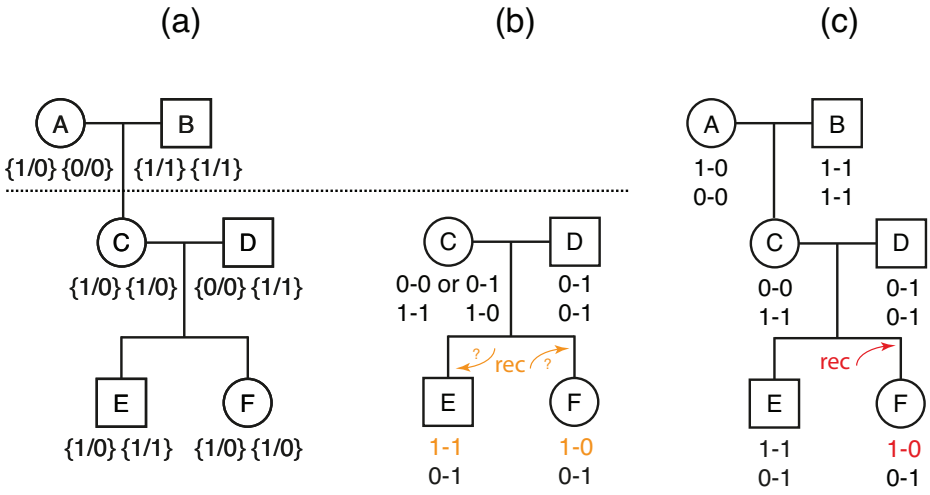


Figure 1.9: Example of allele transmission through a simple pedigree. Simple pedigrees are defined by a strict inheritance structure: breeding individuals have no more than one mate, and at most one mate has ancestry in the pedigree. **(a)** The six individuals, A to F, have been typed at two biallelic loci. The allele observed at each locus is represented by a pair of numbers in curly braces, 1 for the common allele and 0 for the rare allele. The haplotypic phase of the measured genotypes is unknown. **(b)** The haplotypes compatible with transmission through the quartet (C, D, E, F). A cross-over event is detected, however the recipient cannot be determined; the two potential recombined molecules are coloured orange **(c)** The haplotypes compatible with transmission through the entire pedigree. The transmission patterns from the grandparental chromosomes enable the recombination event to be assigned to the maternal chromosome in E and is coloured red.

It is the case in general, if only a single site is analysed, that comparisons of two siblings can be used to detect recombination events, but that the recombinant individual is not determinable. This is resolved by increasing the number of siblings to at least three. However, only 25-40 cross-over events are expected per genome

per meiosis (Coop et al., 2008), hence a single quartet is unlikely to provide much information about recombination in a particular region. To improve resolution it is therefore necessary to observe many meioses, which amounts to collecting information from many families or generations.

As demonstrated in Figure 1.9 large amounts of information can be gained by considering all of the pedigree simultaneously. Importantly, with at least three non-founders recombination events can be assigned to the parental chromosomes, which enables the study of recombination differences between sexes. In order to infer recombination rates from a pedigree, it is desirable to calculate the likelihood of the data as a function of the recombination rate. Naively, the likelihood can be calculated by considering all possible haplotypes within the pedigree. The number of haplotypes for M meioses and L genotyped loci scales as $\mathcal{O}(2^{2M+L})$, and so identifying recombination events through exhaustive search is computationally impractical for contemporary data sets.

To address this limitation, Elston and Stewart (1971) proposed economising the computation by calculating the likelihood as a function of each individuals' inheritance probabilities, which are defined by the inheritance structure of the pedigree. The algorithm only applies to a specific pedigree structure referred to as a *simple pedigree*, defined as a family established by unrelated founders with no inbreeding, consanguinity, or polygamy (e.g Figure 1.9). If the conditions of the simple pedigree are satisfied, the inheritance probability for the haplotype observed in a child, H_c, can be written in terms of the maternal, H_m, and paternal, H_p, haplotypes and the recombination fraction r,

$$\Pr(H_c|H_m, H_p, r)$$

For a pedigree with n individuals, the likelihood conditional on recombination fraction is then calculated by summing over the product of all the offspring-mother-father trios, denoted by $\{c, m, p\}$, for each individual's haplotype pairs that are compati-

ble with the observed genotype and scaled by a prior probability over the founder genotypes,

$$\Pr(D|r) = \sum_{H_1} \cdots \sum_{H_n} \prod_{k \text{ founders}} \Pr(H_k) \prod_{\{c,m,p\}} \Pr(H_c|H_m, H_p, r)$$

The recombination rate can then be estimated by finding the value of r that maximises the likelihood. Despite this improvement, the likelihood calculation still requires summation over all possible genotypes for each individual, requiring $\mathcal{O}(M \times 2^{6L})$ operations (Lander and Green, 1987). Therefore the method has limited impact on the ability to localise recombination events: though many meioses can be studied, the resolution of the genetic map is severely compromised by the limited number of markers that can be feasibly analysed.

Lander and Green (1987) proposed a Hidden Markov Model to efficiently calculate likelihoods in small pedigrees but across large numbers of markers up to entire chromosomes. The method re-defines the likelihood calculation so that the summation occurs over loci rather than individuals. All patterns of inheritance are considered, across all individuals, simultaneously, on a locus-by-locus basis, requiring $\mathcal{O}(L \times 2^{2M})$ operations. Subsequently, heuristics have resulted in performance improvements (Abecasis et al., 2002; Gudbjartsson et al., 2000; Kruglyak et al., 1996).

The likelihood, which relies on an object called the transmission vector, is constructed as follows. Let D_i denote the genotype of all individuals at locus i. The genotypes of the children are determined by the alleles transmitted from the parents, and can be described by a transmission vector v_i,

$$v_i = (p_1, m_1, p_2, m_2, \ldots, p_M, m_M),$$

where the coordinates describe the outcome of the maternal and paternal meioses giving rise to the M nonfounders in the pedigree. Specifically, $p_i = 0$ or 1, according

to whether the grandmaternal or grandparental chromosome was transmitted in the parental meiosis giving rise to the i^{th} nonfounder; m_i contains the information for the corresponding maternal meiosis. Hence for M nonfounders there are 2^{2M} potential inheritance vectors. An example of an inheritance vector compatible with the observed genotypes in Figure 1.9a is

$$v = (p_C, m_C, p_E, m_E, p_F, m_F) \tag{1.1}$$

$$= (0, 1, 1, 1, 1, 1), \tag{1.2}$$

where we follow the chromosome ordering in Figure 1.9c. Inheritance is extremely informative; it specifies the outcome of each meiosis, implies the founder allele carried in each individual, and implies whether a pair of chromosomes are identical by descent. Therefore given v_i, the probability of observing the genotypes can be written as $\Pr(D_i|v_i, H_m, H_p)$. In the absence of recombination and *de novo* mutation, one inheritance vector describe the transmission of alleles to all sites along the non-founder chromosomes. I will use the following notation for the set of possible inheritance vectors for a pedigree with m meioses genotyped at L sites

$$v_{\bullet j} = \{v_{ij}; i = 1, \ldots, m\} \quad j = 1, \ldots, L \tag{1.3}$$

$$v_{i\bullet} = \{v_{ij}; j = 1, \ldots, L\} \quad i = 1, \ldots, m. \tag{1.4}$$

Recombination causes the inheritance vector to transition from one state to the next along the chromosome. However, the current inheritance vector only depends on the inheritance vector at the previous site. In other words, the transitions between inheritance vectors have a Markovian structure. For a single meiosis, the transitions resulting from meiosis can be written as the transition matrix, T, which is a function

of the probability of a recombination event, p_r, between v_{i-1} and v_i,

$$T = \begin{pmatrix} 1 - p_r & p_r \\ p_r & 1 - p_r \end{pmatrix}.$$

Including more meioses requires polynomials of $(1 - p_r)$ and p_r. The transition matrix for M meioses can be written as a M^{th} nested product of T to each of its elements, resulting in a $4M \times 4M$ matrix $\mathcal{T}^M$. The probability of the inheritance vector at v_i can then be written in terms of v_{i-1} and T.

$$\Pr(v_i | v_{i-1}, p_r, M) = \mathcal{T}^M v_{i-1}.$$

The insight of Lander and Green (1987) was that the inheritance vectors (represented by $\mathbf{V}$ in Equation (1.5)) could be considered as a hidden state and summed over in the likelihood calculation to relate the observed genotypes D to the recombination fraction p_r (Figure 1.10):

$$\Pr(\mathbf{D}|r) = \sum_{\mathbf{V}} \left(\Pr(v_{\bullet 1}) \prod_{i=2}^{L} \Pr(v_{\bullet i} | v_{\bullet i-1}, p_r, M) \prod_{i=1}^{L} \Pr(D_{\bullet i} | v_{\bullet i}) \right). \tag{1.5}$$

However, the model has limitations: it assumes no LD between markers, integrates over the founder haplotypes (i.e. has a uniform prior over the founder haplotypes), and is highly sensitive to genotyping errors. This is problematic in pedigrees which have been subject to HTS, since the HTS data are characterised by high variant calling error rates (see Chapter 2).

1.2.2.1 The deCODE genetic map

A pedigree study of particular interest in this thesis is the deCODE pedigree genetic map. Kong et al. (2010) genotyped a large cohort of Icelandic parent-offspring pairs (8850 maternal meioses, 6507 paternal meioses), resulting in sex-specific maps esti-

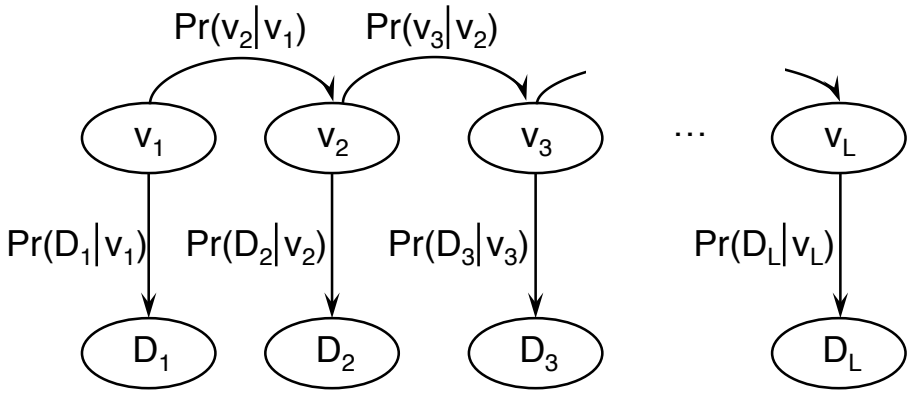


Figure 1.10: Lander-Green Hidden Markov Model for estimating recombination rates from genotype data. The transition probabilities $\Pr(v_i|v_{i-1})$ are a function of the underlying probability of recombination between sites. The symbol v_i represents the inheritance vector at site i, D_i represents that genotype data observed, and $\Pr(D_i|v_i)$ represents the likelihood of the date conditional on v_i and $\Pr(I_i|v_{i-1})$.

mated at a resolution of tens of kilobases. In order to analyse such a large sample, recombination events were identified by phasing both parent and offspring individually based on extensive IBD-sharing within the Icelandic population and then identifying the signal of recombination events where the offspring switched between paternal and maternal haplotypes (later formalised by (Palin et al., 2011)). The high resolution of the map enabled the first fine-scale analysis of the structure of sex-specific variation at genome features. Male-specific and female-specific maps had a correlation of 0.659 (10 kb scale), but there were clear sex differences. Hotspots were defined as intervals where the estimated level of recombination was greater than 10 times the genome average. Of the 4762 male-specific hotspots identified, 704 were estimated to have a rate less than 3 in female (relative to the genome average across bins). Similarly, 14.7% of hotspots in females were observed only in females, however the locations of hotspots with moderate intensity overlapped on average. The most dramatic sex-specific differences were observed at genes, especially for intronic rates (male - female = 0.270, $p = 2.2 \times 10^{-7}$; Figure 1.11). The deCODE data confirmed previous observations from sex-averaged LD-based maps (Myers et al., 2005), in particular rates

were suppressed within transcribed regions.

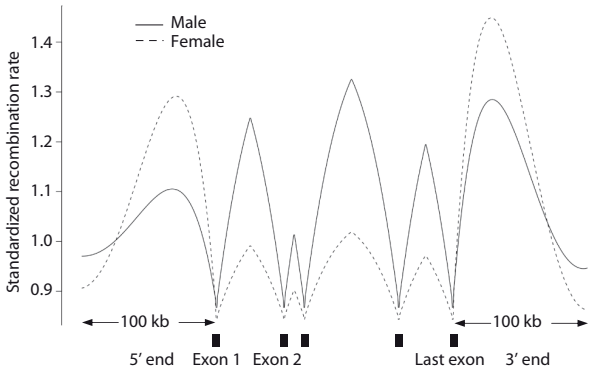


Figure 1.11: Sex-specific recombination rates around genes from deCODE pedigree-based recombination maps. Schematic summarising general structure of recombination rate variation around genes in humans. Taken from (Kong et al., 2010).

Wider comparisons of linkage disequilibrium (LD) patterns from unrelated cohort individuals and the pedigree-based rates confirmed that LD patterns in populations reflected true biological differences. The correlation between deCODE rates and the LD-based rates from the combined-population HapMap map (Consortium, 2007) was 0.729 at 10 kb resolution. A fine-scale rate association scan on the fraction of recombination in new hotspots identified *Prdm9* as the top hit. Difference in the length of the allele explains 60% of heritable hotspot usage variation in males and 40% in females. An inversion on chromosome 17q21 known to associate with female rates and fertility (Stefansson et al., 2005) also appeared to increase the fraction of recombination in hotspots, though the effect is limited to females ($p = 2.9 \times 10^{-5}$ for females, 0.49 for males).

Despite the success of pedigree studies, the inherent requirement for large sample sizes restricts their application due to the sampling cost and the fact that they are impractical for studies in species for which large sample sizes cannot be obtained. However, in Chapter 5 I describe how pedigrees are useful for improving the signal

to noise ratio in HTS.

1.2.3 Linkage disequilibrium

Linkage disequilibrium (LD), or the non-random patterns of association of alleles in haplotypes, provides insight into recombination (Stephens and Donnelly, 2002). Such associations also underlie pedigree-based recombination analyses, but LD refers to the associations within populations of unrelated individuals. In general, chromosomes sampled from unrelated individuals will be more distantly related than those sampled from members of pedigrees, and thus have had more opportunities for recombination to take place. This is precisely what makes LD analyses suitable for fine-scale recombination analyses. Whereas pedigree studies measure recombination events proportional to the number of meioses, LD studies measure historical recombination events proportional to the depth of the genealogy (Nordborg and Tavaré, 2002).

LD is quantified using statistics of association between allelic states at pairs of loci. Though several measures of LD have been proposed, examples considered here are $|D'|$ and r^2.

$|D'|$ is defined as the difference in the frequency of a 2-locus haplotypes and the product of the component alleles normalised to take values between 0 and 1 regardless of the allele frequencies (Lewontin, 1964).

r^2 is defined as the square of the coefficient of association of allele frequencies between two loci (Hill and Robertson, 1968b).

For example, consider two loci with alleles A/a at the first locus and B/b at the second. Let f_A represent the frequency of the common allele at the first locus, and f_{AB} represent the frequency of the AB haplotype, and so on for representing other combinations. Then $|D'|$ and r^2 can be written as,

$$
\begin{aligned}
D &= f_{AB} - f_A f_B \\
|D'| &= \begin{cases} \frac{|D|}{\min(f_A f_B, f_a f_b)} & \text{if } D > 0 \\ \frac{|D|}{\min(f_A f_b, f_a f_B)} & \text{if } D < 0 \end{cases} \\
r^2 &= \frac{D^2}{f_A f_B f_a f_b}.
\end{aligned}
\tag{1.6}
$$

Pairwise LD is highly variable (Pritchard and Przeworski, 2001; Stumpf and McVean, 2003). However, in nature patterns of LD are shaped by many evolutionary processes in addition to recombination: the age of the polymorphism (reflected in their allele frequencies), population history (Golding, 1984), mutation (Calafell et al., 2001), natural selection (Barton, 2000; Charlesworth et al., 1997), and drift (Charlesworth et al., 1997). Patterns of LD depend strongly on the coalescence times of the alleles, which under most models is reflected by their frequencies (Nordborg and Tavaré, 2002). $|D'|$ is normalised to take values between 0 and 1 regardless of the allele frequency. This has the desirable properties that in the absence of recombination or recurrent mutation all loci have a $|D'| = 1$. However, this value will tend to be high if either locus is young. All measures of LD are frequency dependent (Lewontin, 1988).

Hence, despite the statistical relationship between LD measures and distance, the variability arising from the uncertainty of the underlying genealogy is too great for reliable estimates to be made from any particular pair of loci (Stumpf and McVean, 2003). The pattern of association between polymorphic markers is of interest only because it contains information about the patterns of recombination. Therefore population genetic recombination analyses can be thought of as a missing data problem, where the underlying, unobserved genealogy contains all of the information (Griffiths and Marjoram, 1996; Hudson, 1983). This means that LD should be studied with a focus on descent rather than on state, and furthermore that $|D'|$ and r^2 are not likely to be informative about recombination rates. Models are needed to take into

account this additional level of uncertainty. I will consider several models in sections that follow.

1.2.4 Counting recombination events

I will use the four-gamete test (Hudson and Kaplan, 1985) as an example to demonstrate the limitations of model-free recombination estimation approaches. The four-gamete test is a way of spotting historical recombination events by assessing allele combinations at pairs of sites. The test can be explained as follows. Assume that the sampled DNA sequence is well approximated by the infinite sites model where the mutation rate is infinitesimal, and non-recurrent. Then at most one mutation can occur in the history of the sample at that site, and consequently there are only four possible haplotypes in the population. Furthermore, the only way for the four haplotypes to be in the sample is for at least one recombination event to have occurred between the two loci in the history of the sample. This process identifies intervals in which recombination must have occurred. The number of recombination events is then calculated by assuming that all overlapping intervals are the same recombination event. However, not all recombination events in the history of the sample are revealed by this test and it has low statistical power to detect events (Hudson and Kaplan, 1985). The issue of sample size is particularly problematic. Small sample sizes are particularly unlikely to contain the rarest gamete necessary to detect a recombination event, but the rate at which adding chromosomes increases the number of novel haplotypes is extremely slow $\mathcal{O}(\log\log(N))$, where N is the number of chromosomes, (Myers, 2003). Consequently, the four-gamete test underestimates the true level of recombination (Wiuf et al., 2001; Myers and Griffiths, 2003).

1.2.5 Models of recombination

1.2.5.1 The ancestral recombination graph

The ancestral recombination graph (ARG) describes the distribution of genealogies underlying chromosomes sampled from unrelated individuals (Hudson, 1983; Griffiths and Marjoram, 1996). Figure 1.8a represents an ARG. It is interpreted by looking back in time. The ancestral lineages are followed until recombination (where one ancestral lineage splits into two), or coalescence (where two ancestral lineages meet in a common ancestor). Embedded within the resulting ARG are the marginal genealogies for any nucleotide along the sequence (i.e. DNA that is inherited without recombination, Figure 1.8b). By superposing mutations on top of the genealogy, the patterns of variation in the sampled chromosomes can be described. Under models of random mating, constant population size and neutral mutation process, the variation is determined by two parameters:

- The population mutation rate $\theta = 4N_e\mu$, where μ is the per generation mutation rate, determines the rate at which mutations are superimposed on the graph.
- The population recombination rate $\rho = 4N_e r$, where r is the per generation recombination rate, determines the rate at which ancestral lineages undergo recombination.

Time in the ARG is measured in units of the effective number of chromosomes in the present generation $2N_e$.

How do the lineages behave under this model? The scaled time until the first recombination event on a single lineage back in time is exponentially distributed with rate $\frac{\rho}{2}$ in the limit as $2N_e \to \infty$. The probability that the followed lineages recombine with each other is vanishingly small because the sequences overwhelmingly recombine with the (infinitely many) non-ancestral chromosomes in the wider population.

Hence, lineages can be assumed to undergo recombination independently and simultaneous events are ignored. The times at which k-lineages coalesce and recombine are described by a birth-death process with rates λ_c and λ_r, and the rate of mutations, λ_m, are given by:

$$\begin{aligned} \lambda_c &= \binom{k}{2} \\ \lambda_r &= \frac{\rho k}{2} \\ \lambda_m &= \frac{\theta k}{2} \end{aligned} .$$

The quadratic death rate compared to the linear birth rate guarantees that a most recent common ancestor exists within the graph (Griffiths and Marjoram, 1996). The probabilities of the next event are given by:

$$\begin{aligned} \Pr(\text{coalesce}) &= \tfrac{k-1}{k-1+\rho+\theta} \\ \Pr(\text{recombine}) &= \tfrac{\rho}{k-1+\rho+\theta} \\ \Pr(\text{mutate}) &= \tfrac{\theta}{k-1+\rho+\theta} \end{aligned} . \tag{1.7}$$

Stochastic simulation under the model is extremely efficient because at any point in time λ_c and λ_r have a Markovian structure (Hudson, 1983), dependent only on the ancestral lineages present. In contrast, forward simulation requires the entire population to be simulated, and the simulation of genealogies while moving along the sequence (Wiuf and Hein, 1999) depends on all the preceding sequence genealogies and therefore has complex non-Markovian structure. Despite the ease with which coalescent models can be simulated from, using these models for inference remains extremely challenging. In order to perform inference of the the recombination rate within the coalescent, it is desirable to calculate the likelihood of the data as a function of the recombination rate, mutation rate parameters, and the set of genealogies G.

$$L(\theta, \rho|D) \propto \Pr(D|\theta, \rho) = \int \Pr(D|G, \theta, \rho) \Pr(G) dG.$$

However, likelihood approaches are currently restricted because no analytic or numeric expressions have been derived for the likelihood function and the enormous number of genealogies makes MCMC technically challenging. First, the number of recombination events is unbounded, therefore the state space of genealogies G is enormous. Second, the data are not very informative about the actual ARG (unless θ is much larger than ρ; in reality, θ and ρ are of the same magnitude for most organisms surveyed (Nordborg and Tavaré, 2002)), therefore very many genealogies need to be considered to infer the recombination rate.

As a result of these properties, efficient sampling of the diverse contributing ARGs is required. Importance sampling approaches (Fearnhead and Donnelly, 2002) and MCMC (Kuhner et al., 2000; Nielsen, 2000) both increase efficiency by creating bias toward ARGs that make significant contributions to the likelihood. Although such methods have been applied successfully to small regions, for even moderately sized regions they become computationally impractical (Fearnhead and Donnelly, 2001). Nonetheless, the approach of Fearnhead and Donnelly (2001) is currently the most accurate. Note however, it is difficult to extend the model to more general settings such as missing genotype data, genotyping error, and unknown haplotype phase.

Several alternative approaches have been explored, of which I shall describe three (*cf.* for review of other approaches Stumpf and McVean, 2003).

1.2.5.2 Approximate likelihood approaches

Hudson (2001) proposed the *composite-likelihood* (CL) method for estimating ρ over moderate to large genomic regions. The CL method is based on multiplying together likelihoods for every pair of segregating sites (Hudson, 2001),

$$CL(\rho) = \prod_{i,j \in S} L_{ij}(\rho) \quad \text{where } S \text{ are segregating sites,}$$

where the pairwise likelihoods are computed by importance sampling or other approaches.

The CL does not use all available information in the sequence data, and introduces non-independence in the combination of multiple comparisons, but is remarkably accurate and requires orders of magnitude less computation than full likelihood approaches (McVean et al., 2002). As a result, thousands of loci can be analysed at a time. An additional advantage of the CL is the ability to use genotype data. Because only pairs of sites are considered, genotype data is analysed by summing over all possible haplotype phases of each SNP pair. Similarly, missing data can also be incorporated at the expense of computational efficiency and quality of the estimates. Although the CL is not a true likelihood, and therefore the uncertainty in estimates between any pair of sites cannot be quantified, the model corresponds to a well defined genealogical process and it is consistent in that, as the number of loci increases, the CL approaches the true recombination rate (with a constant recombination rate estimated using a windowing approach; Fearnhead, 2003).

The CL-based method, LDhat, developed by (McVean et al., 2002, 2004), extends Hudson's CL to estimate variable recombination rates under a model of recurrent mutation, and has been widely used to estimate recombination rates.

Informally rates are estimated as follows: given a sample of chromosomes from the population, $\boldsymbol{\rho}$ is estimated in a Bayesian setting, placing a prior on the recombination map. To avoid over-fitting, $\boldsymbol{\rho}$ is assumed to be piecewise-constant. The prior probability is a distribution on the number of times $\boldsymbol{\rho}$ changes values, the locations of such changes, and the value of each piecewise-constant segment. Reversible-jump Markov Chain Monte Carlo is used to sample from the posterior distribution over the sample space of step functions when different parts of the space have different numbers of parameters. The pairwise CL is computed for every SNP pair in a region. The pairwise likelihood, L_{ij} is computed under the coalescent with recombination using the

importance sampling algorithm of Stephens and Donnelly (for sample configurations without missing data 2002). Mutation is incorporated under a two-allele model with a transition matrix $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. The product over all pairs give an approximation to the full likelihood,

$$CL(\boldsymbol{\rho})_{\text{LDhat}} = \prod_{i,j \in S} L_{ij}(\rho_{ij}).$$

Missing data is incorporated into the computation by marginalising over appropriate configurations, which scales exponentially with the number of chromosomes with missing data at site i or j.

LDhat is widely used, notably it was applied to the HapMap Phase II to estimate genome-wide fine-scale maps of recombination across 3.1 million SNPs sampled in human populations (Consortium, 2007). I use LDhat to estimate recombination rates in chimpanzees, humans and mice from HTS (Sections 2.2.1.1).

1.2.5.3 Approximate genealogy approaches

Li and Stephens (2003) proposed an approximation to the genealogical process that generates patterns of variation similar to the coalescent, where new haplotypes are constructed as an imperfect mosaic of those previously observed. The model relates the distribution of sampled haplotypes to the underlying recombination rate by exploiting the identity,

$$\Pr(h_1, \ldots, h_n) = \Pr(h_1|\rho) \Pr(h_2|h_1; \rho) \ldots \Pr(h_n|h_1, \ldots, h_{n-1}; \rho) \tag{1.8}$$

where $\{h_1, \ldots, h_n\}$ represent the n sampled chromosomes, ρ represents the recombination rate (which may be a vector of parameters if the rate is allowed to vary across the region). The conditional probabilities on the right hand side of the Equation (1.8) are not computationally tractable for most models. However, the authors proposed approximations for the conditional distributions informed by research into

optimal importance sampling distributions. The approach was named the *product of approximate conditionals* (PAC). The benefits of this approach are that all loci are considered simultaneously (rather than pairwise) and that the PAC likelihoods are true likelihoods, which are computationally tractable for large genomic regions (up to complete chromosomes). However, a disadvantage is that the biological validity of the approximation may be poor. First, chromosomes are no longer exchangeable - they depend in the order in which they are analysed. Second, the approximation does not correspond to a any well-defined genealogical process, and so the estimated recombination rate can only be related to the coalescent through an empirical bias correction. Nevertheless, the approach has good correspondence with alternative approaches if a constant recombination rate is assumed (Li and Stephens, 2003).

1.2.5.4 The African American admixture-based recombination map

Recently, Hinch et al. (2011) applied a PAC-based approach (Hapmix, Price et al., 2009), to polymorphism sampled from African Americans of admixed African and European ancestry. The HapMix approach was used to infer the probability of 0,1, or 2 African or European alleles at each locus in individuals genotyped at hundreds of thousands of SNPs. Positions where the inferred number of African or European allele changes reflect cross-over events that have occurred since admixture. Under the assumption that most hotspots occur in hotspots shared across individuals, recombination rates were then modelled over inter-SNP intervals at the inferred cross-over locations across individuals using MCMC. The approach resulted in a highly accurate map of recombination (Pearson correlation with deCODE-map ~ 0.96 at the 10 Mb scale). However, the approach shares some of the limitations of LD-based maps. The approach does not separately infer male and female recombination rates and requires normalisation by the total map length. Inherently admixed populations, and their ancestral populations, are required to infer recombination events, which may

prevent analyses in rare or endangered species. The map resolution is determined by the marker density and by the level of differentiation between the ancestral groups. In addition, the map may be biased where ancestry deviates from the average, the authors mention an example region on chromosome 8q24 where 10% of the individuals in the study who have prostate cancer have an increased proportion of African ancestry (Hinch et al., 2011).

1.2.6 Typing recombination events in single gamete

There are several experimental approaches that enable exquisite fine-scale detection of recombination events. First, allele-specific PCR can be applied to single sperm or pooled sperm to amplify recombinant sequences above the level of non-recombinant sequences. The method of *sperm typing* was developed by labs of Norman Arnheim and Alec Jeffreys (Hubert et al., 1992; Jeffreys et al., 1991). Recombination events are identified by comparing typed SNPs in sperm samples to their ancestral chromosomes. The approach achieves exquisite precision, limited only by the density of available markers, and resolutions of 500 bp are feasible (Figure 1.3, Jeffreys et al., 2005). Sperm typing has been used to study cross-over and gene conversion events (Baudat and de Massy, 2007). Although it has provided remarkable insights in to how the recombination rate varies at a few locations in human (Jeffreys et al., 2005) and mouse genomes (Guillon and de Massy, 2002), these studies provide no information about rates in females, are extremely technically challenging, and do not scale to genome-wide applications (Paigen and Petkov, 2010).

A second approach, that uses microfluidics to isolate individual chromosomes and nanolitre reaction volumes to improve the accuracy of DNA amplification was used in a single-cell analysis of whole-genome recombination activity and *de novo* mutation rates in human sperm (Wang et al., 2012). This approach enables the study of personal genetic maps, but again their utility in studying female recombination would

be limited. Currently false positives are still a challenge in this approach.

Other base-pair resolution approaches exist, but they require the harvesting of DNA at meiosis by using ChIP-Seq against recombination-specific proteins (Borde et al., 2009; Smagulova et al., 2011; Brick et al., 2012). At any given moment, less than 2% of the cells in the mouse gonads contain DSBs (Meistrich, 1977; Bellvé et al., 1977), necessitating the use of recombination-deficient strains or genetically engineered cell lines. In addition ChIP-Seq is limited by the efficiency of the immunoprecipitation step (Park, 2009).

1.3 Discussion

The aim of this chapter was to provide the reader with an introduction to meiotic recombination. I described the process of recombination and the methods used to study it. I noted that recombination rates vary over the both broad- and fine-scales. The signature of recombination left in patterns of genetic diversity was discussed, as well as commonly-used statistics to measure it. However, relating these patterns to the underlying recombination rate is complex and requires an understanding of the evolutionary process. I reviewed existing methods to calculate the probability of observing a genetic variation sample through the coalescent process, and how, through a composite-likelihood approximation, this is used to infer the underlying recombination rate. It is likely that the study of recombination will benefit from analysis of a diversity of organisms. However, SNP chips are only available for a minority of species. High throughput sequencing technologies provide a mechanism to collect dense population variation information in almost any species of interest, at the expense of higher error rates. In the following chapter I will describe the challenges of estimating recombination rates from HTS population sequence data, and characterise the symptoms of different sequencing error modes, as motivation for

the chapters that follow it.

Chapter 2

Challenges in the analysis of genome sequence data

The ability to generate whole genome sequence data from any sample or species of interest has changed the study of many areas of biology. In particular, population sequencing enables insight into the distribution of genetic variation within and between samples (either individuals or different tissues) which can be highly informative about diverse biological processes. However, the detection and interpretation of genetic variation from high throughput sequencing (HTS) data are far from being solved problems. Rather, error, in the form of false positives and false negatives, has the potential to dominate signals in the analysis of disease or biological process. Using examples from the analysis of chimpanzee, human, and mice population sequencing, I discuss the sources of error in variant discovery and genotyping, and the characteristic symptoms of different error modes. Importantly, I show through analysing haplotype structure that cryptic errors remain in variant calls despite the conservative application of best-practice filtering. Moreover, I demonstrate that even for finished quality genomes (e.g. human) errors result from problems in the reference. Hence, these error structures will impact all HTS analyses and are therefore of interest to the

field. Importantly, many of these sources of error will persist however the sequencing technologies evolve, because they arise from the inherent incompleteness of any reference genome and uncertainty in reconstructing complex mutation events. First I will discuss the quality of reference genomes. Then I will use data from the sequencing of ten chimpanzees to show how reference assembly errors or incompleteness can result in cryptic error structures, the impact of reference mis-assembly in the human reference on variant calls in the 1000 Genomes low coverage pilot data, and how sequence features in sampled wild mouse genomes - which are inadequately represented by the mouse reference - result in errors.

2.1 Reference genomes

Genome sequence assemblies are the foundation of modern genetics. Inherently, their accuracy, fidelity and completeness determine the correctness of contingent analyses and inferences. Though the uses of assemblies are manifold, I shall focus on a particular application - their role in HTS experiments. Mapping sequenced reads to a reference genome is an incredibly powerful and fast strategy to reconstruct the original sample. However, there are both subtle and gross ways in which misleading patterns of variation are realised. From the viewpoint of the sampled genome, errors in the construction of the reference would result in false contractions, expansions and inversions (Figure 2.1).

In addition, divergence from the reference results in reduced sensitivity and specificity of the reconstructed sequence because reads are either mis-mapped or not mapped at all. Some problematic areas are known, yet unresolved: the extensive polymorphism at the HLA (e.g. HLA-B $>$ 1600 alleles) is only represented by seven alternative alleles in the human reference, and in general researchers often do not attempt to map reads to the alternative sequences. Note that even if used, only the

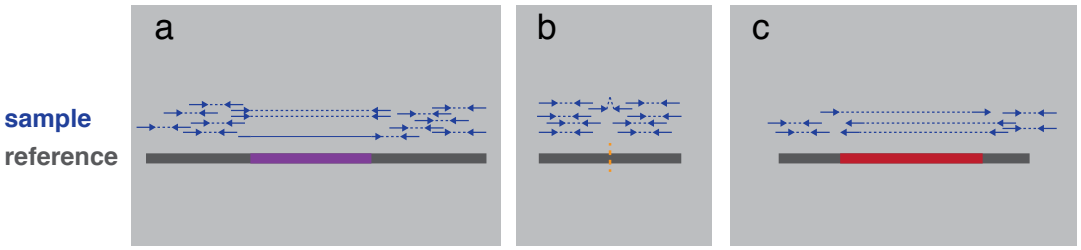


Figure 2.1: Errors or incompleteness in reference assemblies affect mapping. Reference represented by the grey bar, aligned paired-end reads from a sequenced genome are represented in navy blue. **(a)** Mis-placed contig (inserted contig represented by purple bar) results in a false contraction relative to the reference. **(b)** Incomplete contig assembly (missing sequence represented by broken orange line) results in a false expansion relative to the reference. **(c)** Error in contig assembly (inverted contig represented by ochre bar) results in inverted mate pairs.

HLA, 17q21.31, and ABO loci have alternative sequence representations in the human reference. Moreover, even for these limited instances certain representations are only applicable to specific groups (e.g. the 17q21.31 inversion polymorphism in Africans, Steinberg et al., 2012). This covers a variety of length scales. For example, in short heterozygous regions (20 - 35 bp) a reference bias is observed in mapping leading to miscalled genotypes (Degner et al., 2009).

There are several high quality vertebrate reference genomes but, due to the substantial cost, only the genomes of human, mouse and zebrafish are approaching a finished quality (defined as a nucleotide error rate of less than one error in 10^4 bases, and no assembly gaps (Consortium, 2004; Church et al., 2009)). It is likely that other assemblies will remain in their unfinished states until technological improvements substantially reduce the cost of achieving finished genome quality; especially in the ability to accurately sequence and assemble highly repetitive regions.

Given the methodological diversity of sequence generation and assembly, and the importance of high-quality primary data for molecular biology analyses, several measures of assembly quality have been devised. A popular metric is the $N50$ value,

defined as the weighted median contig size so that half the assembly is covered by contigs of size at least $N50$. Although $N50$ quantifies the ability of an assembler to join reads into large contiguous sequences, it fails to capture many aspects of the assembly quality. For instance, the $N50$ value of an assembly could be inflated by lowering the thresholds for combining smaller blocks of reads (often unresolved repetitive sequence) resulting in mis-assembled contigs (Salzberg and Yorke, 2005). Some validation of global assembly accuracy as summarised by $N50$ can be achieved by comparison to genetic maps (as used in the assembly of the first draft of the human reference (Lander et al., 2001), and as discussed in this Chapter) or through alignment to related genomes (utilised in the assembly of the chimpanzee reference (Sequencing and Consortium, 2005)).

In addition to problems in representing assembly quality at large-scales, $N50$ values fail to represent fine-scale inaccuracies such as insertion/deletion (indel) errors. Consequently, fine-scale properties are represented by alternative metrics. The certainty of each nucleotide is quantified as a *Phred*-score (Ewing and Green, 1998) defined as the minus ten times the logarithm base-ten of the error probability for the base. Once assembled, a base is assigned a consensus quality score as a function of its read depth and the quality of each base contributing to that position (Huang and Madan, 1999). An alternative statistical comparative genomics approach, proposed by Meader et al. (2010), estimates the abundance of indel errors between two assemblies relative to the neutral theory prediction. The benefit of this approach is that it is independent of the assembly process, unlike $N50$ or statistics based on read coverage. However, the gold-standard metric remains comparison to bacterial artificial chromosome (BAC) sequence, where disagreement between the assembly and BAC sequences are assumed to reflect errors in the draft sequence, even though a minority of errors arise from the BAC sequence (e.g. Consortium, 2004).

In the sections that follow I will discuss how the quality of reference genomes

impacts downstream analyses. This will be motivated by examples of issues within specific individual human, chimpanzee, and mouse assemblies; and comparison across assemblies of varying quality and construction (Table 2.1). The interplay between HTS analysis methods and reference genome errors is extremely informative in understanding error modes. Therefore, even though I was not directly responsible for the sequencing, mapping, or variant calling in the three presented case studies, I will describe the methods that were used to call variants.

Table 2.1: Properties of *Homo sapiens sapiens* (human), *Pan troglodytes verus* (Western chimpanzee), and *Mus mus domesticus* (mouse) reference genomes from UCSC genome browser (`genome.ucsc.edu`). Size: total bases in the assembly (including N's), % Masked: the genome percentage masked by RepeatMasker and TandemRepeatFinder, % N: the gap percentage, N50: the weighted median of contig lengths, such that half of all the contig lengths are at least N50. Sequencing approaches for the genomes are: whole-genome shotgun sequencing (WGS) and shotgun sequencing of clones (SOC).

Name	Id	Size (Gb)	% Masked	% N	Depth	N50 (Mb)	Sequence
Human	hg19	3.10	51.0	7.60	20.0	38.51	SOC
W. Chimp	panTro2	3.40	49.0	13.0	6.0	8.85	WGS + SOC
Mouse	mm9	2.75	44.0	3.90	20.0	37.68	WGS

2.2 CASE 1: Western Chimpanzee

In this case study I describe the HTS error modes and reference error modes encountered when constructing a genome-wide fine-scale recombination map from HTS of 10 Western chimpanzees, *Pan troglodytes verus* (Auton et al., 2012). The map was created to assess the nature and importance of factors that influence recombination across different scales in apes. Analyses of recombination rate variation at fine-scales are reported in Chapter 4. Consequently, this section also serves as background for following results chapters. Sections 2.5.1, 2.5.1.3, 2.5.1.8, 2.5.1.9, and

2.5.1.10 present the methods used by collaborators to generate the HTS, call and filter variants, estimate the effective population size of Western chimpanzees and identify hotspots. Their work provides the foundation for the construction of the fine-scale map of recombination in Western chimpanzees and can be found at the end of the chapter. Please note that in this thesis *chimpanzees* will be used to refer to *Western chimpanzees* unless otherwise stated.

2.2.0.1 Properties of the chimpanzee reference

DNA for the construction of the chimpanzee reference genome (panTro2) was obtained from whole blood of a captive-born male Western chimpanzee from the Yerkes Primate Research Center. The genome was sequenced using a whole-genome shotgun sequencing approach to an average sequencing depth of 6× (Sequencing and Consortium, 2005) though there are differences in the accuracy and completeness between chromosomes. Notably, chromosome 21 is more complete than other chromosomes. It was refined using additional sequencing from Watanabe et al. (2004), resulting in 273 contigs with an average size of 119,871 bp. The similarly sized chromosome 22 has 3,690 contigs of average size 8,765 bp. In contrast, the assembly of chromosome X assembly is worse than that of the autosomes. It contains 28,366 contigs with an average size 4,616 bp. Reads were assembled using the *de novo* assembler PCAP (Huang et al., 2003) with parameters calibrated to eliminate global mis-assemblies assessed by comparison with the human reference genome (Sequencing and Consortium, 2005). Non-repetitive chimpanzee contigs were mapped to the repeat masked human reference genome using BLATZ (Schwartz et al., 2003) and chimpanzee centromere locations were assumed to match the human centromere locations. BLAT was subsequently used to map more repetitive regions (Kent, 2002). Only regions that satisfied reciprocal-best mapping between the chimpanzee and human reference genomes were retained, achieving an estimated coverage of 97% of the chimpanzee

genome by 192,898 contigs with an N50 length of 44 kb. This left 246,488 gaps in the reference sequence (15.16%), with an average size of 1789 bp, and a range between 1 bp and 144 Mb (UCSC genome browser Fujita et al., 2011). Meader et al. (2010) estimated that the chimpanzee reference genome contains 50% more errors than the human reference (an average density of 0.17 clustered erroneous gaps per kb, compared to 0.11 errors per kb in the human reference).

2.2.1 Recombination map construction

To estimate recombination rates in Western chimpanzees, I applied an existing statistical method, LDhat (the *interval* program, McVean et al., 2004), to haplotypes inferred from whole genome sequencing of 10 unrelated Western chimpanzees (average 9.1× coverage).

Variants and haplotypes were inferred in a manner similar to the Consortium (2010) (See methods 2.5) using best-practice calling and conservative variant filtering procedures. Prior to recombination rate estimation the initial call set was thought to be of high quality as assessed by (1) population-genetic-based checks such as Hardy-Weinberg and adherence to expected transition to transvertion mutation ratio (expected TsTv = 2.10, see 2.5.1.3), and (2) validation of called variants on an independent platform. Note that the quality of mapping to the sex chromosomes was determined to be of poor quality and were therefore removed from the recombination analyses. I do not analyse the sex chromosomes in this thesis. In this section I will demonstrate, that despite best-practice approaches, cryptic error structures remained and were only detected through analysing the structure of fine-scale recombination rates across the chimpanzee genome. Error diagnosis and handling occurred as follows:

- Initial maps estimated from the filtered variants were dominated by large and artefactual increases in genetic distance caused by clusters of mis-mapped SNP

calls, often in large DNA repeats that are systematically under-represented in the chimpanzee reference genome.

- Two approaches were used to guard against errors (1) variant filtering was revised, (2) regional filtering was applied to the estimated recombination rate map (I explain why this was required in addition to SNP filters).

- The validation of the regional filtering protocol and the estimation of the noise arising in maps estimates from HTS of 10 genomes, through analysing maps estimated from HTS of 10 human samples each from populations of European (CEU) or African (YRI) ancestry from the 1000 Genomes Project.

2.2.1.1 LDhat parameters

I applied the LDhat program *interval* (McVean et al., 2004) to estimate recombination rates from the inferred haplotypes. The program uses a reversible-jump Markov Chain Monte Carlo (rjMCMC) algorithm to fit a piecewise constant recombination rate to the observed haplotypes, where the changes in rate can only occur at SNPs, and only cross-over events are considered. The key feature of this algorithm is that rates are estimated using the composite-likelihood (CL) approach of Hudson (2001), which enables computational efficiency through the simplifying assumption of pairwise independence between sites (in reality they are not). McVean et al. (2002) extended this model to include a finite sites model of mutation. Although the independence assumption does not reflect the true correlation structure between sites, studies have shown that these estimates are in good agreement with recombination rates inferred through orthogonal approaches (McVean et al., 2002). Over-fitting is controlled through a block penalty parameter, which is multiplied by the number of change points and added to the log composite likelihood. Hence, larger block penalty values identify positions at which there is very strong evidence for changes in

recombination rate, but at the expense of detail.

I chose a block penalty of 5 which has been shown by simulation to be adequate for the analysis of human data sets (Consortium, 2007). Several choices of population mutation rate θ were assessed. Though $\theta = 0.001$ is biologically justifiable (Sequencing and Consortium, 2005), a larger value of θ could potentially counteract the high error rates in HTS by increasing the relative probability of mutation events versus coalescent and recombination events in the model (Equation (1.7)), however at the expense of fine-scale detail. Rate estimates for $\theta = 0.1$ and $\theta = 0.001$ applied to haplotypes inferred on chromosome 21 were highly concordant. The Pearson correlation coefficient between the two rate estimates was 0.98 at 10 kb scale. For comparison, the correlation at 10 kb scale between the two most accurate human recombination maps, the deCODE pedigree-based map and the human admixture-based map, is 0.75 (Hinch et al., 2011). Figure 2.2 provides insight into the noise between the estimates. The additional variance explained using a model that incorporates a baseline level of noise between the two factors ($y_i = ax_i + b + \epsilon$) compared to a model that did not account for baseline noise ($y_i = cx_i$) was less than 1.2%. Therefore we conclude that there is no systematic large difference in the base-line noise. Hence a value of $\theta = 0.001$ was chosen.

To improve the consistency of the CL (in other words, to enable the property that as the number of loci increases to infinity the CL converges to the true recombination rate), estimates are obtained using a windowing approach where the CL is computed over batches of 50 SNPs (Fearnhead, 2003). Haplotypes were split into chunks of 4000 SNPs, with an overlap of 200 SNPs. For each chunk, the program was run for 60 million iterations and the first 20 million iterations were discarded as burn-in. Following the burn-in the chain was assumed to be in equilibrium and a sample was taken every 40,000 iterations, and a point estimate of the recombination rate between each SNP was obtained as the average across the chain samples. To reassemble each

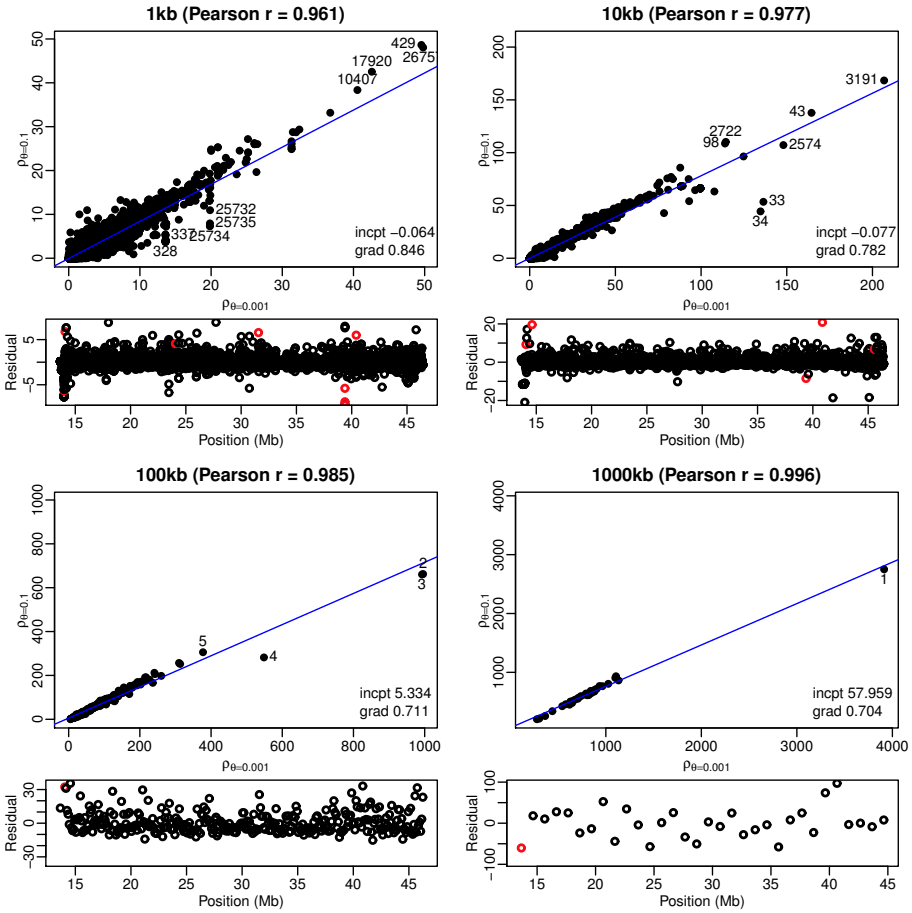


Figure 2.2: The impact of population mutation rate parameter θ on LD-based recombination rates estimated from 10 HTS Western chimpanzee genomes on chromosome 21. Four panels represent scatter plots of recombination rates estimated using $\theta = 0.001$ (x axis) and $\theta = 0.1$ (y axis) at different scales (1 kb, 10 kb, 100 kb, and 1 Mb); and below the scatter plots the residuals for the least squares linear fit. Selected outlier intervals identified on each scatter plot and coloured red in the residuals panel beneath. The Pearson correlation coefficient at each scale is indicated in panel titles. Least-squares linear model coefficients are displayed in the bottom right-hand-side corner of each panel.

chromosome from its overlapping windows the recombination rate estimates for each window were joined at the midpoint of the overlapping region similar to previous studies (McVean et al., 2004; Consortium, 2010).

2.2.2 Detection and removal of errors

Despite best-practice filtering, artefactual LD structure arose at both fine- and broad-scales. Inspection of the factors underlying spurious signals led to the identification of cryptic error structures arising from false-positive and false-negative variant calls. I describe their impact on recombination rate estimates from the chimpanzee sequencing, and describe how these errors were removed.

2.2.2.1 Errors arising from reference mis-assembly driven false-negatives

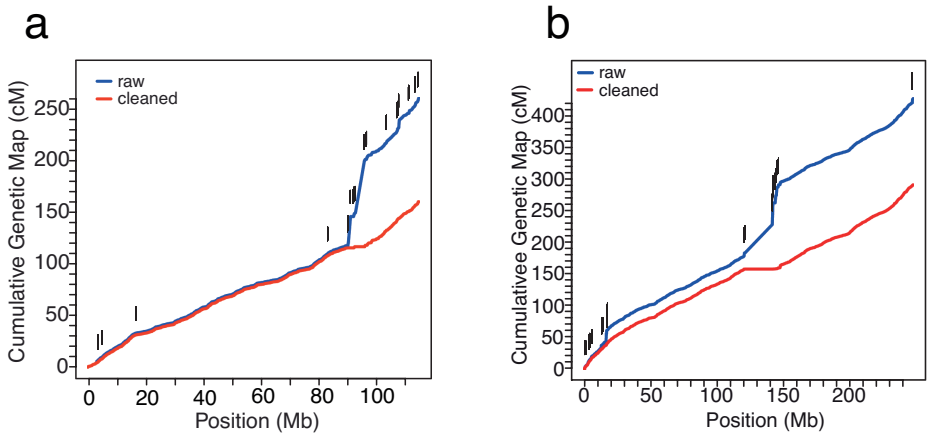


Figure 2.3: Artefacual increases in genetic maps estimated from HTS genomes. **(a)** Cumulative genetic map for chimpanzee chromosome 2a before (blue) and after (red) regional recombination rate filters. **(b)** Cumulative genetic map for chromosome 1 estimated from 10 YRI genomes sampled from 1000 Genomes before and after regional recombination rate filters. Black tick marks represent regions that were identified and corrected by the filters. Recombination rate estimates were scaled by N_e values reported in Section 2.5.1.9.

Initial maps estimated from the haplotypes had large increases in genetic distance characteristically different from recombination hotspot properties observed in humans, mice, and yeast (Jeffreys et al., 2001; Smagulova et al., 2011; Baudat and Nicolas, 1997) in both their magnitude and physical span (Figure 2.3a). Visual inspection of aligned reads at genetic map jumps suggested that jumps resulted from mis-assembly of the chimpanzee reference and not recombination events. For example,

the alignments over a 77 kb region on chromosome 10 showed considerable evidence for local contig mis-assembly (Figure 2.4. Specifically, clusters of inverted mate-pairs were observed in the alignments, with exceptionally large insert sizes; and coverage was elevated at the 'inversion' breakpoints, with high levels of mis-matches (Figure 2.4). This could be explained by two possible hypotheses.

First, this could be an instance of an inversion polymorphism, however, this seems unlikely as all of the 10 chimpanzees show evidence for the inversion (two chimpanzees shown in Figure 2.4). Second, this could represent mis-assembly in the chimpanzee reference. When recombination rate structure is additionally considered, mis-assembly is the most parsimonious explanation. As a result of the mis-assembly, SNPs that appear physically proximal in the reference genome in reality lie on separate haplotypes from different physical locations. The SNPs on misplaced contigs therefore share little LD with proximal contigs in the reference assembly resulting in artefactual signals of intense recombination activity. Note that within these regions recombination rates were consistently larger than the most active human recombination hotspots (as measured in sperm typing studies (Jeffreys et al., 2001, 2005)), and spanned regions an order of magnitude larger than the 1-2 kb expected of recombination hotspots.

To counteract these artefactual signals, I applied regional filters to the estimated recombination rates, setting regions to zero rate if:

- Pairs of adjacent SNPs had an estimated level of recombination rate $\rho > 100$, including the 100 SNPs surrounding identified intervals (50 SNPs in both directions), as these estimates may also carry over signal from the unreliable SNPs.

- Regions containing assembly gaps greater than 50 kb, to avoid regions of likely poor assembly.

Genome-wide 493 SNP intervals had rate estimates $\rho > 100$. These failed SNP intervals were often clustered, and represent approximately 250 distinct regions, a

total of 24,919 SNPs (0.50% of the total) occupied these regions. In addition, 446 gaps in the reference sequence with length greater than 50 kb were masked.

A total of 19,323 SNPs within 50 SNPs of gap break points were removed. Combined, 30,355 SNP intervals were set to have zero recombination rate (0.57% of the total), with 13,887 having failed both filters (46% of filtered intervals).

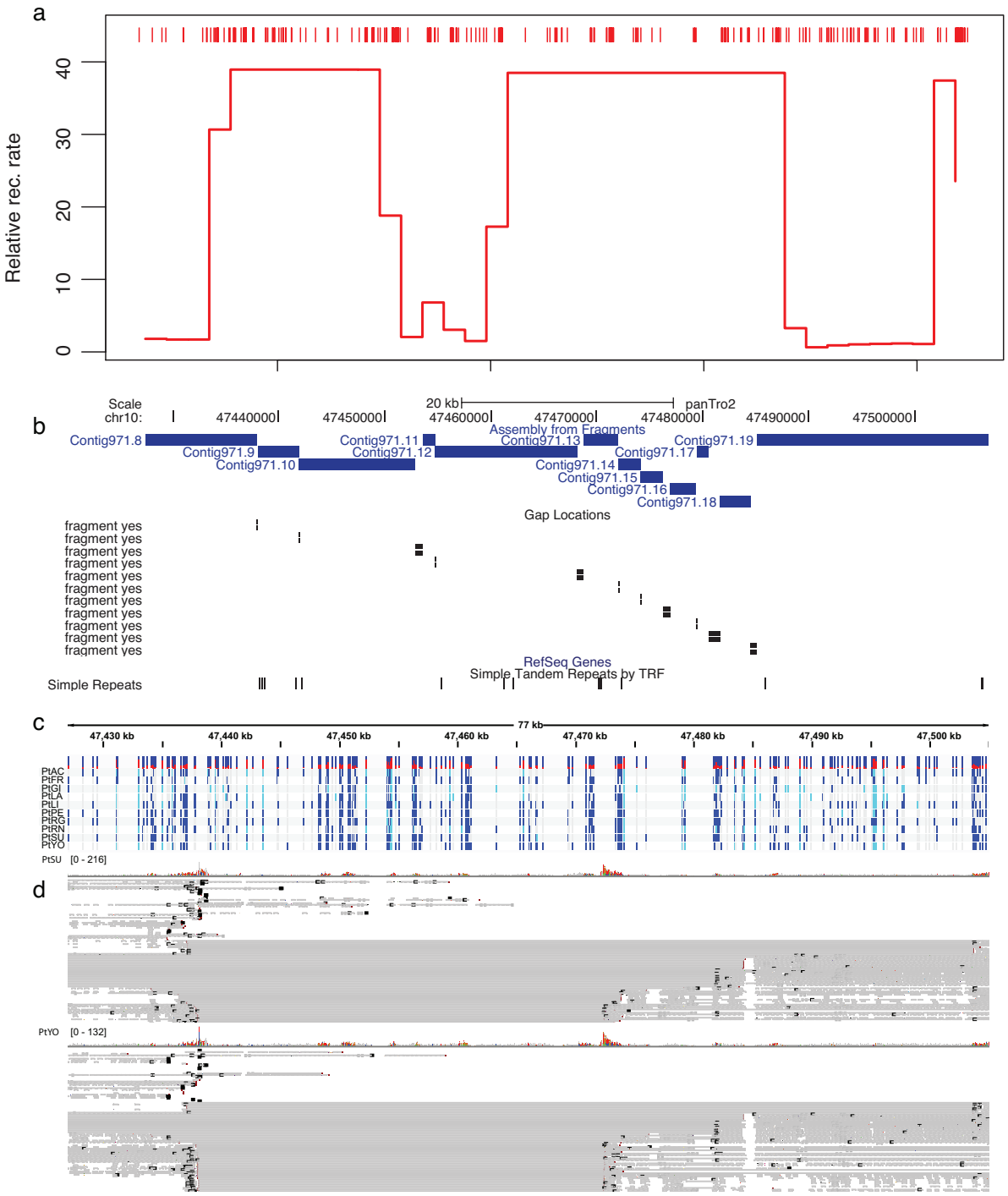


Figure 2.4: Mis-assembled contigs in chimpanzee reference. **(a)** estimated recombination rates relative to the chromosome average rate, tick marks indicate SNP locations. **(b)** UCSC genome browser view of the reference assembly. **(c)** polymorphism called from the 10 HTS genomes represented by rectangles, grey represents homozygous reference, blue heterozygous, and turquoise homozygous alternative genotypes. **(d)** read alignments for two of HTS chimpanzee genomes. Mapped coverage is represented by vertical bars, colours other than grey indicate mismatches with the reference. Horizontal bars represent aligned reads, mate pairs connected by grey lines, red reads indicate mate pairs with inverted orientation. All 10 sequenced chimpanzee alignments show this structure, though only 2 genomes, PtSU and PtYo, are displayed.

2.2.2.2 Errors arising from reference mis-assembly driven false-positives

A prevalent source of genome mis-assembly arises from assembly collapse at repetitive regions (Salzberg and Yorke, 2005). For example in the chimpanzee reference, DNA repeat elements are enriched by 14% around reference gaps (Figure 2.5), but reduced within 28 bp of the gap start/end, indicating that DNA repeats tend to be incompletely assembled in the chimpanzee reference.

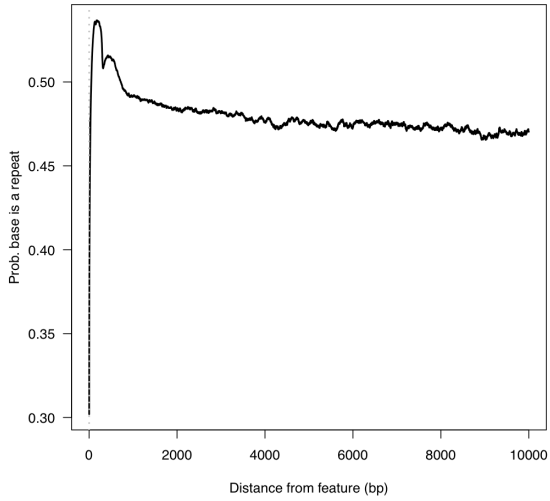


Figure 2.5: Chimpanzee reference genome (panTro2) contigs tend to end in repetitive sequence. Figure shows the probability that a contig base is in a repeat, for all interspersed repeats and low complexity DNA sequence (*RepeatMasker* track, UCSC genome browser) for the 169,088 gaps that break up autosome contigs (*Gap* track, UCSC genome browser). Zero on the x-axis represents the gap start/end point. Probability peaks at 169 bp from gaps. The N_{50} value for the chimpanzee reference genome is 30,568 bp.

Initial fine-scale recombination rate profiles around repeat elements in the chimpanzee genome [1], were dominated by intense localisation signals at Human Endogenous Retrovirus Group-H (HERVH) DNA repeat elements. Out of the 378 element types analysed, HERVH had the 3^{rd} highest recombination activity in the chimpanzee genome with an estimated average recombination at of 15 cM/Mb. Its flanking long

[1] estimated from haplotypes inferred from chimpanzee reads without BAQ recalibration (Li, 2011)

terminal repeat sequence, LTR7, was ranked 7^{th}.

The biology of HERVH initially suggested that the element was potentially a good candidate for localising recombination events. The HERVH provirus is composed of the functional elements 5'-LTR7-*gag-pro-pol-env*-LTR7-3' having reference consensus length of 7713 bp. The LTR units are necessary for viral integration and are known sites of homologous recombination. In humans, a cloning and hybridisation experiment of a sibling-pair identified a NAHR event between the flanking LTR7 elements of a HERVH (Mager and Goodchild, 1989). HERVH is one of the most abundant γ-retroviral transposons in the great apes. However, retroviral transposons are rare relative to other transposable element classes, constituting approximately 7% of the human genome (Smit, 1999). HERV elements as a class are estimated to have incorporated into the genomes of higher primates about 30 MYA at the time of the split between old- and new-world monkeys (Anderssen et al., 1997) and expanded rapidly in the old-world lineage with γ-retrotransposons only observed in Hominidae genomes (Goodchild et al., 1993).

To assess the veracity of the HERVH recombination-activating signal, I tested the consistency of LD structure around all repeat elements. Given the high levels of apparent recombination at HERVH elements, the breakdown in LD should be detectable from the sites neighbouring HERVH elements. However upon estimating recombination rates from only those variants in non-repetitive sequence, the hotspot signals were extinguished (Figure 2.6a). HERVH dropped from 3^{rd} hottest to 362^{nd} and LTR7 dropped from 7^{th} to 365^{th}, where as other repeats were only marginally re-ordered. The average recombination rate at HERVH elements decreased from 15 cM/Mb to 1.5 cM/Mb (see Figure 2.6b). The signals of intense recombination activity at HERVH were not supported by the surrounding haplotype information.

But why would such a signal arise? There is strong evidence that the variation observed in reads corresponding to HERVH sequence contains true segregating

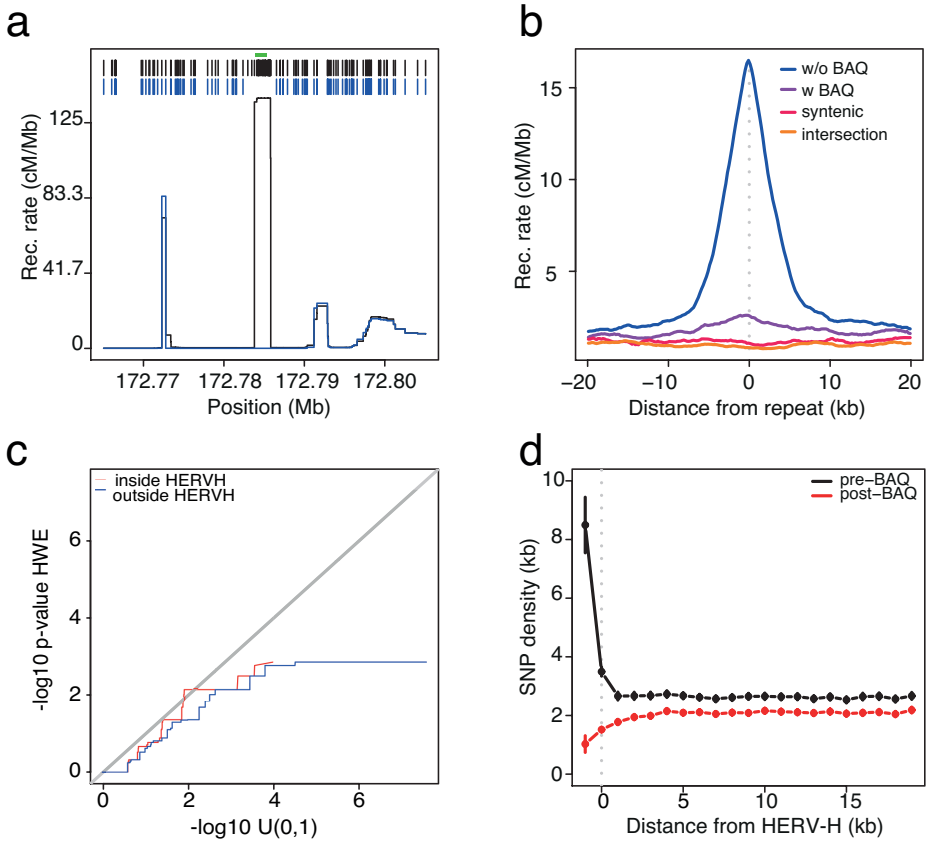


Figure 2.6: Artefacts arising from mis-assembly around HERVH repeat sequences. **(a)** Recombination rate estimates around a HERVH element on chromosome 1 (element represented by green bar). Black line shows recombination rate estimated from haplotypes without Base Alignment Quality (BAQ) recalibration. Blue line shows the recombination rate estimated over the same region with the SNPs inside the HERVH element excluded. Tick marks represent SNP locations. Note hotspot signal at HERVH is extinguished. **(b)** Lines indicate the mean recombination rate for HERVH elements before (blue) and after (purple) application of BAQ. Also shown are post-BAQ estimates for HERVH elements in the syntenic regions (red), and again when estimated using SNPs called in both the chimpanzee and human mappings (orange). **(c)** Quantile-quantile of -$\log_{10}$ p-values for Hardy-Weinberg equilibrium test for HERVH SNPs (red) the rest of the genome (blue) versus uniform distribution without BAQ. **(d)** SNP density as a function of distance from HERVH elements before and after BAQ recalibration.

variation. For instance, the sites have a well-behaved distribution of Hardy-Weinberg equilibrium test p-values (Figure 2.6c). There are three potential explanations. First, HERVH elements may be under-represented in the chimpanzee reference, resulting in mis-mapped variants. Second, a related explanation is that the minisatellite repeat

within HERVH repeats has systematically expanded in the chimpanzee lineage relative to the human lineage. This would be revealed by many reads showing repeat sequence, with subsets subtly different from each other and hence appearing as SNPs. Or, third, they might truly be polymorphic. We can not distinguish between the three alternatives from variant calls alone. For instance, each hypothesis is compatible with the elevated SNP density observed around HERVH elements (Figure 2.6d).

However, properties of the HERVH elements represented in the chimpanzee reference were substantially different from the human reference. First, there are only 4,510 HERVH elements accounting for 0.09% of the chimpanzee reference compared to 5,827 elements accounting for 0.19% of human reference. Second, the individual elements are largely fragmented in the chimpanzee reference (average size of 574 bp in panTro2 versus 962 bp in hg18 and Figure 2.7a). Such rapid and dramatic breakdown of HERVH could only occur through intense recombination activity or through assembly errors. The average distance between the midpoints of HERVH elements and their nearest gap is 1,625 bp compared to an average contig length of 5,892 bp (Figure 2.7b). Indeed, mapped coverage was elevated around HERVH elements with a high proportion of low quality reads (Figure 2.8), a symptom of mis-mapping. This arises for any non-unique sequence under-represented in the reference genome, resulting in high levels of reads being mapped (incorrectly) to the representative instances. Note that we applied depth filters to the variant calls (Section 2.5.1.3), however the stochastic coverage induced through aligning to partially assembled HERVH sequence in the reference results in, by chance, some HERVH loci passing coverage filters.

These observations motivated additional approaches to remove false positive SNPs in large repeats as described in Section 2.5.1.3 methods. Two approaches were used to remove artefacts. First, read qualities were re-weighted using the BAQ program (Li, 2011), which re-scales the quality score of bases based on the potential for local misalignment. Figure 2.6b demonstrates that the artefactual recombination-localisation

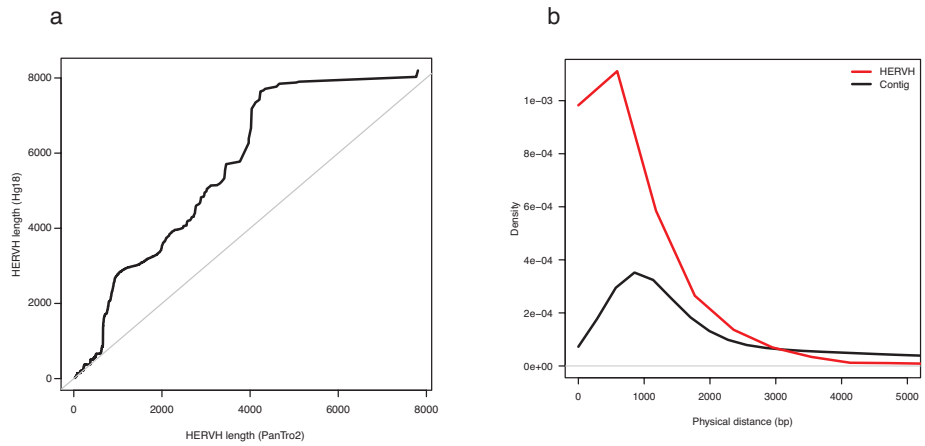


Figure 2.7: The majority of HERVH elements in the chimpanzee reference genome are fragments. **(a)** Quantile-quantile plot of the distribution of HERVH element sizes in the chimpanzee (panTro2, N = 4,510) and human (hg18, N = 5,827) reference genomes. Consensus HERVH reference sequence size is 7713 bp. **(b)** The distribution of HERVH elements within chimpanzee reference contigs versus the density function for sampling positions within the chimpanzee contigs.

signal at HERVH elements is largely removed after BAQ correction. Second, to capture sequence absent in the chimpanzee reference assembly the chimpanzee reads were mapped to the human reference and variants called against the human reference (referred to as the human-mapping, Section 2.5.1.2). Through intersecting the variants called from the human-mapping with those called from the chimpanzee-mapping, potentially problematic SNPs were identified and removed. All the recombination rate analyses used the rates estimated from haplotypes inferred from the BAQ-scaled, recalibrated chimpanzee alignments. To validate our findings, all analyses were checked for consistency against recombination rates estimated from the intersection call set in regions of broad-scale synteny between the two species (referred to as the robust genetic map, Section 2.2.2.3).

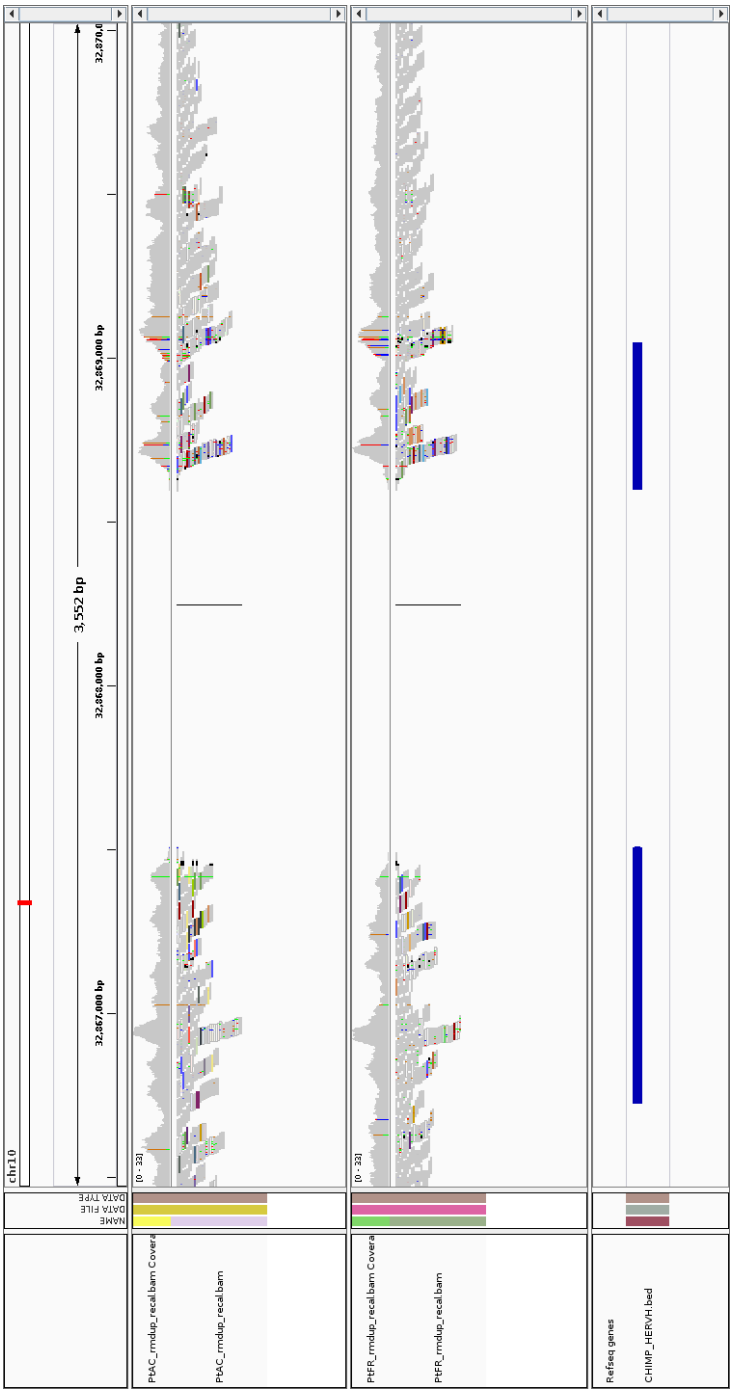


Figure 2.8: An example of reads mapped to a HERVH represented in the chimpanzee reference. IGV visualisation of reads for two HTS chimpanzee genomes aligned to panTro2 chromosome 10. Coverage is represented by vertical bars, colours other than grey alignment mismatches in the reads. Aligned reads are represented by horizontal bars, colours other than grey indicate that the mate pair is mapped to a different chromosome. Small vertical bands of colour within reads represent mismatches with the reference genome. Large blue rectangles at the bottom of the image indicate the locations of HERVH elements. There are no reads between the two HERVH elements because this region is a gap in the reference sequence.

2.2.2.3 The construction of a robust map of recombination

The robust map was generated using SNPs in well-behaved regions of the genome. SNPs were filtered using the following criteria:

1. Sites were called in both the human- and chimpanzee-mappings.
2. Sites passed the uniqueness filters in the human reference genome as defined by the 1000 Genomes low-coverage pilot study (overlapping 35mers that were 2-way unique) (Consortium, 2010),
3. Sites were located in regions of broad-scale synteny.

A total of 3,455,814 of 5,323,301 (65%) sites were retained after applying these criteria, with a TsTv of 2.19. Note that the robust map has lower resolution than the map from chimpanzee haplotypes and also likely underestimates the total recombination rate since non-syntenic sequence was excluded.

2.2.2.4 Generation of genetic maps from 10 human genomes of African and European ancestry

To assess the potential impact of high throughput sequencing and sample size on confounding comparisons between human and chimpanzee genetic maps, I sampled 10 genomes each from European (CEU) and African (YRI) ancestry from the 1000 Genomes low-coverage pilot study (Consortium, 2010). To best match the chimpanzee data, I selected the 10 genomes with highest average coverage that were sequenced using the same technology (Illumina GAII). The average coverage for the 10 genomes was 8.62× for CEU and 7.08× for YRI compared to 9.45× for chimpanzee. Despite efforts to match genomes there were two additional differences. First, some of the sub-sampled genomes had read lengths that differed from the 51 bp read length of the chimpanzee experiment (Table 2.2). Second, the 1000 Genomes haplotypes [2]

[2]1000 Genomes Low Coverage Pilot Project SNP release (July 2010)

were called across all the low-coverage sequenced genomes in the Pilot Study, not just the sampled genomes, and therefore are expected to be more accurate than the chimpanzee haplotypes.

Recombination rates were estimated from the haplotypes through LDhat using the same parameters as for the chimpanzee analysis. The ideal control experiment would require identical experimental conditions, however the 10 human genomes provided a rapid approximation to determine whether 10 HTS samples were adequate for recombination rate estimation. We avoided assessment through simulation because HTS error modes are only partially characterised and would therefore would be contrived.

Maps estimated from the human subsampled genomes also had large artefactual increases in their genetic maps (Figure 2.3). A total of 201 jumps were observed in the CEU and 106 in the YRI. Jumps were identified by comparing genetic distances between the maps from the 10 genomes and the HapMap combined-population map over 20 kb non-overlapping intervals, and were defined as regions where the magnitude of the genetic distance from the 10 genomes was greater than eight standard deviations greater than the average distance for the HapMap map - i.e. rate greater than the top 17% of hotspots over the entire 20 kb interval. Adjacent intervals were merged. Figure 2.9 shows an example of a large jump in the 10 YRI map on chromosome 4 at 69.4 Mb. The 1000 Genomes reads were mapped to the NCBI human reference build 36 (B36). The B36 assembly in this region is complex; composed of eight short contigs (all less than 0.6% of the N50 = 38.51 Mb length), a gap, and annotated with 114 segmental duplications constituting 87% of the bases. As in the chimpanzee example there is strong evidence of mis-assembly: inverted read pairs, indicated by red reads in Figure 2.9. Upstream of the gap, the 5'-end of contig AC147055.2 also has a low density of called variation. Outside of this region elevated rates of variants are called.

Importantly, in this case there appears to be a parsimonious explanation. The ori-

Table 2.2: Subsampling 10 genomes from the 1000 Genomes low coverage pilot genome. Ten Utah residents with Northern and Western European ancestry (CEU) and 10 Yoruba in Ibadan (YRI). Sequences were selected based on the following criteria (1) genomes were sequenced as a part of the low coverage pilot, (2) on Illumina Genome Analyzer II, (3) using paired end sequencing. From this subset the 10 individuals with the highest sequence coverage were selected. Sequence properties obtained from the 1000 Genomes Project Consortium ftp site.

Sample	Pop.	Read count $\times 10^6$	Base count $\times 10^6$	Read length	Coverage
NA11994	CEU	14.3	1,086.7	76	13.6
NA11919	CEU	21.7	1,649.0	76	10.1
NA12249	CEU	13.1	666.5	51	9.1
NA12005	CEU	9.6	728.3	76	8.3
NA12045	CEU	16.8	1,278.0	76	7.9
NA11831	CEU	11.1	841.4	76	7.5
NA12154	CEU	17.2	1,307.8	76	7.3
NA06986	CEU	20.0	1,518.4	76	6.7
NA12003	CEU	8.7	664.2	76	6.5
NA18522	YRI	13.2	999.9	76	9.1
NA19116	YRI	15.4	1,172.3	76	8.0
NA18912	YRI	10.4	531.0	51	7.8
NA19093	YRI	7.0	355.6	51	7.4
NA18516	YRI	8.4	638.9	76	7.3
NA18501	YRI	8.7	445.4	51	7.3
NA18870	YRI	5.2	395.3	76	6.6
NA18498	YRI	13.8	1,048.8	76	6.5
NA18510	YRI	12.6	958.8	76	6.4
NA18499	YRI	5.9	444.8	76	6.3

entation of three contigs change between B36 and subsequent assembly B37 (AC140484.1, AC019173.4, AC093720.2), though the number of called variants and coverage are extremely low across the first two of the three contigs.

To assess whether the disrupted haplotype structure resulted from mis-assembly in the human reference, I lifted the SNPs from reference B36 to B37 using the *LiftOver* program (3 SNPs failed liftover) and re-estimated recombination rates using LDhat (same parameters as described in methods). Figure 2.10 shows the recombination rates for the lifted SNPs. The elevated recombination rate signal over contig AC021146.7 is extinguished suggesting that the incorrect contig ordering in B36 did impact the recombination rate estimates. However a jump remains over contig AC1407055.2, a region of extensive structural variation. We conclude that this signal is likely generated by the human reference providing an inadequate representation of the high levels of variation in human genomes at this locus, also reflected in the breakdown of syntenic nets in this region. Note that this signal remains in the Phase I integrated call set from the 1000 Genomes Project Consortium and that the estimated genetic distance across the region is similar to the estimate from the lifted SNPs, and is 0.11% of the original build 36 estimates (Figure 2.11).

Note, however, that the rate of mismatches in aligned reads for the displayed YRI genome (NA19116) is extremely high, whereas the mismatches are low in CEU genome (NA12249) in Figure 2.9. This is representative of a more general, and interesting observation, that population specific jumps were observed. For example, the jump on chromosome 4 at 69.4 Mb was observed in the map estimated from 10 YRI genomes only. Therefore, despite the problematic assembly in these regions, we are likely identifying signals of population specific structural variation. These observations potentially suggest a high-throughput approach to identify targets of structural variation in between demographic groups. Nevertheless, these jumps in the genetic map do not reflect true recombination events, therefore in order to analyse

the structure of fine-scale variation these artefacts need to be removed. The same regional filters used to clean the chimpanzee genetic map were applied to the genetic maps estimated from the sampled genomes.

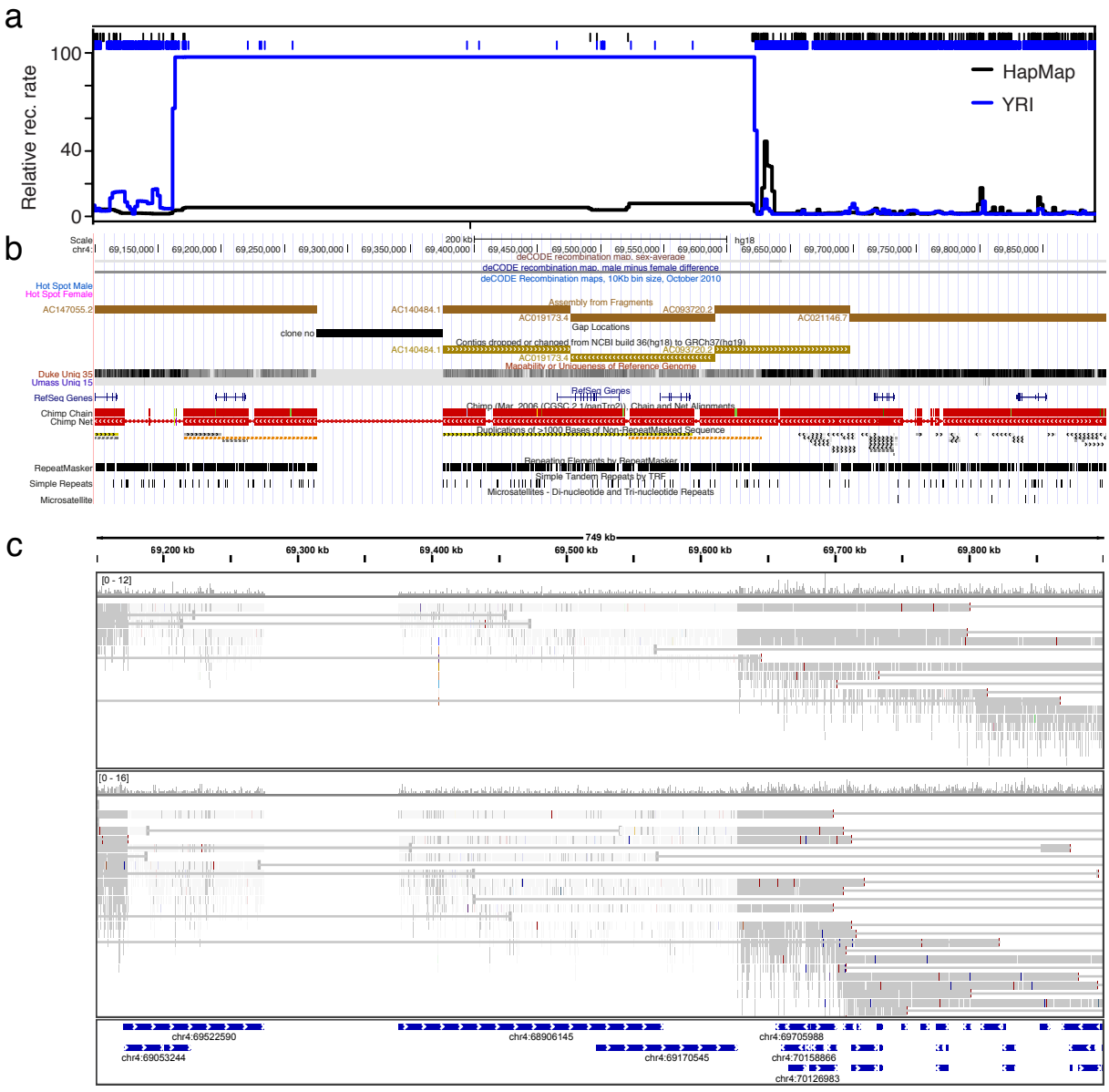


Figure 2.9: Mis-assembled contigs in human reference (B36). **(a)** Recombination rates relative to the average HapMap recombination rate in the region, tick marks indicate SNP locations. **(b)** UCSC genome browser view of (1) pedigree-based recombination rate estimates, (2) contig assembly, gold contigs indicate contigs that have changed orientation between B37 and B36, arrows indicate the direction of the orientation-change, (2) Duke uniqueness mask for 35-mers (grey represents low uniqueness), (4) RefSeq genes, (5) Segmental duplications (Bailey et al., 2002), (6) simple repeats, and (7) microsatellites. Second panel, read alignments for one YRI genome (NA19116, top) and one CEU genome (NA12249, bottom) from the 1000 Genomes Low Coverage Pilot study (Consortium, 2010). Alignments show slightly smaller region than the above panels. Mapped coverage represented by vertical bars, colours other than grey indicate differences from the reference. Horizontal bars represent aligned reads, mate pairs connected by grey lines, red reads indicate mate pairs with inverted orientation. Dark blue bars represent the locations of known segmental duplications. Note that coverage is greatly reduced within the elevated recombination rate region.

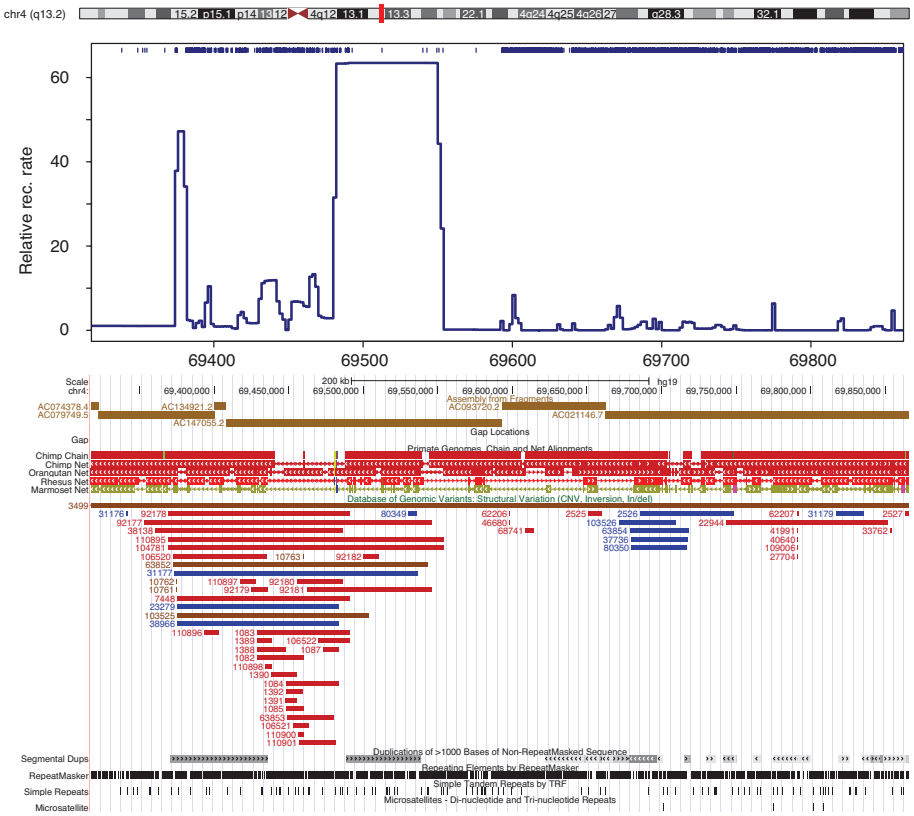


Figure 2.10: Recombination rate estimates for SNPs lifted over from mis-assembled contigs in the human reference (B36, shown in Figure 2.9) to B37. Top panel, recombination rates relative to the average HapMap recombination rate in the region, tick marks indicate the locations of SNPs mapped to B37 coordinates. Second panel, UCSC genome browser view of (1) contig assembly across region, (2) synteny maps (3) Segmental duplications (Bailey et al., 2002), (4) Simple repeats as identified by TandemRepeat masker, and (5) microsatellites.

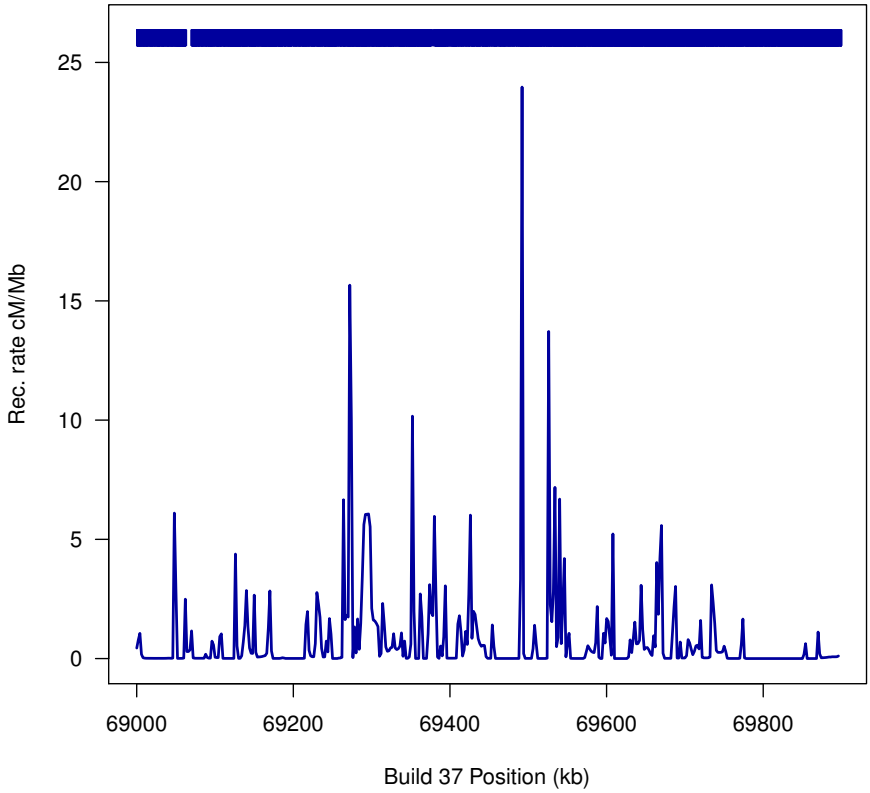


Figure 2.11: Recombination rate estimates from haplotypes inferred for the 10 YRI genomes, but now from the 1000 Genomes Phase I Integrated data set (Consortium et al., 2012) which was aligned to B37 (same region as displayed in Figures 2.9 and 2.10). Note that the recombination peak at 69.5 Mb observed in Figures 2.9 and 2.10) persists.

2.2.2.5 Assessing HTS-map quality by comparison to independent fine-scale maps

To estimate the sampling variance of the genetic map arising from experimental design the correlation between fine-scale genetic maps estimated for humans through different approaches was analysed. It was assumed that the fine-scale genetic maps estimated from admixture history (Hinch et al., 2011), deCODE genetics (Kong et al., 2010), and the HapMap project (Consortium, 2007) are independent estimates of the same underlying fine-scale map with comparable resolution and the same biological variance at each scale of analysis. These assumptions are not valid at fine kilobase resolution (Berg et al., 2010; Hinch et al., 2011), but are good approximations at broader megabase scales. Under this assumption the squared correlation between maps at a given scale can be written in term of the error contribution of each, relative to a chosen genetic map. There are two important caveats to this analysis: (1) the system of equations used to calculate the variance components are over-determined, we motivate our choices based on our viewpoints on the relative accuracy of the maps; (2) the method conflates the true biological variance in maps with the variance arising from errors, however at scales ¿ 100kb we expect that the biological contribution to the variance is negligible relative to the error contribution. We view this analysis as a robust measurement of the relative magnitude of the variance components and not a measurement of the absolute magnitude of the variance components. If absolute measurements were desired, this could be assessed through simulation.

If the vector of recombination rate estimates $\hat{\mathbf{r}}_i$ and $\hat{\mathbf{r}}_j$ for two maps (i and j) is each a linear combination of the true population recombination map and an error component vector, where the error components of the map are not correlated:

$$\begin{aligned} \hat{\mathbf{r}}_i &= \mathbf{r} + \boldsymbol{\epsilon}_i \\ \hat{\mathbf{r}}_j &= \mathbf{r} + \boldsymbol{\epsilon}_j, \end{aligned} \tag{2.1}$$

Then, taken over loci and errors, the first and second moments and the variance are

$$\begin{aligned} \mathrm{E}[\hat{\mathbf{r}}_i] &= \bar{\mathbf{r}} \\ \mathrm{E}[\hat{\mathbf{r}}_i \hat{\mathbf{r}}_j] &= \mathrm{E}[(\mathbf{r} + \boldsymbol{\epsilon}_i)(\mathbf{r} + \boldsymbol{\epsilon}_j)] \\ &= \mathrm{E}[\mathbf{r}^2] \\ \mathrm{V}[\hat{\mathbf{r}}_i] &= \mathrm{E}[(\mathbf{r} + \boldsymbol{\epsilon}_i)^2] - \mathrm{E}[\mathbf{r}]^2 = \sigma_r^2 + \sigma_i^2, \end{aligned} \tag{2.2}$$

where the σ_r term represents the variance in the true recombination rates, and the term σ_i represents the variance due to noise in the recombination rate estimates. Then the Pearson correlation can be written as

$$\begin{aligned} \mathrm{Cor}[\hat{\mathbf{r}}_i, \hat{\mathbf{r}}_j] &= \frac{\mathrm{E}[\mathbf{r}^2] - \bar{\mathbf{r}}^2}{\sqrt{(\sigma_r^2 + \sigma_i^2)(\sigma_r^2 + \sigma_j^2)}} \\ &= \frac{\sigma_r^2}{\sqrt{\sigma_r^2 \left(1 + \frac{\sigma_i^2}{\sigma_r^2}\right) \sigma_r^2 \left(1 + \frac{\sigma_j^2}{\sigma_r^2}\right)}} \\ &= \left(1 + \tfrac{\sigma_i^2}{\sigma_r^2}\right)^{-\frac{1}{2}} \left(1 + \tfrac{\sigma_j^2}{\sigma_r^2}\right)^{-\frac{1}{2}}. \end{aligned} \tag{2.3}$$

To assess the noise contributions to each map relative to a chosen map, we seek to derive a system of simultaneous equations that quantify the ratio of the variance from the noise contribution to the variance in true recombination rate. First, we approximate the correlation term through first order Taylor series expansion of the two $\left(1 + \frac{\sigma_i^2}{\sigma_r^2}\right)^{-\frac{1}{2}} \approx 1 - \frac{1}{2}\frac{\sigma_i^2}{\sigma_r^2} + \mathcal{O}\left(\left(\frac{\sigma_i^2}{\sigma_r^2}\right)^2\right)$ terms under the assumption that the variance due to estimation noise is less than the variance in true recombination rate ($\sigma_i \ll \sigma_r$). The correlation can then be approximated as,

$$\mathrm{Cor}[\hat{\mathbf{r}}_i, \hat{\mathbf{r}}_j] \approx 1 - \frac{\sigma_i^2 + \sigma_j^2}{2\sigma_r^2} + \mathcal{O}\left(\frac{\sigma_i^2 \sigma_j^2}{\sigma_r^4}\right). \tag{2.4}$$

This gives a system of simultaneous equations to solve in order to assess the noise

contributions to each map, given the estimated values of $\frac{\sigma_i^2}{\sigma_r^2}$.

The relative error contributions to the variance in rates were defined in terms of the estimated pairwise correlations between maps (represented by $\rho_{\text{map1 map2}}$) for:

- deCODE (e_D)
- HapMap combined-population (e_H)
- African America admixture (e_A)
- population-specific HapMap ($e_{H_{YRI}}$ and $e_{H_{CEU}}$)
- 10 genomes subsampled from the 1000 Genomes ($e_{10_{YRI}}$ and $e_{10_{CEU}}$).

Because the system of simultaneous equations are overdetermined, we make two modelling assumptions. We assume that the deCODE, HapMap and Admixture maps are more accurate measurements of $\mathbf{r}$ than the population-specific HapMap and subsampled 1000 Genomes maps and therefore calculate error components e_H, e_A, and e_D using these three maps. Note that the Admixture map is an average of European and African maps and therefore may have lower estimates of the error component through reversion toward the mean. For the population-specific LD-based error components, we assumed they are a superposition of the error component measured in the combined-population HapMap map and the population-specific map. This results

in the following equations:

$$
\begin{aligned}
e_D = \frac{1}{2}(1 - \rho_{D\,A} - \rho_{D\,H} + \rho_{H\,A}) &\approx \frac{1}{2}\frac{\sigma_D^2}{\sigma_r^2} \\
e_H = 1 - \rho_{D\,H} - e_D &\approx \frac{1}{2}\frac{\sigma_H^2}{\sigma_r^2} \\
e_A = 1 - \rho_{H\,A} - e_H &\approx \frac{1}{2}\frac{\sigma_A^2}{\sigma_r^2} \\
e_{H_{YRI}} = 1 - \rho_{H\,H_{YRI}} &\approx \frac{\sigma_H^2 + \sigma_{H_{YRI}}^2}{2\sigma_r^2} \\
e_{10_{YRI}} = 1 - \rho_{H\,10_{YRI}} &\approx \frac{\sigma_H^2 + \sigma_{10_{YRI}}^2}{2\sigma_r^2} \\
e_{H_{CEU}} = 1 - \rho_{H\,H_{CEU}} &\approx \frac{\sigma_H^2 + \sigma_{H_{CEU}}^2}{2\sigma_r^2} \\
e_{10_{CEU}} = 1 - \rho_{H\,10_{CEU}} &\approx \frac{\sigma_H^2 + \sigma_{10_{CEU}}^2}{2\sigma_r^2}.
\end{aligned}
$$

The error contribution was calculated between the fine-scale genetic maps at 10 kb, 100 kb, 1 Mb, 5 Mb (Table 2.2.2.5). For each scale, the genetic distance was interpolated from the cumulative genetic map across non-overlapping windows equal to the scale size. The first 5 Mb of telomeric regions of each chromosome, and the 5Mb regions either side of the centromere were excluded to match the limited extent of the deCODE map (Kong et al., 2010). Note that at the resolution of hotspots the assumption ($\sigma_i \ll \sigma_r$) could well be violated therefore caution should be used when interpreting estimates at the 10 kb scale. In addition, the limit of resolution in the deCODE map is 10 kb, potentially further limiting analysis at the 10kb scale. The Pearson correlation was calculated over the aggregate of all the interpolated genetic distances for the autosomes. The intervals in the 10 YRI and 10 CEU genetic maps that had their rate set to zero by the map cleaning procedure (rate-zeroed regions) were corrected by scaling their genetic distance by the remaining non-zeroed physical distance in that window, and if the rate-zeroed region accounted for more than 30% of the physical distance of the window, that window was excluded. This had the

effect of removing sequence from the HTS-maps where we were not confident in the recombination rate estimates.

Table 2.3: Fraction of variance in rate estimates at a given scale arising from estimation noise. Admixture refers to the admixture map from Hinch et al. (2011), deCODE the sex averaged genetic map from Kong et al. (2010), HapMap the combined population map from Consortium (2007), HapMapYRI the YRI population specific map from HapMap, HapMap CEU the CEU population specific map from HapMap, 10YRI and 10CEU the ten genomes sampled from the 1000 Genomes Low Coverage Pilot Consortium (2010).

Scale	Admixture	deCODE	HapMap	HapMap YRI	HapMap CEU	10YRI	10CEU
10kb	0.086	0.166	0.105	0.103	0.050	0.336	0.411
100kb	0.039	0.062	0.064	0.063	0.044	0.240	0.380
1Mb	0.013	0.019	0.028	0.024	0.022	0.119	0.181
5Mb	0.005	0.010	0.015	0.012	0.013	0.066	0.110

To assess the contribution of (i) constructing maps from a particular population and (ii) the estimation noise arising from analysing sequencing data from ten individuals, we compared (i) the combined-population HapMap estimates to values obtained from a single HapMap population and (ii) the population-specific HapMap genetic maps to those obtained from the 10 individuals from the same population in the 1000 Genomes Project (Consortium, 2010). Using the values presented in Table 2.2.2.5, the fraction of variance in rate estimates at a given scale arising from estimation error can be calculated, which we refer to as the estimation noise.

These values are likely to be overestimates as the contribution of biological divergence between populations and the increased resolution of genetic maps arising from sequencing data has been ignored.

Nevertheless, they suggest that for the map from 10 YRI genomes (10YRI) the correlation between the truth and our estimate is over 0.9 at the 1 Mb scale and between 0.5 and 0.75 at the 10 kb scale (depending on what fraction of the divergence between HapMap combined population map (HapMap) and HapMap map from YRI

only (HapMapYRI) is due to biological divergence; Table 2.3). Note that the major contributor to variance arises from fine-scale estimation error at recombination hotspots. Estimates from the map from 10 CEU genomes are noticeably worse, perhaps reflecting the lower diversity and higher linkage disequilibrium in the population.

Table 2.2.2.5 Pearson correlation between human genetic maps over 10kb,100kb,1Mb,5Mb intervals across autosomes.
Variance of recombination rates are reported along the diagonal.

10kb	D	HA	HC	HY	10C	10Y	A
D	0.001	0.728	0.697	0.637	0.461	0.473	0.749
HA		0.001	0.929	0.879	0.589	0.626	0.806
HC			0.001	0.699	0.568	0.521	0.719
HY				0.001	0.503	0.643	0.764
10C					0.001	0.597	0.503
10Y						0.001	0.574
A							0.001
100kb	D	HA	HC	HY	10C	10Y	A
D	0.015	0.872	0.835	0.782	0.512	0.625	0.901
HA		0.016	0.915	0.889	0.578	0.707	0.891
HC			0.022	0.750	0.548	0.605	0.794
HY				0.016	0.504	0.684	0.826
10C					0.035	0.701	0.532
10Y						0.016	0.678
A							0.022
1Mb	D	HA	HC	HY	10C	10Y	A
D	0.452	0.956	0.927	0.888	0.646	0.790	0.971
HA		0.558	0.945	0.923	0.663	0.817	0.953
HC			0.657	0.847	0.636	0.745	0.884
HY				0.589	0.603	0.765	0.874
10C					0.935	0.842	0.628
10Y						0.549	0.789
A							0.638
5Mb	D	HA	HC	HY	10C	10Y	A
D	4.310	0.978	0.963	0.927	0.759	0.883	0.985
HA		7.350	0.978	0.975	0.778	0.899	0.983
HC			8.236	0.947	0.767	0.862	0.959
HY				7.377	0.760	0.898	0.956
10C					8.403	0.882	0.755
10Y						6.278	0.878
A							8.287

Key:		
Decode (sex-combined)	(Kong et al., 2010)	D
HapMap (combined population)	(HapMap Consortium, 2007)	HA
HapMap (CEU population only)	(HapMap Consortium, 2007)	HC
HapMap (YRI population only)	(HapMap Consortium, 2007)	HY
1000G Pilot 1 (10 CEU genomes)		10C
1000G Pilot 1 (10 YRI genomes)		10Y
Admixture	(Hinch et al., 2011)	A

2.2.2.6 Comparison of sampled genetic maps to reference maps at Winckler et al. (2005) regions

To estimate the accuracy of the fine-scale genetic maps I calculated the correlation between the genetic maps estimated from the 1000 Genomes sampled genomes and independent genetic maps at the regions studied by Winckler et al. (2005). These estimates rely on the same assumptions mentioned in the previous section.

The structure of fine-scale variation at the (Winckler et al., 2005) regions is shown in Figure 2.12. Note that the maximum resolution of the deCODE genetic map is 10 kb and is therefore more prone to reduced correlation values due to end effects of the interpolation relative to the other maps. The 10 sampled genomes had correlation greater than 60% with the HapMap combined-population map for all regions (Table 2.2.2.6).

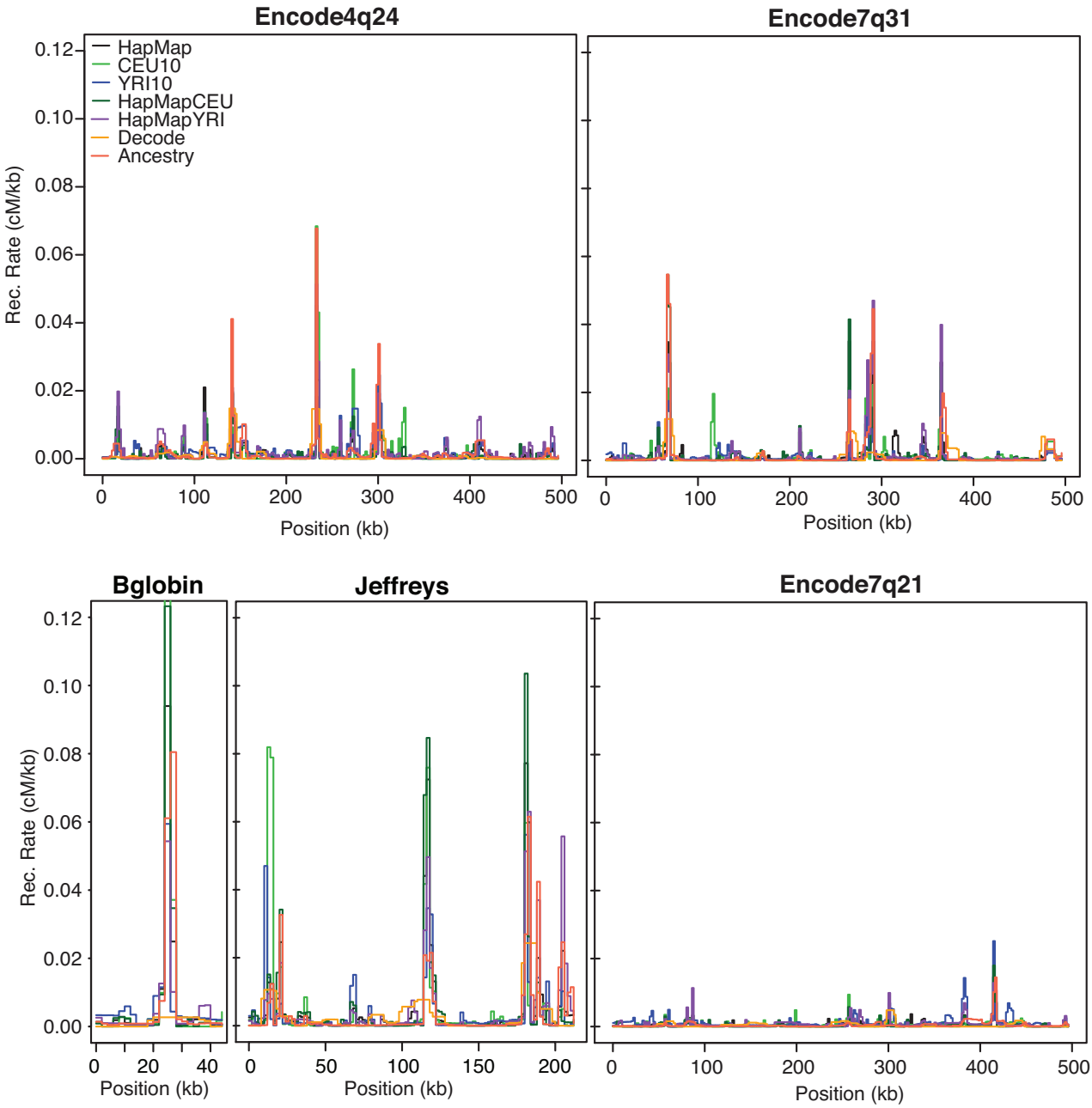


Figure 2.12: Comparison of estimates of human fine-scale recombination rates at the (Winckler et al., 2005) regions, estimated using different approaches: LD-based from the HapMap combined-population map and population specific maps (HapMap, HapMapCEU, HapMapYRI; Consortium, 2007), and population specific maps from the 10 subsampled 1000G individuals (CEU10, YRI10); the sex-averaged deCODE pedigree-based map (deCODE; Kong et al., 2010)); and African American admixture-based map (Ancestry; Hinch et al., 2011). All rate estimates have been smoothed at the 5 kb scale to aid visual comparison.

Table 2.2.2.6 Pearson correlation between human genetic maps over 10kb partitions regions studied by Winckler et al. (2005).

Variance of recombination rate over partions are reported along the diagonal.

Encode4q24	HA	10C	10Y	HC	HY	D	A
HA	1.446E-04	0.859	0.625	0.869	0.947	0.806	0.870
10C		3.186E-04	0.506	0.683	0.786	0.646	0.740
10Y			1.509E-04	0.769	0.563	0.501	0.591
HC				7.689E-05	0.814	0.705	0.749
HY					2.030E-04	0.741	0.849
D						1.378E-04	0.841
A							2.474E-04
Encode7q21	HA	10C	10Y	HC	HY	D	A
HA	1.597E-05	0.638	0.711	0.919	0.902	0.074	0.858
10C		8.234E-06	0.411	0.625	0.619	0.399	0.481
10Y			4.690E-05	0.655	0.731	0.191	0.663
HC				1.874E-05	0.803	0.022	0.889
HY					2.661E-05	0.109	0.763
D						1.074E-05	-0.103
A							2.301E-05
Bglobin	HA	10C	10Y	HC	HY	D	A
HA	3.994E-03	0.998	0.998	0.999	0.990	0.672	0.999
10C		7.816E-03	0.998	0.999	0.980	0.639	1.000
10Y			1.336E-03	0.999	0.980	0.624	0.998
HC				6.645E-03	0.985	0.654	1.000
HY					1.381E-03	0.766	0.985
D						4.716E-05	0.660
A							5.488E-03
Jeffreys	HA	10C	10Y	HC	HY	D	A
HA	1.896E-03	0.664	0.904	0.971	0.922	0.802	0.916
10C		2.139E-03	0.862	0.734	0.524	0.619	0.511
10Y			7.174E-04	0.906	0.804	0.799	0.784
HC				2.141E-03	0.815	0.732	0.819
HY					1.686E-03	0.764	0.966
D						6.038E-04	0.815
A							1.156E-03
Encode7q31	HA	10C	10Y	HC	HY	D	A
HA	2.303E-04	0.826	0.832	0.839	0.935	0.662	0.926
10C		1.603E-04	0.722	0.738	0.832	0.587	0.864
10Y			5.809E-05	0.743	0.821	0.475	0.795
HC				3.032E-04	0.704	0.756	0.882
HY					2.698E-04	0.551	0.907
D						9.169E-05	0.748
A							3.750E-04

Key:		
Decode (sex-combined)	(Kong et al., 2010)	D
HapMap (combined population)	(HapMap Consortium, 2007)	HA
HapMap (CEU population only)	(HapMap Consortium, 2007)	HC
HapMap (YRI population only)	(HapMap Consortium, 2007)	HY
1000G Pilot 1 (10 CEU genomes)		10C
1000G Pilot 1 (10 YRI genomes)		10Y
Admixture	(Hinch et al., 2011)	A

2.3 CASE 2: Wild Mus musculus domesticus

This experiment was not a central component of my thesis and I contributed to the project as a consultant. Through my involvement in the project, errors were identified in the initial variant call set. The errors that I characterised are informative about the challenges of calling variation from HTS and are therefore included in this chapter.

In this section I will discuss the quality of the filtered calls from the GATK produced by collaborators. We observed large regions of homozygosity in individual mice samples (Figure 2.13), representing recent inbreeding. Within the homozygous regions (identical by descent; IBD regions), more heterozygous sites were observed than are expected to have arisen through mutation since the inbreeding events. Inspection of mapped reads containing heterozygous variants in homozygous regions led to the identification of four error classes:

1. Sequence complexity driven false-positives arising from analysis methods failing to distinguish between true variation and false positives in regions of low sequence complexity, especially at tandem repeat elements.

2. Local-mapping driven false positives around true variants, particularly around small indels. Figure 2.14 provides an example of false heterozygous calls around small homozygous alternative indels.

3. Reference-genome incompleteness driven false positives in segmental duplications that are polymorphic. Regions are characterised by high or variable read coverage with high levels of reference mismatches (Figure 2.15).

4. In IBD regions, homozygous alternative sites were called as heterozygous sites despite very little, if any, evidence for the reference allele in the reads (Figure 2.16). This error-class arises from the Hardy-Weinberg equilibrium assumption in the GATK genotyping algorithm which inherently causes heterozygous calls

to be imposed in regions of extensive IBD sharing.

IBD regions provided unprecedented insight into error-modes and the opportunity quantity error rates in HTS. Over a 1 Mb region of chromosome 19 in mouse 234, thirty putative errors in homozygous regions were inspected: four resulted from mis-called indels in repetitive sequence (three were tandem repeats, one was a poly-G), five resulted from HWE-based genotyping error, and 21 errors at a segmental duplication. In addition, error rates were assessed through calculating the relative rates of heterozygosity for each mouse. The relative heterozygosity was defined as the ratio of the rate of heterozygosity within an individual mouse to the whole population. Figure 2.13 shows relative heterozygosity over chromosome 7 for four mice. Under the assumption that the heterozygous calls within homozygous regions were all errors (a negligible number represent true mutation events), the genotyping error rate, defined as the ratio of heterozygous calls in homozygous regions to the total number of variants in the region, was calculated. The error rate was estimated to be 13.6% for the heterozygous calls from the 75 Mb homozygous region in mouse 329, versus the heterozygous call rate on the rest of mouse 329 chromosome 7.

To counteract these errors I implemented two approaches. First, I took the intersection set of the variants calls as reported by the GATK (using settings and subsequent filtering specified in methods) and Platypus (`http://www.well.ox.ac.uk/platypus`; using default settings). The Platypus calls were used to identify those GATK calls that were of high quality. Platypus was chosen for its capacity to accurately call variation at indels. Poor quality genotype calls in the intersection set were set to missing using two criteria:

1. Variants with conflicting genotype assignments between the GATK and Platypus.

2. Variants with a GATK reported genotype likelihood difference less than 10

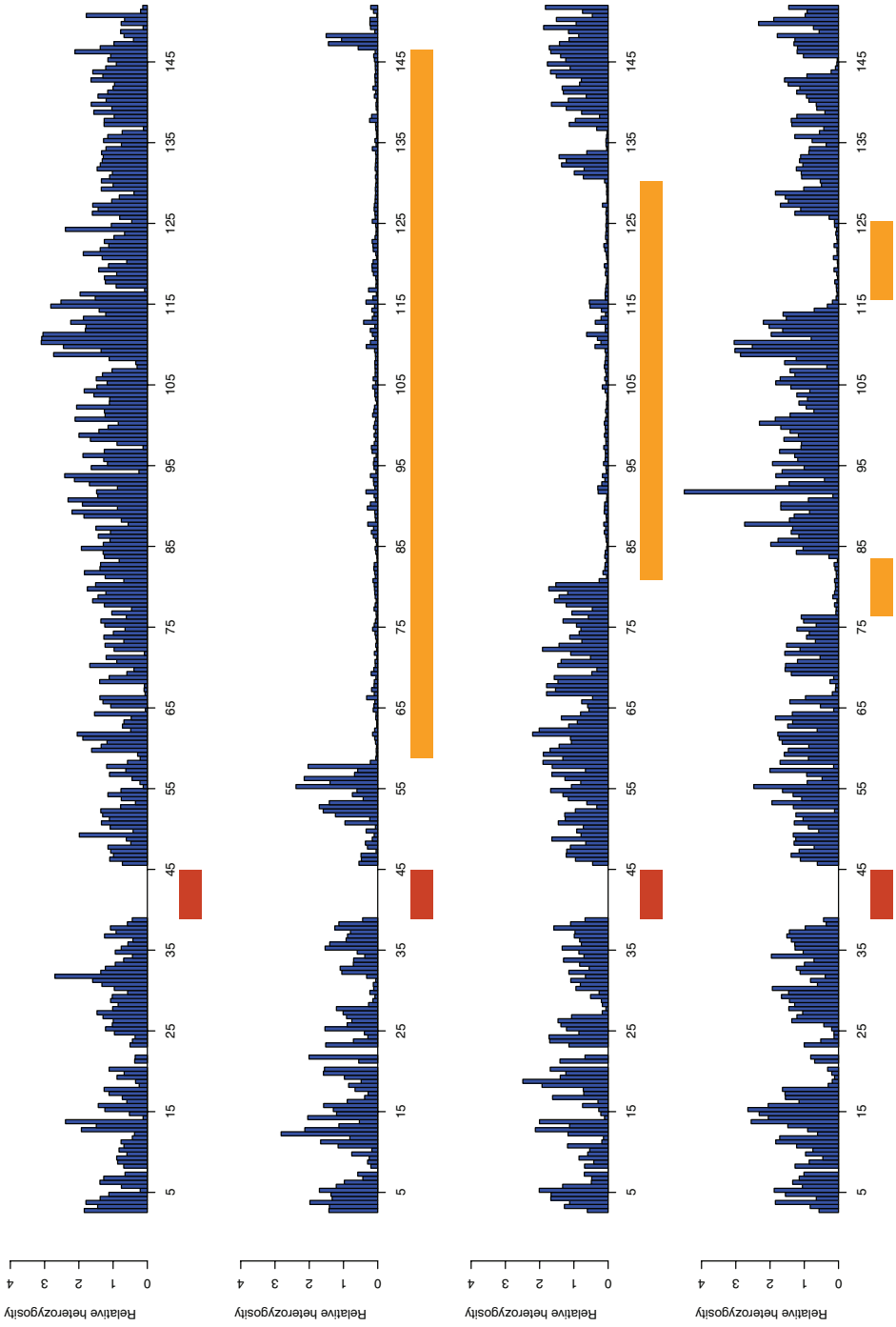


Figure 2.13: Estimated heterozygosity across chromosome 7 for four of the 20 HTS mice genomes. The relative heterozygosity was calculated by taking the ratio of the individual fraction of heterozygous calls to the sample average for the filtered GATK call set. A spectrum of homozygous block sizes is observed across the four mice, represented by orange rectangles. Ochre rectangles represent a gap in the mouse reference genome.

Phred units between the most likely and second best biallelic genotype model were set to missing.

The intersection approach reduced the error rate to 1.6% and 14% of the genotypes were set to missing by this process. This demonstrates that no variant caller is perfect and that a consensus based approach, as advocated by the 1000 Genomes, is very powerful. Importantly however, this approach does not identify errors arising from reference errors or incompleteness.

Second, I generated a mask for each sampled genome independently based on the properties of its alignments. The idea behind the masking process is the that the reference genome is an inherently inadequate representation of all the genomes in a population, and that divergence between the reference and the sample causes errors. Based on the properties of the erroneous heterozygous calls in homozygous regions, several heuristics were implemented to identify problematic regions. The masking process can be explained by the following algorithm. For a chromosome of length e start at the position, x, the first nucleotide on the chromosome that is covered by one or more reads,

1. if x > e, end.
2. advance x by 1 nucleotide.
3. Count the number of reads covering x. If the coverage is less than 1 or greater than 22, mask x and return to 1.
4. If any read covering x contains an indel, mask x and return to 1.
5. Count the number of reads covering x that fall into each of the following classes:
   (a) That have an insert size less than l_l or greater than l_u. If more than one read is counted, mask x and return to 1.

(b) That have more than five mismatches with the reference. If more than one read is counted, mask x and return to 1.

(c) That have unmapped mate pairs. If more than one read is counted, mask x and return to 1.

(d) That have incongruent mate pair orientation. If more than one read is counted, mask x and return to 1.

The mask filters were motivated by the analysis of the erroneous heterozygous calls. Mask 5a was applied to identify regions of the genome that were potentially mis-assembled or sites of structural variation. Differences in insert size between libraries necessitated custom insert size thresholds. A lower bound of 100 bp was applied for all samples. An upper bound of 600 bp was applied to mouse 246 and mouse 227, and 500 bp for the other samples (Figure 2.17). Thresholds on coverage for Mask 3 were chosen to exclude alignments with coverage in the bottom 0.01^{th} percentile or in the top 99.99^{th} percentile assuming that coverage followed a Poisson distribution with mean coverage $10\times$. Mask 5b was introduced to counteract false-positive calls resulting from local mis-mapping, especially at tandem repeat elements and segmental duplications. Reads with more than 5% mismatching nucleotides were flagged. Mask 4 was implemented to avoid erroneous SNP calls around indels, which we observed frequently occurred from local mis-mapping. Mask 5d identified regions where the sampled genome was inverted relative to the reference, which we assumed reflected mis-assembly, though some instances may represent segregating inversions.

The filters masked an average of 8% of the autosomal sequence per sample. The rate of heterozygous calls in homozygous regions decreased to 0.98% after application of the alignment masks, this approach set to missing an additional 4% of the genotypes.

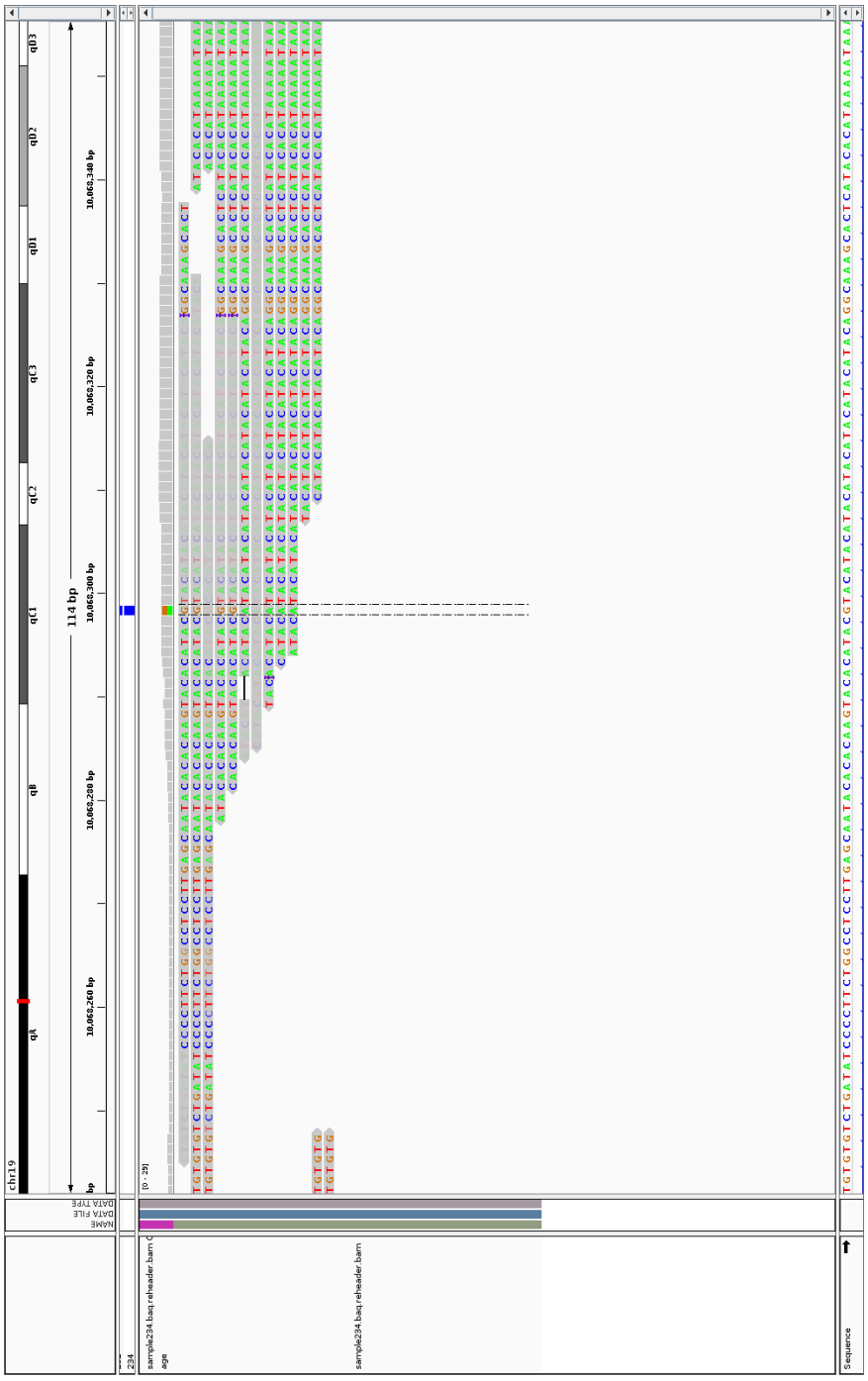


Figure 2.14: Erroneous heterozygous genotype calls at small homozygous alternative indels in the 20 HTS mice genomes. IGV view of reads mapped to 114 bp of chromosome 10 for one of the 20 sampled mouse genomes (mouse 234). The variant site resides in a homozygous region for this mouse. Failure to report 6 bp homozygous alternative insertion TACGTACA on chromosome 19 (mm9) results in artefactual heterozygous SNP call. Vertical bars represent read depth for the alignment. Horizontal bars containing nucleotides represent aligned reads (opaque letters represent bases with Phred quality < 5). The purple *I* character represents insertions, black lines represent deletions. Reference sequence displayed at the bottom of the image. Note this regions contains an abundance of (1) reference mis-matches, (2) indels, (3) reads with low base qualities.

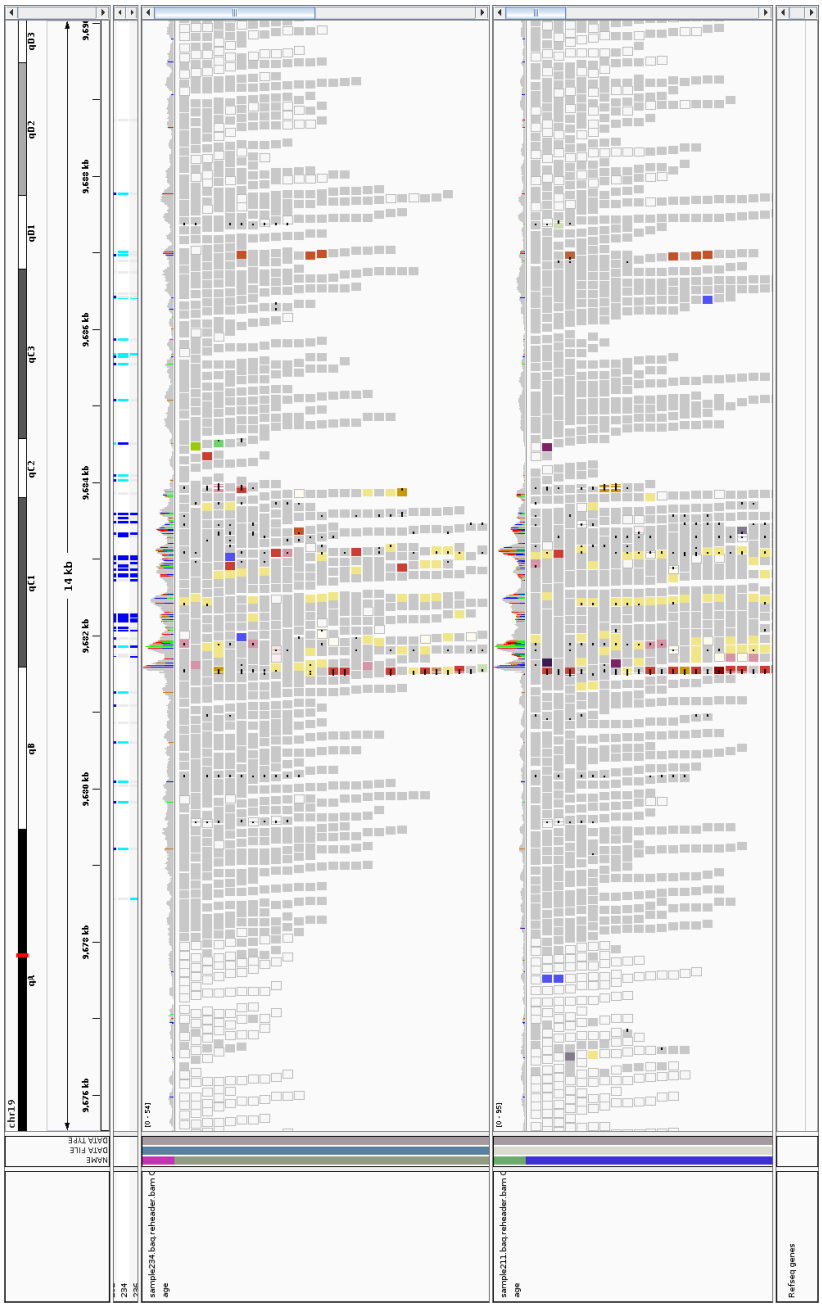


Figure 2.15: Mis-mapping at a segmental duplication in the 20 HTS mouse genomes. IGV view of reads mapped to 14 kb of chromosome 19. Two mice 234 (top) and 211 (bottom) are displayed. Mouse is 234 is homozygous across this region. Turquoise and blue bars represent homozygous and heterozygous alleles called within the 20 HTS mice genomes. Below, vertical bars represent depth of aligned reads, colours other than grey represent mismatches to the reference observed in more than 20% of the covering reads. Horizontal bars represent aligned reads, coloured reads indicate that mate pairs were mapped to a different chromosome. Note that the reads contain high levels of (1) mis-matches (not all called as variation), (2) coverage, (3) reads whose mate is mapped to an alternative chromosome, (4) indels.

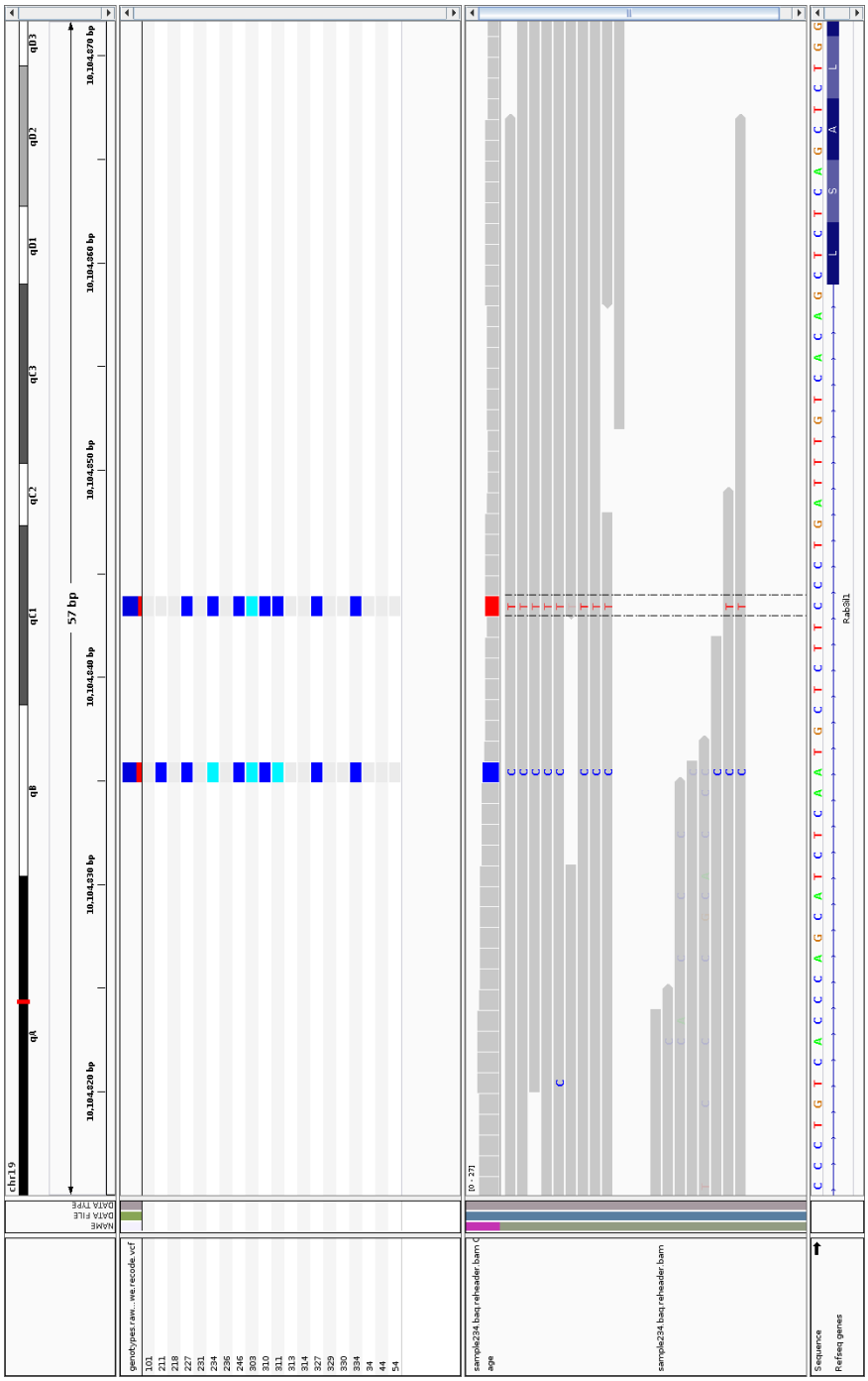


Figure 2.16: Genotyping error arising from the GATK genotype model assumptions, which are inappropriate for the 20 wild mice. Top panel: represents genotype calls. Turquoise, blue and grey blocks represent homozygous alternative (AA), heterozygous (RA), and homozygous reference (RR) genotypes. Note that in this region several mice are identical by descent. Bottom panel: represents alignments for mouse 234. Vertical bars represent mapped coverage, colours other than grey represent reference-mismatches in more than 20% of the reads. Horizontal bars represent individual reads, coloured nucleotides indicate reference-mismatches, the differential transparency of the letters reflects differential Phred scores. At the second SNP, despite no evidence for the reference allele in any of the 10 reads covering the site, and the genotype likelihood of the AA genotype $\sim$ 4 times more likely than RA genotype, RA is called by GATK. 83

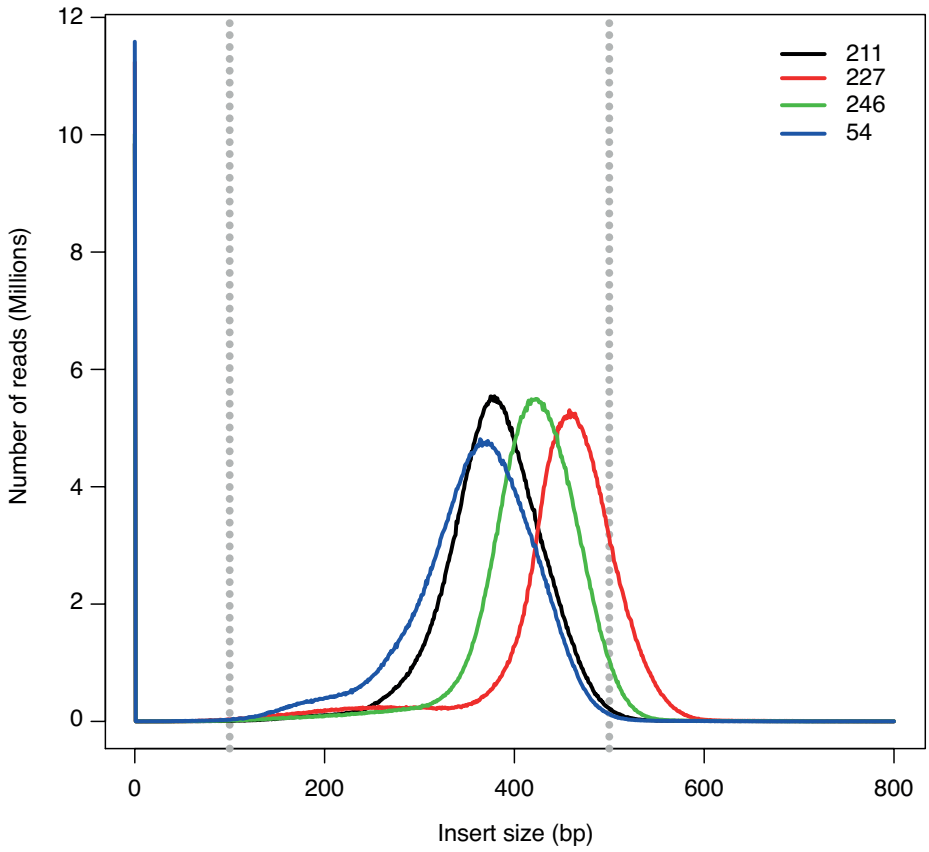


Figure 2.17: Considerable variation in the shotgun-library insert sizes for four of the 20 sequenced mice. Lines represent the insert size distributions for four mice 211 (black), 227 (red), 246 (red), 54 (blue) measured in millions of reads. Dotted lines represent 5th and 95th percentile of the insert size distribution observed across all 20 mice. The spike of reads with insert size zero result from PCR amplification artefacts.

2.4 Discussion

This chapter presents a detailed deconstruction of the anatomy of HTS errors. Investigation of suspicious LD patterns in chimpanzees led to the discovery of cryptic errors in variant calls that were missed by state-of-the-art variant filters. Their symptoms present at different scales. Gross-scale artefactual signals of intense recombination were detected arising from reference errors or incompleteness in regions of structural variation and highly repetitive sequence. Fine-scale errors were detected, arising from local read mis-mapping around regions of low-complexity sequence, polymorphic indels, and repeat element families under-represented in the reference sequence. In the inbred wild mice, heterozygous calls in IBD regions enabled the categorisation of error modes suggesting an error rate of ∼10-15% in the heterozygous calls. Importantly, errors were observed across all reference genomes, from the draft-quality chimpanzee reference (assembled from fragments and limited BAC sequencing), to the finished-quality human genome (constructed almost exclusively from BAC sequencing), though analyses in genomes with poor references are much more problematic. These underlying error modes impact all analyses of variation from HTS. For example errors can impact studies in the following ways:

Association studies Systematic differences between treatment of cases and controls result in false association. In some species it might be that phenotype is correlated with divergence from reference.

The discovery of new mutations False negative errors within samples result in false *de novo* mutations or other changes (for instance in trio-sequencing experiments).

The discovery of complex or rare events Random sequencing errors overwhelm rare variant analyses which are hard to correct through using population-based appraoches.

The analysis of haplotype structure or recombination False-positive or incorrectly mapped variants result in artefactual disruption of haplotype structure leading to erroneous recombination rate estimates.

2.4.1 The anatomy of HTS-errors

Broadly six types of error were detected arising from the inference of genome sequence from HTS data of chimpanzees and mice.

Reference-genome mis-assembly driven false negatives Arise in the construction of reference genomes due to variation in coverage and inability to resolve repeat structures correctly. A related problem is the reference genome being inappropriate for the sample (for example 17p21.31 inversion in humans (Stefansson et al., 2005), or translocation events). A symptom of this error is inconsistency among the mapped reads from the same or subsequent individuals to the assembled genome. This leads to the inability to correctly-assemble variation from new samples.

Coverage-driven false-negatives Arise from fluctuations in mapped-coverage. Loci have symptoms of low coverage, which leads to false *de novo* mutations in trio comparisons arising from the variant site in the parent being missed. Addressed by (1) increasing sequencing depth, (2) imputation particularly in population isolate settings or pedigrees, or (3) incorporating uncertainty.

Complexity-driven false negatives Arise from analysis methods failing to report, or reliably distinguish true variation from false positives in regions of either high sequence diversity or low sequence complexity. Loci have symptoms of high variant rate in reads but not in variant calls.

Read-error driven false positives Arise from technical artefacts and processing or analysis errors (base miscalls, PCR duplicates, indel errors). Symptoms of

(1) sporadic distribution between samples (however, note that the experimenter must be careful not to conflate this with true variation for example generated through geography), (2) low-frequency variants within a sample and population, often with technical covariates such as read cycle. Note that these error modes are likely to change as sequencing technologies evolve.

Local-mapping driven false positives around true variants Arise from correct global mapping, but incorrect or inconsistent local mapping of reads to a reference assembly. Symptoms of clustered variants and inconsistent local mapping often arising in regions of low local sequence complexity (e.g. Figure 2.14). Longer read lengths and more accurate reads will tend to help this.

Reference-genome mis-assembly or incompleteness driven false positives Arise from inaccuracy in, or incompleteness of the reference, for example polymorphic segmental duplications (e.g. Figure 2.15), or under-represented repeats (e.g. Figure 2.8). Symptoms of (1) inconsistent local haplotype structure, (2) sometimes unbalanced variant evidence (for instance all variants arise on one strand) and (3) sometimes high or variable read coverage (e.g. HERV-H pseudo-hotspots in chimpanzees). Longer read lengths and more accurate reads will not tend to help this, better reference genomes are required.

2.4.2 Approaches to guard against error modes

Why do errors occur in HTS at such a higher frequency than SNP-chips? A potential hypothesis for the lower error rate in SNP-chips is that the limitations of oligo design preclude assaying sites in regions with low levels of uniqueness. So why not rely on SNP-chips if they are more accurate and cheaper than sequencing? First, SNP-chips are only available for a limited number of species. Second, SNP-chips provide a limited view of the changes impacting genomes. Indeed, some areas that are particularly

challenging to sequence, but that are not covered well by SNP-chips, are known to be important in biological processes. For example, minisatellites are known to be sites of intense recombination activity (e.g. Myers et al., 2008). Therefore, the challenge of sifting signal from the noise in HTS sequencing is a major barrier to leveraging one of the most important potential advantages of HTS. What can the researcher do to help guard against different error modes?

Incorporating uncertainty. Errors are inevitable, but if it is possible to measure the certainty with which DNA sequences are detected, mapped, re-aligned and assigned to haplotypes, this can be helpful in identifying true signal. Uncertainty is measured using several approaches. (a) Base quality scores measure the probability that the nucleotide is incorrect. Base quality scores can be used to calibrate SNP calling, and in the chimpanzee sequencing project the value of base quality re-calibration was demonstrated. (b) Indel-error scores, measure the probability of an indel error between base calls. (c) Mapping qualities provide an estimate of correct read-pair placement. Mapping qualities can be used in combination with (a) and (b) to estimate the quality of variant calls. (d) Genotype likelihoods measure the probability of the data given the different potential genotypes (considered in Chapter 3). Genotype likelihoods can be used to measure likely false-negative rates and prioritize *de novo* mutations. (e) Genome masks measure the ability to access variants in different parts of the genome.

Assembly and reassembly. Mapping reads to a reference genome is a very powerful and fast way of working out how sequence fragments fit together. But methods that try to assemble variation *de novo* (either from the start or from an initial round of mapping) have the potential to avoid many of the drawbacks of rigidly assuming that mapping (and mapping qualities) is correct. These approaches are now incorporated into the GATK and Platypus.

Combining information from different data sources. There is an abundance

of prior information available about genetic variation (and false positive variants). This should be used when possible, though care must be taken not to bias results or propagate errors. For example, the utility of aligning reads to alternative references was discussed, and the utility of integrating LD information in the refinement of genotype calls has been convincingly demonstrated in the 1000 Genomes (Consortium, 2010).

Data visualisation. No algorithm is perfect (as demonstrated by the GATK's propensity to call false heterozygous sites in IBD regions). Common sense is extremely powerful and there are good approaches to visualise HTS data (IGV (Robinson et al., 2011) alignment views have been used extensively in this chapter) that enable the researcher to assess whether a variant is credible. Likewise using an independent technology to validate calls is very useful, however no technology is free of errors. Importantly, the analyses in humans and chimpanzees show that error rates vary greatly depending on sequence context; most notably polymorphic segmental duplications have higher error rates than unique sequence.

2.4.3 Case study: estimating recombination rates from reference genome free variant calling

To investigate whether variant calls from *de novo* assembly could help to guard against false-positive calls arising from local mis-mapping and reference genome incompleteness, I compared recombination rate estimates on chimpanzee chromosome 21 for (1) the alignment-based HTS haplotypes (henceforth referred to as the *alignment-map*), and (2) the GATK-genotypes for sites also called by Cortex, a *de novo* multiple genome assembler (henceforth referred to as the *intersection-map*). The Cortex program (Iqbal et al., 2012) is capable of assembling multiple genomes simultaneously through an extension of classical de Brujin graphs - the nodes and edges in the graph are coloured by the samples in which they are observed. Variation is detected as

bubbles in the graph where two nodes are connected by more than one edge.

SNPs and indels were called by applying Cortex to sequence data from the 10 chimpanzees (Iqbal et al., 2012). I provide a short description of the methodology used by the authors. A k-mer size of 31 was chosen for the chimpanzee samples because the graph complexity was expected to reduce significantly between 21 bases and 31 bases, and any higher kmer size would risk completely removing singletons. A 12-colour graph was then assembled (including the chimpanzee reference, and a colour for the pooled-population graph), cleaned, and variants called on the pooled-population colour. Variants were then annotated using the individual-specific colours. A total of 3,583,205 variants were identified, of which 2,961,307 were SNPs and 34,667 were complex.

On chromosome 21 32,528 SNPs (0.70 SNP/kb, TsTv=1.97) were identified through Cortex, and 75,581 SNPs (1.63 SNPs/kb, TsTv = 2.23) were identified through the GATK after filtering (described in Section 2.5.1.3). Intersection contained a total of 28,448 SNPs (0.61 SNP/kb, TsTv = 2.20). Recombination rates for the intersection map were then estimated from the GATK-genotypes at intersection set sites. The Cortex-genotypes were not used so as to avoid potential genotype errors at heterozygous variant sites that can arise in low-coverage situations, where heterozygous sites are called as homozygous sites because only one allele is represented in the multicolour graph.

Note that to remove artefactual jumps in genetic distance, the alignment-map required regional filtering (Section 2.2.2.1). Six large jumps were filtered in chromosome 21 alignment-map. In contrast, the intersection-map required no filtering.

To assess the quality of the recombination rate estimates for the intersection-map, I calculated the Pearson correlation between the HapMap combined-population map (Consortium, 2007), and the alignment-map, and the intersection-map. Figure 2.18 shows that comparable correlation is observed for the intersection-map without the

requirement for regional filtering. This suggests that *de novo* assembly variant calling provides a high-quality set of variants and removes the requirement for *ad hoc* filtering of the genetic map.

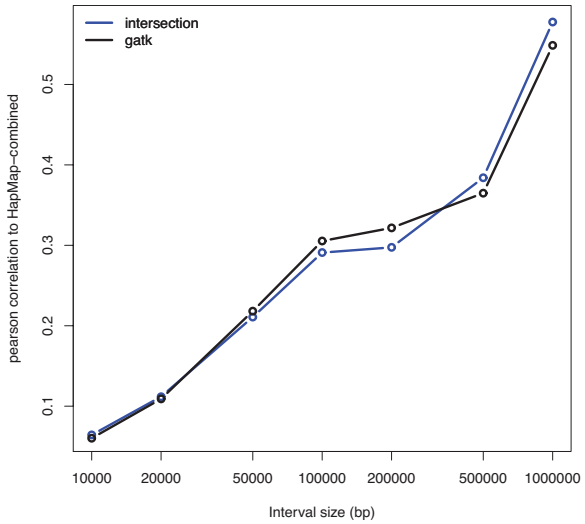


Figure 2.18: Pearson correlation coefficients at different scales between recombination rates estimated from the combined-population HapMap map, and the *filtered* alignment-based map in chimpanzees (black; regional filters zeroed 6 artefactual jumps in the estimated map), and the *un-filtered* map from the genotypes called by the GATK at sites called by Cortex in chimpanzees (blue; no regional filtering required).

What is the future for the analysis of HTS data for the discovery of variation? It is important that we ensure the lasting value of genome sequences collected today in the face of rapidly changing technologies. A pressing need in the field is to develop standards that, at one end, enable research and discovery and, at the other end, provide clinical grade information about genome sequence. These two objectives should not be separate strands, rather there should be a single approach, but with different thresholds on certainty (focus of Chapters 3 and 5). Such an objective would be bolstered by improvements in how known variation and incompleteness of the reference sequence is represented. In this sense it is important to measure confidence in the potential to access variation at each site in the genome, not only at called sites. As demonstrated here, fully integrating *de novo* assembly and mapping

is a powerful approach and, furthermore, is an attractive approach to study more complex variation.

2.5 Methods

This methods section provides aspects of the two case studies that I was not directly responsible for, but are important background for my work.

2.5.1 Chimpanzee background methods

2.5.1.1 Sample preparation and sequencing

DNA from genomes of 10 unrelated Western chimpanzees were obtained from Epstein-Barr Virus immortalized cell lines. Each genome was sheared by sonication, size selected for 500 bp fragments, and constructed into paired-end shotgun libraries with a 450 bp insert size. Each library was sequenced using the Illumina Genome Analyzer II platform (Illumina, CA) and their FASTQ files obtained from raw images using the default Illumina software.

2.5.1.2 Read mapping

Reads were mapped to the panTro2 reference genome using Stampy (version 0.95 Lunter and Goodson, 2010), using the default BWA-based (version 0.5.9 Li and Durbin, 2009) pre-mapping step to improve mapping efficiency. Before mapping, the reference was curated: all assembled chromosome sequences, unresolved contigs, and the EBV reference sequence (NC007605.1) were included, and the pseudoautosomal region 1 (PAR1) on chromosome Yp was masked out to avoid mapping ambiguities between the sex chromosomes. After mapping, read duplicates arising from PCR amplification artefacts were identified and removed using SAMtools (Li et al., 2009) yielding an average 9.45× coverage for the sampled genomes (Table 2.4).

Read base qualities were recalibrated using two procedures. First, read base qualities were recalibrated using the Genome Analysis Toolkit (GATK) (McKenna et al., 2010, version 1.0.3471), on technical covariates for read group, reported base quality, machine cycle, and dinucleotide context. Second, base qualities in low-complexity regions were down weighted using the SAMtools Base Alignment Quality (BAQ) algorithm (Li, 2011).

To capture sequence potentially absent from the chimpanzee reference, reads were aligned to the human build 36 reference (UCSC genome browser Fujita et al., 2011). The human reference was curated to include all chromosomes, unresolved contigs, and the EBV reference sequence (NC007605.1). Variants called from the human-mappings were used to verify the chimpanzee-mappings calls.

Variants were called separately for each alignment and then compared. The procedure followed for variant discovery and genotyping is described next.

2.5.1.3 Variant discovery, genotyping and filtering

Variants were identified using GATK *UnifedGenotyper* using default parameters. Additionally, filters were applied through GATK *VariantFiltration* as follows (SNPs that

Sample ID	Name	Sex	Mapped coverage$^{\gamma}$
PtAC	Annaclara	Female	9.96
PtFR	Frits	Male	9.87
PtGI	Gina	Female	10.33
PtLA	Lady	Female	8.80
PtLI	Liesbeth	Female	9.72
PtPE	Pearl	Female	9.20
PtRG	Regina	Female	11.29
PtRN	Renee	Female	5.57
PtSU	Susie	Female	10.12
PtYO	Yvonne	Female	9.63
Average			9.45

Table 2.4: Western chimpanzee sample properties. DNA material was obtained from Eptein-Barr Virus immortalized cell lines. γ: Mapped was coverage calculated after duplicate removal and only counting non-inverted read pairs mapped to the same strand. A genome size of 3.06Gb was assumed.

exceeded these thresholds were removed):

1. Depth of coverage. The total coverage across all samples was < 30 or > 180.
2. Strand bias. Fisher exact test on the counts of variants observed on each strand.
3. Potential reference errors. All individuals homozygous for a non-reference base.
4. Low complexity regions resulting in potential sequencing errors. The longest continuous homopolymer run of the variant allele was > 2.
5. Read quality. The number of reads with zero mapping quality was $> 40\%$.
6. SNP quality. The SNP quality score divided by the total depth of coverage of reads with the non-reference genotype was < 1.0 or > 27.0.
7. Read quality. The root mean squared mapping quality across all samples at that position was < 25.
8. Missing information. The number of chromosomes with missing genotypes was > 4.
9. Potential Mis-mapping. The site is covered by reads that contain a deletion.
10. Clustering. Number of SNPs within a 10 base window was ≥ 3.
11. Haplotype incongruence. The read bases within 10 bases of the variant are consistent with less than 3 haplotypes.

In addition, a test for HWE was applied using VCFtools (version 0.1.3.2 Danecek et al., 2011). Sites with a p-value less than 0.01 were excluded.

We estimated that the TsTv value in chimpanzees is approximately 2.10 by limiting the ratio calculation to substitutions on the chimpanzee lineage at bases that had a base quality of at least 90 Phred units.

For the chimpanzee-refernce-mapping, 5,323,301 autosome SNPs remained after filtering with an aggregate TsTv of 2.18. Eighty-six percent of the calls were novel compared to dbSNP release 125, and of the dbSNP calls 51.1% were rediscovered. For the human-reference-mapping 5,827,490 autosome SNPs remained after filtering with an aggregate TsTv of 2.35. The intersection of the two call sets contained 4,005,230 SNPs with an aggregate TsTv of 2.25. The variants called in the sex chromosomes were of substantially lower quality than the autosomes [FDR of 17.0% (95% C.I. = 6.87% - 27.09%) on chromosome X, FDR of 2.7% (95% C.I. = 0.08% - 5.2%) on autosomes; Section 2.5.1.6], and therefore were excluded from the recombination rate analyses.

2.5.1.4 Genotype refinement and phasing

Genotype calls were refined using an approach similar to the 1000 Genomes Project (Consortium, 2010). The program BEAGLE (Browning and Browning, 2007) was used to impute genotypes from the genotype likelihoods estimated by GATK. This imputation procedure has two benefits. First, the inferred haplotypes are used to try to impute missing genotypes. A total of 390,797 genotypes (0.73%) were imputed. Second, genotype calls with a high level of uncertainty are potentially refined using local LD information. A total of 594,629 genotypes (1.12%) were modified.

The phase of haplotypes inferred by BEAGLE were refined using the program PHASE (Stephens et al., 2001), which has been demonstrated to have better phasing accuracy for small sample sizes than other approaches (Browning and Browning, 2007). PHASE is computationally intensive and only hundreds of bases can be analysed simultaneously. Calls were split into windows each containing 401 SNPs tiled to overlap at 100 SNP between adjacent windows. PHASE was run for 200 iterations, with a burn-in of 300 iterations. The phased haplotypes were spliced together by choosing the combination of haplotypes with the minimum Hamming distance.

2.5.1.5 Human-chimpanzee syntenic regions

To reduce the potential impact of sequence divergence on comparisons of recombination rates between humans and chimpanzees, regions of synteny between the two species' reference genomes were identified. Pre-computed genome mappings between reference genomes are available for fine-scales (e.g. UCSC genome browser liftOver tool). However, at larger scales the identification of synteny is more challenging. Synteny is disrupted by large structural events that may be ancestral to a single lineage, or that arise through reference sequence mis-assembly. In both cases it is important to identify these regions.

To identify regions of broad-scale synteny, SNPs called against the chimpanzee reference were mapped to the human reference using the program liftOver (Kent et al., 2002). To avoid ambiguities at translocations only the SNPs that mapped to the primary homologous human chromosome were retained. These sites were then mapped back to the chimpanzee reference genome. Only those SNPs that mapped to their original position were retained. Next, groups of SNPs separated by no more

than 50 kb whose relative order was preserved were identified. Groups with numbers of SNPs less than 5000 were iteratively removed until the remaining regions contained more than 5000 SNPs.

This scheme identified 160 blocks of synteny, containing a total of 4,892,640 SNPs (8% loss). The smallest synteny block was 1.5 Mb, the largest 103.8 Mb, and the median size 10.3 Mb (Figure 2.19).

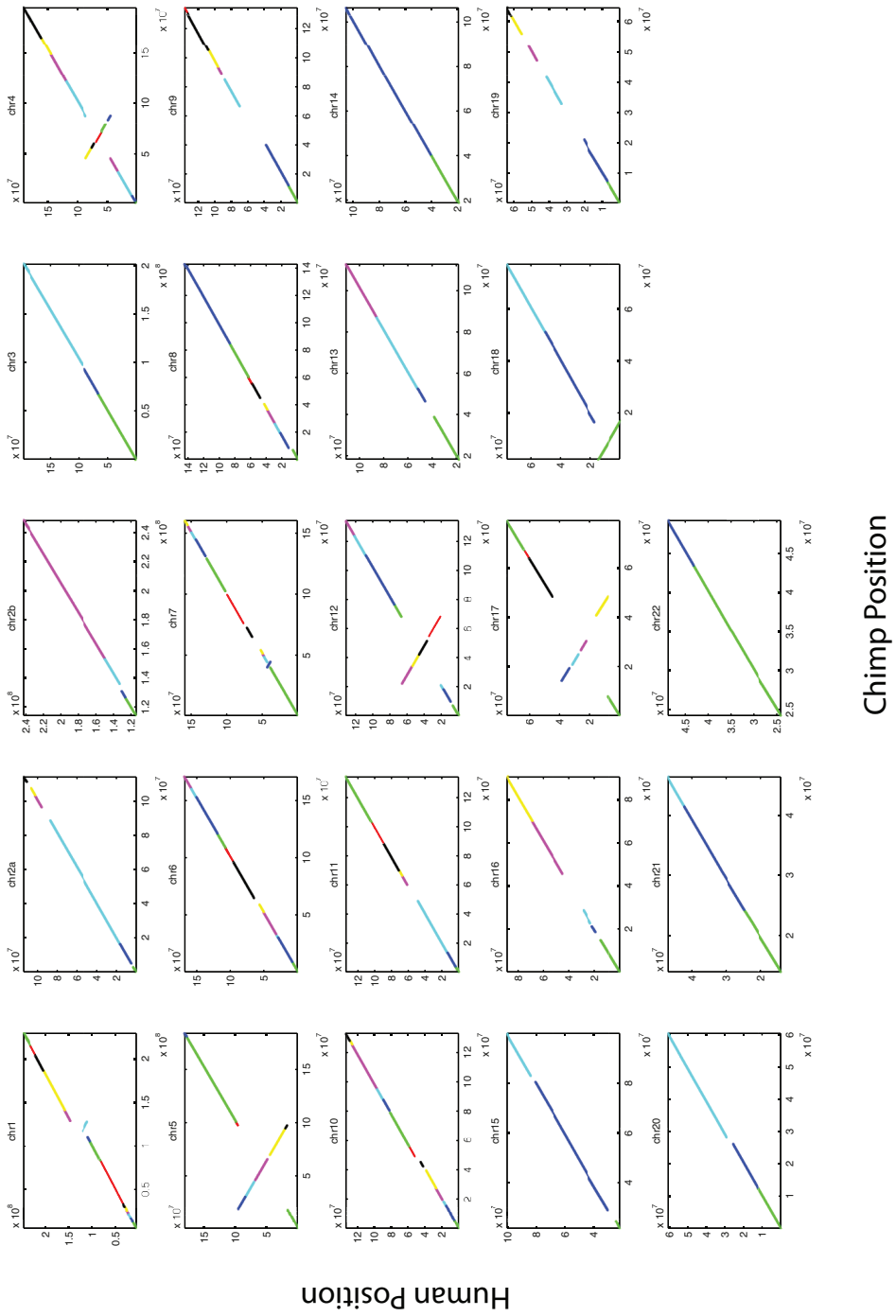


Figure 2.19: Regions of broad-scale synteny between chimpanzee and human reference genomes. SNPs are represented by their panTro2 coordinate (x-axis), and their hg18 coordinate on the primary homologous human chromosome (y-axis). The SNPs on the $y = x$ diagonal indicate synteny between the species. Large-scale inversions on chromosomes 4, 5, 12, 17 and 18 are indicated by the presence of SNPs along the $y = -x$ diagonal. Different syntenic blocks are represented by seven alternating colours.

2.5.1.6 False discovery rates and genotype accuracy

To assess SNP false discovery rate (FDR), a subset of the SNP calls were validated using mass spectrometry on the MassArray Sequenom platform. Variants were sampled uniformly from the initial call set as follows: 300 SNPs were sampled from the autosomes requiring that the SNPs were separated by at least 1 Mb, and 115 SNPs were sampled from chromosome X separated by at least 100 kb. The SNPs were verified to have a similar site frequency spectrum and TsTv to that of the entire call set.

The validation SNPs were then filtered on the following criteria:

1. SNPs were excluded if a missing base was found within 200 bp, or the sequence within 200 bp contained more than five SNPs compared to the union of dbSNP and our discovered SNPs.

2. SNPs were excluded if the Sequenom PCR primer design software failed to identify an appropriate primer.

3. SNPs were excluded if the designed primer mapped to multiple locations in the PanTro2 reference.

Criteria 1 removed four percent of autosome SNPs and 16% of the chromosome X SNPs, criteria 2 removed three percent of the autosome SNPs and four percent of the chromosome X SNPs, and criteria 3 removed 35% of the autosome SNPs and 40% of the chromosome X SNPs. This left 180 SNPs on the autosomes, and 56 SNPs on chromosome X.

The selected validation SNPs were arranged on multiplexes using Sequenom software, and the multiplexes were applied to a panel of 93 chimpanzee genomes that included the 10 HTS chimpanzees in our study. To assess call concordance, three chimpanzees (PtFR, PtPE, and PtRG, Table 2.4) were genotyped twice.

Autosomes. The genotype concordance was 98.5 % for PtFR, 98.2 % for PtPE, and 99.5 % for PtRG. For the 170 SNPs in the initial call set, ten were identified as monomorphic in the panel of 92 chimpanzees, implying a discovery-FDR of 5.9% (95% C.I. = 2.4% - 9.4%). Assay results were then used to calibrate the SNP filtering procedure to control FDR. The final call set overlapped 156 assays (6 failed assays). Seven of the assays were called as monomorphic, implying a FDR of 2.7% (95% C.I. = 0.08% - 5.2%). In addition, the intersection of the human-mapping and chimpanzee-mapping calls overlapped 125 assays (6 failed). Four of the assays were called as monomorphic, implying an FDR of 2.6% (95% C.I. = 0.00% - 5.29%).

Chromosome X. The implied FDR of SNPs on chromosome X was considerably worse than the autosomes. One of the 54 SNPs from the initial call set assays failed, and eight assays were called as monomorphic implying an FDR of 17.0% (95% C.I. = 6.87% - 27.09%). We hypothesise that the high FDR of the chromosome X SNPs is caused by the low quality of the reference assembly relative to the autosomes.

2.5.1.7 SNP discovery power

SNP discovery power was calculated by comparing SNP calls to two studies that had also typed the 10 HTS chimpanzees. First, SNPs were compared to the Myers et al. (2010) dataset consisting of 694 sites genotyped in 36 Western, 20 Central and 17 Vellorosus chimpanzees. The reference coordinates of the Myers et al. (2010) data were obtained from dbSNP. We were unable to obtain the coordinates for two SNPs because they were not catalogued in dbSNP. Second, SNPs were compared to the Winckler et al. (2005) dataset, consisting of 1,081 sites genotyped in 28 Western chimpanzees (Winckler et al., 2005). The reference coordinates of the Winckler et al. (2005) data were converted from hg15 to panTro2 using the UCSC *liftOver* program, with the loss of 12 SNPs. Subsequently each call set was filtered as follows: if the SNP genotypes were fixed, or if more than 50% of the genotypes at a SNP were missing they were excluded, leaving 593 of Myers et al. (2010) sites, and 866 Winckler et al. (2005) sites. The discovery power for variants present at least once in the samples from our study is 86.83%. The estimates are comparable between the two studies (Winckler 85.0% - 89.4%, Myers 87.9%).

2.5.1.8 Assessment of genetic map filters by simulation

We were concerned that the regional filtering approach could remove true signals of recombination.

First we can assess the impact of the filters by reasoning about what we expect to observe and through an understanding of the LDhat algorithm. The properties of human hotspots have been extensively studied through sperm-typing, pedigree analyses, and LD-based analyses. Through choosing a value for N_e in humans we can convert between units of cM measured at human hotspots and 4Ner. We chose the estimated effective population size of CEU which is around $10,000$ (Consortium, 2007) for the unit conversion calculation. In the human genome, the median hotspot cross-over rate is 0.043 cM, or 17.2 4Ner (Consortium, 2007). The largest recombination hotspot observed in the human genome has an estimated cross-over rate of 1.2 cM (Consortium, 2007), or 480 4Ner. If all the recombination events were concentrated at a single site for this hotspot, this hotspot would fail the regional filter.

However, we expect this profile to be rare in LDhat estimated recombination rates because the algorithm inherently penalises situations where all the recombination occurs in a single SNP interval. Hence overwhelming evidence for a complete linkage breakdown would be required to observe such a signal. Therefore, prior experience indicates that our filtering approach is reasonable, assuming chimpanzee recombination is similar to human.

Nevertheless, to ensure that this filter was not removing substantial amounts of true recombination rate variation, a simulation study was conducted by collaborators. To match the chimpanzee sequencing study design data was simulated for 20 chromosomes over a 300 kb region containing a single hotspot of magnitude 1.2 cM at the central position. SNPs were simulated to have an average density of 1.5 SNPs/kb and singletons were excluded. Note that this is slightly less than 2 SNPs/kb density

observed in the chimpanzee data. Errors were then introduced: a genotype error rate of 1%, a switch error of 1%, and 1% of sites were substituted with an erroneous SNP where the genotypes were randomly shuffled across individuals.

LDhat was then applied to the simulated data using the same parameter settings as were applied in the chimpanzee rate estimation. The simulation procedure was repeated 300 times. Three of the simulations (1%) contained a hotspot that would fail our filter, indicating that very few, if any, true recombination hotspots were removed by filtering. Given that I was interested in the structure of fine-scale variation at different genomic features in the chimpanzee genome, a 1% false-negative rate was acceptable.

2.5.1.9 Estimates of effective population size

The LD-based recombination rate estimation method *LDhat* returns estimates of the population recombination rate, ρ, which is related to the per-generation recombination rate via the relation $\rho = 4N_e r$, where r is the per-generation recombination rate and N_e is the effective population size. In order to convert population recombination rate estimates to per-generation rate estimates, it is therefore necessary to obtain an estimate of the effective population size.

In previous human studies, this has been achieved assuming a single effective population size for the genome, via comparison of the LD-based maps to those obtained from pedigree studies (Myers et al., 2005; Consortium, 2007; McVean et al., 2004), which provide broad-scale estimates of the per-generation recombination rate, r (Kong et al., 2010; Consortium, 2007; Kong et al., 2002). The correlation between LD-based and pedigree-based maps has been shown to be extremely high (e.g. Pearson r = 0.977 at 3 Mb scale (Kong et al., 2010)). However, in chimpanzees, there exist no pedigree-based genetic maps. We therefore assumed that the total genetic map lengths of humans and chimpanzees are approximately equal. Errors in this assumption would result in systematic down- or up-scaled recombination rate estimates depending on whether the N_e value was too large or too small. Using the syntenic regions described previously, we estimated the recombination rate in both the human (in per-generation units, obtained from HapMap and chimpanzee (in population-based units) at the 5 Mb scale.

At this scale, we observed the Pearson correlation coefficient between the rate estimates of 0.766. We used a robust linear regression (without intercept) to fit a line to the binned rate estimates from the two species, the gradient of which can be used to obtain an estimate of N_e. The use of robust regression ensures that local deviations in the correlation between human and chimpanzee do not bias the N_e estimate. Using this method, we obtained an estimate of N_e in chimpanzees of 11,413, which is within the range of previous estimates from studies of genetic diversity levels at a limited number of loci, which vary from 9,100 (Caswell et al., 2008) to 13,000 (Yu et al., 2003). This estimate is largely robust to moderate changes in bin size, with N_e = 10,463 at the 1 Mb scale (Pearson correlation coefficient = 0.597), rising to a maximum of N_e = 11,671 at the 10 Mb scale (Pearson correlation coefficient = 0.798). For comparison, we estimated values of N_e for the maps estimated from 10 CEU and 10 YRI. The

Pearson correlation between the binned rate estimates from the 10 genomes and the HapMap 2 estimates at the 5 Mb scale in the syntenic regions are 0.931 for CEU and 0.953 for YRI. The estimated effective population sizes for CEU was 10,040 and 19,064 for YRI, which are consistent with previous estimates from LD data in larger sample sizes (Consortium, 2007).

2.5.1.10 Recombination hotspot inference

The *LDhat* method uses an approximation of the likelihood of the data under a coalescent with recombination model, a composite-likelihood estimate (CL), which treats information from all pairs of SNPs in the data independently (in reality they are not). Though accurate point estimates of recombination are obtained through this approach (McVean et al., 2002), a consequence of the non-independence that is introduced through the CL is that estimates of uncertainty around recombination rate estimates cannot be obtained reliably (McVean et al., 2002), hence it is difficult to determine whether large peaks represent true hotspots or noise in the data. Alternative statistical approaches have been used to identify hotspots (McVean et al., 2004). We used coalescent simulations to assess the significance of recombination rate peaks in the data. The method detects hotspots through a pseudo likelihood ratio test where the composite likelihood is used instead of the full likelihood. The method compares a model where recombination within a test-window is proportional to the background recombination rate (null model), to a model where the recombination rate is allowed to take any value (alternative model). To account for the CL approximation, coalescent simulations are used to determine the null distribution of the test statistic in the absence of a hotspot. The null distribution is then compared to the distribution of the test statistic obtained from the observed data. A hotspot is called if the test statistic obtained from the observed data is in the extreme tail of the distribution.

The genome was interrogated for evidence of hotspots over windows of 200 kb, where the central 2 kb was tested for the presence of a recombination hotspot. After each test the 200 kb window was advanced along the chromosome by 1 kb. The average recombination rate was calculated from the estimated genetic map across the 200 kb window, excluding the 2 kb test region. The null distribution was obtained through simulations of a coalescent with recombination model assuming an infinite sites mutation model along the lines of McVean et al. (2004). The rate was specified as the rate estimated from the observed data with the test 2 kb set to the chromosome average rate. The number of mutations was adjusted to match the number of SNPs in the observed data, and mutations were placed at the SNP locations. Five-thousand simulations were performed for each test region, although simulations were terminated if there was no evidence for significance ($\Pr > 0.1$) after 100 simulations

Once the genome was tested, hotspots were identified through the following scheme.

1. Window had significance $\Pr < 0.01$.

2. Adjacent windows were merged. The minimum p-value and maximum rate estimate was chosen to represent the merged windows.

3. Hotspot candidates were discarded if:
   (a) peak rate $< 5.0\frac{4N_e r}{kb}$
   (b) merged width > 5 kb
   (c) overlapped a phasing boundary
   (d) more than 5% of the reference nucleotides were N (i.e. gaps).

For each hotspot, we attempted to identify a nearby region with no evidence for recombination, we refer to as a coldspot, matched for GC content and SNP density. Coldspots were identified as follows:

1. Maximum CLE equal to zero.
2. Candidate was discarded if it:
   (a) overlapped a phasing boundary,
   (b) more than 5% of the reference nucleotides were N,
3. was within 100 kb of the hotspot,
4. had GC content within 1% of the hotspot content.
5. If more that one candidate remained, the coldspot with closest SNP density to the hotspot was retained.

A total of 6,920 hotspots and 5,038 matched coldspots were identified from the chimpanzee data, 11,910 hotspots with 11,910 coldspots for the 10 CEU, and 15,471 hotspots with 15,471 coldspots for the 10 YRI. Though human results from the human data only detect a third of the known historical recombination hotspots (Consortium, 2007), and therefore imply that we have low power to detect chimpanzee hotspots. The identified hotspots do provide an opportunity to assess the structure of fine scale recombination rate variation around hotspots in humans and chimpanzees. Nevertheless the inferred hotspots do provide enough instances to reliably assess the profile of fine-scale rates around these features (Chapter 4)

2.5.2 Wild mice sequencing

2.5.2.1 Data generation

Collaborators collected a total of 239 mice from a single locality within a 30 km-radius from Ambert, France, allowing at least 2 kms between capture sites. From this collection 20 Mus musculus domesticus mice were sequenced to an average depth of 10× on Illumina Hiseq 2000 with 100 bp paired-end reads and a target insert size of 500 bp (Table 2.5, Figure 2.20). Reads were aligned to the mouse reference genome (mm9, UCSC genome browser database) using Stampy (version 1.0.16).

Table 2.5: Properties of wild mice sampling. Gender determined by PCR.

Identifier	Gender	Sample location
34	M	Le Perrier
44	M	Le Moulin de Collange
54	M	Les Ballays
101	M	Les Serves
211	F	Thiolières
218	F	Marioux
227	F	Chabrol
231	F	Lossedat
234	M	Le Garet
236	F	Bosson
246	M	Sauvanis
303	M	Boisseyre
310	M	Champsiaux
311	M	La Grange
313	M	Germalange
314	F	Chadernolles
327	F	La Thiolière
329	F	Eglisolles
330	F	Sauvessanges
334	F	Saint-Bonnet-le-Chastel

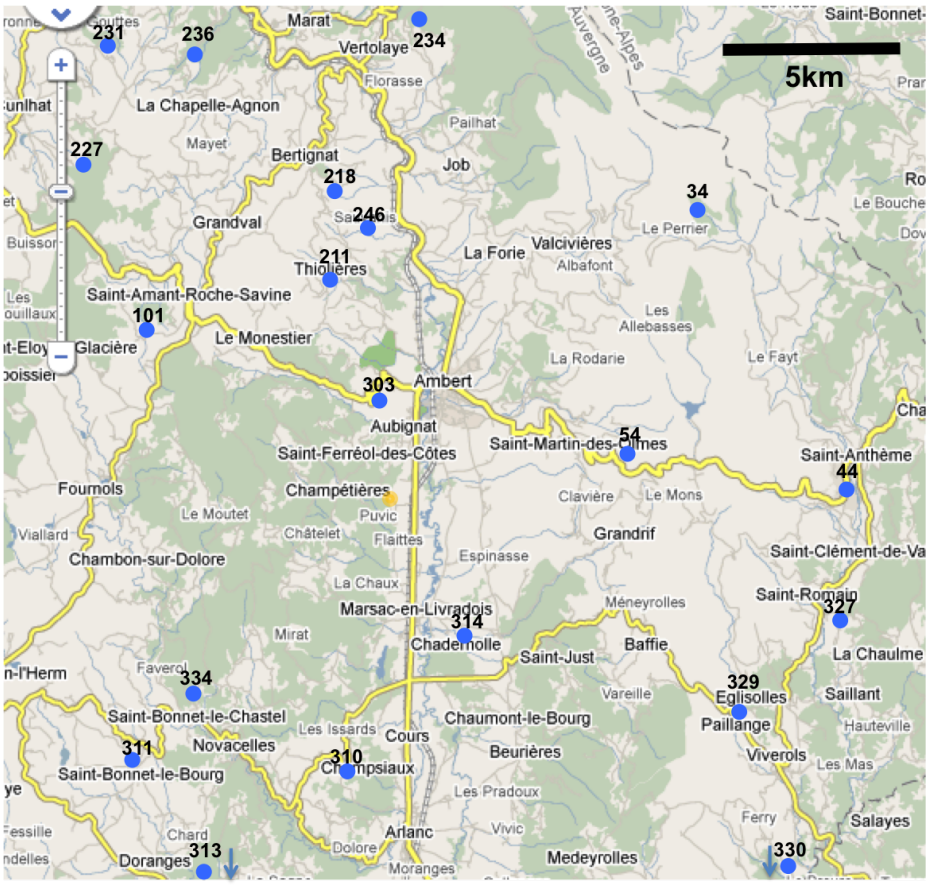


Figure 2.20: Map of the capture sites for the 20 sequenced mice located near the town of Ambert, France (`www.maps.google.fr`). Blue dots and numeric identifiers indicate the capture site for each of the 20 mice. Arrows for mice 313 and 330 indicate that sample locations are located outside of the displayed map.

Variants were called on the autosomes using two approaches (1) the GATK (version 1.5-21, (McKenna et al., 2010)) was applied with parameters *strand call confidence* and *strand emit confidence* set to 30.0 and 10.0 respectively, (2) Platypus (version 0.1.8, `http://www.well.ox.ac.uk/platypus`) was applied using default parameter settings. Methods (1) and (2) were applied to all samples jointly.

The GATK variants were filtered using the following parameters:

1. Depth of coverage. The total coverage across all samples was or < 60 or > 340.
2. Low complexity regions resulting in potential sequencing errors. The longest continuous homopolymer run of the variant allele was > 4.
3. Read quality. The number of reads with zero mapping quality was $> 40\%$.
4. SNP quality. The SNP quality score divided by the total depth of coverage of reads with the non-reference genotype was < 1.0.
5. Read quality. The root mean squared mapping quality calculated across all the sampled chromosomes was < 25.

6. Potential Mis-mapping. The site is covered by reads that contain a deletion.
7. Clustering. Number of SNPs within a 10 bp window was ≥ 3.
8. Haplotype incongruence. The read bases within 10 bp of the variant are consistent with less than 3 haplotypes.

Chapter 3

Methods for estimating LD and testing for recombination in HTS

Many population-scale sequencing studies rely on low-coverage designs (defined here as mapped coverage less than 5× per individual; e.g. Consortium, 2010; Consortium et al., 2012) . Under such circumstances variant calling and genotyping is challenging and there is often considerable statistical uncertainty arising from stochastic variation in coverage and sequencing errors. At low coverage the uncertainty arises in distinguishing between putative homozygous and true heterozygous sites, and missing sites with no coverage. One approach to reduce the uncertainty arising from low coverage is to sequence target regions to high depth (more than 20× coverage). However the desire to sequence large collections of samples, both to catalogue the genetic variation in populations (e.g. Liti et al., 2009) and to increase mapping power in association studies (Kim et al., 2010), has motivated a sequencing design where many samples are sequenced to medium or low coverage. Projects such as the 1000 Genomes (Consortium, 2010) pioneered low-coverage whole-genome population sequencing designs to catalogue variation in humans. In this chapter I explore methods to incorporate genotype uncertainty into estimating measures of LD ($|D'|$ and r^2)

and a pair-wise test for recombination through a likelihood ratio test (LRT). I assess the performance of three estimation approaches applied to HTS (i) the marginalised maximum a posteriori probability (MAP) estimate genotypes at each site (hard calls, hGT), (ii) haplotypes inferred through best practice, i.e. haplotypes estimated from LD-refined genotype likelihoods [1] (LD-refined haplotypes, rHT), and (iii) genotype likelihoods (uncertainty aware, $uaGT$). The uncertainty aware approach is evaluated against the hard calls and refined haplotypes by comparing their estimated LD and LRT profiles against the profile obtained from genotypes measured by SNP-typing in the same 10 Western chimpanzees. Inherently the SNP-typing data has a non-zero genotyping error rate, and consequently the true genotypes are not known with certainty, however an empirical-based evaluation was chosen over simulation for three reasons. First, an empirical approach enables performance assessment under *bona fide* conditions. Second, the SNP-typing was estimated to be more accurate than the HTS genotyping for two reasons (a) 38 Western chimpanzees samples were assayed in SNP-typing enabling improved genotyping [I use only the 10 individuals sequenced in the HTS], (b) the discordance error rate in the genotyping calls is estimated to be an order of magnitude lower than the HTS. Third, since we evaluate the three methods relative to the SNP-typing, errors in the SNP-typing should not affect the comparisons. I conclude the chapter by discussing challenges and potential approaches to improve the inference of recombination from LD.

3.1 Inferring haplotype frequencies through Expectation Maximisation for LD estimation

In this section I apply the Expectation Maximisation (EM) algorithm (Dempster et al., 1977) to estimate 2-locus haplotype frequencies under the assumption of Hardy-

[1] as described in Section 2.5.1.4

Weinberg equilibrium. The estimated frequencies are then used to calculate r^2 and $|D'|$ statistics. The application of EM to estimate haplotype frequencies is widely used. Hill (1974) suggested using the equivalent of an EM with 2-locus 2-allele systems, which was generalised by Excoffier and Slatkin (1995). I extend the EM algorithm to incorporate HTS genotype uncertainty into the haplotype frequency estimation using genotype likelihoods (GLs, see Section 3.1.3.1 for definition).

3.1.1 Deriving the likelihood of haplotype frequencies

In outcrossing organisms, apart from regions like the male X chromosome, or through custom molecular techniques, the haplotypic phase of variants identified from HTS are not known *a priori*. Here I will use *observed diplotype* to denote a pair of diploid SNP genotypes. I denote the potential observed diplotypes as G, which corresponds to the nine unphased pairs $\{00, 01, 02, 10, 11, 12, 20, 21, 22\}$ where 0 denotes two copies of the common genotype, 1 the heterozygous genotype combination, and 2 two copies of the rare genotype. For 2-locus 2-allele systems there are four haplotypes $\{00, 01, 10, 11\}$. I will refer to a *phased diplotype* as a particular combination of two 2-locus haplotypes. The EM algorithm has been derived using a multinomial probability distribution on the genotype counts. I briefly describe its derivation, then I show how GLs can be incorporated into the algorithm.

Given a sample of n sequences, the probability of the data, D, conditional on population diplotype frequencies $\mathbf{p} = (p_1, p_2, ..., p_9)$ is calculated through counting the number of sequences that correspond to each diplotype, and hence is described by the multinomial probability distribution

$$\ell = M(\mathbf{n}) \left(\mathbf{n} \cdot \log(\mathbf{p})\right) \tag{3.1}$$

where the n diplotypes are observed with counts $\mathbf{n} = (n_1, n_2, \ldots, n_9)$ and $M(\mathbf{n})$

represents the log of the multinomial coefficient. The maximum likelihood solution is the values of diplotype frequencies that maximise the likelihood, and the expected value of the log-likelihood, is the product of the expected counts multiplied by the diplotype frequencies.

$$\mathrm{E}[\ell] = \mathrm{E}[\mathbf{n}] \cdot \log(\mathbf{p}). \tag{3.2}$$

Under the assumption of Hardy-Weinberg equilibrium, the frequency p_i of the observed diplotype combination i is given by the sum of the population frequencies of the haplotypes, $\{q_1, q_2, q_3, q_4\}$, of each of the compatible haplotype pairs. I define $P(j,k)$ as the multiplied population frequencies of haplotypes j and k,

$$P(j,k) = \begin{cases} 2q_j q_k & \text{for heterozygous diplotypes} \\ q_j q_k & \text{for homozygous diplotyes,} \end{cases} \tag{3.3}$$

and f as the function that maps a diplotype to the set of compatible ordered haplotypes. For example, applying f to the diplotype $\{01\}$ would index the tuple $(00, 01)$ and this would represent the set of all compatible haplotypes $\{(01, 00), (00, 01)\}$. The frequency for diplotype i is calculated by summing over the compatible tuples

$$p_i = \sum_{(j,k) \in f(i)} P(j,k), \tag{3.4}$$

and the unknown values are the assignment of the double heterozygote counts.

Substituting Equation (3.4) into (3.1), the likelihood can be written as a function of the haplotype frequencies

$$\ell = M(n_i) \sum_i n_i \log \left(\sum_{(j,k) \in f(i)} P(j,k) \right). \tag{3.5}$$

Under this framework the values of the observed diplotype frequencies can be

optimised to maximise the likelihood (and concomitantly the expected value of the likelihood) by optimising the unknown allocation of double heterozygotes across the diplotypes (Section 3.1.3).

3.1.2 Estimating haplotype frequencies through Expectation Maximisation applied to diplotypes

The EM algorithm is an iterative method used to maximise a likelihood. I will use EM to compute successive sets of haplotype frequencies, q_m, starting with initial values $q_m^{(0)}$. The initial values are used as if they were the true frequencies to calculate the likelihood of the expected diplotype frequencies (expectation step). In turn, the haplotype frequencies are maximised conditional on the observed diplotypes $q_m^{(1)}$ (maximisation step). This new value then serves as the initial value for the next iteration until convergence is reached.

In the implementation of the EM algorithm several sets of initial values for haplotype frequencies can be envisioned. We estimated the haplotype frequencies from the counts of the diplotypes for each sample, and allocated the double heterozygous diplotypes uniformly across compatible haplotypes.

The diplotype frequencies for the $(i+1)$th expectation step are calculated from the products of the haplotype frequencies through Equation (3.4).

In the maximisation step, a new estimate of the haplotype frequencies are obtained from the expected diplotype frequencies, where double heterozygotes (the unknown value in the maximisation) are allocated according to the previous population haplotype frequency estimates through the scaling $B = \frac{q_1^{(i)} q_4^{(i)}}{q_1^{(i)} q_4^{(i)} + q_2^{(i)} q_3^{(i)}}$,

$$
\begin{aligned}
q_1^{(i+1)} &= \frac{1}{A}(2p_1^{(i)} + p_2^{(i)} + p_4^{(i)} + Bp_5^{(i)}) \\
q_2^{(i+1)} &= \frac{1}{A}(2p_3^{(i)} + p_2^{(i)} + p_6^{(i)} + (1-B)p_5^{(i)}) \\
q_3^{(i+1)} &= \frac{1}{A}(2p_7^{(i)} + p_4^{(i)} + p_8^{(i)} + (1-B)p_5^{(i)}) \\
q_4^{(i+1)} &= \frac{1}{A}(2p_9^{(i)} + p_6^{(i)} + p_8^{(i)} + Bp_5^{(i)}),
\end{aligned}
\tag{3.6}
$$

where A is a normalisation factor. This process is repeated until the incremental improvement in likelihood falls below a set threshold (I used $\Delta L < 10^{-3}$), or a maximum number of iterations are exceed (I used $N_{\text{iter}} > 100$). If the algorithm converges, $|D'|$ and r^2, are calculated from the final haplotype frequencies (defined in Section 1.2.3). As with any convergence scheme based on non-global optimisation, potential problems are that the algorithm does not converge, or converges to the wrong value. However, Weir (1997) notes that the violation of Hardy-Weinberg at either site is a more frequent source of errors. Since the chimpanzee variant calls were filtered for HWE, it is unlikely that departure from HWE is a concern in this analysis. Furthermore, testing LD estimates under multiple random starting points indicated that the starting procedure and convergence were robust and failures in convergence were not observed across the analyses.

3.1.3 Incorporating genotype uncertainty

To incorporate genotype uncertainty into the EM scheme, I re-derived the likelihood calculation in the EM scheme to operate on genotype likelihoods. In this section I will define genotype likelihoods, and then describe our method to incorporate genotype likelihoods into haplotype frequency estimation through EM.

3.1.3.1 Measuring genotype uncertainty using genotype likelihoods

Most HTS genotyping algorithms use a probabilistic framework to quantify the uncertainty associated variant calling and genotyping in HTS. This is achieved through calculating the GLs at variant sites. First variant sites and their alleles are identified in the mapped reads. Next the GLs are inferred from the alignments under specified allele and error models, and can be calculated marginally at sites, or over haplotypes. In this chapter only 2-allele sites will be considered. GLs will be reported per genotype per site per individual and represent the likelihood of observing the aligned reads, $\Pr(D|G)$, for each of three biallelic genotypes $\{\{0/0\},\{0/1\},\{1/1\}\}$, where 1 represents the alternative allele and 0 the reference allele.

GLs are reported on a $-10\log_{10}$ scale relative to the genotype that maximises the likelihood, denoted G^*_{igk}, where i indexes the site, g the genotype and k the individual. The GL γ_{igk} is written,

$$\gamma_{igk} = -10 \times \log_{10}\left(\frac{\Pr(D|G_{igk})}{\Pr(D|G^*_{igk})}\right), \tag{3.7}$$

If $\gamma_{i\bullet k}$ took values $\{0, 10, 20\}$, the probability of the data under a homozygous reference genotype model $\{0/0\}$ is 10 times more likely than $\{0/1\}$ and 100 times more likely than $\{1/1\}$.

Under a specified prior probability on the expected genotypes, the genotype with the highest posterior probability is generally chosen as the estimated genotype for the individual, and the posterior probability is used as a measure of confidence. The advantages of a well calibrated probabilistic approach are that it provides a measure of the statistical uncertainty underlying the genotype calls, leads to higher accuracy of genotype calls (Nielsen et al., 2011), and provides a natural framework to propagate measured uncertainty through subsequent analyses.

Here HTS genotypes are reported under three approaches. First, the genotypes

that maximise the posterior probability at the site, or that maximise the genotype likelihood, are used (hard calls). Second, LD is used to refine the genotype likelihoods across samples. There are several methods to refine genotype likelihoods using LD (e.g., Browning and Yu, 2009; Howie et al., 2009; Li et al., 2010), but the focus of this chapter not on evaluating imputation methods. Current best-practice (e.g. Consortium, 2010; Consortium et al., 2012) is to infer haplotypes from genotypes called from GLs imputed and refined through statistical imputation (LD refined haplotype calls). Third, the raw genotype likelihoods can be propagated through analyses without hard calling (uncertainty aware calls). To date uncertainty propagation approaches have not been fully investigated or widely used, despite the potential benefit for inferences that are sensitive to genotyping error such as analyses of haplotype structure. In this chapter I will evaluate the uncertainty aware approach versus the other two approaches.

3.1.3.2 Likelihood calculation with genotype uncertainty

In the setting where genotypes are known, the EM algorithm can be used to calculate the maximum likelihood estimates of two-locus haplotype frequencies (Section 3.1.1). In the setting where we only have genotype likelihoods (i.e. there is uncertainty in the sample's genotype), we wish to ascertain whether a similar EM approach can be used to infer haplotype frequencies.

Under the EM algorithm there are two steps. In the first step, we need to calculate the expectation of the log-likelihood of the data under the current parameter values (here the haplotype frequencies). In the second step, we have to find the updated parameter values that maximise this expected likelihood.

Given a current set of haplotype frequencies, it is simple, using Bayes rule, to calculate the posterior distribution of the underlying haplotypes given a sample's genotype likelihoods. Informally, we could imagine sampling from this distribution

for each sample. The log-likelihood for the haplotype frequencies under such a sample is given by the multinomial distribution. Critically, the observed number of each haplotypes appears as a linear term (multiplied by the log of the frequencies of the haplotype). Hence, calculating the contribution from each sample to the expected log likelihood is equivalent to obtaining expected posterior haplotype counts for each sample (i.e. the average over an infinite number of draws from the posterior distribution). It follows that the haplotype frequencies can be updated in a simple fashion as the posterior mean frequency, which follows from the multinomial likelihood. This process is iterated until the process is assessed to have converged (change in expected log likelihood of less than 1×10^{-6}).

3.2 Inferring evidence for recombination activity through pair-wise likelihood ratio tests

To test for evidence of recombination activity, or inactivity between sites, I implemented a likelihood ratio test based on the likelihoods calculated as the importance weights of simulated genealogies under the method of Fearnhead and Donnelly (2001). First I describe how the pair-wise likelihood ratio test is computed for hard genotype calls, through querying LDhat likelihood lookup tables, and then I explain how the approach can be modified to incorporate genotype uncertainty.

3.2.1 Pair-wise likelihood ratio test on hard calls

The likelihood ratio test consists of three stages. We estimate ρ under the LDhat model of sequence evolution, which is that all sites in the sequence conform to a two-allele model with reversible, symmetric mutation, such that the rate of mutation per site per generation is μ and is constant across sites. The first stage is to assign every pair of segregating sites with two alleles into equivalent sets of unordered alleles.

For, example under the assumed model of mutation, if a pair of sites segregating in 4 individuals had the ordered data {CC, GC, CG, GG} and another site {AA, TA, TT, AT} these are equivalent to the unordered sequence {00, 10, 11, 01}, where the number 1 represents the rare allele at each site and 0 the common allele. The number of equivalent sets, and hence the execution time of the calculation, is affected by the number of samples and their levels of diversity. In the limit, where every possible combination of segregating sites occurs in a sample of n sequences, the number of equivalent sets scale as $\mathcal{O}(n^3)$ (McVean et al., 2002).

The second stage is to estimate the likelihood of each equivalent set under the estimated value of population mutation rate parameter θ and a range of recombination rates (typically $0 \leq \rho \leq 100$) using the importance sampling method of Fearnhead and Donnelly (2001) under a symmetric, reversible mutation model. The likelihoods are calculated as the average of the importance weights of the simulated genealogies. These likelihoods are pre-calculated and stored to improve computational efficiency, reducing subsequent calculation of the likelihood to a repeated table look-up and summation operation of log-scaled likelihoods.

In the final stage, the pair-wise likelihood ratio is computed as the likelihood ratio of the value of ρ that maximises the likelihood to a specified null model. The elements of the likelihood are indexed to test for evidence of recombination activity (null model: $\rho = 0$, alternative model: $\rho > 0$), or, conversely, evidence for low recombination activity (null model: $\rho = 100$, alternative model $\rho < 100$). When displaying results I will plot the minus one times log-likelihood difference, with the values for the null model $\rho = 0$ above the $y = x$ diagonal, and the values for the null model $\rho = 100$ below the diagonal. Values close to 0 are plotted in blue, large positive values are plotted in red.

3.2.2 Constructing compatible sets of genotypes from genotype likelihoods

To accommodate uncertainty a strategy is required to integrate over compatible diplotypes. Here I will describe an approach to incorporate genotype uncertainty through the construction of a set of equivalent sets compatible with the reported GLs (the *compatible set*), where each included equivalent set is assigned a relative weight within the compatible set. A challenge in constructing the compatible set is that exhaustive construction of all the potential DL configurations is computationally expensive; for a sample of n chromosomes the number of potential diplotype configurations scales as $\mathcal{O}(9^n)$ per site. Hence a greedy algorithm was implemented to obtain those diplotypes, and their relative likelihoods (henceforth referred to as weights), compatible with the data up to a specified level of tolerance, which I will refer to as α. This approach has the desirable property that the size of the compatible set scales with the level of uncertainty in the called genotypes. The resulting equivalent sets and their weightings can then be used to generate a combined likelihood through indexing, summation and scaling operations on the pre-computed table of LDhat log-likelihoods.

Figure 3.1 provides an example of the compatible set approach. The scheme is initiated with the first individual, the nine diplotypes are calculated through the Kronecker product of the GLs at the two sites (Equation (**??**)), and only those diplotypes with weight greater than α are retained (remember weights are relative likelihoods). In Figure 3.1a, for the first individual only genotypes {00} and {01} have weights greater than α in the first individual. The next individual is introduced, and the two surviving diplotypes from the first individual are augmented with the potential DL combinations from the second individual. In Figure 3.1b, 18 configurations are constructed from the combination of the two filtered diplotypes with the potential diplotypes of the second individual. Duplicated configurations are identified, their weights combined, and normalised. The resulting set of configurations is pruned us-

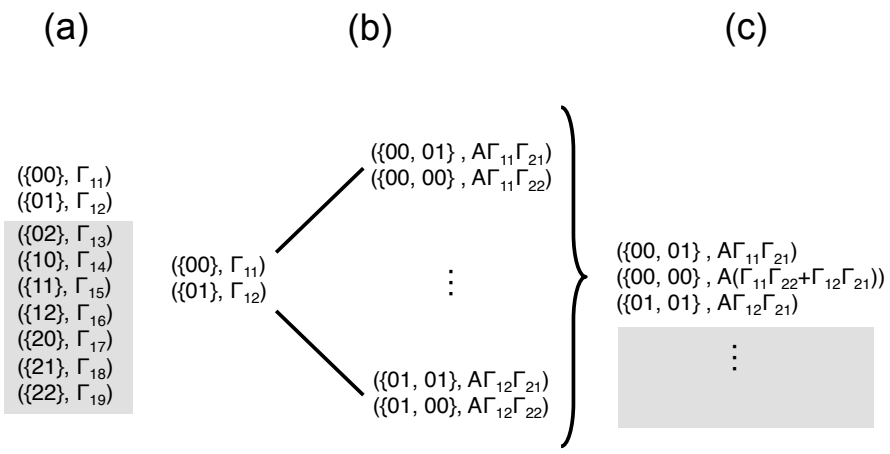


Figure 3.1: Exemplar of method to incorporate uncertainty in to equivalent set construction. (a) Initiation: the nine 2-locus genotypes of the first individual and the associated diplotype likelihood (DL) values $\{\Gamma_{11}, \ldots, \Gamma_{19}\}$, the grey box represents DLs that fall below the minimum relative likelihood threshold (α in text). (b) Augmentation: the second individual is included, the relative likelihood of each new configuration calculated, and rescaled using $A = \Gamma_{11}(\Gamma_{21}, \Gamma_{22}, \ldots, \Gamma_{29}) + \Gamma_{12}(\Gamma_{21}, \Gamma_{22}, \ldots, \Gamma_{29})$. (c) Reduction: The set of compatible configurations is collapsed to remove redundancies, as in (a) the grey box represents DLs that fall below α.

ing threshold α. In Figure 3.1c, only diplotype sets $\{00, 01\}$, $\{00, 00\}$ and $\{01, 01\}$ have relative weight greater than α. After the contributions from the last individual have been introduced, the likelihoods for each diplotype configuration in the compatible set are then corrected for the size of the permutation sets by dividing each weight by the multinomial coefficient calculated from the configuration's diplotype counts. The corrected likelihoods are then filtered against α to generate a set of plausible diplotype configurations that is compatible with the data. This method will be referred to as *compatible set approach*. A downside of this algorithm is that it is influenced by the order in which samples are processed. I shall address this concern through a simulation study in Section 3.2.2.2, but first I assess the performance of the approach.

3.2.2.1 Simulation study 1: Assessment of the compatible set approach for different HTS paradigms

To investigate the performance of the compatible set approach I performed a simulation study to assess how many times the true genotype was contained in the set of weight-ordered genotypes that contribute 95% of the compatible set likelihood.

GLs were simulated under five simplifying assumptions:

1. The sample comes from a large population, and hence allele counts can be described by the multinomial distribution.

2. All the sites are segregating.

3. The base calling has a fixed error rate, ϵ, that is symmetric across the four nucleotides.

4. The base quality scores are well calibrated and all nucleotides have the same quality $Q = -10\log_{10}(\frac{\epsilon}{1-\epsilon})$.

5. The mapped coverage follows a Poisson distribution.

The GLs for a site are simulated for n sampled chromosomes at coverage x, and population mutation rate θ through four stages.

First, allele counts, denoted by c, are simulated by sampling from the multinomial distribution, where the allele proportions are obtained from an allele frequency distribution defined by

$$f_i = A\left(\frac{\theta}{i}\right) \quad i \in [1, 2\text{n-}1],$$

where θ is the mutation rate parameter, and A represents a normalisation factor such that the frequencies sum to one. Two of the four possible nucleotides are then uniformly sampled to represent the alleles at the site, and are assigned counts c

and $n - c$. Each variant is uniformly shuffled into the n chromosomes, producing n genotypes with unknown haplotypic phase. The allele count for each of the four nucleotides in each individual is calculated and stored in the $n \times 4$ matrix denoted by C.

Second, to simulate reads with error, errors are introduced at rate ϵ to generate a new matrix of counts, C^{error}, given by,

$$C^{error} = \frac{x}{2}\left((1-\epsilon)C + \frac{\epsilon}{4}I\right),$$

where I is the identity matrix.

Third, reads are sampled from a Poisson distribution with rate specified by the appropriate element of C^{error} to produce observed counts C^{obs}.

Fourth, the probability of the data given the genotypes is calculated,

$$\ell = \log_{10}(C^{obs}\, p_G),$$

where p_G is the matrix of probabilities of observing a nucleotide conditional on the 10 biallelic genotypes,

$$p_G = (1-\epsilon) \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccccccccc} AA & AC & AG & AT & CC & CG & CT & GG & GT & TT \end{array} \\ \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 1 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 1 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 1 \end{pmatrix} \end{array} + \frac{\epsilon}{4}\mathbf{1}.$$

The $n \times 9$ GL matrix, γ, is obtained by,

$$\gamma_{\bullet gk} = -10\big(\ell_{\bullet gk} - \max_{g}(\ell_{\bullet gk})\big).$$

To assess the behaviour of this algorithm over potential experimental design parameters, I simulated 100 GLs for two sites over combinations of following parameter ranges:

1. mapped coverage $x \in \{1, 3, 5, 10, 15, 20, 25, 30, 35\}$,
2. sample size $n \in \{5, 10, 20, 30, 40, 50\}$,
3. sequencing error rate $\epsilon \in \{0.01, 0.05, 0.1, 0.15, 0.2\}$.

I then applied the compatible set approach to the simulated GLs to identify the compatible 2-locus genotypes using a threshold $\alpha = 0.05$, and recorded three summary statistics:

1. The number of samples with at least one covering read at both sites, to assess the impact of fluctuations in coverage.

2. The number of equivalent sets, ordered by their weights, that account for 95% of the combined likelihood in the compatible set ($s95$), to assess computational efficiency.

3. The fraction of the probability that was assigned to the true genotype in the compatible set (w), to assess precision.

Results were summarised by calculating the mean value of the three summary statistics across simulations and are presented in Figures 3.2 and 3.3. The method shows considerable impact in improving the accuracy of the diplotype estimates compared to marginal maximum likelihood estimates. Below 10×, given the variability in mapped coverage, even at low sequencing error substantial genotype uncertainty

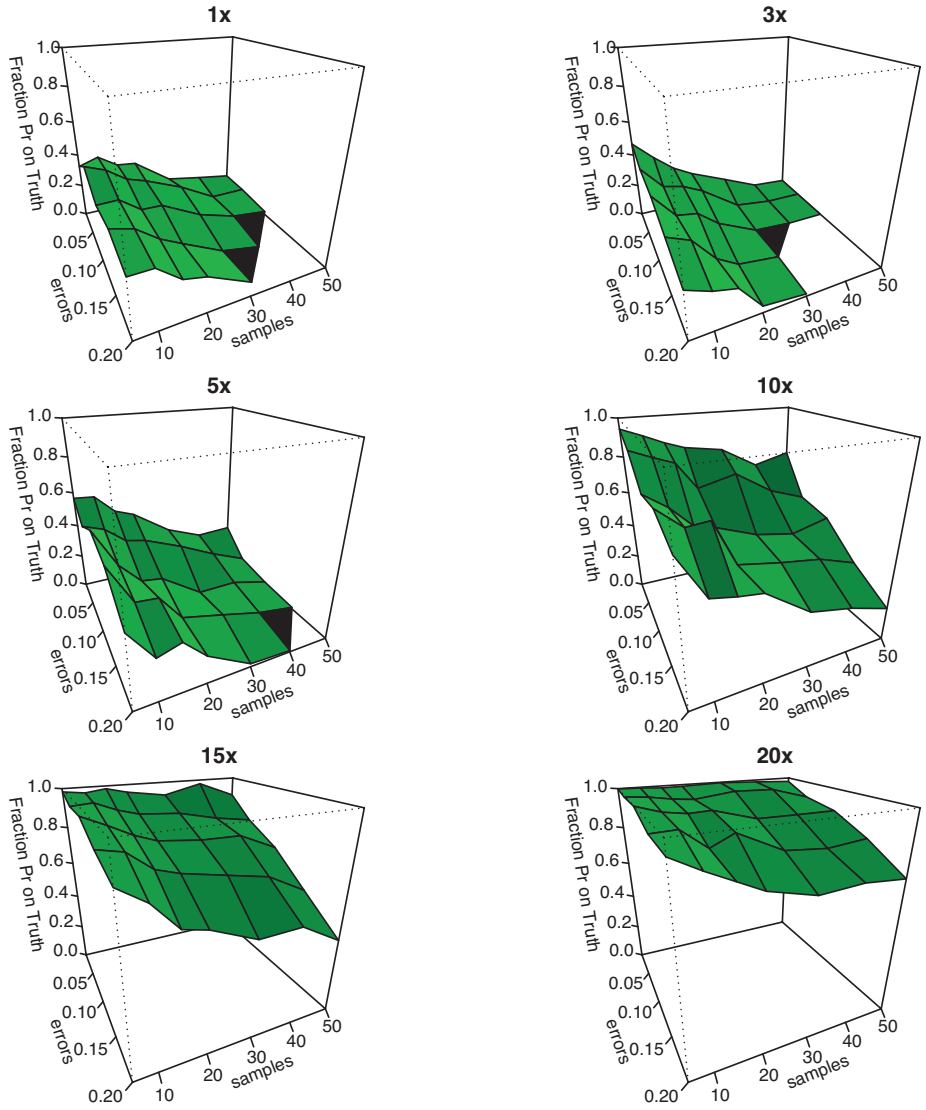


Figure 3.2: Precision of the compatible set approach. Surfaces represent the mean weighting assigned to the true 2-locus genotype within the set of compatible genotypes (z-axis) for varying levels coverage (the six surfaces), sequencing error rate (x-axis), and number of sequenced samples (y-axis). The average value was calculated across 100 simulated GL pairs using a threshold of $\alpha = 0.05$ on the configurations.

is observed. For the simulation of 10 samples sequenced to 3× with an error rate of 0.05 a mean value of 21 compatible genotypes were required, only 0.27 of the weight was on the correct genotype, and an average of one sample had coverage at both sites. In contrast, above 10× there is little uncertainty in the genotypes; only three configurations are required to constitute 95% of the weight, and 83% percent

of the weight is assigned to the correct genotype. Increasing the simulated coverage to 20× for the same number of individuals at the same sequencing error rate of 0.05 results in the true genotype being included in the compatible set 98% of the time with an average weight of 95%, in contrast the maximum likelihood 2-locus GL estimate corresponds to the true genotype only 77% of the time. Even with an error rate of 0.15 the true genotype is identified 95% of the time with an average weight of 87% (the maximum likelihood matched the correct genotype 24% of the time). This suggests that sequencing samples at 20× would provide near perfect variant calls under the simulation assumptions, including the inherent assumption that the reference genome sequence is an error-free and faithful representation of the sampled genomes (I will comment on the appropriateness of this assumption at the end of the chapter). The simulations suggest that 10× is the point at which improvements in reducing genotype uncertainty is first gained. Hence if an experimenter sought to maximise the number of samples sequenced, a design of 10-15× coverage per sample combined with an uncertainty aware analysis is attractive. Though the impact of incorporating uncertainty into analyses is reduced above 20×, the compatible set approach has little overhead computational cost since the set of configurations collapses to a single genotype at each iteration, thus increasing the computation time by $\sim$ 9-fold.

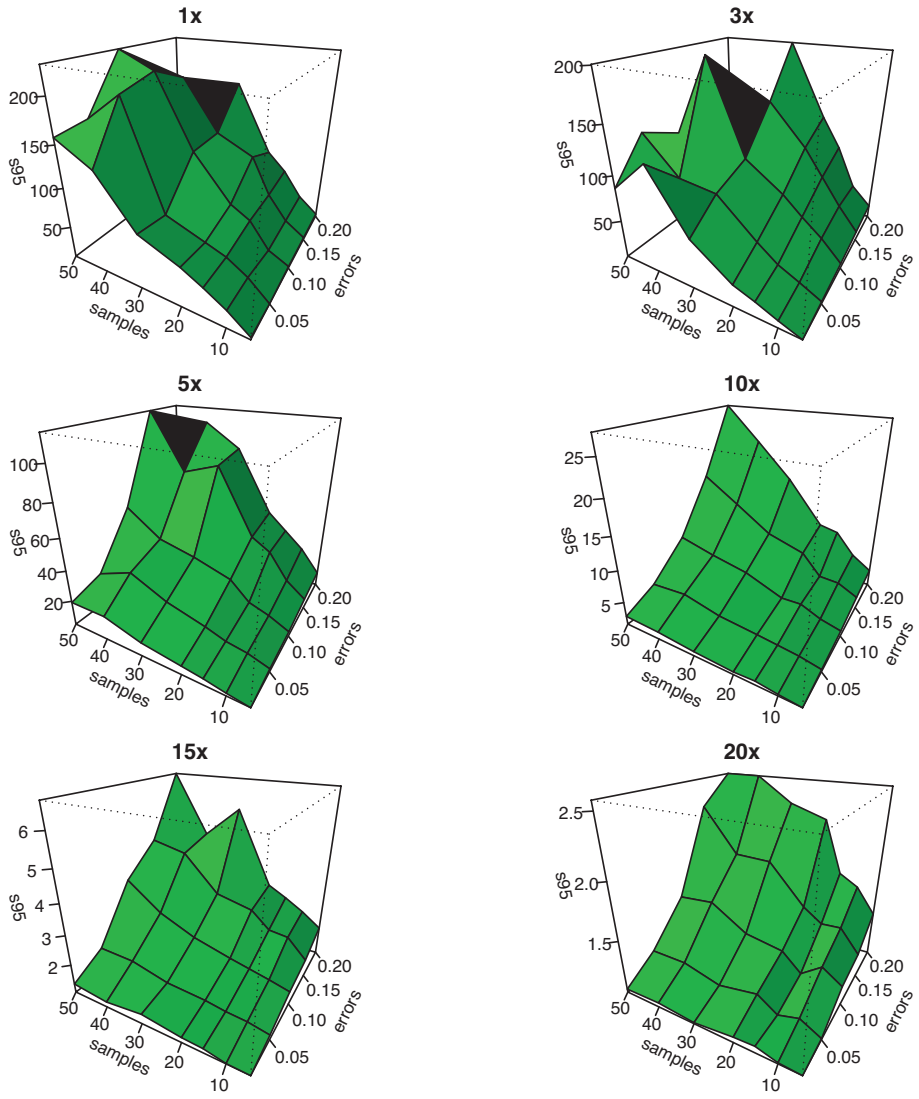


Figure 3.3: Computational cost of the compatible set approach. Surfaces represent the mean number of configurations that account for 95% of the weight in the compatible set (z-axis) for varying levels of coverage (six panels), sequencing error rate (y-axis), and number of sequenced samples (x-axis). The average size of the sets were calculated across 100 simulated GL pairs using a threshold of $\alpha = 0.05$ on the configurations.

3.2.2.2 Simulation study 2: Impact of order dependency in compatible set construction

To assess how the compatible set approach is affected by order dependency I uniformly shuffled the sample order for each of 100 simulated GL pairs and recorded standard deviation in the relative likelihood assigned to the true simulated genotype (w) and the size of the compatible set that contained 95% of the probability ($s95$) across the permutations. Twenty-seven experimental paradigms were explored:

- coverage, $c \in \{5\times, 10\times, 15\times\}$,
- sample, $n \in \{5, 10, 15\}$,
- errors, $\epsilon \in \{0.05, 0.1, 0.15\}$.

For simulations with $n > 5$, the GL ordering was shuffled 2500 times and for simulations with $n = 5$, ordering was shuffled 120 times. The standard deviation in the relative likelihood assigned to the true genotype was then calculated across simulations. For $n = 5$, at low to moderate coverage ($c < 10\times$) the standard deviation in w across simulations is less than 10%, whereas at 15× the standard deviation of w is less than 5% (Table 3.1). The standard deviation in $s95$ decreases from 10% at low to moderate coverage, to less than 3% at 15× suggesting that the compatible set reduction procedure is robust to GL ordering. Inspection of simulations that had high standard deviation in w revealed that the dispersion in estimates was driven by sites that had low genotype probabilities, especially sites that had uniform likelihoods across the three alternative genotypes (Figure 3.4). This explains the trend of increasing standard deviation with increasing sample size because the frequency of low quality genotypes encountered is proportional to the sample size. Hence, under appropriate filtering order dependency should not impact the compatible set approach.

Table 3.1: Standard deviation in the relative likelihood assigned to the true genotype through the compatible set approach under genotype likelihood (GL) sample permutation (N = 2500 for 10× and 25× , N = 120 for 5×) for 100 simulated GL pairs. The percentage of the observed standard deviation is reported over 10% intervals.

ϵ	c	$N = 5$						$N = 10$						$N = 15$					
		10	20	30	40	50	> 60	10	20	30	40	50	> 60	10	20	30	40	50	> 60
	5×	98	0	0	2	0	0	87	4	2	4	2	1	88	3	4	4	1	0
0.05	10×	100	0	0	0	0	0	100	0	0	0	0	0	100	0	0	0	0	0
	15×	100	0	0	0	0	0	100	0	0	0	0	0	100	0	0	0	0	0
	5×	95	1	0	2	2	0	88	3	3	3	3	0	89	5	2	2	2	0
0.1	10×	99	0	0	0	1	0	100	0	0	0	0	0	100	0	0	0	0	0
	15×	100	0	0	0	0	0	100	0	0	0	0	0	100	0	0	0	0	0
	5×	97	1	0	1	1	0	92	2	2	3	1	0	89	3	4	3	1	0
0.15	10×	100	0	0	0	0	0	100	0	0	0	0	0	100	0	0	0	0	0
	15×	100	0	0	0	0	0	100	0	0	0	0	0	100	0	0	0	0	0

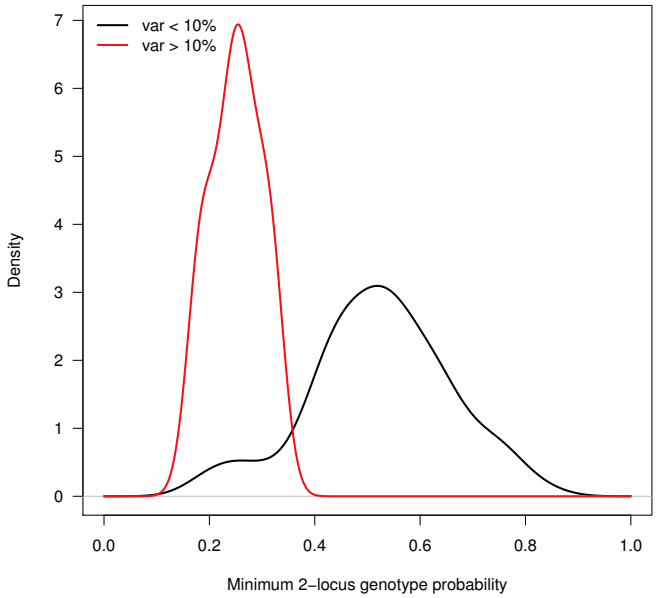


Figure 3.4: Density estimates for the minimum 2-locus genotype probability for the genotype model with maximum likelihood in each sample (N = 100 simulations).

3.2.3 Incorporating compatible sets into the pair-wise likelihood ratio test

To incorporate genotype uncertainty into a pair-wise psuedo-likelihood ratio test (LRT) for recombination, the compatible set is used to construct a likelihood surface, ℓ_{combined}, through weighted products of the pre-computed LDhat likelihood tables. Missing genotypes are addressed by integrating over the likelihoods indexed by the compatible haplotypes.

To generate the surface, the potential haplotypes underlying each compatible genotype configuration are calculated accounting for the combinatorics of the indexing. Informally, to calculate the counts for each haplotype we enumerate all unambiguous diplotype to haplotype mappings, and then consider the set of all potential haplotype counts generated by the double heterozygous diplotype. We then correct for the potential haplotypes.

More formally, given the allele counts for a compatible configuration, where $n_{G(h_i,h_j)}$ represents the count of the unphased diplotype G corresponding to haplotype pair $\{h_i,h_j\}$, the potential counts of the haplotype h_i denoted by $n_{h_i,k}$ are enumerated,

$$n_{h_i,k} = 2n_{G(h_i,h_i)} + \sum_{(G \sim h_i) \setminus \breve{G}} n_G + f(k).$$

The statement $(G \sim h_i) \setminus \breve{G}$ denotes the subset of diplotypes compatible with haplotype h_i excluding the double heterozygous diplotype $\breve{G}$. The variable k, which takes values between 0 and $n_{\breve{G}} - 1$, indexes the possible contributions of $\breve{G}$. The function $f(k)$ specifies the allocation of $\breve{G}$ diplotype counts to each haplotype,

$$f(k) = \begin{cases} k & \text{if } h_i \in \{00, 11\}, \\ n_{\breve{g}} - k & \text{otherwise.} \end{cases}$$

The likelihoods corresponding to each of the potential haplotype counts are extracted from the pre-computed likelihood surface, and are corrected for the double heterozygous genotypes. Two rounds of integration, over the haplotypes compatible with each potential genotype, and the weighted potential genotypes compatible with the data produce a new likelihood surface for the site that now incorporates the potential underlying genotype uncertainty

$$\ell_{combined} = \log \left(\frac{1}{A_2} \sum_i \frac{w_i}{A_1} \sum_{k=0}^{n_{\breve{G}_i}-1} k! \times \exp(\ell_c(\{n_{h_{11,k}}, n_{h_{01,k}}, n_{h_{10,k}}, n_{h_{11,k}}\})) \right).$$

The values A_1 and A_2 represent normalising constants such that $\ell_{combined}$ takes values between 0 and 1. The two constants are included to represent the rounds of normalising that occur to prevent numerical underflow at each integration.

3.3 Evaluation of uncertainty aware method performance

To assess the impact of sequencing and variant calling errors in HTS on LD estimates, I compared patterns of LD estimated from the genotypes identified in HTS of 10 chimpanzees (Chapter 2) and in SNP-typing.

3.3.1 Data

I compared LD estimates at two of the SNP-typing regions, ENCODE 4q26 and ENCODE 7q21, which are each $\sim$ 500kb in size. The four call sets were curated as follows:

SNP-typing data. The SNP-typing data was collected by Winckler et al. (2005) using single-base primer extension chemistry followed by mass spectroscopy (Sequenom, CA), henceforth referred to as the *SNP-typing data*. Winckler et al. (2005) estimated fine-scale recombination rates across 19 localised regions in 38 Western chimpanzees (of which the 10 HTS chimpanzees are a subset). As described in Section 2.5.1.7, the SNP-typing data were lifted from their original human reference coordinates (Build 34) to the chimpanzee reference coordinates (panTro2) to enable comparison with the panTro2 aligned HTS calls and were thresholded to have an estimated minor allele frequency greater than 0.001. The SNP-typing data are reported as hard-called genotypes, measures of genotype uncertainty were not available.

Hard-called HTS genotypes and associated GLs. For the HTS genotypes, I used the filtered genotypes called through the GATK. The variants and genotypes were called in the panTro2-mapped, PCR duplicate cleaned, BAQ-recalibrated, realigned HTS reads. I used the GLs estimated through the GATK.

HTS Haplotypes from LD-refined GLs. For the best-practice HTS haplotypes, I used the haplotypes inferred through PHASE (Stephens et al., 2001) applied

to the genotypes called from the GLs imputed and refined through BEAGLE (Browning and Yu, 2009).

The SNP-typing covered fewer variant sites than were identified in the HTS: 314 versus 903 on 4q26, and 382 versus 1,039 on 7q21. For analyses I used the intersection set between the two call sets for those sites that had less than 3 missing genotypes in both call sets, $\sim$ 250 SNPs for each region.

To compare LD estimates between studies, the SNP-typing data were assumed to represent the true genotypes and the root mean squared error (RMSE) was calculated between their LD estimates and those from HTS. In reality a small fraction of the SNP-typing calls are errors, but this is likely to be negligible compared to the HTS calls. Linkage disequilibrium estimates were calculated from haplotype frequencies estimated through EM (with or without incorporating genotype uncertainty; described in Section 3.1.1), or directly from the haplotype frequencies, and missing genotypes were excluded.

In this chapter I present several figures showing the estimated LD structure in 4q26 and 7q21. Estimated values of $|D'|$ statistics are plotted below the $y = x$ diagonal, and r^2 above the diagonal. The LD measures were calculated from haplotype frequencies inferred through EM. In the LD figures, each cell represents the pairwise LD between two sites, blue represents weak LD and red represents strong LD.

3.3.2 Empirical study 1: LD

I applied the uncertainty aware EM approach to GLs estimated regions 4q26 and 7q21. The improvement in accuracy through incorporating uncertainty is modest relative to the hard called genotypes; $\sim$ 4% in 4q26 and equal in 7q21 (Table 3.2). Note that the RMSE in LD estimates are 4-fold higher for the $|D'|$ statistic compared to the r^2 statistic, as evident in Figure 3.5. In contrast, the RMSE values for the refined haplotypes are appreciably different than the other two approaches; differing

by more than 15 percent. The patterns of LD in Figure 3.6 estimated from the refined haplotypes show, qualitatively, better consistency within resolved LD blocks, the reader should note that we would expect stronger LD block signals which arise from the PHASE algorithm's prior distribution (Stephens et al., 2001). I will discuss the potential reasons underlying the only modest improvements in estimates in the Discussion section.

		4q26			7q21		
		hGT	rHT	uaGT	hGT	rHT	uaGT
Absolute	$\|D'\|$	0.335	0.437	0.321	0.366	0.423	0.366
	r^2	0.128	0.162	0.122	0.135	0.162	0.135

Table 3.2: Root mean squared error (RMSE) in LD estimates, on $|D'|$ and r^2 scales respectively, from hard called HTS genotypes (hGT), HTS haplotypes from LD-refined genotype likelihoods (rHT), and from HTS genotype likelihoods (uaGT). Measures of LD r^2 and $|D'|$ were calculated from haplotype frequencies through Expectation Maximisation on hGT, directly from the rHT, or through incorporating genotype uncertainty in to Expectation Maximisation on uaGT. The RMSE in output from the three approaches was calculated relative to SNP-typing data, where the SNP-typing data was assumed to represent the true genotypes. The higher RMSE value of the $|D'|$ estimates for both 4q26 and 7q21 likely reflects the statistic's sensitivity to extremes in allele frequency (Lewontin, 1988).

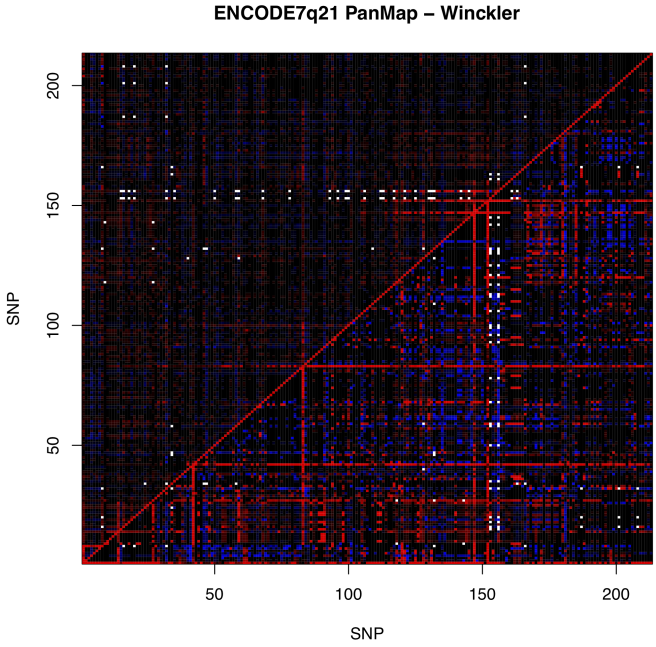


Figure 3.5: Differences between $|D'|$ (below $x = y$ diagonal) and r^2 (above diagonal) estimates for PanMap hard called genotypes and SNP-typing genotypes. Blue represents small values, black intermediate values, and red large values. $|D'|$ shows greater disagreement between call sets.

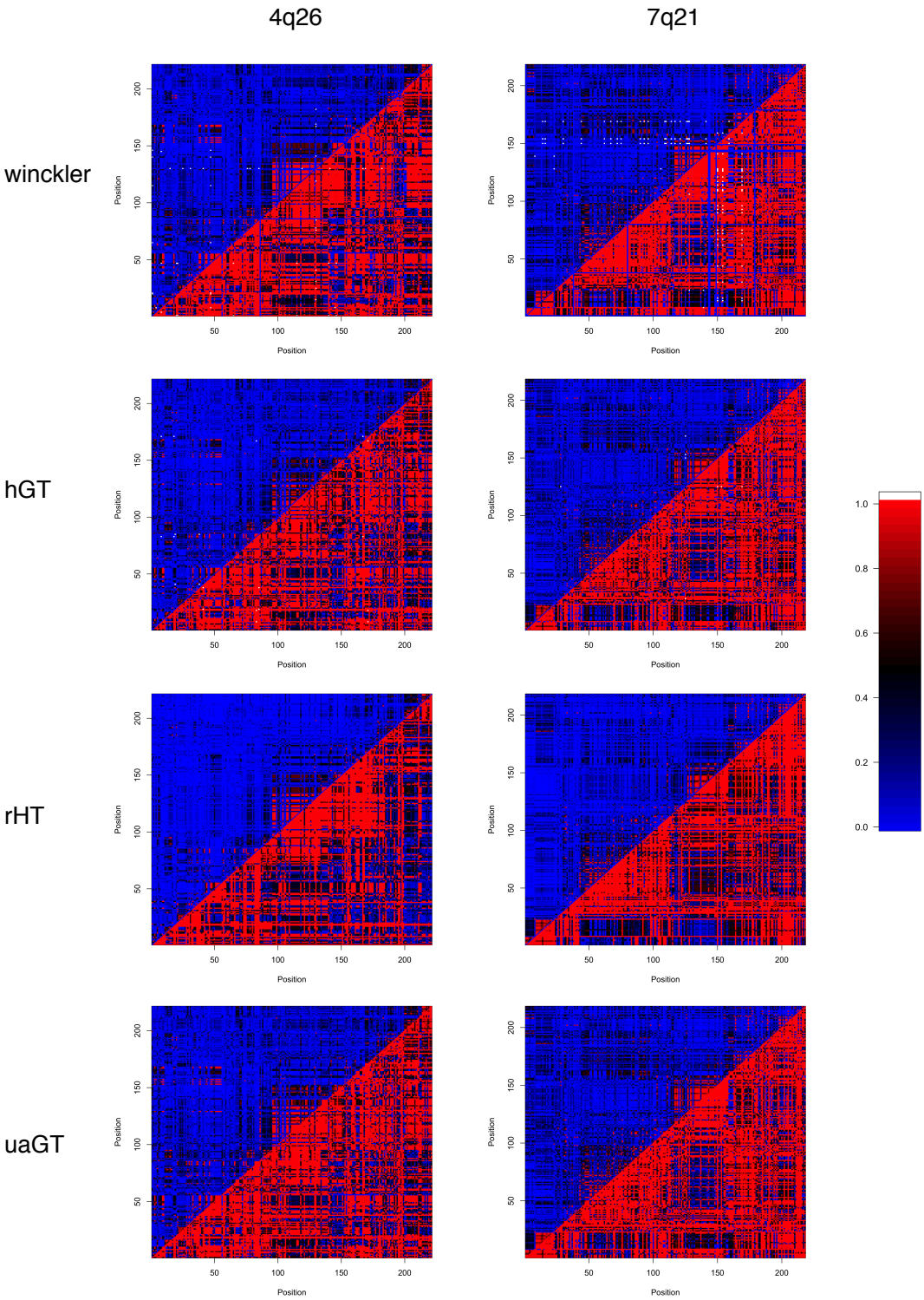


Figure 3.6: LD structure estimated from SNP-typing genotypes (Winckler), hard called HTS genotypes (hGT), haplotypes from LD-refined genotype likelihoods (rHT), and through incorporating uncertainty into haplotype frequency estimation (uaGT). LD measures, $|D'|$ below and r^2 above the $y = x$ diagonal, estimated from haplotype frequencies in regions 4q26 and 7q21.

3.3.3 Empirical study 2: pair-wise likelihood ratio test

To assess whether the compatible set approach improved the ability to detect recombination events, I applied the LRT using a LDhat generated likelihood lookup table generated for $\theta = 0.1$ to the four data sets. Figure 3.7 shows results for 4q26 and 7q21. Region 7q21 shows evidence for three potential recombination hotspots. Region 4q26 has potentially two hotspot signals, however the signals are inconsistent at 4q26 across all 4 data sets. Similar to the first empirical study the improvements through incorporating uncertainty are modest, however local consistency in the LRTs appears to be improved (Figure 3.7).

		4q26			7q21		
		hGT	rHT	uaGT	hGT	rHT	uaGT
Absolute	$\rho = 100$	1.296	1.229	1.194	1.859	1.909	1.752
	$\rho = 0$	3.144	3.086	3.155	2.824	2.843	2.884

Table 3.3: Root mean squared error (RMSE) in likelihood ratio test (LRT) statistics on minus one times logarithm scale from hard called HTS genotypes (hGT), HTS haplotypes from LD-refined genotype likelihoods (rHT), and from HTS genotype likelihoods (uaGT). The RMSE of output from the three approaches was calculated relative to SNP-typing data, where the SNP-typing data was assumed to represent the true genotypes.

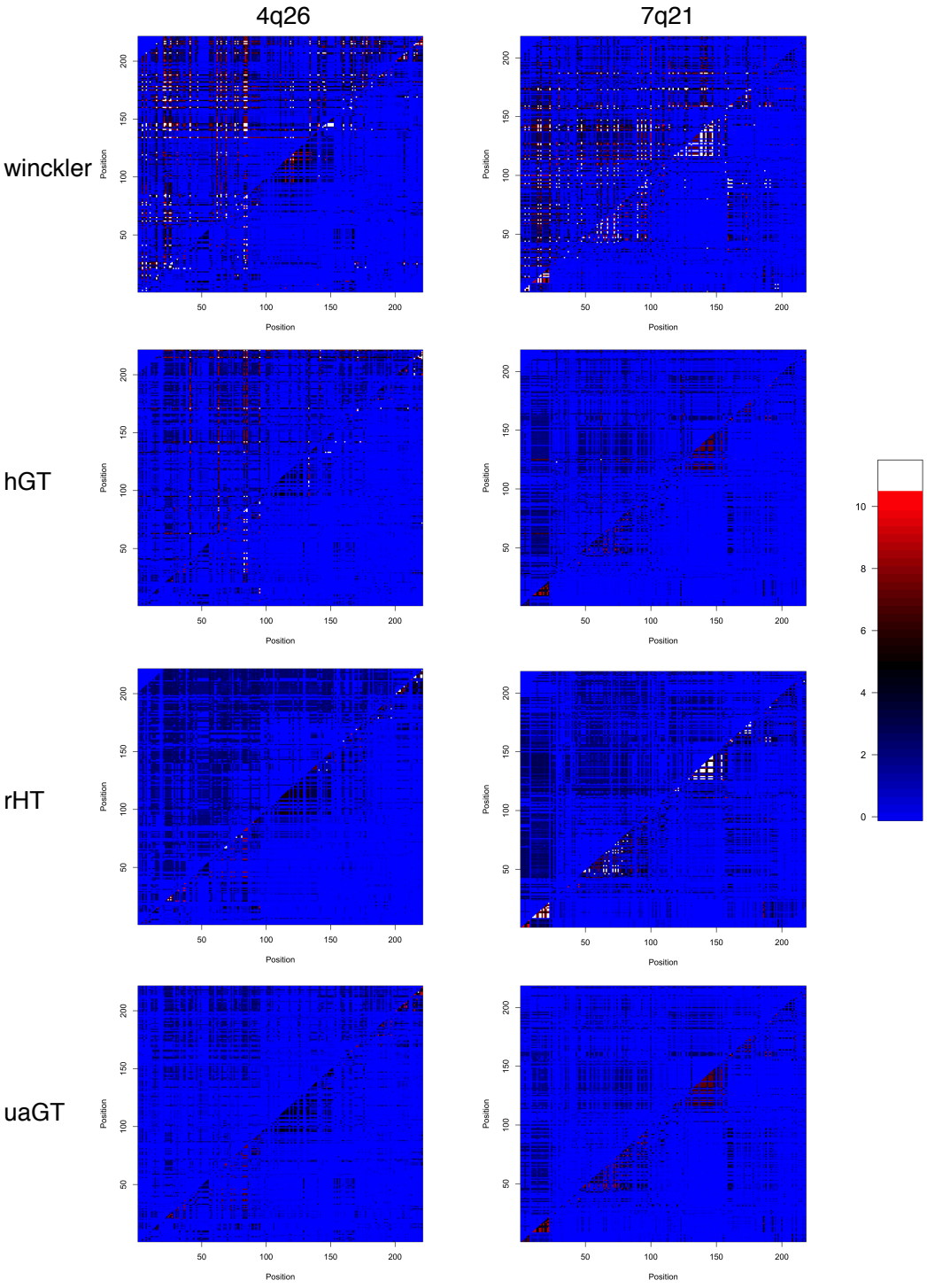


Figure 3.7: Evidence for recombination events inferred by pair-wise pseudo-likelihood ratio test. Applied to SNP-typing genotypes (Winckler), hard called HTS genotypes (hGT), haplotypes from LD-refined genotype likelihoods (rHT), and through incorporating uncertainty into haplotype frequency estimation (uaGT). Test for evidence of recombination activity (H_0: $\rho = 0$) above the $y = x$ diagonal, test for evidence of low recombination activity (H_0: $\rho = 100$) below the diagonal. Fixed colour scale over the range of minus one times log-likelihood ratios between 0 (blue) and 10 (red).

3.4 Challenges

Why does the uncertainty aware approach have such a modest impact in improving LD estimates and the LRT? There are at least three issues in HTS genotyping that overwhelm the impact of genotype uncertainty. First, I will demonstrate that the genotype likelihoods are poorly calibrated. Second, I will assess the impact of false-positives on patterns of LD. Third, I will show that even if we incorporate reference assembly summary statistics such as N50 or Phred scores, holding repeat and missing sequence constant, we are unable to predict which regions of the reference genome will be problematic.

3.4.1 Assessment of genotype likelihood calibration

If GLs are to be an effective representation of genotype uncertainty they need to be well calibrated. I assessed GL calibration by calculating the genotype probabilities of the genotypes called from HTS that were discordant with the genotypes from SNP-typing at ENCODE regions 4q26, 7q21, and 7q31. For a called variant, the probability of each genotype was calculated from the GLs using Bayes' rule, and using a prior on the observed frequency of each genotype in the genotyping data (The posterior probability was changed on average by less than 10^{-17} when a uniform prior was used). The probability of the genotype G_i is then the quotient of its likelihood over the sum of the likelihoods for the locus,

$$\Pr(G_i|D) = \frac{\Pr(D|G_i)\Pr(G_i)}{\sum_j \Pr(D|G_j)\Pr(G_j)}. \tag{3.8}$$

To avoid apparent errors that could result from inconsistent reference allele assignment between call sets, I restricted the analysis to heterozygous genotypes in the SNP-typing data that were discordant with the HTS-genotypes, and only analysed sites with concordant observed alleles.

Across the three regions a genotype discordance rate of 4.1% was observed in agreement with independent estimates (See 2.5.1.6). However, we found the genotype probabilities are poorly calibrated: sites with high certainty on the SNP-typing genotype have higher than expected erroneous assignments, likely arising through mis-mapping, accounting for 31% of the discordant genotypes. Conversely, less certain genotypes were rarely discordant (Figure 3.8).

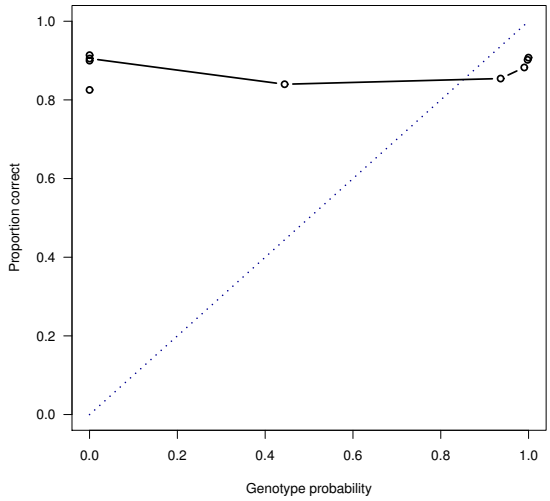


Figure 3.8: The proportion of correct HTS genotypes stratified by their estimated genotype probabilities. The black line represents the proportion of correct genotypes in the chimpanzee HTS samples compared to SNP-typing. To control the variance, genotype probability bins were constructed using 10% percentile intervals. They navy dotted line represents the expected proportion of correct genotypes. The genotype probabilities were calculated from genotype likelihoods (GLs) estimated through the GATK and a prior on the each genotype. N = 4244 genotypes.

Why would these likelihoods be poorly calibrated? First, I will describe how GLs are calculated by the GATK, then I will point out the potential limitations of the GATK approach. The approach used in the GATK is a natural starting point for calculating genotype likelihoods from re-sequencing data, however it is simplistic and misses some features of the data.

For a site in a sampled genome, the GL for genotype G is calculated as the product

of the likelihoods for each covering read, where the read i is denoted by D_i,

$$\Pr(D|G) = \prod_i \Pr(D_i|G). \tag{3.9}$$

The likelihood for each of the three biallelic genotypes is calculated from the contributions of the two alleles $\{a_1/a_2\}$ constituting genotype G,

$$\Pr(D_i|G = \{a_1/a_2\}) = \frac{\Pr(D_i|a_1) + \Pr(D_i|a_2)}{2}. \tag{3.10}$$

The likelihood for allele j, $\Pr(D_i|a_j)$, depends on the probability that the read is miscalled, reported as the quality score of the base denoted by ϵ_i,

$$\Pr(D_i|a_j) = \begin{cases} 1 - \epsilon_i & D_i = a_j \\ \frac{1}{3}\epsilon_i - \Pr\left(D_i \text{ is miscalled}\right)) & \text{otherwise.} \end{cases} \tag{3.11}$$

The probability that the read is miscalled is a pre-specified empirical error rate. The error rates are calculated for each sequencing technology as the frequency of observed mismatches in reads aligned to regions of expected homozygosity in the human reference genome. There are three potential drawbacks to inference under this approach. First, there is an implicit assumption of independence among reads in Equation (3.9), which may be violated in the presence of read-mapping errors or PCR artefacts. Second, the base qualities are assumed to be well calibrated. Third, the empirical confusion matrix values may not apply equally to the entire reference genome, especially in regions of poor assembly, or for multi-copy sequences that are under represented in the reference genome.

Figure 3.9 indicates that there is a subset of errors in the distribution of estimated genotype qualities at mismatches where the mis-called genotypes have high confidence. This likely results from a failure to model the underlying sequencing and mapping errors. Therefore, it is not surprising that we only observe modest

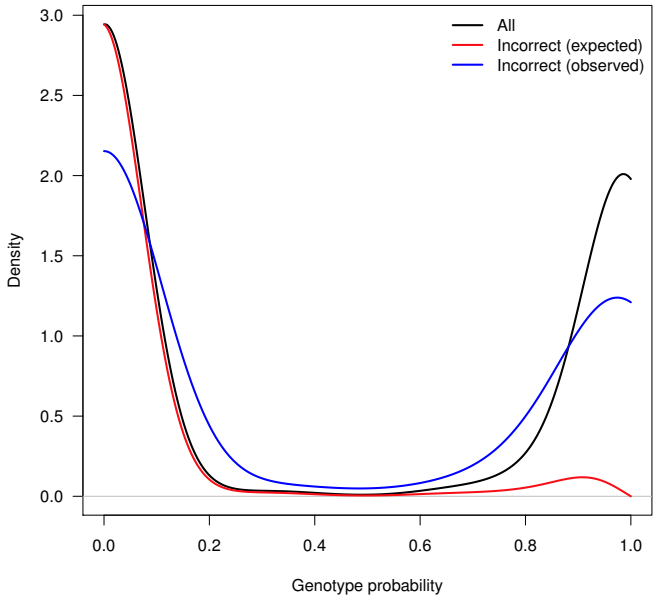


Figure 3.9: Distribution of GATK genotype quality scores for HTS genotypes. The posterior genotype probabilities (calculated using using Equation (3.8)) is the confidence that the true genotype is the one provided. The distribution of genotype posterior probabilies, $\phi(p)$, induces the distribution on expected errors which is proportional to $(1-p)\phi(p)$. $N = 4244$ genotypes.

improvements through incorporating uncertainty.

3.4.2 Examples of artefactual LD patterns at SNP clusters in mapped HTS

A motivation for HTS is the ability to sample greater levels of variation. I now analyse all sites detected in the HTS for region 4q26 and 7q21, and demonstrate the impact of false-positives on estimates of LD from HTS.

Figure 3.10a provides an example of a strong estimated rate peak at 90.08 Mb on 7q21 is driven by six false-positive sites that lie inside a 5334nt LINE element. Four of the six SNPs lie in a 1914nt L1MA8 fragment at the 3'-end of the LINE. Within this region there are symptoms of potential read mis-mapping: reads have

elevated coverage ($> 20\times$, 99.99^{th} percentile) and many mate-pairs map to alternative chromosomes. Note that the apparent hotspot at 90.08Mb is not estimated in the HTS haplotype map (Section 2.2.1.1) indicating that surrounding variants do not support the hotspot signal.

An additional difference arises between the rate estimates from SNP-typing and HTS. First, there is a background of peaks in the HTS rate estimates, at least some of which may be artefactual (Figure 3.11). For example, the weak peak at 121.2Mb on 4q26 corresponds to 13 SNPs in an L1PA15 element (Figure 3.10b). The reads mapped to this region contain indels and high rates of mis-matches relative to the chimpanzee reference sequence. In general, SNPs in these problematic sequence contexts often, but not always, seem to produce weaker peaks.

Hence it is not surprising that after the application of the uncertainty aware approach the false-positive driven errors at the L1PA7 element remain (the six SNPs all have a Phred-scaled quality score greater than 650), it would be unexpected for the uncertainty aware approach to eliminate signals that arise through mis-mapping.

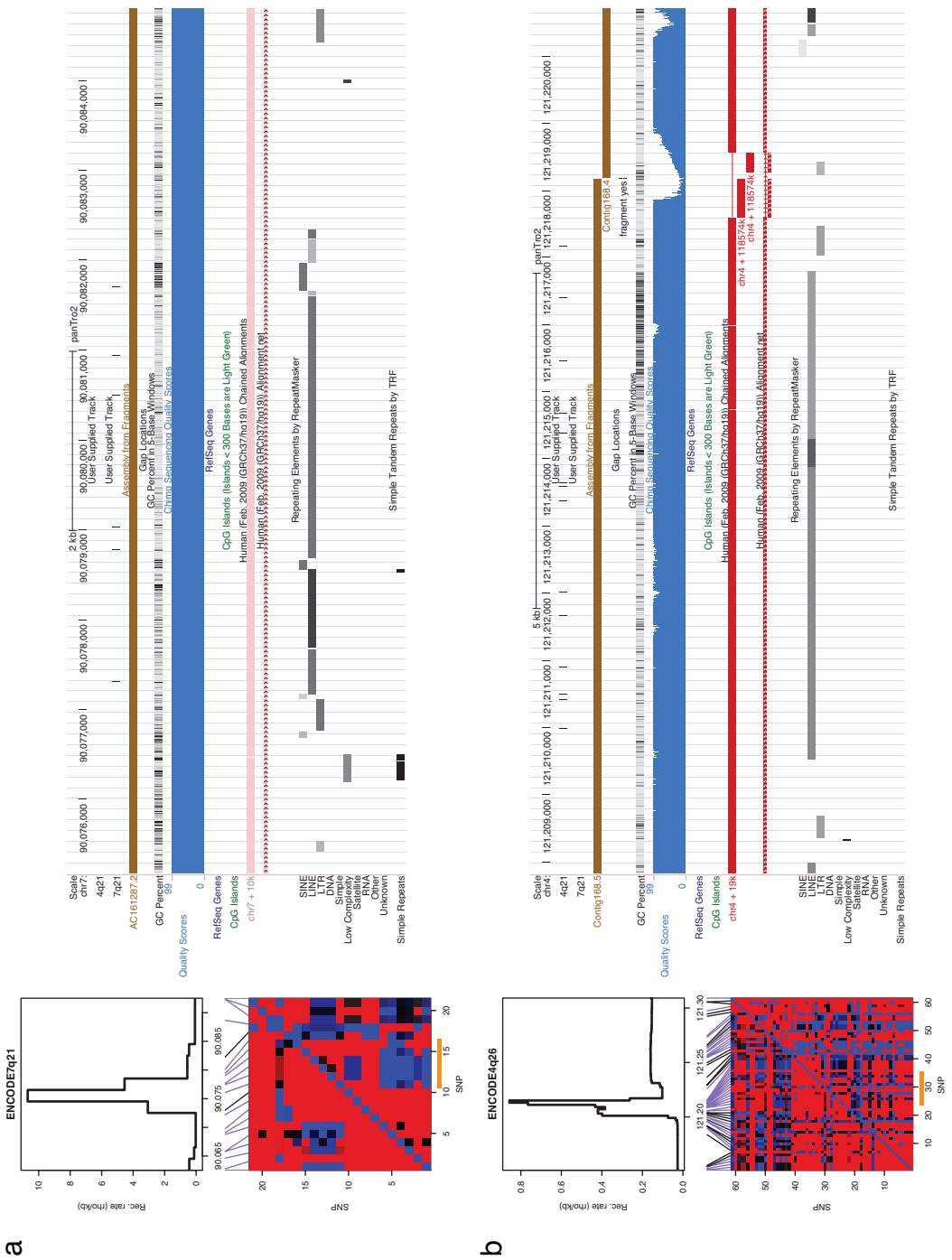


Figure 3.10: Fine-scale estimates of recombination rates and LD in chimpanzees at problematic HTS SNPs. The two panels on the left show recombination rate profiles and LD profiles estimated from HTS genotypes at Winckler et al. (2005) typed sites. Tick marks represent the locations of SNPs with respect to their physical location in the reference, purple lines represent sites in repetitive sequence. The orange rectangles indicate SNPs that are plotted in UCSC genome browser view on the right. (a) A strong hotspot signal in 7q21. (b) A weak hotspot signal in 4q26.

3.4.3 Challenges in predicting errors from assembly sequence statistics

An important motivation for studying the two Winckler regions was that they have different assembly characteristics, and therefore through comparing the regions it was anticipated that we could assess the impact of local assembly on the quality of HTS variant calls by comparing the regions.

The assembly across 4q26 is composed of 34 short unconnected contigs (Table 3.4). In contrast, region 7q21 is spanned by four large, connected contigs. The average sequence quality and repeat quality of the two regions are above the genome average but are of similar magnitude. The Phred scores in 7q21 are significantly higher than 4q26 ($P < 2 \times 10^{16}$, unpaired t-test).

Table 3.4: Properties of the Western chimpanzee reference genome (PanTro2) at ENCODE regions 4q26 and 7q21. Summary statistics were obtained from the UCSC Genome Browser. *Phred* represents the average Phred-scaled sequencing quality of the contigs.

Region	Pos. start (bp)	Pos. end (bp)	N_{repeat}	% repeat	N_{contig}	N_{gap}	Phred
4q26	120,872,889	121,395,272	749	54.64	34	33	90
7q21	89,778,740	90,283,322	855	53.34	4	0	98
Genome	1	3.4×10^9	5.1×10^6	48.98	246,918	197,548	81

Given the reported assembly statistics for the regions we might expect more accurate mapping to 7q21 and therefore higher quality genotype calls. However the assembly statistics are misleading. We expect to obtain equal total genetic distance values for rates estimated from the SNP-typing data and the HTS data, with the only difference being improved resolution in rates estimated from the HTS. The total genetic distance for the HTS genetic distance across 4q26 was 7% greater than the distance estimated from SNP-typing, within expected levels of estimated rate variability. In contrast the genetic distance across 7q21 was 70% greater for the HTS than

the SNP-typing (Figure 3.11). Hence, this implies that the assembly Phred scores are not well-calibrated, and demonstrates that the local impact of inadequacies in the reference genome assembly cannot be predicted from the assembly statistics alone.

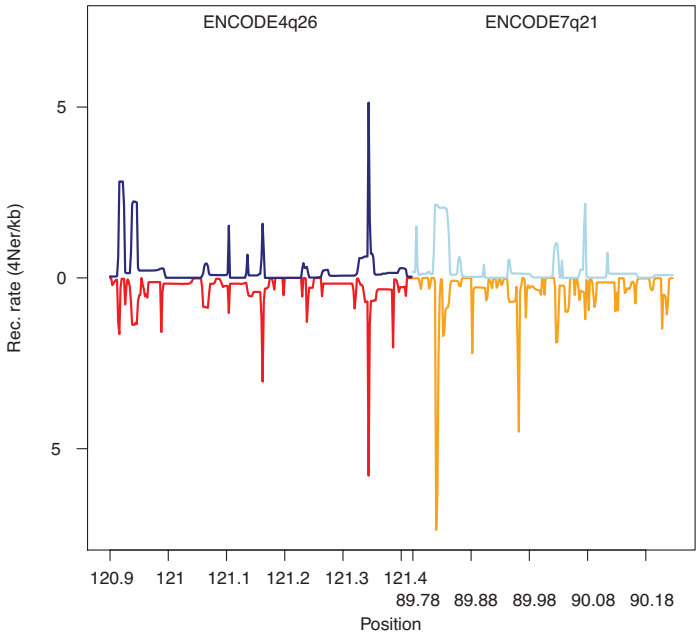


Figure 3.11: Fine-scale recombination rate estimates of 10 chimpanzees from SNP-typing genotypes, and from genotypes inferred from HTS. Above the x-axis: dark and light blue lines represent the fine-scale recombination rate estimates from the SNP-typing genotypes for ENCODE regions 4q26 (left) and 7q21 (right). Below the x-axis: red and orange lines represent rates estimated from genotypes inferred from the HTS. Rate estimates were averaged over non-overlapping 2kb windows. All the called HTS variants were used to estimate recombination rates.

3.5 Discussion

In this chapter I investigated how the inherent uncertainty in low-coverage HTS data can be incorporated into the study of recombination. I demonstrated that uncertainty in HTS genotypes inflates the noise in LD estimates, and how this can be addressed by incorporating GLs into the EM estimation of haplotype frequencies. I conclude that incorporating uncertainty through GLs is an effective approach to guard against genotyping errors, however, the application to real data is severely limited by the

poor calibration of GLs and the impact of false-positives (though the impact of false-positives does not interfere with the comparisons of LD and LRT in this Chapter). These two additional error modes dominate the signal.

I highlighted potential inadequacies in the way that GLs are computed which leads to incorrect genotype calls. There are two ways in which erroneous GLs may affect inference. First, at well-behaved SNPs the GLs may not be well calibrated, which likely arises from overdispersion. It has been suggested that a weighting scheme should be used to take correlated errors into account (Li et al., 2008). Martin et al. (2010) also suggested estimating error rates directly from the mapped read data for each site independently, instead of using quality scores. This would potentially guard against reference assembly driven errors. However, Nielsen et al. (2011) point out that such an approach discards the considerable information regarding errors gained through the base calling and mapping process. Hence, a potentially powerful approach would be to estimate error rates jointly from the data and the information obtained during base-calling and mapping.

A second, and more problematic, situation is that at false-positive variant calls GLs are meaningless and incorporating uncertainty is ineffective. False-positive errors appear to dominate in many situations (Figure 3.9) and improving GLs will not help in this instance. Although improvements can be made by refining the probabilistic models to call variation, it is likely that to improve the accuracy of HTS-based analyses increased coverage and improved reference genome assemblies are required.

Chapter 4

Features of fine-scale recombination rates in chimpanzees

In Chapter 2 I evaluated the feasibility of using HTS data from 10 genomes for the analysis of recombination rates at different scales. Here I describe analyses of the fine-scale map of recombination estimated from the 10 Western chimpanzee genomes. This enabled investigation of (a) what factors influence recombination in chimpanzees, (b) whether these factors are shared with humans, (c) how these factors have evolved between the species.

4.1 Methods to evaluate cross over activity at genome features

Common, dispersed, genome features such as genes and DNA repeat elements provide many replicates across the genome, and are therefore a powerful resource for investigating the impact of sequence and/or epigenetic factors on recombination activity.

However, the study of genome features in their aggregates is potentially problematic. For instance,

- The average recombination profile may not reflect the true activity of a feature because of small sample sizes and/or high variance in the recombination rates.
- Clustered features introduce dependence structure in the average profile.

4.1.1 Sampling to remove clustering

The dense 2 SNP/kb genome-wide polymorphism we observed in the 10 chimpanzee genomes enables fine-scale estimates of recombination rate variation at the resolution of individual hotspots, which are estimated to be 1-2 kb in humans (Jeffreys et al., 2001). An approach to study the recombination-promoting activity of a particular feature is to estimate its the average recombination rate profile around these features. Such an approach assumes that the features are independent and identically distributed, however, features are often clustered across the genomes and therefore not independent. Consider the toy example in Figure 4.1, we wish to observe the recombination activity at a feature in the genome (represented by red tick marks). The local effect of the feature is to suppress recombination activity, however a subset of the features are located in regions of high recombination activity. If we calculate the average profile across all feature locations, the regions of elevated recombination dominate the signal; we would conclude, incorrectly, that the feature activated recombination (crimson profile). Therefore a procedure to remove clustering is required.

There are several potential thinning procedures, however, it is crucial that the thinning procedure does not itself introduce dependencies. To remove overlapping signals, I devised a sampling procedure control for clustering. The intuition behind the approach is to uniformly sample features across non-overlapping windows, such that impact of a feature on recombination rates over the scale of the window size is

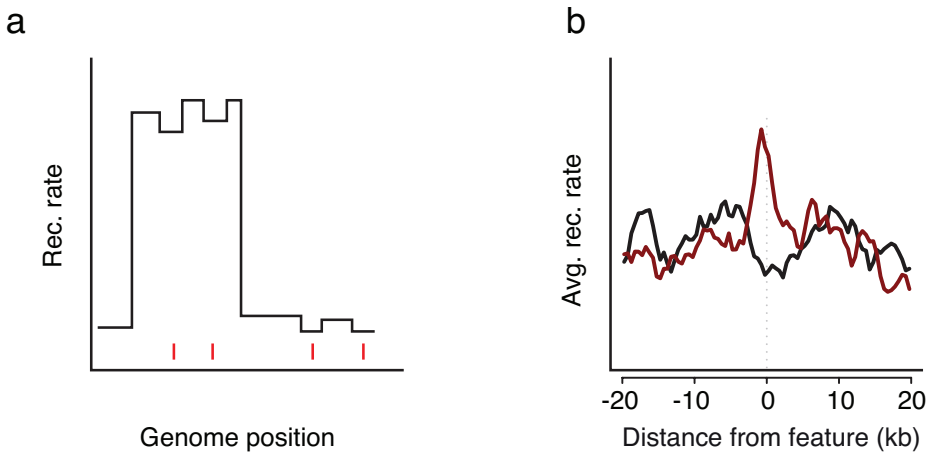


Figure 4.1: Impact of feature clustering on average fine-scale recombination rate variation profile. **(a)** Toy example of the recombination rate variation across a chromosome region. Feature copies are represented by red tick marks. Note that the recombination rate is locally suppressed at features. **(b)** Average recombination rate profile around feature elements calculated without thinning to remove clustering (crimson), and thinning through uniform sampling within windows (black, see methods).

analysed. The sampling procedure for a specified window size l is described as follows:

1. Given chromosome G of n nucleotides with positions $\{1, 2, 3, \ldots, n\}$, and a matrix X of the start and end feature coordinates along G.

2. Generate the partition set, S, of chromosome segments each with length l,

$$S = \big\{\{1, 2, \ldots, l\}, \{l+1, l+2, \ldots, 2l\}, \ldots, \{n-l+1, n-l+2, \ldots, n\}\big\}$$

3. For each segment in S, calculate the coordinates of the set of central windows W of length $c < l$,

$$w_j = \big\{\lfloor \tfrac{l \times j}{2} \rfloor - \lfloor \tfrac{c}{2} \rfloor, \lfloor \tfrac{l \times j}{2} \rfloor + \lfloor \tfrac{c}{2} \rfloor\big\} \quad j \in [1, \lfloor \tfrac{n}{l} \rfloor].$$

4. Calculate the midpoints M of the feature locations in X.

5. Find the features whose midpoints fall in the central windows

$$M^* = M \cap W.$$

6. If more than one feature midpoint falls in a single central window, choose one uniformly and exclude the others.

Informed by previous recombination rate estimates in chimpanzees at limited genomic regions (Ptak et al., 2004; Winckler et al., 2005; Watanabe et al., 2004; Ptak et al., 2005), I assumed that chimpanzee recombination signals are similar to the 1-2 kb widths of hotspot signals in humans and therefore looked for effects at the 5 kb scale in chimpanzees. Hence, a central window of 10 kb was selected ($c = 10$ kb) to assess the impact of the feature on rate variation (constituting a quarter of the window; $l = 40$ kb), flanked by an additional 15 kb of peripheral sequence either side to observe the decay profile back to genome-average levels. This approach has attractive properties (1) it is fast, which is important for scanning through the numerous genome annotations, and (2), has direct biological interpretation. However, it is only robust to confounding signals that operate over the length scale of the windows. Consequently, I validated observations using the robust map (See 2.2.2.3). An additional down-side of the approach is that the thinning procedure can potentially remove too much data. Note that there are several alternative approaches to address this issue. For instance, block bootstrap can be used to guard against systematic clustering, or wavelet decomposition would allow for control over several length scales. Nevertheless, the approach has the attractive property in that it can be interpreted as the recombination localising effect of introducing a target feature into 40 kb of genomic sequence.

4.1.1.1 Comparisons between human and chimpanzee

To reduce the potential impact of genome rearrangement on comparisons of recombination rate we identified regions of broad-scale homology between the species (synteny regions). Informally, we identified syntenic regions from SNP sites that were consistent under reciprocal mapping between the human and chimpanzee reference assembly, and then recursively identified blocks of consistent homology (see Section 2.5.1.5).

4.1.2 Quantifying recombination-influencing activity

There are three major challenges in comparing the fine-scale structure of recombination rate variation between human and chimpanzee at genome features. One, there are different levels of noise in the fine-scale estimates of recombination (see Section 2 and Figure 4.2a,b). Two, the genome average recombination rate (the background rate) is different between species (Figure 4.2a,b and 4.19c). Three, the number of feature occurrences vary widely (there are over 1.1×10^6 Alu elements in the chimpanzee genome, but less than 50×10^3 Tigger transposable elements). The stochasticity introduced by these three factors compromise signal. Therefore to enable accurate comparison between the species and/or features, a statistical approach is required. To quantify the recombination-influencing activity of features, the enrichment of cross over rates in the average profile around a feature relative to the local background rate was calculated, by using a maximum likelihood approach to fit a Gaussian curve to the average rate profile. For an observed average rate profile $r(\mathbf{x})$, the Gaussian curve with peak rate A, variance σ^2, and background rate B, was specified as

$$r(\mathbf{x}) = A \times N(\mathbf{x}, 0, \sigma^2) + B + \varepsilon, \tag{4.1}$$

where x_i is the distance from the the midpoint of feature and ranges between -20

kb and 20 kb (to math the thinning window).

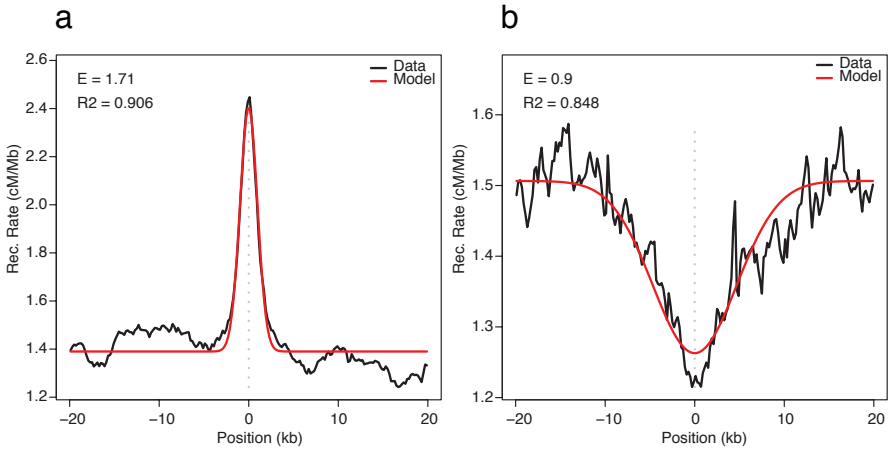


Figure 4.2: Maximum likelihood model to quantify fine-scale recombination-influencing activity at genome features. A Gaussian profile (red) is fit to the average recombination rate profile (black) using least squares regression. The model coefficients and R^2 of the model fit are reported in top left hand corner. **(a)** Average recombination rate profile calculated from the combined population recombination map from HapMap around THE1B elements in regions of synteny (N = 7746, thinned across autosomes). **(b)** As for (a), but for chimpanzee genetic map. Values of R^2 differ by 6% between the maps, however, larger disagreement is observed for repeats with fewer counts. The background rate for the chimpanzee profile is 21% larger than Hapmap.

The average recombination rate profile at a particular feature was calculated as follows. The set of features were selected from regions of human-chimpanzee synteny and were thinned using the protocol outlined above. For each feature the genetic distance across the 40 kb region was calculated in 200 bp windows. For each window the arithmetic average of recombination rates was calculated, and the 95% confidence interval was calculated under the assumption that rates across windows were Normally distributed. Windows in reference sequence gaps were excluded.

To identify genome features that had a local impact on fine scale rates, the model was fixed to be symmetric around the feature midpoint, and σ^2 was allowed to vary such that the Gaussian curve had a full width half maximum between 0.59 kb (approximately minimum resolution for hotspot detection in our data) and 2.94 kb (the

average hotspot width of human hotspots (Consortium, 2007)). As a result of these parameter choices, features with a broad recombination profile over the 40 kb window would not be captured well by the model.

The fold enrichment of recombination around at the genome feature was calculated the ratio of maximum likelihood estimate of the rate at the feature midpoint ($x = 0$) to the background rate,

$$E = \frac{\hat{A}}{\sqrt{2\pi}\hat{B}}. \tag{4.2}$$

Values of E greater than one show evidence for activating recombination, whereas values of E less than 1 show evidence for suppressing activity. To enable robust comparisons between the profiles of different maps, the quality of each fit was assessed by calculating the R^2 of the model fit and the number of observations recorded.

4.2 Recombination hotspots

Previous studies of localised synteny regions suggest that recombination hotspots are typically not shared between humans and chimpanzees (Myers et al., 2010; Ptak et al., 2004; Winckler et al., 2005; Wall et al., 2003) likely due to the function of the zinc-finger protein PRDM9 (Myers et al., 2010; Baudat et al., 2010; Parvanov et al., 2010), which binds motifs associated with hotspot activity (Baudat et al., 2010; Grey et al., 2011) and is highly diverged between the human and chimpanzee reference genomes (Myers et al., 2010; Oliver et al., 2009). In humans, sequence variation within the PRDM9 zinc-finger array leads to differential activity at both allelic and non-allelic cross-over hotspots (Baudat et al., 2010; Berg et al., 2010, 2011) and alleles found only in individuals of African ancestry lead to population-specific hotspots in patterns of both linkage disequilibrium (LD) and admixture (Hinch et al., 2011). However, in humans the PRDM9-bound 13 bp motif is detected clearly only in 40%

of hotspots (Myers et al., 2008), although activity at some hotspots with no clear match is PRDM9-dependent (Baudat et al., 2010; Berg et al., 2010). Nevertheless, there could exist different classes of hotspot in humans, some of which are PRDM9-independent and hence potentially shared between species.

To assess whether different classes of hotspot evolve in different ways, I compared the average profile of human and chimpanzee fine-scale rate variation at recombination hotspots identified in each species (see 2.5.1.10 for methods); those from 10 genomes from 1000 Genomes of CEU and YRI ancestry, 10 chimpanzee genomes, and HapMap. Each set of hotspots was thinned to remove hotspot clustering (see Section 4.1 methods).

As expected, estimated recombination rates were markedly increased (15-fold) in chimpanzees at inferred chimpanzee hotspots in syntenic regions (Figure 4.3a). At orthologous locations in human, however, the estimated recombination rate did not deviate from background levels (Figure 4.3a). Similarly, there was no evidence for sharing between species at hotspots identified in the 10 YRI genomes (Figure 4.3b). Furthermore, there is no evidence for hotspot-sharing even for those HapMap hotspots that are estimated to have low probability of motif based activity (Figure 4.4, probabilities provided by Simon Myers). Note that differences in recombination intensity between human populations of different ancestry (Figure 4.3b,c) likely reflect the asymmetric partial dominance effects between PRDM9 alleles and their relative frequencies in human populations (Berg et al., 2011; Ségurel et al., 2011, and Chapter 6).

Three potential artefacts could erroneously cause lack of hotspot correspondence in the inferred locations. First, population structure within the sampled genomes. Second, comparisons in regions with low rates of sequence similarity. Third, the data and analytic methods provide insufficient statistical power to detect hotspots even when present.

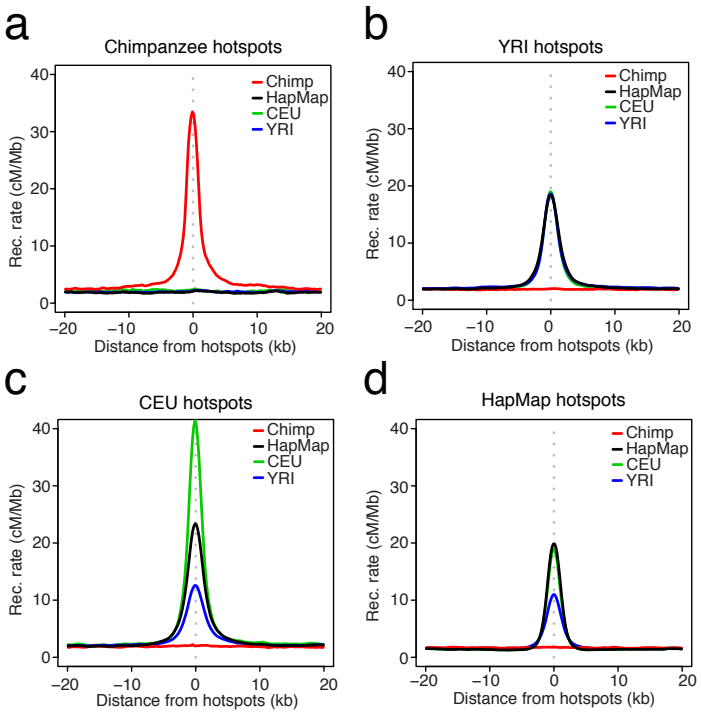


Figure 4.3: Fine-scale recombination rate variation around inferred hotspots in chimpanzees and humans. **(a)** Recombination rates around hotspots identified in chimpanzees (red, N=2268) at syntenic regions in CEU (green), YRI (blue), and HapMap (black). **(b)** Hotspots called from 10 YRI 1000 Genomes individuals (N=6950). **(c)** Hotspots called from 10 CEU 1000 Genomes individuals (N=5293). **(d)** HapMap hotspots (N=10718). Hotspot coordinates were thinned to remove hotspot clusters.

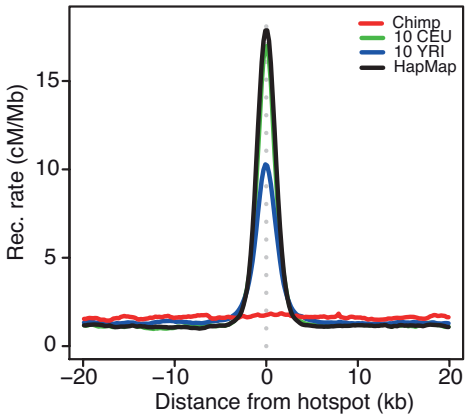


Figure 4.4: Recombination rates around human hotspots that have no clear match to the defined PRDM9 binding motif (N=3284) at syntenic regions in CEU (green), YRI (blue), and HapMap (black).

We addressed, the potential impact of population structure by sequencing chimpanzees that were previously confirmed as unrelated and derived from a single population (Winckler et al., 2005). We minimised the impact of sequence divergence through comparing rate estimates only at syntenic regions. Low power to detect hotspots, or a high rate of false positives, could cause a lack of overlap in the observed hotspots between species, though the sensitivity and specificity would have to be extremely poor to explain the nearly complete lack of correspondence across hotspots in both directions of comparison. Reassuringly, rate estimates from both subsampled human controls, have good correspondence with independent human estimates at several ENCODE study regions including hotspots previously validated by sperm typing (see Sections 2.2.2.5 and 2.2.2.6).

Note, that the experiment was not designed to exhaustively catalogue chimpanzee hotspots. Indeed, less than 50% of the potential hotspots found by HapMap were detected in the 10 subsampled human genomes ($\sim 12,000$ in the 10 CEU, $\sim 15,500$ in the 10 YRI, and ∼20,000 in the HapMap (Consortium, 2007)). However, we do identify sufficient instances to enable accurate assessment of the fine-scale rate variation at hotspots in each species.

4.3 The distribution of fine-scale recombination rates across the chimpanzee genome

Given a map of recombination, the concentration of recombination across a genome can be assessed by calculating the proportion of sequence that contains a given proportion of the observed recombination (Myers et al., 2005). If recombination occurred uniformly across the genome, such a relationship would be linear with gradient one, however, if recombination were to be concentrated at specific locations the relationship arcs away from the diagonal.

In chimpanzee, recombination rates have a similar distribution to human with subtle differences across populations with different ancestry (Figure 4.5a). For Europeans, 80% of recombination events occur over 10% of the sequence. However, for Africans and chimpanzees, 20% of the sequence is required. It is evident that the landscape of recombination is similar between the species, but raises the question: if Europeans and Africans share some but not all hotspots why is their level of concentration different?

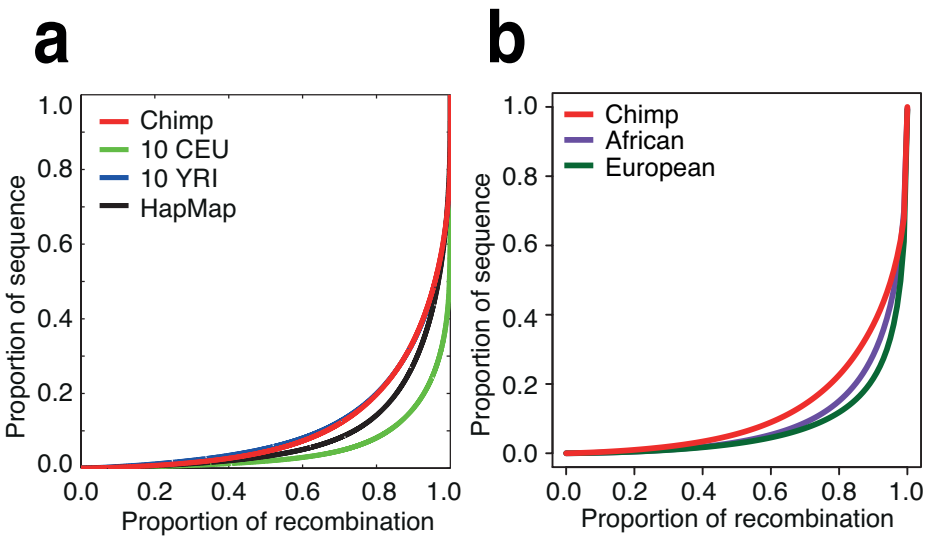


Figure 4.5: The concentration of recombination rate in fine-scale genetic maps estimated for chimpanzee and human populations. **(a)** Concentration in chimpanzee (red), 10 human genomes from populations of European (CEU, green) and African (YRI, blue) ancestry, and the combined-population HapMap map (black). **(b)** As for (a), but recombination maps from African (purple) and European (dark green) components in admixed African-Americans Hinch et al. (2011).

There are three potential explanations for this observed difference in the concentration of recombination. First, the difference lies at the level personal genetic maps (the frequency recombination events across an individual's gametes). Under such a model (Figure 4.6a), the average European genome (purple) has a greater concentration of recombination than an African (blue) or chimpanzee genome, which is also realised as differences a the population-averaged LD-based map.

Second, the differences may be generated by personal maps with comparable con-

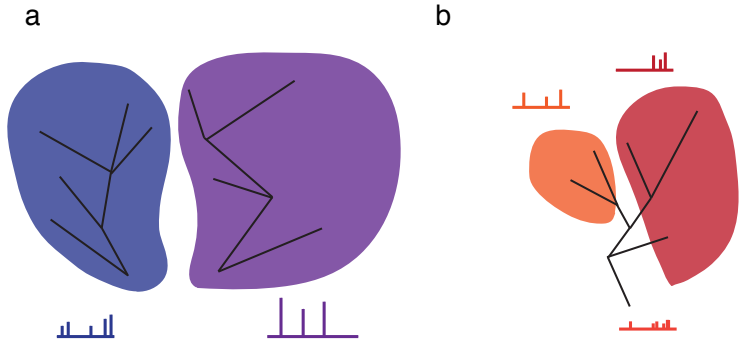


Figure 4.6: Models of the impact of PRDM9 on recombination hotspots. Coloured regions represent populations with different PRDM9 alleles and recombination rate profiles represented by coloured lines. **(a)** Model 1: PRDM9 alleles generate different distributions of hotspot intensities. Individuals with purple ancestry have more intense hotspots than individuals with blue ancestry. **(b)** Model 2: individuals with different PRDM9 alleles have similar distributions of hotspot intensities (orange and red populations) and individuals with multiple PRDM9 ancestries have a superposition of the PRDM9-allele specified maps, resulting in an averaged map with lower hotspot intensities.

centration, but variation in the locations of hotspots between individuals within each population. Under such a model (Figure 4.6b), variation in the personal genetic maps of African and chimpanzee populations would be greater than Europeans, realised as decreased concentration in the population-averaged LD-based map.

Third, differences in population history between the ancestral groups could potentially impact recombination rate estimates (e.g. Stumpf and Goldstein, 2003).

It is not possible to separate true differences between populations and individuals from population-based maps alone. However, these models can be assessed by comparing LD-based maps with maps estimated through independent approaches.

I compared the concentration of recombination in the human LD-based genetic maps to the African and European genetic maps estimated from patterns of admixture in African-Americans by Hinch et al. (2011). The African and European maps were inferred by comparing the African-American admixture map to the deCODE European pedigree-based map (Kong et al., 2010), the African map was identified as

the component that could not be explained by the deCODE map. Hence by comparing the concentration of recombination for the African and European components, recent recombination events (within the last six generations since admixture) can be assessed in contrast to the deeper historical events inferred from unrelated population variation data. The African and European components largely reflect activity of the C and A PRDM9 allelic variants respectively with each allele localising different hotspot positions in the human genome (Hinch et al., 2011).

Figure 4.5b indicates that the African and European components have closer similarity in the distribution of the concentration of their events compared to the population-level LD-based maps.

The increased agreement relative to LD-based maps supports the hypothesis that individual PRDM9 alleles generate similar densities and intensities of hotspots, but the genetic map estimated from variation data represents an average over the alleles that have been active over evolutionary time. Hence, the greater diversity of PRDM9 alleles in Africa results in a less concentrated African population genetic map than the European map.

The averaging process is potentially reinforced by the way that PRDM9 alleles function in heterozygous combinations. Two studies suggest partial dominance relationships in hotspot usage between human PRDM9 alleles. LD-based analyses of the Hutterite community revealed that individuals heterozygous for the main A variant and the minor I variant (which bind distinct sequences) show about a 70% decrease in hotspot usage as compared with AA individuals (Baudat et al., 2010). In addition, (Ségurel et al., 2011) analysed sperm-typing data to demonstrate that the C-type variant partially dominates the A-type variant in heterozygous combinations (which bind distinct sequences). In addition to the authors' observations, I note that the variance in recombination activity for the heterozygotes is greater than the variance observed for either homozygote combination which is compatible with the averaging

hypothesis.

Does this averaging over distinct maps affect the chimpanzee map? There is the possibility that the averaging process may underly the chimpanzee map, but an assessment of the distribution of chimpanzee and human recombination alone cannot separate this from alternative hypotheses such as differential localising activity in chimpanzee PRDM9 alleles. However, insight can be gained through the following analyses. First, the role of PRDM9 in localising hotspots can be assessed through the study of hotspot locations between the two species. Second, the sequence specificity of PRDM9 can be assessed through observing profiles of recombination at genome features and through a search for DNA motifs that activate recombination. Third, a survey of the allelic diversity of chimpanzee PRDM9 would provide additional insight into the evolution of PRDM9 in apes.

4.4 Fine-scale recombination variation structure at genome features

Despite recent advances, the precise relationship between PRDM9 DNA binding and hotspots is unknown. In addition to specific PRDM9-binding DNA motifs, sequence context such as base-composition and the location of genes have been identified as influencing recombination in humans (Spencer et al., 2006; Myers et al., 2005). Indeed, the activity of the degenerate 13 bp binding motif CCNCCNTNNCCNC for human PRDM9 A-allele is influenced by the extended sequence background and the bases at the degenerate sites. For example, the presence of the core 13 bp bound PRDM9 A-allele motif CCTCCCTNNCCNC on the THE1A background led to a detected hotspot 73% of the time, whereas in unique DNA it led to a detected hotspot 10% of the time (Myers et al., 2008). The high predictive power on the THE1A background is in part caused by the immediate context surrounding the repeat element, which

leads to the presence of the most recombinogenic nucleotides outside the motif (Myers et al., 2008). However, when considered across the entire genome, such highly recombinogenic motif-background combinations drive a minority of hotspots (the location of 3.5% of hotspots is determined by motif-background combinations with relative risk greater than 10, (Myers et al., 2008)). Indeed, two sperm typing studies showed that the activity of several hotspots were all influenced by PRDM9 genotype, even when the hotspot did not contain an apparent match to the two known recombination motifs (Berg et al., 2011, 2010). Therefore one might postulate that many potential hotspot locations are influenced by factors other than sequence context alone.

4.4.1 Genes and CpG islands

To assess the influence of genes and CpG islands on recombination I estimated the fine-scale structure of recombination rates around these features. Previous studies in humans have shown that the profile of recombination at genes is suppressed in transcribed regions, but elevated at the transcription start and end sites (Consortium, 2007; Myers et al., 2005; Kong et al., 2010). I wanted to determine whether chimpanzees share this profile and if this signal is explained by CpG content.

Given the limitations in the chimpanzee reference assembly, transcript locations were obtained from the Ensembl gene set for the autosomes of the human reference build 36 totalling 58,159 transcripts (Hubbard et al., 2002, Release 63). I required that gene coordinates mapped reciprocally between chimpanzee and human references, I removed exact duplicates and partially overlapping transcripts that had opposing strand orientations, and thinned the remaining partially overlapping transcripts by selecting one from each of the overlapping transcripts uniformly. This process left a total of 15,241 non-overlapping transcripts. I then calculated recombination rates over non-overlapping 500 bp intervals between the 5' transcription start site (TSS), transcript, and the 3' end (TES) of the thinned transcripts thereby uniformly av-

eraging over the remaining transcripts excluded through thinning. Intervals were then ordered by whether they reside inside transcripts and are nearer the 5' TSS or 3'end, or upstream or downstream of their nearest 5 TSS or 3 TES, and then by their distance from this nearest TSS/TES with respect to the strand orientation of the transcript.

I then plotted the average recombination rate as a function of distance from TSS/TES, over bins around the TSS or TES selected to each contain 2.5% of observations in order to reduce the variance in the profile. For each bin, I plotted the relevant mean distance on the x-axis versus the average recombination rate on the y-axis and its associated 95% confidence interval (Figure 4.7a). The number of intervals at any location decreases as a function of distance from the nearest TSS/TES with a maximum of 15,241 at the TSS/TES, a minimum of approximately 2,945 at a distance of 100 kb from the TSS/TES, and a minimum of 1,656 30 kb from the TSS/TES inside transcripts.

This revealed a profile of recombination where rates are on average lower within genes and increase with distance from genes for $\sim$ 30 kb before decreasing to genome average levels (Figure 4.7a). In both species recombination seems to preferentially occur near genes (within 50 kb) but outside of the transcribed domain. Recombination increases on average by about 20% around transcription start and end sites and decreases on average by about 30% within the transcribed region .

Possibly reflecting a similar effect, I found recombination to be elevated around CpG islands (CGI; Gardiner-Garden and Frommer, 1987) in both species (Figure 4.7b), although the effect is stronger in chimpanzees (increase of nearly 50% in rate relative to background compared to 15% in humans). Such concordance suggests that features affecting chromatin state (for example, nucleosome occupancy) which is destabilised around CpG islands and promoters (Ramirez-Carrozzi et al., 2009) may similarly shape the propensity for recombination at these sites in humans and

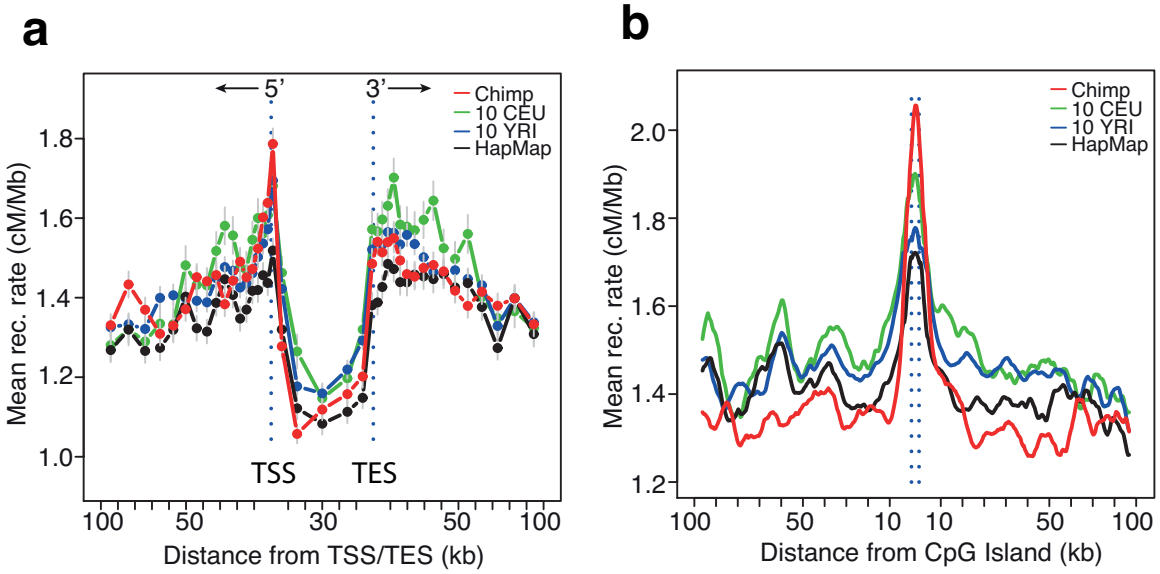


Figure 4.7: The fine-scale profile of recombination rate variation around genomic features in chimpanzees and humans. **(a)** Average recombination rate as a function of distance to nearest transcription start site (TSS) and transcription end site (TES) in chimpanzee (red), YRI (blue), CEU (green), and HapMap (black) for 15,241 homologous transcripts across autosomes. Grey bars represent 95% CI, each point represents the average recombination rate over 2.5% of the recombination intervals (see text for details). **(b)** Average recombination rate as a function of distance to nearest CpG island (N = 19,809); colours as for panel (a). Dashed lines indicate start and end of elements, estimates smoothed using running average with a 7.5 kb window.

chimpanzees (Smagulova et al., 2011; Pan et al., 2011; Petes, 2001).

Why might recombination be targeted towards promoter regions despite the potential risk of disrupting the function of downstream genes? Approximately 70% of annotated gene promoters are associated with a CGI, making this feature the most common promoter type in the vertebrate genome (Saxonov et al., 2006) and it is likely that all CGIs are sites of transcription initiation. Vertebrate CGIs are short interspersed DNA sequences that deviate significantly from the average genomic pattern by being GC-rich, CpG-rich, and predominantly nonmethylated. These shared properties have allowed CGIs to be isolated as a relatively homogeneous fraction of the genome despite variation in their individual nucleotide sequences (Bird et al., 1985), and therefore potentially provide a common, constant, and dispersed set of

sequences. Moreover, the same properties of CGIs that enable promoter function, destabilising nucleosomes and attracting proteins that create a transcriptionally permissive chromatin state, are compatible with the factors known to be important in recombination (Petes, 2001).

Normally, CpG dinucleotides are rare in vertebrate DNA because the cytosine tends to be methylated. However, over evolutionary time, methylated cytosines mutate to thymidine nucleotides through the process of spontaneous deamination. There are approximately 80k hypomethylated regions in human and chimpanzee sperm with a mean size of 1.8 kb (Molaro et al., 2011). Given its importance in mitosis, evidence for suppressed recombination through methylation (in *Ascobolus*; Colot and Rossignol, 1999), and observations that spermatogenesis in *Dnmt3L*$^{-/-}$ mice terminates before pachytene stage and is characterised by widespread non-homologous synapsis and aberrant synaptonemal complex structures (Bourc'his and Bestor, 2004), one might ask whether Cytosine methylation impacts recombination activity in primates. To investigate the impact of Cytosine methylation on recombination at promoters I analysed genome-wide Cytosine methylation profiles of human and chimpanzee sperm samples from Molaro et al. (2011), and used these tracks to stratify transcripts by their mean Cytosine methylation proportion (as measured over a 1 kb window surrounding their TSS). Note that these data have their limitations since Cytosine methylation patterns vary dramatically during gametogenesis (Sanford et al., 1987), and preimplantation development. Nevertheless, the methylation profiles in sperm may potentially reflect the meiotic process. The methylation profile in sperm is different than other tissue considered - an embryonic stem cell (ESC) line. The methylation profiles between sperm samples in each species were highly consistent and are enriched for gene ontologies associated with meiosis. Additionally, the inter-species correlation between sperm sample profiles exceeded the intra-species profiles between sperm and ESC; the human-chimpanzee sperm profile correlation at promoters is greater than

91%, versus less than 73% for comparisons of sperm and ESC promoter profiles within species (Molaro et al., 2011).

Average recombination profiles were calculated for those transcripts with low levels of '5 TSS methylation (bottom 50th percentile across the thinned transcripts), and separately for those with high levels of methylation (top 50th percentile). In humans recombination activity tended to be higher at genes that have a high rate of CpG methylation in sperm. This is surprising given that nucleotide methylation has been demonstrated to impair Cys2His2 ZF protein binding in vitro (Zweidler-Mckay et al., 1996). In contrast, in chimpanzees, activity tended to be higher around genes with low rates of sperm CpG methylation. This revealed discordant structure between human and chimpanzee recombination profiles at genes (Figure 4.8).

I also assessed recombination profiles at independently identified CpG regions with little, or no, methylation identified by capture-sequencing on genomic material from human liver and testes using an engineered KDM2B histone H3 lysine 36 demethylase with biotin moiety (Blackledge et al., 2012, data provided by Rob Klose, Hanna Long and Dave Sims). The average recombination profile at these hypomethylated CpG islands (HMRs) shows the best recombination localisation signal around any feature in chimpanzees with an increase over background recombination rates of over 170%, compared to the nearly 50% increase at CpG islands (Figure 4.9b versus Figure 4.7b). Together, these results are compatible with chromatin state influencing recombination in both species, with greater impact on chimpanzee recombination.

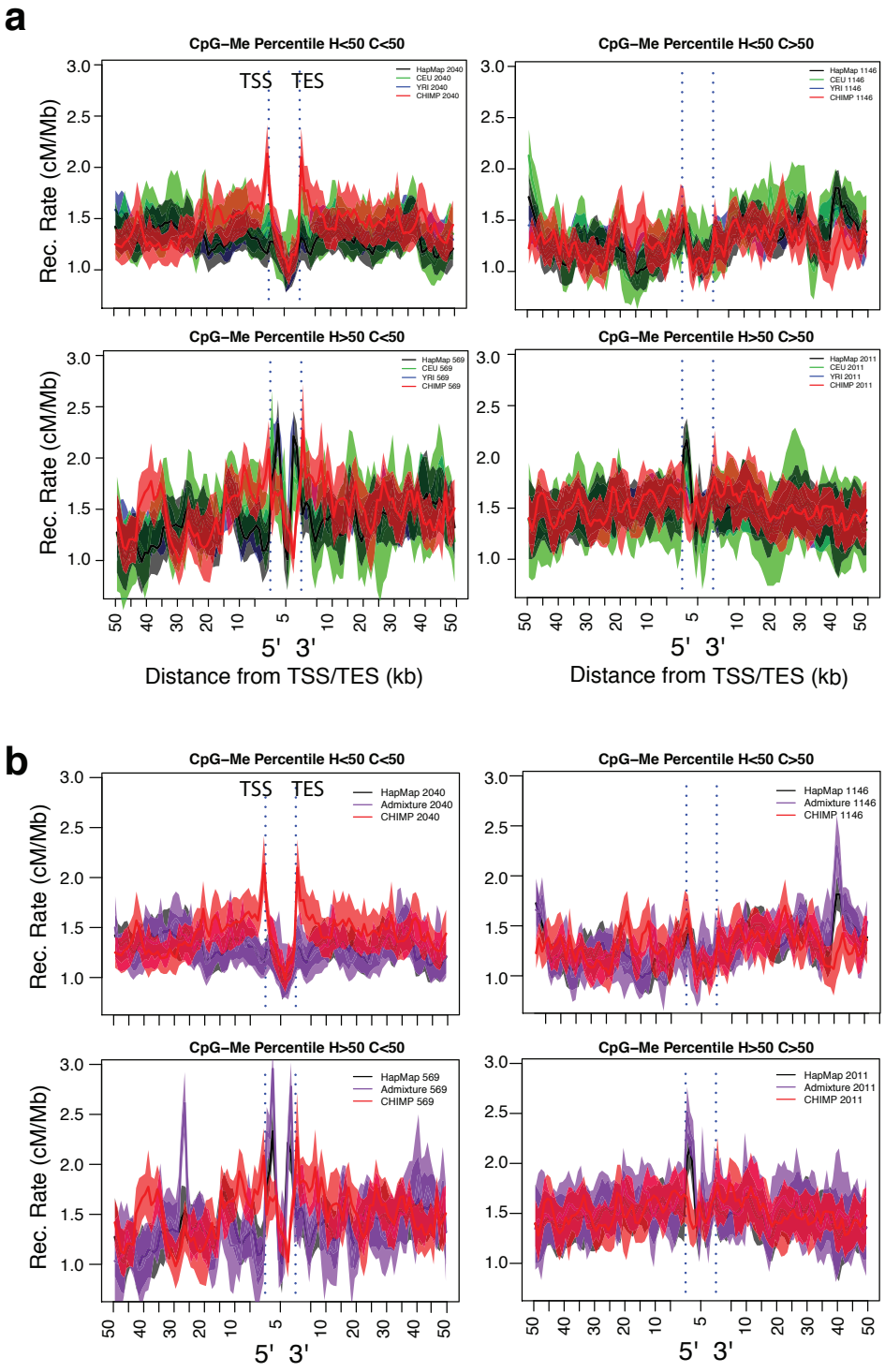


Figure 4.8: The fine-scale profile of recombination variation at genes in human and chimpanzee stratified by Cytosine methylation proportion measured in sperm (Molaro et al., 2011) over the 1 kb surrounding the transcription start sites (TSS). **(a)** Average recombination profile at genes for combined population HapMap genetic map (black), 10 CEU (green), 10 YRI (blue), and chimpanzee (red). **(b)** Average recombination rate profile at genes for combined population HapMap (black), from admixture (purple), and chimpanzee (red). Coloured bands represent 95% confidence intervals. The number of transcripts in each class is indicated in the legend. Methylation percentiles indicated in figure headings for human (H) and chimpanzee (C).

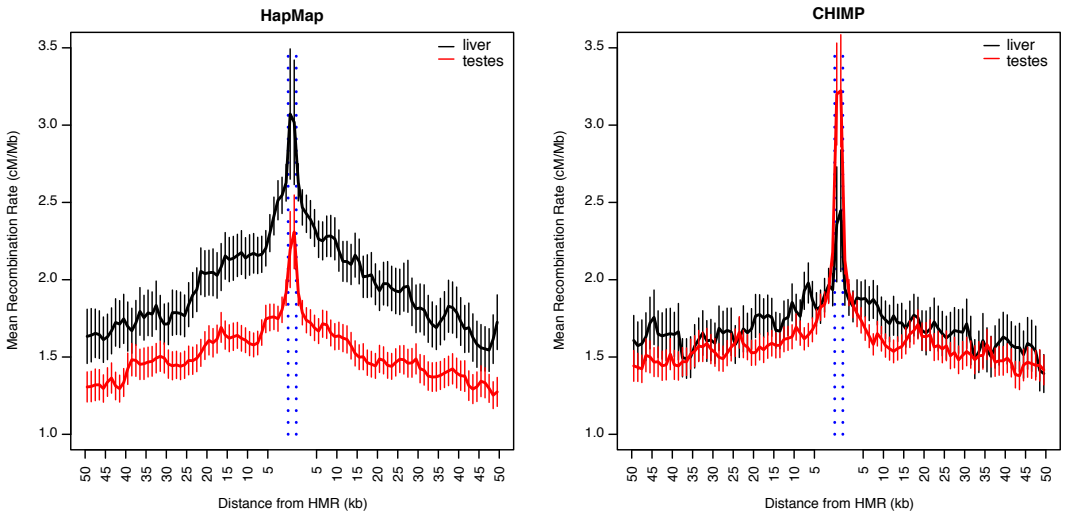


Figure 4.9: The fine-scale profile of recombination rate variation around hypomethylated regions (HMRs) in human testes (red, N = 13,731) and liver (black = 8,483) samples for **(a)** human recombination rates (HapMap combined-population map) and **(b)** chimpanzee recombination rates.

4.4.2 DNA repeat elements

A natural approach to study the impact of sequence context on recombination is the evaluation of fine-scale rates at DNA repeat elements. In humans, recombination has different activity on repeat backgrounds (Myers et al., 2008). Despite the potential fitness cost of permitting recombination at repeat elements through non-allelic homologous recombination, cross-over rates at some repeat elements such as THE1B and Alu elements are elevated. McVean (2010) proposed that short repeat sequences, particularly families of extinct transposable elements, may provide a common, dispersed, and non-functioning source of potential hotspot motifs, and that as the selective pressure to acquire hotspots becomes large, the cost associated with NAHR-mediated mutation arising from using motifs present in repeat DNA is overwhelmed.

To evaluate cross-over activity at transposable elements in chimpanzees, I calculated the fine-scale structure of rate variation on different repeat element backgrounds.

I obtained repeat element annotations from the UCSC genome browser database for individual repeat elements (as characterised by RepeatMasker) and tandem repeat elements (as characterised by Tandem Repeats Finder). I thinned the elements (described in methods 4.1) and removed classes that had less than 200 occurrences in syntenic regions after thinning. I calculated the average recombination rate over 200 bp intervals for the 20 kb around elements and calculated the ratio of the peak rate to background rate after fitting a Gaussian profile (Section 4.1). To filter out repeat elements that had low signal-to-noise profiles, I required that the profile fit had a Pearson R^2 value greater than 0.55 for either the robust chimpanzee genetic map, or the combined population HapMap genetic map.

Repeat element families showed no overall correlation in recombination-localising activity between humans and chimpanzees (Pearson $r = 0.16$, $p = 0.22$, Figure 4.10a). The strongest activating repeats in humans (LTR49, THE1A, and THE1B), which

all contain the human PRDM9 A-allele 13 bp binding motif CCTCCCTNNCCAC, suppress recombination in chimpanzees (Figure 4.10b, top). A second class of elements, typically low-complexity (CT-rich, GA-rich, and G-rich) was found to be weakly activating in both species (Figure 4.10b, middle left), whereas a few elements (e.g. L1PA2) suppress recombination in both species (Figure 4.10b, middle right). Only a few elements (notably $(GGAA)_n$ and MER92B elements) showed activation only in chimpanzees (Figure 4.10b bottom and Figure 4.11), and very few elements showed strong recombination activity in chimpanzees.

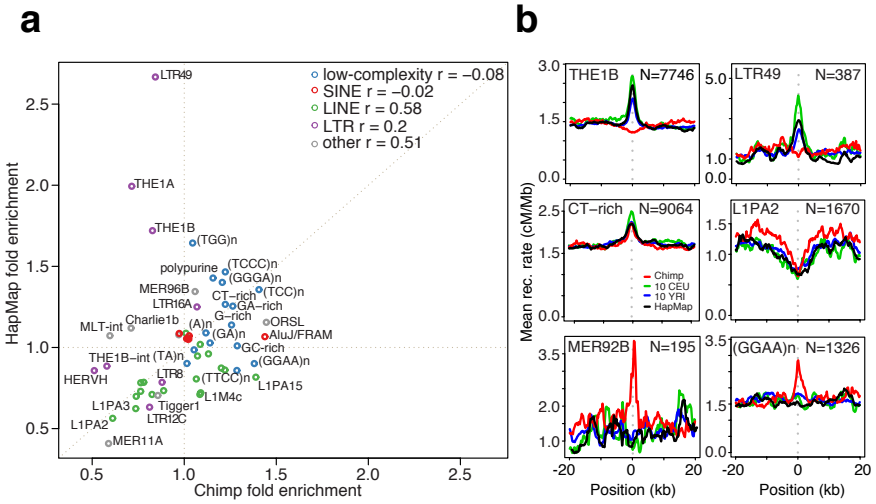
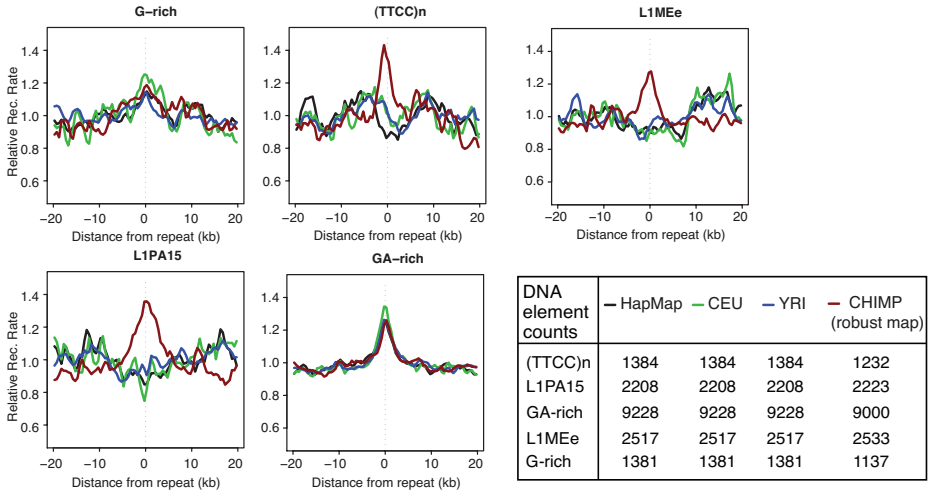


Figure 4.10: Recombination rates around DNA repeat elements in chimpanzees and humans. **(a)** Recombination-influencing activity of repeat element families in chimpanzees and humans (HapMap). The value reported is the ratio of the peak rate to background rate, as estimated from the robust genetic map after fitting a Gaussian profile using maximum likelihood (Sections 4.1.2 and 2.2.2.3). Selected repeat elements are labelled. (b) Recombination rate profiles around selected repeat elements, as estimated in chimpanzees. Top: two elements (THE1B and LTR49) that are recombination-promoting in humans only. Middle: Elements that are recombination-promoting (CT-rich repeats) or recombination suppressing (L1PA2) in both humans and chimpanzees. Bottom: Two elements ($(GGAA)_n$ and MER92B) that are recombination-promoting in chimpanzees only.



DNA element counts	HapMap	CEU	YRI	CHIMP (robust map)
(TTCC)n	1384	1384	1384	1232
L1PA15	2208	2208	2208	2223
GA-rich	9228	9228	9228	9000
L1MEe	2517	2517	2517	2533
G-rich	1381	1381	1381	1137

Figure 4.11: Recombination rate estimates around selected DNA repeats that have signals of recombination activity in chimpanzees. Estimates are shown for chimpanzee (dark red) as estimated in the robust map (see 2.2.2.3), HapMap (black), 10 CEU (green), and 10 YRI (blue). Also shown is a table of the repeat element counts used to calculate average recombination rates. Additional repeat elements can be viewed online at `http://www.well.ox.ac.uk/~oliver/repeat_rates/repeat_rates.html`

4.5 Detecting DNA sequence motifs associated with recombination

I used two orthogonal approaches to search for DNA sequence motifs associated with recombination in Chimpanzees. First, I used a *de novo* motif search to assess the impact of DNA motifs on the recombination profile at DNA repeat elements. Second, I looked for signals of recombination activity at matches to the Position Weight Matrix (PWM) motif predictions for the DNA sequence bound by chimpanzee PRDM9 alleles.

4.5.1 Impact of DNA motifs on recombination in DNA repeats

A major challenge in testing DNA motifs for association with high recombination rates is controlling local sequence context across comparisons. Fortunately, DNA re-

peat elements can be used as a constant background. Such approaches were used by Myers et al. (2005) to identify the core 7-bp human recombination motif and its extended motif (Myers et al., 2008) associated with the PRMD9 A-allele. However, in human populations with high allelic diversity, such as Africans, the approach did not identify the African/C-allele motif potentially due to averaging of A-allele and C-allele maps (see Figure 4.20c,d for motivation for why this is difficult). Indeed, the African motif was only discovered by testing for motifs enriched in African-specific cross-overs (Hinch et al., 2011). To look for recombination-activating motifs, I calculated the association of 4-9 bp DNA motifs with recombination rates at major repeat element classes, individual elements, and simple tandem repeats as annotated by RepeatMasker (`http://www.repeatmasker.org`) for the human reference (B36) and chimpanzee reference (panTro2). Repeats with less than 200 thinned occurrences in synteny regions were excluded.

To assess the impact of motifs on repeat backgrounds, I first thinned elements to avoid non-independence between motif counts, then I ranked elements by their genetic distance as measured across their central 5 kb interval and identified those elements with high recombination (in the top 10^{th} percentile, referred to as the *hot class*) and low recombination (in the bottom 50^{th} percentile, referred to as the *cold class*). I recorded the number of motif occurrences for each class by counting at most one instance of the kmer on each tested element in the class. To the counts of motifs observed, and the total number of kmers tested in each class at that length scale, I applied a one-sided Fisher exact test and calculated the relative risk (RR). To guard against false positives arising from systematic length differences between hot and cold element classes (i.e. identifying differential DNA sequence between hot and cold classes due to length differences), I tested for significant length differences through the Wilcoxon rank-sum test and excluded elements where length differences were significant. Figure 4.12 demonstrates that elements in the L1P family (but not

all) have strong systematic length differences, compatible with non-retroviral retrotransposition. Indeed, it should be noted that element-based FET significance tests will also be affected by systematic length differences. Fortunately, only a minority of elements are affected by length differences and therefore we restrict analyses to element backgrounds where the null cannot be rejected.

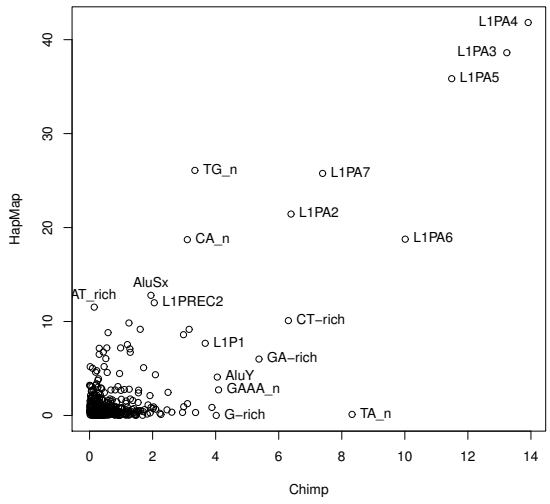


Figure 4.12: P-values for systematic length differences between hot and cold element backgrounds. Presented are the P-values for hot and cold repeat element length distributions for Chimpanzee (x-axis) and HapMap (y-axis). L1 elements are enriched for length differences.

I then prioritised motifs by their p-value on motif enrichment (Bonferroni corrected for the number of distinct motifs tested at that length scale) and their RR, and looked for overlap between motifs on the assumption that signals for cross over activity would be shared across motifs and sub-motifs. The top motif candidates with RR > 1 and length p-value > 0.05 are presented in Table 4.2. To evaluate the biological consistency of the motif candidates, I measured the number of nucleotides changed between the motif candidate and the consensus reference sequence by mapping potential motifs to their consensus sequence using the Stampy mapper (Lunter and Goodson, 2010).

Three CT-rich/CG-rich motif candidates were observed. The first candidate was a 10 bp motif CTCTCNTTCT which matched exactly on AluSq and AluJo backgrounds and had a one base change on the THE1B background (Figure 4.13). The sharing across multiple backgrounds can be viewed as independent replication of the putative signal, as are the non-overlapping hits on THE1B conditional on the background.

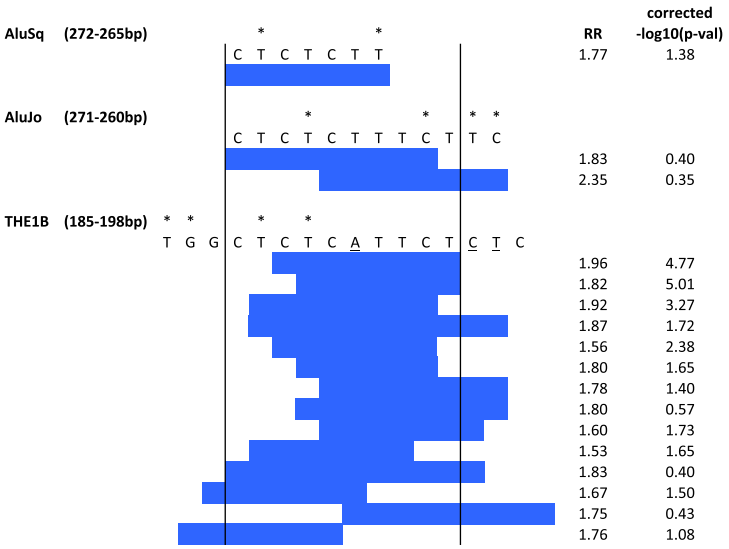


Figure 4.13: Recombination-activating motif candidates on AluSq, AluJo and THE1B repeat backgrounds. Coordinates represent location of the extended motif. The extended motif sequence is listed for each background, with underlined bases representing mismatches between motifs discovered on the three backgrounds, and stars representing mismatches with the element consensus sequence. Note that base changes away from the reference sequence are shared on two backgrounds at position 2 and 4 in the core 10 bp shared motif. Blue rectangles represent the location of the motif hit in the extended motif. Relative risks (RR) and Bonferroni-corrected p-values are listed for each hit. Note the motifs are located on the reverse complement of the consensus sequence.

The second candidate is a degenerate CG-rich motif enriched on AluJo and L2. In humans motifs on the L2 background have been observed to be degenerate (Simon Myers, personal communication). The CG-rich signal is a wider recombination localisation feature in chimpanzees. When length restrictions on motifs were relaxed, motifs with high GC fraction and CpG dinucleotide content were observed to lead to local rate increases in chimpanzees. For example, on Alu elements, especially AluY

(length p-value = 8.9×10^{-5}), the motif CGGGCGC showed significant hotspot enrichment ($p_{corrected} = 2 \times 10^{-4}$, RR = 1.2), but the effect was better explained by CpG content within Alu elements (Figure 4.15). The third and apparently distinct candidate is an extended motif, TCCTCACCCT, that shows localising activity on L1PA15 (Table 4.2).

		RR	corrected -log10(p-val)
AluJo	C C C G G C	1.33	3.04
L2	C C C A G G C	1.35	3.16
	C C C T C C A G G	2.08	1.05
	C C C C A G G C	1.63	0.78
	A G G C C C C	1.40	1.07
	C C G C A G G	2.08	1.18
	C T C C A G G	1.23	1.29
	C A G G A C G	1.98	1.49
	G G A G A C G	1.74	1.53

Figure 4.14: Recombination-activating motif candidates on AluJo and L2 repeat backgrounds. Motifs are stacked to highlight overlaps, underlined bases represent mismatches between the motifs. Relative risks (RR) and Bonferroni-corrected p-values are listed for each hit.

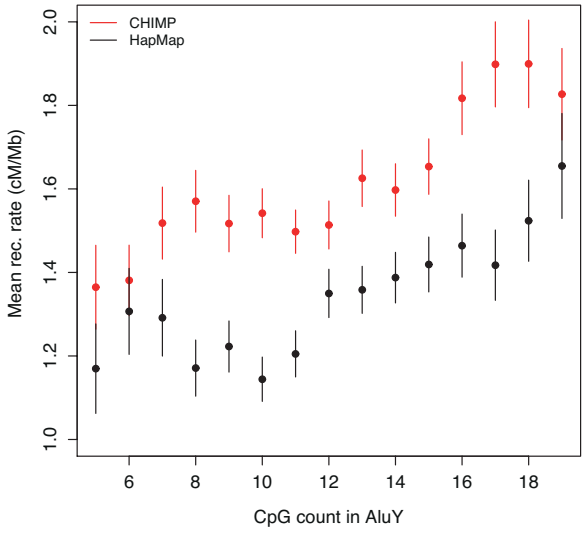


Figure 4.15: Recombination rate in AluY elements as a function of CpG count within the repeat sequence. Mean rate estimates are shown for chimpanzee (red), and HapMap (black), with error bars representing 95% CI.

To investigate the veracity of the candidates we compared the motifs to other analyses. First, we note that the three motifs are compatible with the best ranking

cross-over localising activity for $(GGAA)_n$, L1PA15 and AluJ/FRAM in chimpanzees (Figure 4.10), which is consistent with the signals of activation in human repeat backgrounds that contain the human recombination motif. Second, there is no clear match in the motifs to chimpanzee PRDM9 predictions (though it may be argued TCCTCACCCT matches the W2-allele motif prediction at 20 bp to 29 bp [Figure 4.17], but the allele has an estimated frequency of only 2.1% suggesting that it would have a small impact on shaping motif activity). However, the lack of agreement between the motif candidates and motif predictions from *PRDM9* sequencing, potentially suggests inaccuracy in chimpanzee PRDM9 motif predictions. Third, we note that the overlap with motifs found in non-repeat DNA is weak though suggestive (see reproduced Supplementary Table 8 from (Auton et al., 2012) at page 173). Indeed the weak evidence for motifs in non-repeat DNA may indicate that several motifs operate in chimpanzees. For example, the mismatches evident between THE1B and AluJo/Sq suggest that signal may be lost by combining motifs together. In general, of those activating motifs, values of RR were observed for the motif tests were lower than in humans. Analysing the human data identified the known 7mer recombination motif, CCTCCCT (Myers et al., 2005), on THE1B with $p_{\text{corrected}} < 10^{-19}$ with RR of 2.99 (Figure 4.16). The fold enrichment around motif candidates was also weak in chimpanzees. For example, the 7 bp motif and background combinations showing the largest effect in humans was CTCCCTA on THE1B which had a fold increase of 3.92, whereas the strongest signal in chimpanzee was CCGCAGG on L2 which had a fold increase of 2.08. We therefore conclude that in chimpanzees there is evidence for fine-scale DNA sequence motifs associated with recombination, which are likely CT-rich motifs, however there is little evidence for a single DNA motif (like the Myers motif in human). Another likely possibility is that there may exist longer DNA motifs that localise recombination events in chimpanzee.

Supplementary Table 8 - Reproduced from Auton et al. (2012) with permission

Motifs showing enrichment in hotspots

Motif Length	Motif	Rev Comp	Hotspots w/ Motif	Hotspots w/o Motif	Coldspots w/ Motif	Coldspots w/o Motif	Relative Risk	p-val (All)	p-val (Repeats)	p-val (NonRepeat)
4	CGCG	CGCG	1369	3669	1046	3992	1.42	4.77E-14	3.09E-09	1.79E-05
4	GCGC	GCGC	2632	2406	2393	2645	1.21	1.92E-06	0.000758	0.000927
5	TCGAA	TTCGA	2214	2824	1956	3082	1.24	1.80E-07	0.000219	0.000248
5	ACCCG	CGGGT	2410	2628	2134	2904	1.25	3.28E-08	0.000548	0.000292
5	CCGCG	CGCGG	861	4177	639	4399	1.42	5.19E-10	1.39E-07	0.000319
5	CGCGC	GCGCG	949	4089	706	4332	1.42	6.41E-11	1.32E-07	0.000811
6	CCCGGC	GCCGGG	1648	3390	1355	3683	1.32	1.75E-10	7.58E-06	6.98E-06
7	ACCTCGT	ACGAGGT	702	4336	525	4513	1.39	6.97E-08	7.17E-05	0.000132
8	AAAAAGAA	TTCTTTTT	2812	2226	2537	2501	1.25	4.03E-08	8.76E-05	0.000274
9	AAGAAGGAA	TTCCTTCTT	623	4415	379	4659	1.73	4.44E-16	2.22E-16	0.000698
9	CTTTCCTTC	GAAGGAAAG	459	4579	285	4753	1.67	3.39E-11	5.17E-08	0.000398
10	AAAGAAGGAA	TTCCTTCTTT	346	4692	173	4865	2.07	6.33E-15	3.91E-14	0.000749
11	GAAGGAAAGAA	TTCTTTCCTTC	175	4863	71	4967	2.52	1.90E-11	8.19E-09	0.000494

Motifs with $p < 0.001$ in both repeat and non-repeat DNA
P-values shown before Bonferroni correction

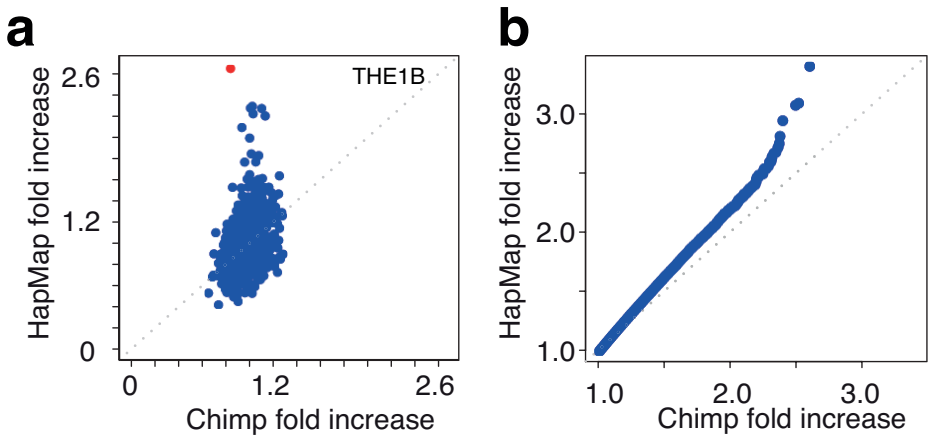


Figure 4.16: Motif-based recombination-activity is systematically much stronger in humans than chimpanzees. **(a)** Recombination-activating effect of 7-mer motifs with greater than 100 occurrences on the THE1B background in humans and chimpanzees calculated by fitting a Gaussian curve with maximum likelihood. The motif CCTC-CCT (Myers et al., 2005), known to be part of a larger PRDM9-binding motif in humans, is shown in red. (b) Quantile-quantile plot comparing the recombination-activating effects of motifs on different repeat element backgrounds in humans and chimpanzees (estimated fold increase associated with presence of motif on a particular background).

Table 4.2: Candidate recombination-activating DNA sequence motifs on DNA repeat backgrounds in chimpanzees. Presented are the top ranking motifs (lengths 4 to 9 bp), ordered by their Bonferroni-corrected p-value (FET) each with relative risks (RR) greater than 1. To guard against false positives arising from systematic length differences between hot and cold backgrounds, only elements with $P > 0.05$ (Wilcoxon rank-sum test) between the hot and cold length distributions were interpreted.

Element	Kmer size	P-value on length	Sequence	RevComp	Hot Kmers	Cold Kmers	Hot Motif	Cold Motif	RR	-$\log_{10}(P)$	-$\log_{10}(P)$ corrected
L2	4	0.61	CCCG	CGGG	847843	4238568	2033	8104	1.25	18.38	15.00
L2	4	0.61	CCGC	GCGG	847843	4238568	1548	6135	1.26	14.86	11.50
L2	4	0.61	ACGG	CCGT	847843	4238568	2273	9397	1.21	14.86	11.50
L2	4	0.61	CGGC	GCCG	847843	4238568	1509	5964	1.26	14.82	11.40
L2	4	0.61	CGTC	GACG	847843	4238568	1915	7808	1.23	14.34	10.90
L2	4	0.61	ACCG	CGGT	847843	4238568	1769	7159	1.24	14.19	10.80
L2	4	0.61	CGCC	GGCG	847843	4238568	1637	6599	1.24	13.66	10.30
L2	4	0.61	CCGG	CCGG	847843	4238568	1009	3858	1.31	12.80	9.39
L2	4	0.61	CACG	CGTG	847843	4238568	2672	11500	1.16	11.21	7.80
L2	4	0.61	AACG	CGTT	847843	4238568	2286	9814	1.16	9.99	6.58
L2	4	0.61	CGCA	TGCG	847843	4238568	1703	7147	1.19	9.76	6.35
L2	4	0.61	CCGA	TCGG	847843	4238568	1709	7200	1.19	9.38	5.98
L2	4	0.61	AGCG	CGCT	847843	4238568	2013	8640	1.17	8.92	5.51
AluJo	4	0.87	CCGG	CCGG	470971	2364006	1030	4197	1.23	8.39	4.98
L2	4	0.61	GCGC	GCGC	847843	4238568	666	2563	1.30	8.35	4.94
L2	5	0.61	CCACG	CGTGG	1505334	7584909	1081	4117	1.32	14.77	10.80
L2	5	0.61	CCGTC	GACGG	1505334	7584909	672	2370	1.43	14.55	10.50
L2	5	0.61	CACGG	CCGTG	1505334	7584909	983	3772	1.31	12.89	8.88
L2	5	0.61	CACCG	CGGTG	1505334	7584909	756	2833	1.34	11.65	7.64
L2	5	0.61	CGGGC	GCCCG	1505334	7584909	536	1896	1.42	11.59	7.58
L2	5	0.61	CGGGA	TCCCG	1505334	7584909	787	2974	1.33	11.55	7.54
L2	5	0.61	CGGCC	GGCCG	1505334	7584909	499	1741	1.44	11.54	7.53
L2	5	0.61	CCCCG	CGGGG	1505334	7584909	802	3041	1.33	11.52	7.51

Element	Kmer size	P-value on length	Sequence	RevComp	Hot Kmers	Cold Kmers	Hot Motif	Cold Motif	RR	-$\log_{10}(P)$	-$\log_{10}(P)$
L2	5	0.61	CGCAG	CTGCG	1505334	7584909	716	2668	1.35	11.51	7.50
L2	5	0.61	CCGCC	GGCGG	1505334	7584909	597	2167	1.39	11.20	7.19
L2	5	0.61	ACGGC	GCCGT	1505334	7584909	561	2021	1.40	11.04	7.03
L2	5	0.61	CCCGG	CCGGG	1505334	7584909	611	2240	1.37	10.88	6.87
L2	5	0.61	CCGCA	TGCGG	1505334	7584909	579	2121	1.38	10.36	6.35
L2	5	0.61	AGGCG	CGCCT	1505334	7584909	762	2916	1.32	10.32	6.31
L2	5	0.61	CGTCC	GGACG	1505334	7584909	613	2271	1.36	10.24	6.23
L2	6	0.61	CCCAGG	CCTGGG	2010203	10189147	1651	7019	1.19	9.57	4.96
L2	6	0.61	CGTCCC	GGGACG	2010203	10189147	211	643	1.66	9.04	4.43
L2	6	0.61	CCCACG	CGTGGG	2010203	10189147	312	1053	1.50	8.91	4.30
L2	6	0.61	CCGCAG	CTGCGG	2010203	10189147	218	678	1.63	8.75	4.14
L2	6	0.61	CGTGGC	GCCACG	2010203	10189147	312	1059	1.49	8.68	4.07
L2	6	0.61	CCCAGC	GCTGGG	2010203	10189147	1745	7526	1.18	8.66	4.04
L2	6	0.61	CCCGTC	GACGGG	2010203	10189147	224	705	1.61	8.59	3.97
L2	6	0.61	CCACCG	CGGTGG	2010203	10189147	249	810	1.56	8.35	3.74
L2	6	0.61	CAGGGG	CCCCTG	2010203	10189147	1450	6176	1.19	8.32	3.71
L2	6	0.61	CGGGAG	CTCCCG	2010203	10189147	306	1055	1.47	7.97	3.36
L2	6	0.61	GGCCCC	GGGGCC	2010203	10189147	762	3049	1.27	7.93	3.32
AluJo	6	0.87	CCCGGC	GCCGGG	1057422	5339046	510	1933	1.33	7.65	3.04
L2	6	0.61	CGCAGC	GCTGCG	2010203	10189147	189	589	1.63	7.63	3.02
L2	6	0.61	CGGCTC	GAGCCG	2010203	10189147	196	617	1.61	7.57	2.95
L2	6	0.61	CCTCCC	GGGAGG	2010203	10189147	1968	8668	1.15	7.53	2.92
THE1B	7	0.13	AGAATGA	TCATTCT	275097	1457378	178	517	1.82	10.22	5.01
L2	7	0.61	CCCAGGC	GCCTGGG	2223343	11289462	518	1950	1.35	8.37	3.16
L2	7	0.61	CCCCAGG	CCTGGGG	2223343	11289462	562	2155	1.32	8.12	2.91
THE1B	7	0.13	CTCATTC	GAATGAG	275097	1457378	223	759	1.56	7.59	2.38
AluJo	7	0.87	AGCCGGG	CCCGGCT	1147100	5792942	265	903	1.48	7.21	2.00
THE1B	7	0.13	CATTCTC	GAGAATG	275097	1457378	179	591	1.60	6.94	1.73

Element	Kmer size	P-value on length	Sequence	RevComp	Hot Kmers	Cold Kmers	Hot Motif	Cold Motif	RR	-$\log_{10}(P)$	-$\log_{10}(P)$
THE1B	7	0.13	AATGAGA	TCTCATT	275097	1457378	212	733	1.53	6.87	1.65
L2	7	0.61	CGTCTCC	GGAGACG	2223343	11289462	131	382	1.74	6.74	1.53
THE1B	7	0.13	GCTCTCA	TGAGAGC	275097	1457378	151	480	1.67	6.72	1.50
L2	7	0.61	CAGGACG	CGTCCTG	2223343	11289462	89	228	1.98	6.71	1.49
AluSq	7	0.10	AAGAGAG	CTCTCTT	954666	4764704	121	341	1.77	6.59	1.38
L2	7	0.61	CCTGGAG	CTCCAGG	2223343	11289462	789	3257	1.23	6.50	1.29
L2	7	0.61	CCGCAGG	CCTGCGG	2223343	11289462	76	186	2.08	6.39	1.18
THE1B	7	0.13	GAGAGCC	GGCTCTC	275097	1457378	117	353	1.76	6.29	1.08
L2	7	0.61	AGGCCCC	GGGGCCT	2223343	11289462	303	1098	1.40	6.29	1.07
THE1B	8	0.13	AGAATGAG	CTCATTCT	284444	1486387	151	403	1.96	10.59	4.77
THE1B	8	0.13	GAATGAGA	TCTCATTC	284444	1486387	135	367	1.92	9.09	3.27
L1PA15	8	0.49	GGTGAGGA	TCCTCACC	233572	1130798	83	183	2.20	7.73	1.91
THE1B	8	0.13	GAGAATGA	TCATTCTC	284444	1486387	131	381	1.80	7.46	1.65
THE1B	8	0.13	AGAGAATG	CATTCTCT	284444	1486387	132	388	1.78	7.22	1.40
L2	8	0.61	CCCCAGGC	GCCTGGGG	2298349	11672749	161	503	1.63	6.59	0.78
THE1B	8	0.13	ATTCTCTC	GAGAGAAT	284444	1486387	118	352	1.75	6.24	0.43
AluJo	8	0.87	AAAGAGAG	CTCTCTTT	1186772	5981121	102	281	1.83	6.21	0.40
AluJo	8	0.87	CTTTCTTC	GAAGAAAG	1186772	5981121	56	120	2.35	6.17	0.35
L1PA15	8	0.49	AGGGTGAG	CTCACCCT	233572	1130798	82	200	1.99	6.15	0.33
L1PA15	8	0.49	CCTCACCC	GGGTGAGG	233572	1130798	76	181	2.03	6.11	0.29
THE1B	8	0.13	ATGAGAGC	GCTCTCAT	284444	1486387	133	416	1.67	6.05	0.23
AluJo	8	0.87	CCGGCTAA	TTAGCCGG	1186772	5981121	165	529	1.57	5.96	0.14
THE1B	8	0.13	GAGAGCCA	TGGCTCTC	284444	1486387	97	279	1.82	5.83	0.02
THE1B	9	0.13	AGAATGAGA	TCTCATTCT	283087	1405001	129	342	1.87	8.14	1.72
L2	9	0.61	CCCTCCAGG	CCTGGAGGG	2311476	11700759	90	219	2.08	7.46	1.05
THE1B	9	0.13	AGAGAATGA	TCATTCTCT	283087	1405001	123	340	1.79	6.99	0.57
THE1B	9	0.13	CTCATTCTC	GAGAATGAG	283087	1405001	114	310	1.82	6.80	0.38

4.5.2 Refinement of predicted PRDM9 binding motifs

To assess whether any of the PRDM9 variants localised recombination events, I applied the bioinformatic model of Persikov et al. (2009) and used the pre-trained polynomial model to infer the predicted DNA sequence bound by each of the 11 Western chimpanzee, 2 Eastern chimpanzee, 1 Bonobo, and 2 human binding variants (Figure 4.17).

The amino acid sequence for PRDM9 alleles were inferred from sequenced alleles. The zinc finger (ZF) array of 48 Western chimpanzee PRDM9 alleles, including alleles from the ten HTS genomes, were PCR amplified, cloned, and sequenced independently by collaborators at the University of Chicago and the University of Oxford (Auton et al., 2012). Fifteen unique allelic variants were found corresponding to 11 binding variants, as defined by the three positions making direct contact with the primary DNA strand and one position interacting with the complementary strand. These sequences not only varied greatly in their number of ZFs (6 to 18), but also at the identity of their DNA contacting residues (Figure 4.19a).

For chimpanzee and bonobo sequences, the model predicted that the first ZF in the array was unlikely to bind DNA and were therefore excluded from the ZF array for subsequent binding predictions. In contrast, full-length ZF sequences were used to generate PWMs of human PRDM9 alleles. Logo plots for the ZF binding motifs were generated using the R-package seqLogo (`http://bioconductor.org/biocLite.R`). Figure 4.17 shows the PWMs for the Western chimpanzee PRDM9 variants as observed in 48 chimpanzee sequences (W1 - W11). Also shown, are PWMs of two variants (A1 and E1) present at high frequency in Eastern chimpanzees, but absent from all 48 sequences obtained from Western chimpanzees. One of these alleles (A1) was also observed in Bonobo, and hence is a potential ancestral variant. Finally, the PWMs of the major Human variants (A- and C-allele (Berg et al., 2010, 2011)) are also shown. Note that the chimpanzee predictions have lower total information

than the human predictions; 20% of the positions in the chimpanzee PWMs have information content below 1 in contrast to 10% for the A allele and 15% for the C allele (Figure 4.18).

I evaluated whether there was evidence for association between predicted motifs and recombination activity by calculating the average profile of recombination rates at the 10,000 genome locations that had the highest likelihood for matches to the prediction. The likelihood of binding was calculated by applying the PWM, M, across the reference genome (remember that M has five rows representing the alphabet (A,C,G,T,N) and l columns representing the length of the binding-prediction):

1. Require that each element of M has a binding probability above 0.01. If required redistribute probability from the other elements in that column to satisfy the threshold, then log-transform the elements of M.

2. Generate the reverse complement of M, denoted by M', by permuting both the rows and columns of M [note $\rho(A, C, G, T, N) \rightarrow (T, G, C, A, N)$], so that both forward and reverse strands can be scanned in one pass of the genome.

3. Calculate the log-likelihood of the PWM match (for both M and M') at each location in the genome by taking the sum of the position-specific scores for each symbol in the set of substrings $\{(b_i, b_{i+1}, \ldots, b_{i+l})\}$ where $i \in [1, n - l + 1]$.

I applied this scheme to each of the alleles and their matches reference assembly autosomes.

While Western chimpanzees have high PRDM9 sequence diversity, many of the alleles share predicted binding motifs (Figures 4.19a, 4.17). The three most common motifs (green, purple and blue boxes) are all observed in at least eight of the 11 predictions, and were analysed for evidence of recombination rate elevation, however no signal was found (Figure 4.19c). To assess whether a larger motif within the predictions was responsible for binding, exact matches to the nucleotides corresponding

to the most informative Cys2His2 amino-acid combinations were identified yielding the prediction $CN_2CN_2CCN_7CCC$. This 18 bp motif is present exactly in four of the 11 predictions and in an additional two predictions if one mismatch at position 3 in the prediction is allowed. Despite conditioning on the most informative bases, no recombination-localising activity was observed for the 18 bp motif prediction (Figure 4.19c, light blue motif). The localisation signal in the human data can be attributed to overlap with the degenerate human A allele 13mer motif CCNCCNTNNCCNC over nucleotides 7 to 18.

To identify which, if any, of the ZF are important for localising recombination events in chimpanzee, I split the PWMs into sub-matrices and analysed each sub-matrix separately. I chose to explore two submatrix sizes, 15 bases and 19 bases, that matched the lengths of the two DNA motifs bound by human PRDM9 variants A and C (Baudat et al., 2010; Hinch et al., 2011), and were larger than the motif lengths searched for on repeat element backgrounds. Each submatrix was evaluated for recombination activity in non-repetitive DNA using the following algorithm:

1. Identify the 2,500 matches with best likelihood, as described above, in non-repeat DNA.

2. Calculate the average recombination rate over 200 bp intervals for the 20 kb around matches.

3. Calculate the recombination-activity in the recombination rate profile (Section 4.1.2).

No chimpanzee submatrices show evidence for recombination localisation. For example, W6, a common western chimpanzee allele observed at 25% frequency in our samples showed little, if any, evidence for recombination activity through this analysis (Figure 4.20b). The 19 base submatrix with the strongest enrichment signal (1.06 fold increase) at position two in the W6 PWM had an average recombination rate profile

similar to that of the potential ancestral allele A1 (Figure 4.20a). In contrast, the same method applied to human PRDM9 alleles A and C identified the submatrices that match the two human recombination motifs (Figure 4.20c, d). However, the signal is stronger for the A than C variant, even when trained on the map from 10 YRI genomes, potentially reflecting the greater diversity of PRDM9 alleles in African populations (Figure 1.7).

To address the possibility that the predicted PWMs were too degenerate to identify true PRDM9 binding sites, I sought to refine the predicted PWM submatrix predictions using DNA sequences identified as strong binding candidates that also showed elevated recombination rates. This was motivated by the observation that PWM predictions can be improved around the human degenerate 13mer recombination motif, CCNCCNTNNCCNC.

This was achieved using the following algorithm:

1. Identify the 2,500 matches with best likelihood across 15 of the autosomes (uniformly sampled).

2. For each match, calculate the recombination rate over the 3 kb interval centred on the match (the approximate width of hotspots found in our data).

3. Identify the matches that rank in the top 85^{th} percentile of the matches.

4. Generate a refined PWM based on the frequency of each nucleotide at each position in the matches.

5. Cross-validate the improvement by comparing the localisation signal at the top 2,500 matches to the refined versus original PWM on the remaining seven chromosomes.

No refined chimpanzee submatrices show evidence for recombination localisation. However, when I applied this method to the human A variant PWM, the method

converged to the human degenerate 13-mer motif with more than 0.2 ($\log_2$) improvement in information content at 4 of 8 cytosine nucleotides in the 13-mer motif (Figure 4.20e,f). One shortcoming of this approach is that it assumes a strict model of zinc finger serial, pairwise cooperativity, however it has been noted that motif recognition varies based on DNA helix forms (Klug, 2010).

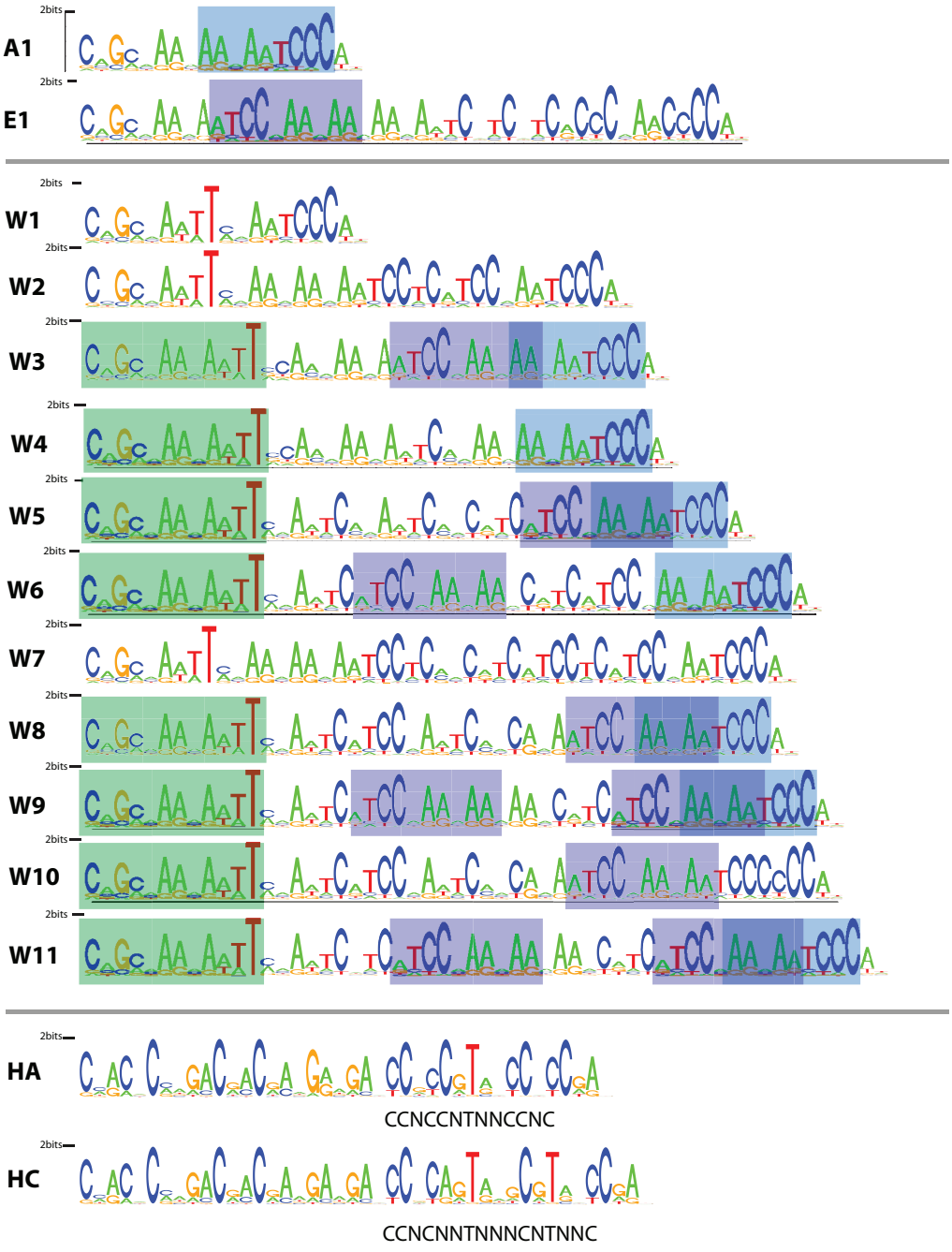


Figure 4.17: Logo plots (Schneider and Stephens, 1990) showing predicted binding sequences for PRDM9 Zinc Finger (ZF) variants predicted for Bonobo (A1), Eastern (A1, E1) and Western (W1-W11) chimpanzee, and human A- and C-alleles (HA, HC). Western chimpanzee PRDM9 ZF alleles as shown in Figure 4.19a. Position weight matrix (PWM) binding predictions were obtained using method of Persikov et al. (2009) under the polynomial pre-trained SVM model. Many of the Western chimpanzee PRDM9 alleles share predicted binding motifs. The three most common shared motifs are highlighted by coloured boxes (green, purple, and blue). Motifs under the human predictions (black) represent the DNA motifs enriched in A and C allele-driven hotspots

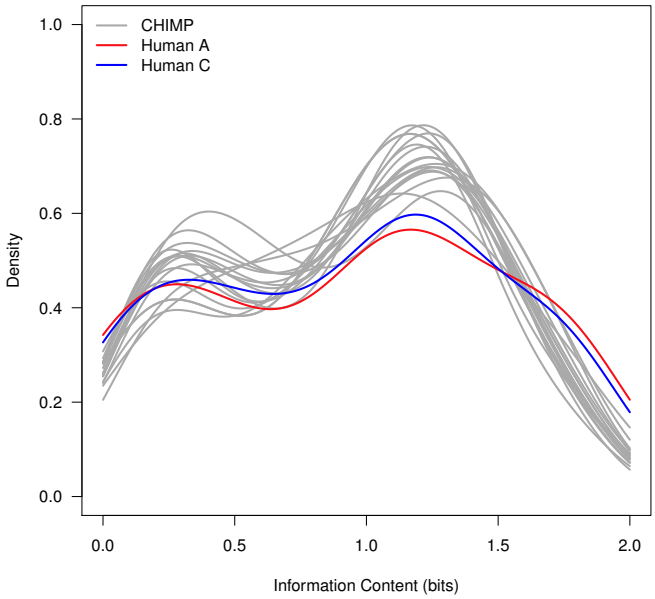


Figure 4.18: Information content in PRDM9 allele position weight matrix predictions for humans and chimpanzees. Distribution of expected self information (logarithm base 2) for each element for Human A- (red), C-allele (blue), and the 11 Western chimpanzee alleles (W1 - W11).

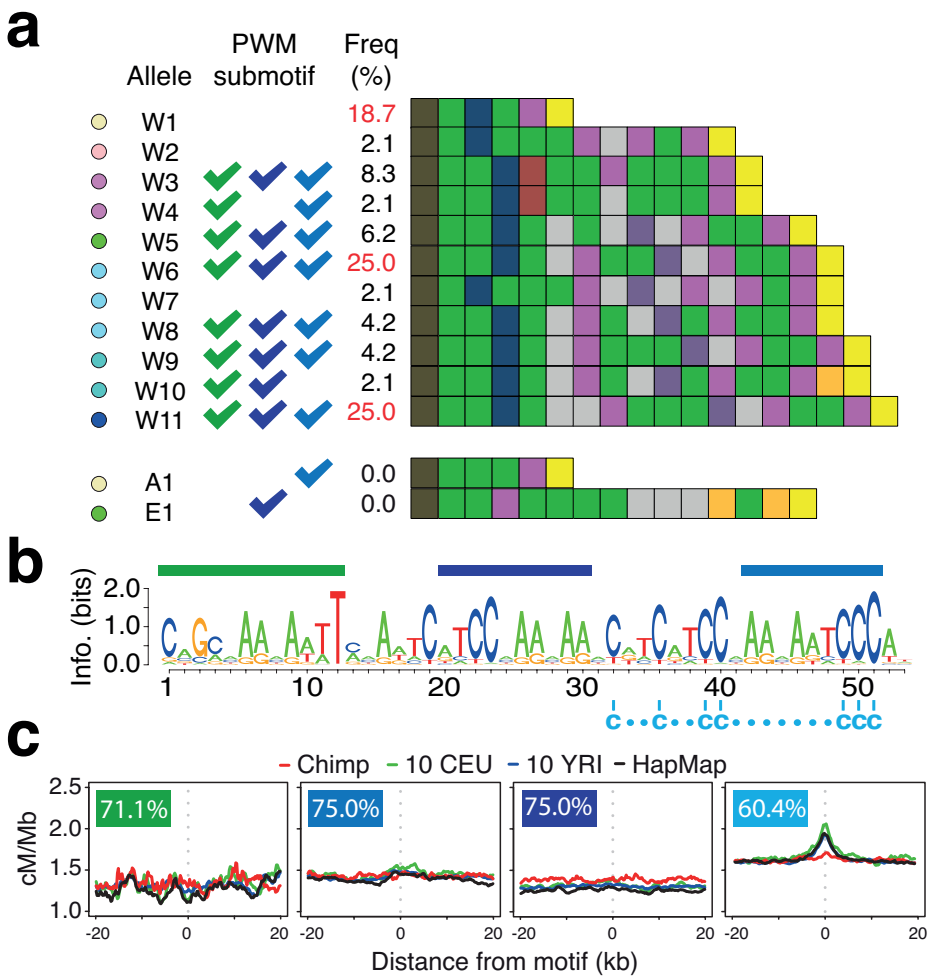


Figure 4.19: Sequence and structural variation in chimpanzee PRDM9 and implications for hotspot motifs. **(a)** Schematic representations of the zinc-finger arrays found in chimpanzee PRDM9 alleles with colours representing unique combinations of DNA-contacting amino acids within zinc fingers (amino acid sequence provided by Laure Segurel and Simon Myers). Western chimpanzee alleles are labelled W1 through W11. Also shown is the putatively ancestral allele shared between Bonobo and Eastern chimpanzee (A1), and the remaining Eastern chimpanzee allele (E1). Tick symbols indicate binding specificity to motifs indicated in panel (c). Allele frequencies estimated from 48 Western chimpanzees alleles. **(b)** Predicted binding motif for the chimpanzee reference PRDM9 allele (W6) showing positions of shared sub-motifs referred to in parts (a) and (c) and a shared set of cytosine residues (below sequence). **(c)** Recombination rates around shared predicted sub-motifs for chimpanzee PRDM9 alleles in non-repeat DNA (percent of alleles predicted to bind indicated).

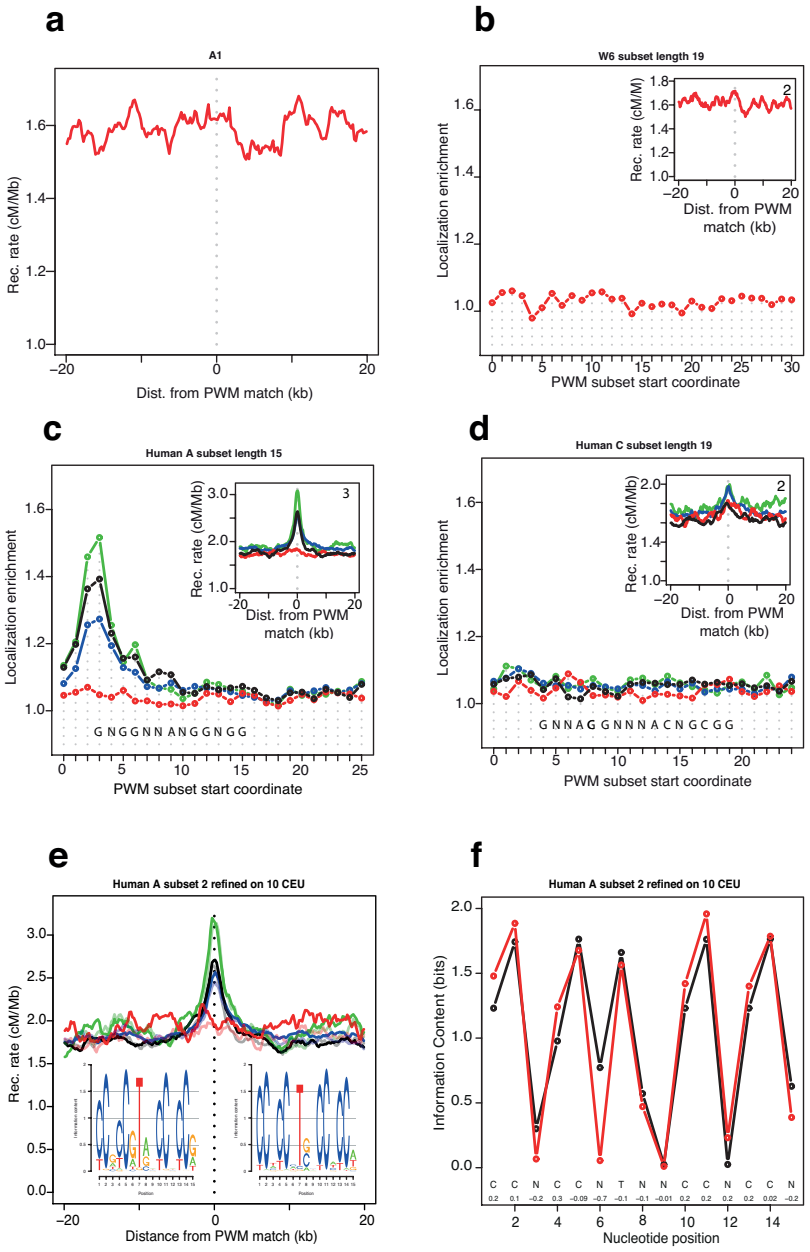


Figure 4.20: Predicted chimpanzee PRDM9 binding alleles show little, if any, evidence for activation in non-repeat DNA. Red, green, blue, and black correspond to chimpanzee, 10 CEU, 10 YRI and HapMap maps (except (f)). **(a)** Recombination rate profile around matches to the position weight matrix (PWM) with best likelihood for allele A1, a putative ancestral allele observed in Eastern and Bonobo but not Western chimpanzees. (**b-d**) Recombination-influencing activity of matches to PRDM9 PWM submatricies in chimpanzees (W6) and humans (A, C). The value reported is the ratio of the peak to background rate, as estimated from the robust genetic map after fitting a Gaussian profile using maximum likelihood. Insets show fine-scale recombination profile around submatrices with best recombination-localizing activity. **(e)** Recombination rate profile for locations matching 15 bp submatrix of PWM for human allele A before (opaque) and after (solid) refinement. **(f)** Information content for A allele submatrix before (black) and after (red) refinement.

4.6 Discussion

This work demonstrates how fine-scale genetic maps can be obtained though the analysis of patterns of genetic variation from population sequencing that should enable further informative studies of recombination across a wide range of species. Studying humans and Western Chimpanzees, I found no hotspot sharing between the two species, consistent with earlier reports based on limited data (Myers et al., 2010; Ptak et al., 2004; Wall et al., 2003; Winckler et al., 2005). The complete lack of hotspot sharing is consistent with the hypothesis that in humans PRDM9 plays a critical role in localising cross-over activity at all hotspots, not just those that contain clear matches to previously-identified motifs bound by PRDM9. Despite the dramatic shift in hotspot locations between the two species, some fine-scale patterns, particularly the average profile of recombination rate around genes and CpG islands, remain similar, pointing to the importance of chromatin state in influencing where double strand breaks occur (Pan et al., 2011) or to additional levels of control acting on broader scales (Pan et al., 2011; Paigen and Petkov, 2010).

A notable difference between the species is that in chimpanzees no repeat elements, simple DNA motifs, or predicted PRDM9 binding sites are strongly or consistently associated with hotspot locations. There are three possible explanations. First, PRDM9 may have lost its role in specifying hotspot locations in chimpanzees, as has occurred in dogs (Axelsson et al., 2011), although collaborators found no evidence for inactivating mutations (Auton et al., 2012). Second, PRDM9 alleles may each have similar specificity to target DNA sequences, but the substantial allelic diversity and their possibly recent origin may obscure signals for individual alleles in population based maps. However, this hypothesis cannot explain why, when the density and strength of hotspots at the population level are similar in African populations and Western Chimpanzees (Figure 4.3), we can recover known PRDM9-binding motifs in humans but no comparable motif in chimpanzees. However, the evidence for the mo-

tif bound by human PRDM9 C-allele is far more subtle than the A-allele as assessed through the motif finding methods described above, motivating the hypothesis that PRDM9 alleles show potentially wide variation in sequence specificity. The forces that generated such strong sequence specificity in the A variant are an open question potentially driven by a combination of demographic history and dominance relationships with other variants. Third, chimpanzee PRDM9 may play the same role as in humans and mice, but individual PRDM9 alleles may bind to a much greater variety of target sequences than do the predominant human alleles. If so, hotspot localisation in chimpanzees may be more strongly driven by other factors, such as chromatin state as motivated by observations around genes and CpG islands. Brick et al. (2012) observed that DSBs are directed away from function genomic elements in PRDM9$^{+/+}$ mice, whereas in PRDM9$^{-/-}$ DSBs were shifted to gene-promoter regions and other functional elements marked by H3K4Me3.

Chapter 5

Sequencing of an extended chimpanzee pedigree

The analysis of variation data from the HTS of 10 unrelated Western chimpanzees gave us insight into the patterns of recombination rate variation at fine- and broad-scales. However, fundamental questions remain that cannot be addressed through coalescent-based approaches. For instance, what is the total genetic distance in chimpanzees? How do recombination rates vary between the sexes? And what are the features of complex cross-over events? To address these questions we deep sequenced ($\sim$ 30$\times$) a nine individual three-generation Western chimpanzee pedigree (Figure 5.1).

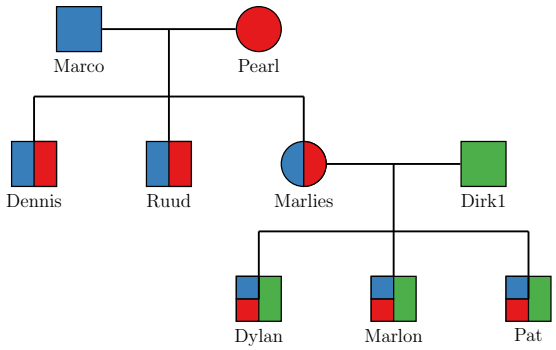


Figure 5.1: The sequenced Western chimpanzee pedigree.

We also aim to detect and characterise the spectrum of rare events that impact genomes. This has three important challenges, detection, prioritisation of candidates, and specificity. In principle, sequencing can resolve genome changes from single base changes up to large, megabase events at single base resolution. However, we have demonstrated that HTS is impacted by many error modes arising through sequencing error, reference driven artefacts, stochastic levels of coverage, and variant calling and genotyping inaccuracy. Furthermore, a substantial proportion of errors are not solved through increased coverage, or filtering. We boost power to distinguish signal from noise through inherent Mendel transmission consistency checks. In addition, under the assumption of Mendel transmission, we develop model based approaches to prioritise candidate genome change events, which we expect to be rare.

To demonstrate the utility of transmission for distinguishing signal from noise, I provide an example of three different classes of encountered genotype vectors. Consider Figure 5.2, class-I sites represent instances of *de novo* mutation. The three founders have the same genotype 0/0, hence defining a 0/0 genotype throughout the pedigree. However, individuals 6, 7, and 9 have genotypes 'inconsistent' with the expected transmission. These Mendel inconsistencies (MICs) could represent genotyping error, or the signal of *de novo* mutation. The consistent transmission of the candidate *de novo* mutation in individual 6 to 7 and 8 is strong evidence that we have observed a *de novo* mutation. In contrast, class-II sites represent MIC with inconsistent marginal transmission within the pedigree indicating genotyping error. Class-III sites represent an additional error mode, where systematic error results in Mendel consistent transmission conditional at the site, but which is not consistent with the local transmission pattern and additionally is only described by a restrictive mode of inheritance. In general, we can identify sites with MIC genotypes by enumerating the potential genotypes transmitted by the founders (Section 5.2.4).

Additionally, the specific extended pedigree structure was chosen for three rea-

sons. First, as motivated in Section 1.2.2 at least three siblings are required to assign events (cross-over, gene conversion and mutation) to individuals. Second, we chose to sequence two generations to provide the opportunity for natural validation of candidate events through transmission. Third, we maximised the number of non-founders in each generation to capture *de novo* events, under this pedigree 7/8 of each founder genome is transmitted at least once. In this thesis I will only report results for cross-overs detected in the pedigree, though we have also characterised gene conversion and *de novo* mutation. Characterising gene conversions using a whole-genome, high-coverage sequencing approach is attractive because identifying gene conversion through population genetic approaches is challenging (e.g. Gay et al., 2007; Padhukasahasram et al., 2006) and experimental approaches do not scale to genome-wide application (Jeffreys and May, 2004).

Consequently, sequencing extended pedigrees provides not only information on meioses, but also more information to distinguish signal from noise through transmission to the F2s. To our knowledge this is the first analysis of extended pedigree HTS, we therefore also report methodologies to analyse the data and to prioritise regions of interest.

Given the features of cross-over studied through LD-based approaches in Chimpanzees, one might postulate that humans and chimpanzees have similar genome change rates. If we assume that cross-over, gene conversion and mutation rates are similar between the species we can estimate what we might observe from 12 meioses:

- **Cross-over (CO)**. Under an assumed genome-wide cross-over rate of 25 CO per meiosis, we expect (12 meioses)(25 CO/meiosis) = 300 crossovers,
- **gene conversion (GCO)**. Under an assumed track length of 50 bp, diversity of 0.001 bp/pair, and 10 GCO/CO, we expect (300 CO)(10 GCO/CO)(50 bp/GCO)(0.001 bp/pair)(12 haploid genomes) = 480 detectable gene conversions,

- ***de novo* mutation**. Under an assumed point mutation rate of 1.2×10^{-8} mutations/bp/gen, we expect (12 meioses)$(1.2 \times 10^{-8}$ mutations/bp/gen$)(2.5 \times 10^9$ bp$) = 360$ point mutations.

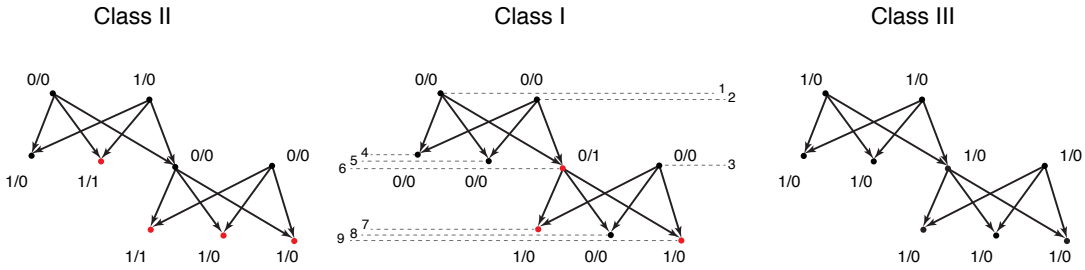


Figure 5.2: Three example transmission classes. The three biallelelic genotypes are $\{0/0, 0/1, 1/1\}$, where 1 represents the reference allele and 0 represents that alternative allele. **Class I** Mendel inconsistent genotypes (inconsistencies indicated by red dots), consistent with *de novo* mutation. **Class II** Mendel inconsistent genotypes, consistent with a false positive. **Class III** marginal Mendel consistent genotypes, but not consistent with local inheritance patterns (i.e. the inheritance at neighbouring sites).

5.1 Acknowledgements

Natasja de Groot collected blood samples and extracted DNA. The WTCHG sequencing core prepared and sequenced the DNA libraries. Iain Mathieson called variants through the Platypus (`www.well.ox.ax.uk/platypus`) program. Isaac Turner called variants through the Cortex program (Iqbal et al., 2012). Gilean McVean and I worked on developing the methods to analyse the pedigree variant calls.

5.2 Methods

In this section I describe properties of the sequencing experiment, variant calling, and methods to phase the inferred genotypes using transmission structure (represented

through inheritance vectors, described previously in Section 1.2.2) and identify recombination events.

5.2.1 Sample preparation and sequencing

Blood samples were supplied by the Biomedical Primate Research Centre (Rijswik, Netherlands) under the EUPRIM-Net EU contract RII-026155 and CITES approval. DNA was extracted from peripheral blood mononuclear cells using affinity bead selection (Qiagen, Netherlands), and 100 bp paired-end shotgun libraries constructed. Libraries were sequenced on Illumina HiSeq instruments using standard chemistry and base calling versions. A total of 0.9 Tb of sequence was generated (Table 5.2.2).

5.2.2 Alignment and variant calling

Reads were aligned to the CGSC 2.1.3/panTro3 build of the chimpanzee genome using Stampy (Lunter and Goodson, 2010), resulting in an average mapped coverage of 35×, and an average of 90.8% of the aligned genome covered by at least 15 reads (Table 5.2.2).

Individual	Bases (Gb)	Mean coverage (×)	% aligned > 15×
Marco	185.0	63.81	97.8
Pearl	166.9	57.54	97.0
Dennis	103.6	35.73	94.7
Ruud	66.1	22.78	82.6
Marlies	90.2	31.11	94.0
Dirk1	74.2	25.59	87.6
Dylan	70.8	24.42	86.0
Marlon	72.1	24.85	86.8
Pat	84.9	29.28	90.6

Table 5.1: Western chimpanzee pedigree sequencing properties. DNA material was obtained from peripheral blood mononuclear lymphocytes. Mapped coverage was calculated after removing PCR duplicates, and against the panTro3 reference genome.

Initial variant calls for the autosomes and the X chromosome were made using Platypus. These were then filtered additionally by removing calls in any of the following categories:

- More than two alleles at the same position.
- In a homopolymer run of length 5 or greater.
- In a tandem repeat run of length 10 or greater.
- Five or more variants in ± 25 base window.
- Every chimpanzee was homozygous non-reference at that site (i.e. most likely an error in the reference).
- In an indel hotspot.

After filtering 4,182,960 SNPs remained with an overall transition/transversion ratio of 2.14 and 169,387 indels. In this thesis I analyse the SNP calls, which will be referred to as the *Platypus calls*. For the autosomal SNPs, 0.94% (34,452 sites) had genotype calls that were not compatible with Mendel transmission. On the male X chromosomes 3.2% of the variants (PAR excluded) were called heterozygous, which must arise, overwhelmingly, from errors given that males are hemizygous for the X chromosome.

We use a Lander and Green (1987) based approach to phase genotypes calls and therefore our approach, despite incorporating genotype uncertainty, is sensitive to mis-specified genotypes and false-positive errors. To guard against errors we identified a high quality subset of calls to define the transmission patterns across chromosomes (Section 5.2.5), performed analyses conditional on the high quality scaffold (Section 5.2.10]), and furthermore attempted to counteract mis-specified genotypes by analysing GLs.

5.2.2.1 Identifying a robust call set: intersection with variants called through Cortex

To identify a high quality subset of calls, we intersected the Platypus calls with the variants called through Cortex. As described in Chapter 1, Cortex has a low false discovery rate, we therefore used these sites to identify Platypus calls with low false-positive rate. The intersection identified 1,595,502 sites with an overall transition/transversion ratio of 2.24, the total MIC genotypes increased to 1.5% of sites (34,067 SNPs), but the number of heterozygous calls on the male X chromosome decreased to 0.7% of intersection sites (2,895 sites, PAR excluded).

We verified the sex of the individuals in the pedigree by calculating the total heterozygous genotype call frequency on the X chromosome for male and female samples (P-value 0.007, unpaired t-test; Figure 5.3). We also evaluated relatedness between individuals in the samples by calculating the Pearson correlation between individuals' genotype vectors excluding monomorphic sites, or sites with missing genotypes. The observed pairwise correlation between individuals is concordant with their expected relatedness (Figure 5.4).

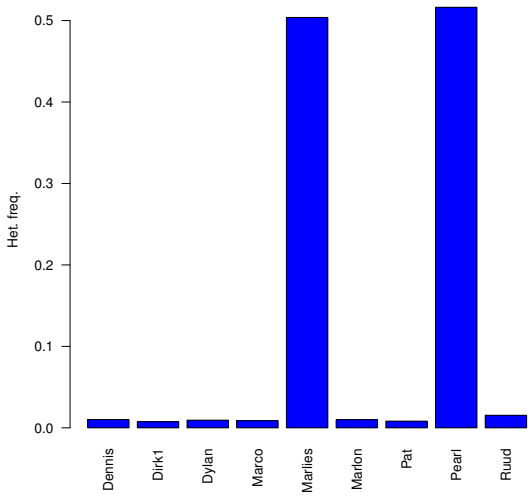


Figure 5.3: Proportion of heterozygous genotype calls on chromosome X.

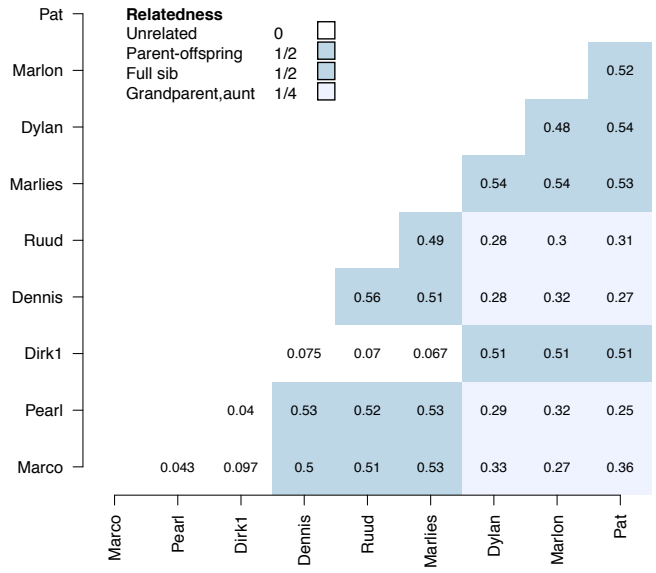


Figure 5.4: Relatedness between individuals in the sequenced pedigree. Values represent the Pearson correlation between individuals' genotype vectors, cells are coloured by their expected relatedness.

5.2.3 Estimates of genotype-error in chimpanzee HTS

The founder Pearl was sequenced in both the unrelated chimpanzee sequencing experiment and the extended pedigree sequencing experiment (referred to as the *PanMap call set*). To estimate genotyping error, we assessed concordance between the 2 HTS data sets (PanMap and pedigree) and SNP-typing data from Myers et al. (2010). We analysed genotype concordance at heterozygous sites to reduce the bias from differential ascertainment in the two experiments and the impact of reference bias (since PanMap HTS were aligned to the panTro2 reference and then lifted over to panTro3). For the pedigree genotypes in the 3-way intersection, we observed one disagreeing pedigree-genotype, where Pearl was genotyped as homozygous reference in the pedigree. We checked whether the homozygous genotype assignment was the result of mis-genotyping in the pedigree data, but we found that all 54 reads that cover the site support the homozygous reference assignment. In addition, the panMap assign-

ment is also homozygous reference. We therefore interpret this disagreeing genotype as a genotyping error in the SNP-typing. To calculate the genotyping error rate in the HTS samples, we then removed sites where both HTS genotypes disagreed with the SNP-typing; for the 141 sites that pass this criteria 100% of the pedigree genotypes are concordant with SNP-typing, and 93.4% are concordance between panMap genotypes and the SNP-typing. We estimate that the genotyping error at heterozygous sites in panMap is $\sim 6.2\%$ (CI 2.3%, 10.1%; normal approximation to binomial proportion), because the concordance rate between the pedigree and panMap is also 93.4%. This analysis suggested that the error rate in the pedigree genotyping was ~ 1 in 141. To better characterise this value we analysed sites in the 2-way intersection between the pedigree calls and the SNP-typing. This increased the number of sites to 162 heterozygous sites. Under this more liberal call set, we discover one genotyping error in the pedigree (verified by inspecting the alignments) which suggests an error rate of 0.62% (CI 0.0%, 1.8%).

5.2.4 Inferring inheritance vector scaffolds

To detect recombination events, the inheritance vector pattern across chromosomes is required. To estimate inheritance we infer the marginal likelihood of inheritance vectors per site conditioned on the observed genotype likelihood. Inheritance vectors are binary vectors of length two times the number of non-founders that describe which grandparental chromosome is inherited for each non-founder. Hence each individual is represented by a pair of binary digits, the first digit represents the chromosome inherited from the father, the second digit represents the chromosome inherited from the mother. Zero represents grand-paternal inheritance, and 1 represents grand-maternal inheritance. To aid rapid interpretation, I pair inheritance vector elements to reflect diploidy, under this notation an inheritance vector contains a pair of binary digits for each non-founder. Figure 5.5 provides an example of a potential inheritance

vector $\mathbf{v} = (00, 10, 00, 10, 01, 10)$ in the chimpanzee pedigree.

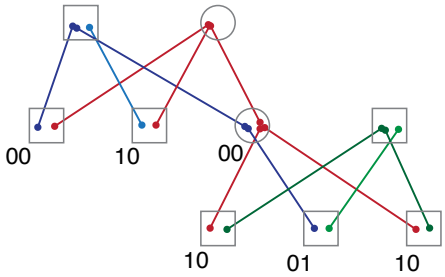


Figure 5.5: Visual representation of an inheritance vector $\mathbf{v} = (00, 10, 00, 10, 01, 10)$. The number 1 represents grand-maternal inheritance, 0 represents grand-paternal inheritance. The first digit in each pair represents the chromosome inherited from the father, and the second digit represents the chromosome inherited from the mother.

For a simple pedigree structure there exists a finite set of inheritance vectors V, and a related finite set of genotypes G, compatible with Mendelian transmission. In order to explain the algorithm for calculating the marginal likelihoods of inheritance vectors from GLs, an extension of previous work on hard-called genotype vectors (e.g. Thompson, 2000), I define some notation.

Let the pedigree structure be represented as a directed, labeled tree denoted by $S = (T, U)$. Each vertex in T represents an individual, and the directed edges in U represent the ancestry. We chose to enumerate vertices to reflect ancestry over two levels. We stratified labels by founder and non-founder individuals, starting with the most ancestral founder, their mate (if a founder), the next founder, and so on. We then labelled non-founders starting with the most ancestral generation and ordering siblings by their age within each generation (Figure 5.6).

The number of founders is denoted by n_f and the number of non-founders by $n_{f'}$. The $2n_{f'}$ edges delineate relatedness (Figure 5.6). Given S, the set of inheritance vectors V is induced by allowing each offspring to choose a chromosome from each parent independently. Since chimpanzees are diploid, $2^{2n_{f'}}$ inheritance vectors are generated

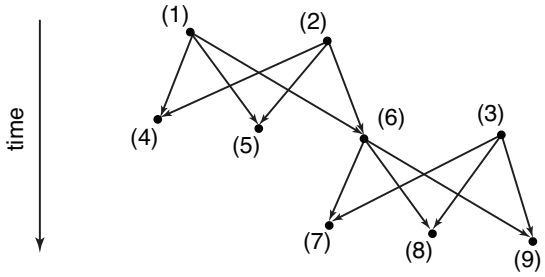


Figure 5.6: Labelling strategy for a simple pedigree. Directed edges indicate relatedness between individuals. $T = \{1, 2, \ldots, 9\}$, $U = \{(1, 4), (1, 5), \ldots, (3, 9)\}$.

for the autosomes. However, this process introduces redundancies. Consider the inheritance vector $\mathbf{v} = (00, 10, 00, 10, 01, 10)$, because the observed genotypes are not phased we can generate 2^3 isomorphic inheritance vectors by permuting each of the founder chromosomes independently (Figure 5.7).

Hence, to ensure identifiability and encode the dependency structure of transmission, one inheritance vector is selected to represent its family of isomorphic inheritance vectors. We do this by fixing the transmission from each founder to its first child to be the grand-paternal chromosome, in other words the inheritance vector element for that offspring is set to zero. Hence, the first child of two founder parents is fixed to inherit (0,0) [For the pedigree in Figure 5.7 this is individual 4], and the first child with one founder parent could inherit (0,0) or (0,1) [individual 7], their siblings are free to inherit any of the four possible transmission combinations [individuals 5, 6, 8, and 9]. For the analysed chimpanzee pedigree, the nine member HTS pedigree has $2^{2\times6-2\times1-1\times1} = 512$ identifiable inheritance vectors.

Once V is constructed, we can enumerate the set of genotypes G compatible with Mendelian transmission by applying the transmission vectors in V to the set of potential founder genotype vectors and, reciprocally identify MIC genotypes as those genotypes not in G (Figure 5.8).

In summary, we use the *onto* functions to generate the set of inheritance vectors

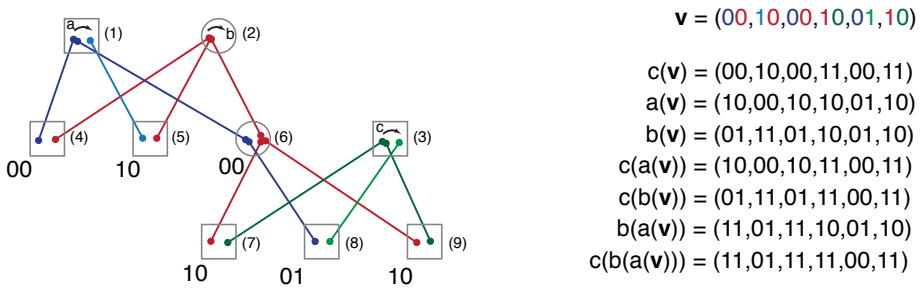


Figure 5.7: Example of redundant inheritance vectors. Coloured lines show the transmitted alleles. Alleles inherited from the grand-paternal chromosome are labelled 0, and grand-maternal alleles are labelled 1. The vector v represents the inheritance vector depicted on the left-hand-side. Symbols a, b, and c represent symmetric permutation functions that applied to v generate isomorphic inheritance vectors (right-hand-side) through switching founder chromosome ordering. Hence the eight inheritance vectors on the right-hand-side can be represented by v alone.

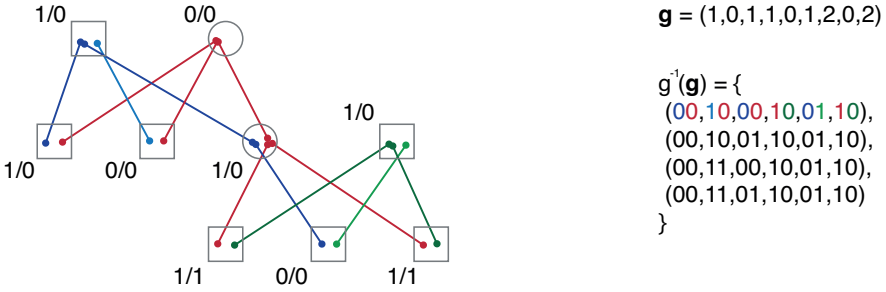


Figure 5.8: Example of the *onto* mapping from an observed genotype vector $\mathbf{g}$ (Equation (5.2)) to the set of compatible inheritance vectors through the inverse of the mapping function g^{-1}. The first inheritance vector in the mapping is represented in the left-hand-side panel. Each inheritance vector corresponds to a different combination of phased genotypes compatible with $\mathbf{g}$.

V, and genotypes G given a pedigree tree S.

$$f\colon S \mapsto V \tag{5.1}$$

$$g\colon V \mapsto G \tag{5.2}$$

$$g \circ f\colon S \mapsto G. \tag{5.3}$$

The function f maps the directed tree S onto the set of compatible inheritance vectors V, fixing the inheritance of the first child of each generation. The function g maps V onto the set of genotypes for the possible founder genotype vectors, this can be thought of as an allele dropping experiment. We can then apply these functions to transverse between inheritance and genotype vector spaces, but because the mapping is not bijective, there is uncertainty in the assignment at each site. The Sections that follow describe methods to deal with this uncertainty.

5.2.5 Calculating the likelihood of an inheritance vector at a site

To incorporate the uncertainty underlying HTS genotypes the genotype likelihoods (GLs) are analysed. Let the set of genotype likelihoods for the measured genotypes be denoted by X, which is the tuple of genotype likelihoods for each individual for each site. I use tensor notation in this chapter, hence to index the ith row of an object O, I write O^i, whereas to raise O to its nth power I write $(O)^n$.

The likelihood of each inheritance vector is calculated as the sum over the genotype likelihoods for the compatible genotypes scaled by a prior probability over the founder genotypes - we assumed that the prior is uniform. For the site k, the likelihood of the ith inheritance vector is:

$$L(\mathbf{v}_i|D^k) \propto \Pr(D^k|\mathbf{v}_i) = \sum_{\mathbf{g}\in G_i} \Pr(D^k|\mathbf{g}) \Pr(\mathbf{g}), \tag{5.4}$$

where G_i is the set of unique genotype vectors induced by projecting the set of ordered founder alleles (i.e. distinguishing between 0/1 and 1/0) along $\mathbf{v}_i$, and $\Pr(D^k|\mathbf{g})$ is extracted from the GLs. Since each inheritance vector induces a set of potential genotype vectors given the founder genotypes, we can use linear algebra and pre-calculated tables for G and V to rapidly index the GLs for the likelihood

calculation.

5.2.6 Generating a scaffold of inheritance vectors across genomic segments

Given a procedure to calculate likelihoods, we would like to do inference on the pattern of inheritance across chromosomes. As described in Chapter 1, Hidden Markov Models (HMMs) (e.g. Lander and Green, 1987) are an attractive approach to infer underlying cross-overs from high density polymorphism data in small pedigrees, where cross-overs are identified as a transition in the hidden state of a non-founder. However, a major downside in this approach is that HMM approaches are extremely sensitive to genotype errors (e.g. Abecasis et al., 2002), which cause artefactual switches between states. In addition, to enable factoring, the model assumes that founder genotypes are independent along the chromosome, which is not the case.

To overcome the potential artefacts arising from mis-assembly and genotyping errors in the HTS, we implement a two-step approach. First, to guard against mis-assembly or genotyping error driven artefacts, we generate an inheritance vector scaffold across non-overlapping 1 Mb chromosome segments, which given LD-based cross-over estimates would be expected to contain at most one recombination event. Second, to improve the resolution of inferred cross-overs, we refine the breakpoints for each event separately.

To generate the scaffold we calculate the likelihood of each inheritance vector in V across the chromosome segments. The site-based likelihoods are calculated through Equation (5.4), then sites are combined assuming that no cross-over event has occurred within the segment. Therefore, if L represents a segment, s represents the full vector of observed SNPs, and V the ordered set of inheritance vectors, then the vector of inheritance vector likelihoods is,

$$L(V|D,L) \propto \Pr(D|V,L) = \prod_s \Pr(D^s|V), \tag{5.5}$$

where $\Pr(D|V,L)$ is the column vector of site likelihoods for each inheritance vector in V and s . Note that implicitly we are are calculating the product of each inheritance vector's likelihoods over sites. However, to improve the efficiency of the calculation, we pre-calculate the mapping for each inheritance vector to its compatible genotype vectors, and then use linear algebra to index the GLs simultaneously for all the inheritance vectors per site. For each segment the inheritance vector that maximises the likelihood is easily identified,

$$\mathbf{v}^* = \arg\max \Pr(V|D,L). \tag{5.6}$$

Note that individually per site likelihoods are uninformative since the mapping between V and G is surjective. By combining information across sites we can resolve the uncertainty in the underlying inheritance vector and additionally guard against local ($<$ 1Mb) mis-assembly. Figure 5.9 shows the log-likelihood difference between the maximum likelihood and the remaining 511 inheritance vectors on chromosome 21 demonstrating that the chunk likelihoods are well-differentiated under this approach.

In addition to inheritance vector likelihoods, we calculate the proportion of Mendel consistent sites for each inheritance vector by counting the proportion of hard called genotype vectors compatible with each inheritance vector. This statistic also reflects the underlying inheritance vector scaffold. The sequence of vectors across chromosome chunks are used to construct the inheritance vector scaffold (Section 5.2.8). After constructing a robust scaffold we detect recombination events through switches in the inferred inheritance vector between chunks (Section 5.2.8), and investigate local events at fine scales conditional on the scaffold (Section 5.2.10).

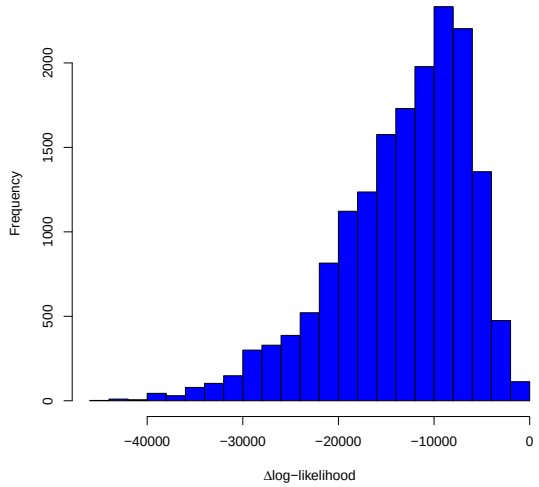


Figure 5.9: Difference in log-likelihood between the aggregated window maximum likelihood inheritance vector and the other inheritance vectors over chromosome 21.

5.2.7 Model for recombination

Inheritance vectors provide a natural framework to identify recombination events; once the sequence of inheritance vectors across a chromosome have been inferred, candidate recombination events are identified as transitions between inheritance vectors along the sequence. We chose to model the recombination process as single cross-over transitions between the 1 Mb segments, since the probability of coincident cross-overs is extremely low between the six non-founders. Conceptually, single cross-over events are equivalent to permuting an individual edge in $\mathbf{v}$, hence the set of inheritance vectors reached through one recombination event is equal to the number of edges (there are $2n_{f'}$ such moves). Hence, we generate the outcomes of recombination events through the function $r(\mathbf{v}, i)$, which maps a given inheritance vector $\mathbf{v}$ to the inheritance vector arising from a recombination event at inheritance vector element i (Figure 5.10).

To preserve identifiability we enumerate all possible single cross over events on V, but map those inheritance vectors generated by recombination events on chro-

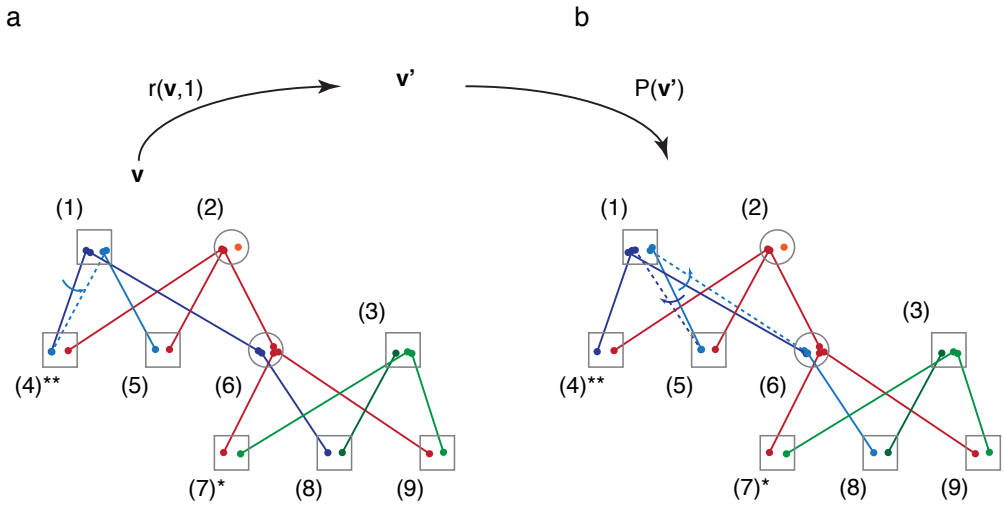


Figure 5.10: Example of how recombination events are enumerated. **(a)** the impact of a single recombination event (curved arrow) on an inheritance vector (straight arrows), and **(b)** the re-mapping of the recombination event in 4 to the chromosomes transmitted to individuals 5 and 6. The stars represent chromosomes that we require to have fixed transmission to avoid generating isomorphic inheritance vectors.

mosomes transmitted to the 'fixed' non-founders to their siblings. As an example consider the inheritance vector $\mathbf{v} = (00, 10, 00, 10, 01, 10)$ in Figure 5.10, we introduce a recombination event in the chromosome transmitted from individual 1 to individual 4 by permuting the chromosome transmitted from 1 to 4. This is reflected in the inheritance vector by switching the first element the binary pair. However, individual 4 is the first progeny of 1 and 2 and is hence fixed for both inheritance vector elements, to preserve identifiability we apply the projection matrix P to the recombination product that maps the recombination event to siblings 5 and 6:

$$P(r(\mathbf{v}, 1)) = (00, 11, 01, 10, 01, 10)^T. \tag{5.7}$$

In order to enable recovery of the generative cross-over event, we assign a unique label to each inheritance vector and record the $2n_{f'}$ single recombination products for each inheritance vector in a $\mathcal{O}(V) \times 2n_{f'}$ matrix C of cross-over transitions:

$$C_j^i = \sum_k \left[r(V_j, i)\right]_k \, 2^{2n_{f'}-k}, \tag{5.8}$$

note that the inheritance vectors are in practice vectors of length $2^{n_{f'}}$, hence this sum encodes the inheritance vector by recording the ordered sum of grandmaternally inherited chromosomes. Hence, the fixed inheritance vectors and recombination events can be de-convoluted through indexing the relevant elements of C. This is robust when aggregating information across sites.

5.2.8 Hidden Markov Model to infer a broad-scale inheritance vector scaffold

We implemented two alternative approaches to infer the underlying inheritance vector scaffold; a Hidden Markov Model (HMM) and a Finite State Machine (FSM). To limit the impact of reference genome mis-assembly and genotyping error, we analysed information combined across chromosome segments (described below). The hidden states are the inheritance vectors, which we infer across the chromosome segments. A transition in the hidden state of a child represents a change in chromosome inherited from a parent and therefore represents a cross-over. As mentioned before, transitions in the maternal and paternal chromosomes are independent of each other and of other meioses. To avoid bias in the inferred location of towards non-founder map, we used a uniform transition probability.

The models differ over both their observation probability distributions and their transition probabilities.

Model 1 To incorporate the measured uncertainty over genotypes, we implemented a HMM that takes the inheritance vector likelihoods across chromosome segments (Equation (5.5)) as the observation distribution. For this calculation we assume a uniform prior over the founder genotypes, however this could be extended

to incorporate allele frequency information from the 10 unrelated HTS chimpanzee genomes. The transitions were modelled assuming a genetic distance of 1 cM per Mb. We therefore assume that the probability of more than one cross-over between chromosome segments is negligible. Therefore if i and j represent the full state matrix of inheritance vectors then the transition probability matrix A_1 is written,

$$A_1(i \to j) = \begin{cases} -\log(\frac{1}{1000}) & \text{if } j \text{ is reached within 1 cross-over applied to } i \\ -\log(1 - \frac{11}{1000}) & \text{if } i = j \\ & \text{other transitions not allowed.} \end{cases} \tag{5.9}$$

The first case represents a single cross-over, the second case represents no cross-over, and other transitions are not allowed. The initial emission distribution is specified as the likelihood vector for the first segment.

Model 2 To incorporate transmission information, we implemented a FSM that takes the proportion of Mendel consistent sites compatible with each inheritance vector across chromosome segments as the observation distribution. Costs on transitions were imposed using the matrix A_2 to limit switches,

$$A_1(i \to j) = \begin{cases} -0.25 & \text{if } j \text{ reached within 1 cross-over applied to } i \\ 0 & \text{if } i = j \\ & \text{other transitions not allowed.} \end{cases} \tag{5.10}$$

The values of A_2 were chosen heuristically, and we note that switches were insensitive to different penalty values. The initial emission distribution is specified as the Mendel consistent proportion for the first segment.

5.2.9 Inheritance vector path across segments

To generate the broad-scale inheritance vector scaffold across chromosomes, we used the Viterbi algorithm (Viterbi, 1967) to identify the most likely inheritance vector sequence across the segments.

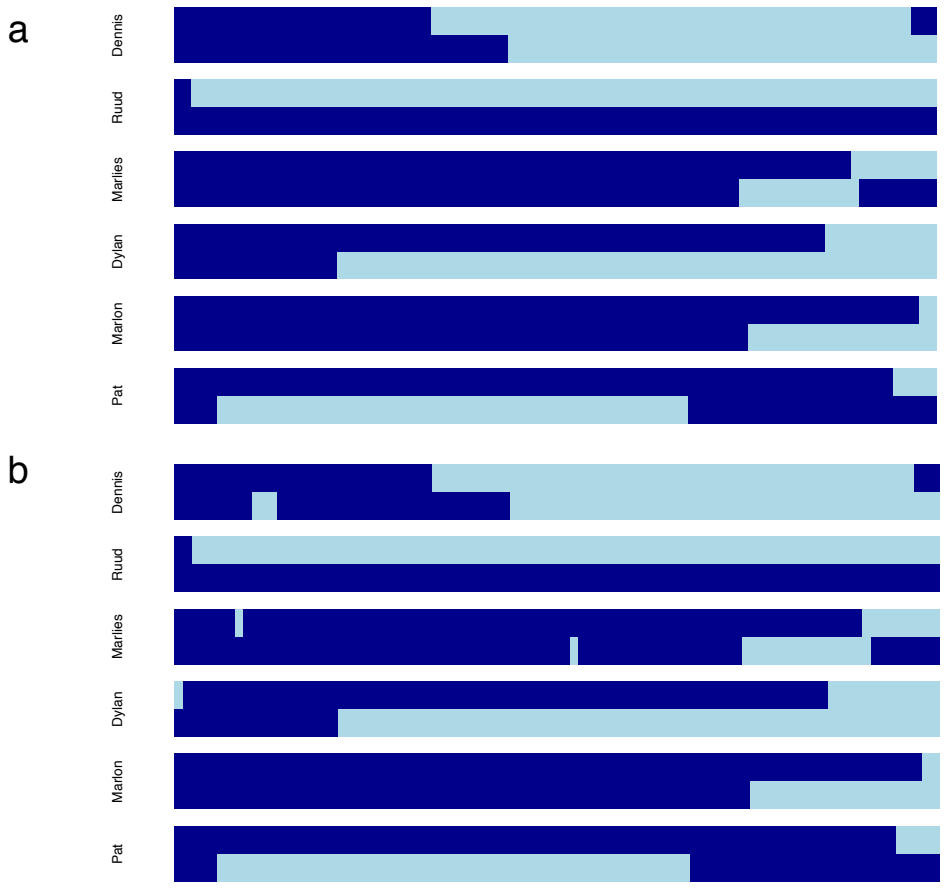


Figure 5.11: Inferred inheritance vector chunks across chromosome 14 for **(a)** the Mendel consistent proportions, and **(b)** the inheritance vector likelihoods. Recombination events in each of the non-founders are allocated to the inherited chromosomes and are indicated by colour changes. The top chromosome represents the paternally inherited chromosome, the bottom the maternally inherited chromosome. The abundance of cross-overs inferred in (a) not replicated in the more conservative (b) demonstrates the impact of genotype errors on confounding cross-over inference, even if uncertainty is incorporated. Note this figure does not represent phased founder haplotypes, but the switches in inheritance vectors across each individual non-founder.

Since, the FSM does not calculate likelihoods, we used empirical observations to compare the two models. We found that Model 2 has lower levels of artefactual transitions than Model 1 (Figure 5.11). This is likely results from Model 2 being more robust to false positives generated by assembly errors. Note that under that we do not estimate the founder haplotypes these approach, therefore switch conflicts between Model 1 and Model 2 should not be interpreted as conflicts in the founder haplotype phase assignment. Indeed phase the founder haplotypes, we need to consider the ensemble across non-founders (i.e. all the rows in Figure 5.11), which results in consistent founder haplotypes between the approaches.

5.2.10 Mapping recombination events at fine-scale through posterior decoding

To improve the resolution of the cross-overs inferred across chromosome segments, we implemented a modified HMM to use all available SNPs and analysed each event separately (others have used similar constrained fine-mapping approaches (e.g. Hinch et al., 2011)). We used the Forward-Backward algorithm applied to SNPs in the 1Mb surrounding cross-over regions identified at the inheritance vector scaffold level. To guard against genotyping errors we used the inheritance vector likelihoods (Model 1) as our observation distribution (See Section 5.4). We incorporated the scaffold information by constructing the initiating and terminating state space distributions to reflect the state in the Viterbi path for the left- and right-hand-side neighbouring chunks. Specifically, we set these distributions to place all probability on the initiating, or terminating Viterbi state. We then applied posterior decoding to determine the location of the candidate cross-over event, assessed as the switch in posterior probability dominance for the two Viterbi states. Cross-over breakpoints were defined to be the 95% credible interval between inheritance vector change points. The resolution of pedigree maps depends on the number of meioses and genetic diversity

within the individuals. For the 375 cross-over events identified, the median credible interval span was 856 bp spanning 2 sites (Figure 5.12).

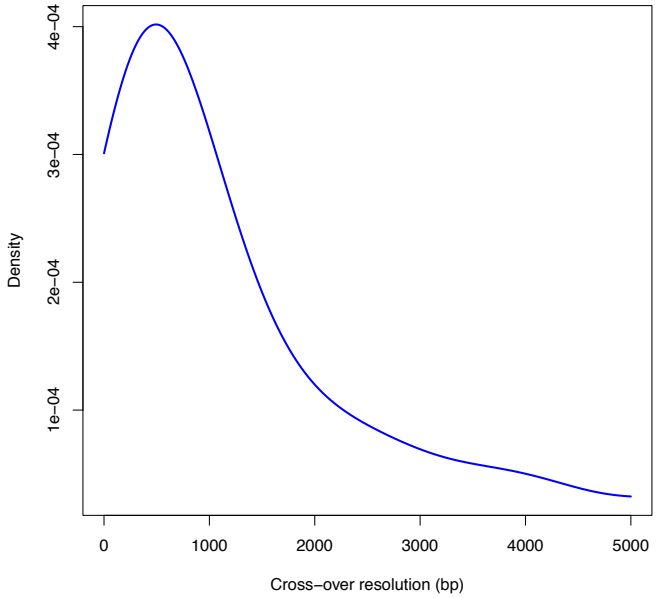


Figure 5.12: The resolution of identified cross-over breakpoints defined as the 95% credible interval between inheritance scaffold changes.

5.2.11 Genome-wide map of cross-over events

We applied Model 2 to the autosomes and identified cross-over events genome wide (Figure 5.14). We were reassured that the regional recombination inference approach (i.e. inferring an inheritance vector scaffold across chunks), eliminated the need for masking recombination events: no individual (Figure 5.16), nor chromosome exhibited unexpected levels of cross-over (Figure 5.14).

To check for false positives, we assessed regions that have been problematic for other studies (Broman et al., 1998; Kong et al., 2010, 2002), specifically centromeric and telomeric regions, for double cross-over events. We infer 3 double cross-over events (defined as two cross-overs on the same inherited chromosome within 2 Mb).

Additionally, we validate the LD-based map by detecting a cross-over localisation in rate estimates at the pedigree-inferred cross-over breakpoints (Figure 5.13).

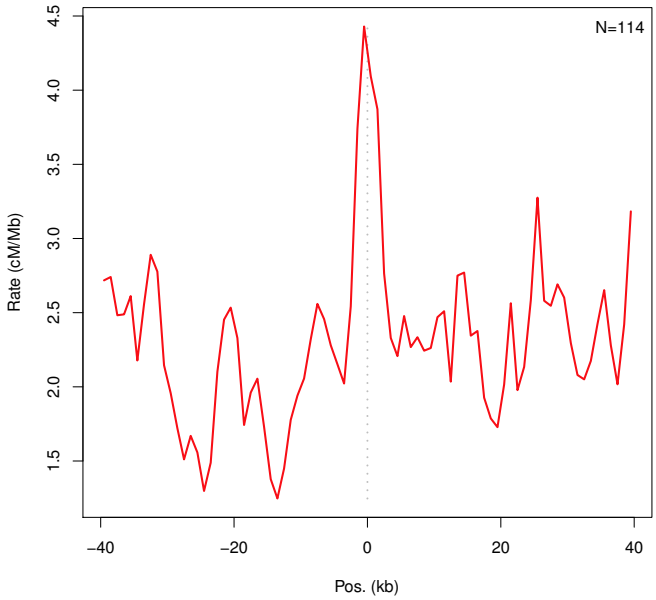


Figure 5.13: Fine-scale recombination rate variation in LD-based estimates around cross-over breakpoints identified in the pedigree subsampled such that events separated by at least 50 kb.

However, there are some examples of problematic regions. In particular, there are examples of poorly assembled regions with high levels of inferred recombination (e.g. the first 10 Mb of chromosome 15). These could be potentially addressed by generating individual specific masks such as those described in Chapter 2. However, not all regions of high activity are necessarily artefactual, for example the clustered activity at the telomeric region of chromosome 12q has sex-directionality. Given that the number of female events dominate male events, it is exceedingly unlikely that such a distribution would be generated through artefact (i.e. we would expect a 60:40 ratio in female to male events in the cluster). Across chromosomes male-specific cross-over increases with proximity to telomere (Figure 5.15).

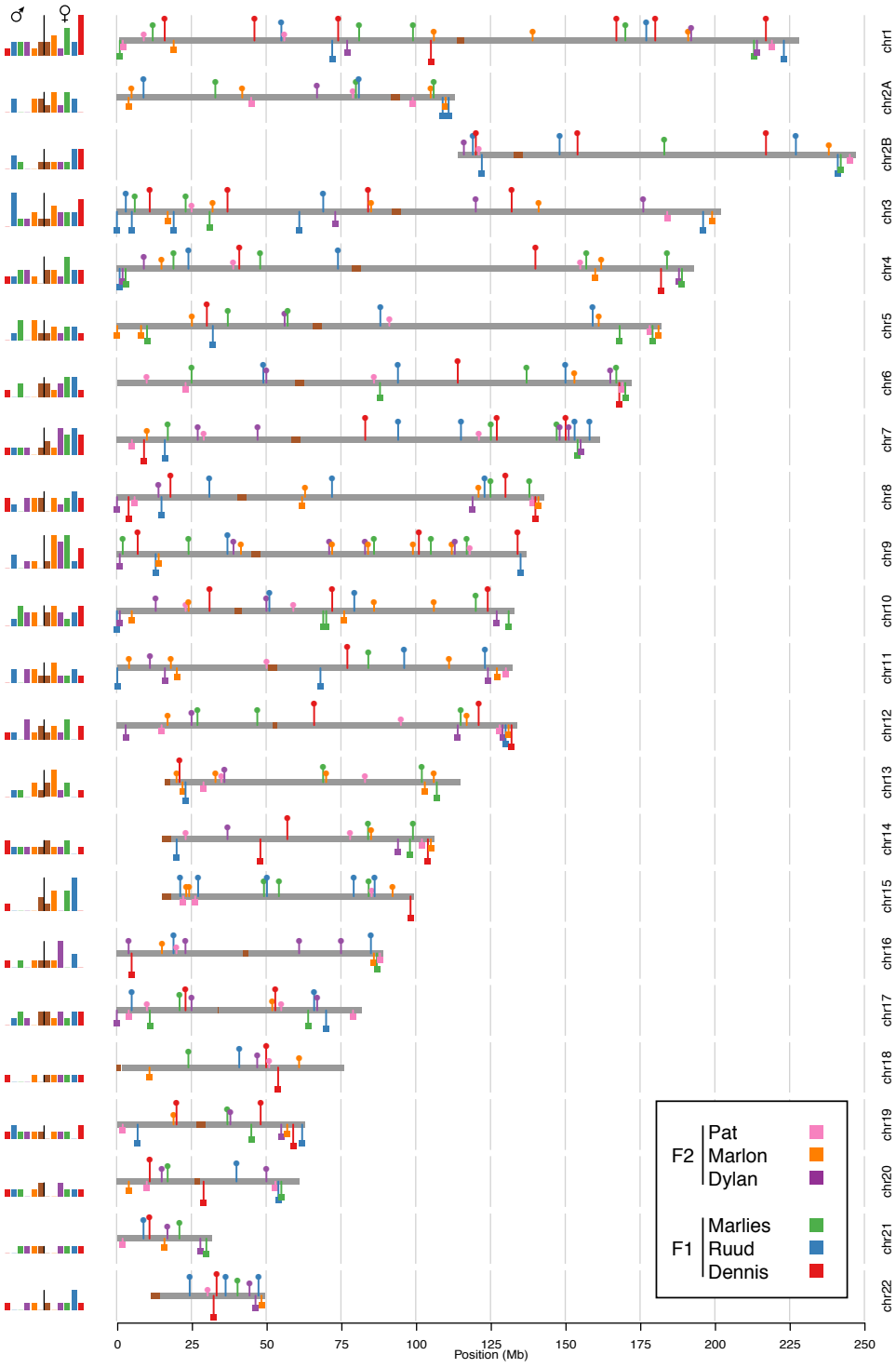


Figure 5.14: Map of cross-over events across the genome. Chromosomes are represented by grey rectangles, and brown rectangles represent the estimated location of centromeres. Events are represented by coloured ticks (see key), squares represent events that occurred in the paternally inherited chromosome, circles represent events that occurred in the maternally inherited chromosome. Bar chart on left hand side represents the total counts for each non-founder.

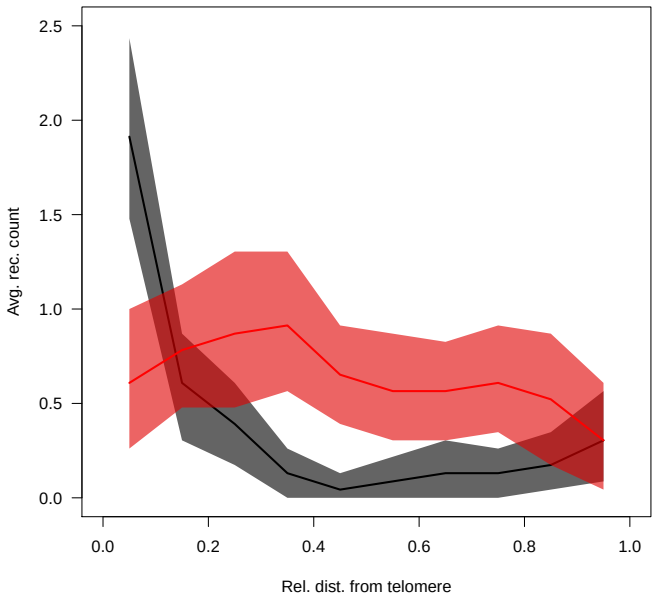


Figure 5.15: The average distribution of recombination events across chromosomes for male-specific (black) and female-specific (red) cross-overs. Bands represent 95% confidence intervals and were calculated via bootstrap (N=1000).

5.3 Estimates of genetic map length in male and female chimpanzees

To compare estimates of genetic distance estimates in chimpanzees, we converted the population-scaled, LD-based estimates from the 10 unrelated Western chimpanzees to centi-Morgans using an effective population size used by Auton et al. (2012), $N_e = 11,413$.

The total genetic distance estimated from the Western chimpanzee pedigree (3,448.4 cM; Table 5.2) is very close to humans (3,435.7 cM; Kong et al., 2002). Given that our estimate potentially contains some errors, the current estimate likely represents an upper bound on total genetic distance in chimpanzees. Note that the LD-based estimate of the total genetic distance in chimpanzees, using $N_e = 11,413$, is 16% larger than the pedigree estimate (Table 5.2). This suggests that the effective population size of Western chimpanzees is closer to 9,000 (in agreement with previous estimates; Caswell et al., 2008). At the chromosome level, there is strong sex-specific differentiation in chimpanzee recombination rate variation ($P = 8.32 \times 10^{-5}$, paired t-test and Figure 5.17).

In total, male chimpanzees have 60% of the estimated genetic distance of females, which we believe is a robust observation. In contrast, noise arising from the low number of meioses likely drives the lower correlation between sexes in chimpanzees compared to human (Table 5.3).

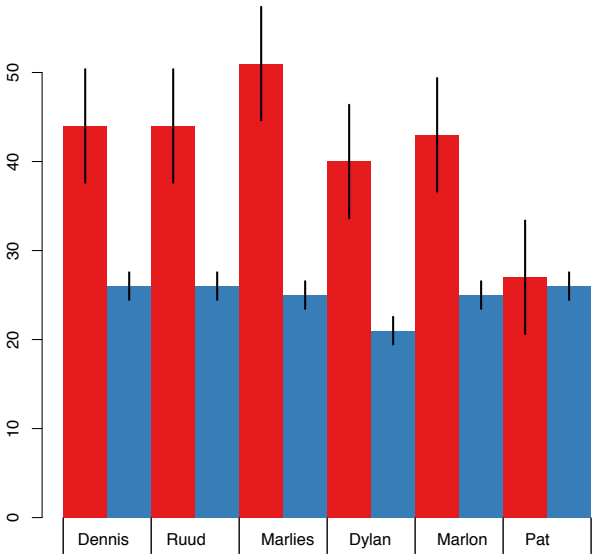


Figure 5.16: Total number of cross-over events per individual. Red represents female-specific cross-overs, blue represents male-specific cross-overs. black lines represent 95% CI for female and male cross-over counts across the individuals.

	Human pedigree			Chimp. pedigree			Chimp. LD
Chr.	Male	Female	Avg.	Male	Female	Avg.	Avg.
1	195.1	345.4	270.3	259.9	311.9	285.9	305
2	189.6	325.4	257.5				
2A				69.3	207.9	138.6	160
2B				69.3	190.6	130.0	183
3	160.7	275.6	218.2	173.3	242.6	207.9	260
4	146.5	259.1	202.8	138.6	242.6	190.6	239
5	151.2	260.2	205.7	173.3	156	164.6	223
6	137.6	241.6	189.6	86.6	259.9	173.3	229
7	128.4	230.3	179.3	121.3	277.3	199.3	262
8	107.9	209.9	158.9	156	173.3	164.6	186
9	117.3	198.2	157.7	104	329.2	216.6	181
10	133.9	218.1	176	138.6	225.3	182.0	188
11	109.4	195.5	152.4	121.3	156	138.6	172
12	135.5	206.6	171.1	138.6	156	147.3	202
13	101.3	155.9	128.6	86.6	173.3	130.0	147
14	94.6	142.4	118.5	121.3	121.3	121.3	122
15	102.6	155	128.8	52	294.6	173.3	147
16	108.1	149.6	128.9	69.3	138.6	104.0	145
17	108.6	161.5	135	104	173.3	138.6	150
18	98.6	142.6	120.6	69.3	121.3	95.3	110
19	92.6	126.8	109.7	121.3	86.6	104.0	144
20	74.7	122	98.3	104	86.6	95.3	111
21	47.3	76.4	61.9	69.3	69.3	69.3	55.9
22	49	82.8	65.9	34.7	121.3	78.0	82.2
Total	4,450	2,590	3,164	2,634	4,263	3,448	4,004

Table 5.2: Total map length in centi-Morgans estimated from human Icelanding pedigrees Kong et al. (2002), the Western chimpanzee pedigree, and estimated through LD-based inference on 10 unrelated Western chimpanzees. For the pedigree maps the Kosambi map function was used to convert recombination fractions to genetic distance, and an effective population size of 11,413 was used to convert the LD-based rate estimates.

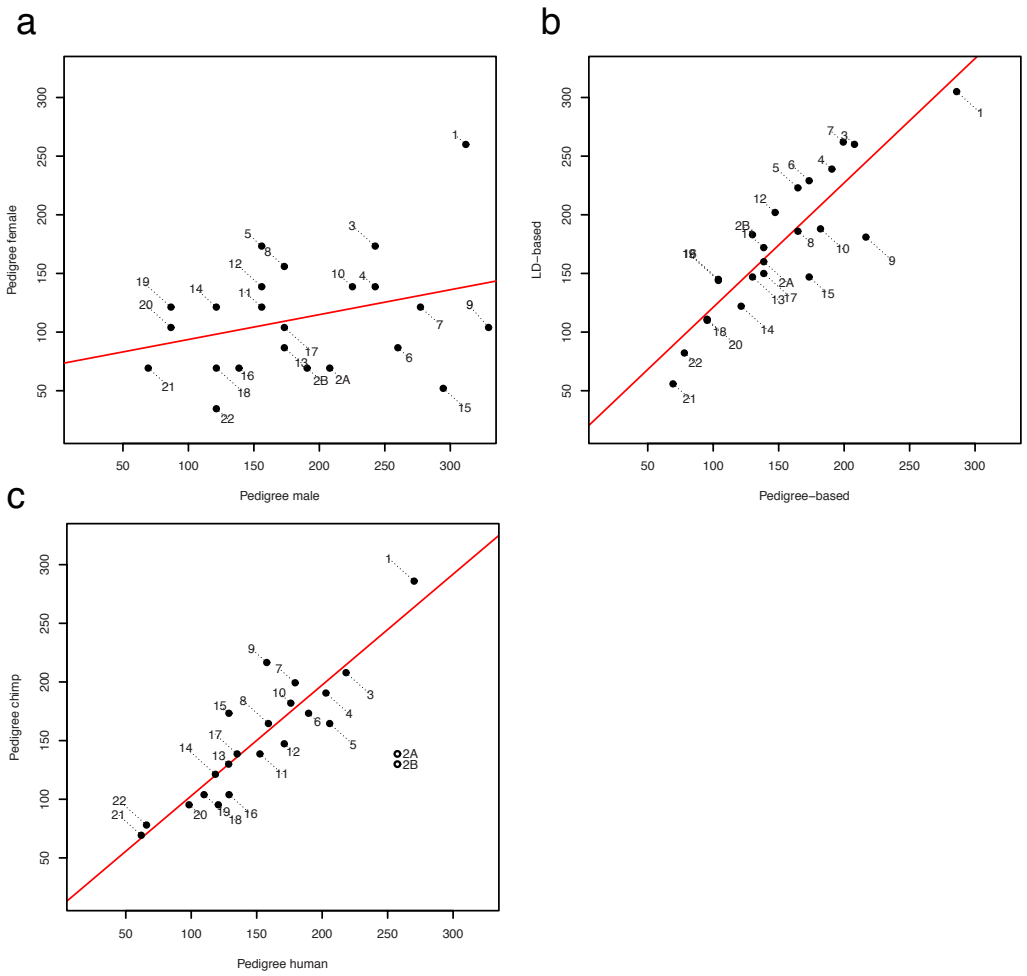


Figure 5.17: Comparisons of total genetic distances for **(a)** male versus female recombination rates in the chimpanzee pedigree, **(b)** the sex-averaged genetic distance in chimpanzee pedigree versus LD-based estimates from the HTS of the 10 unrelated chimpanzees, **(c)** the sex-averaged genetic distance for the chimpanzee pedigree and the sex-averaged human map from Kong et al. (2002). The two hollow circles represent the genetic distances of chimpanzee chromosomes 2A and 2B versus human chromosome 2.

		HP		CP			CU
		Female	Avg.	Male	Female	Avg.	
	Male	0.98	0.99	0.82	0.74	0.92	0.95
HP	Female			0.83	0.76	0.94	0.97
	Avg.			0.83	0.76	0.94	0.97
	Male				0.39	0.76	0.77
CP	Female					0.89	0.79
	Avg.						0.93

Table 5.3: Correlation between total map lengths for syntenic chromosomes. Abbreviations, *HP* human pedigree map from Kong et al. (2002), *CP* chimpanzee pedigree map, *CU* chimpanzee LD-based map from 10 unrelated chimpanzees. Chromosomes 2, 2A, 2B, 16, 9 excluded from calculation (see text).

5.3.1 Anatomy of a single and a double cross-over event

The resolution of our HTS experiment allows fine-scale mapping of cross-overs. An informative example is the region on chromosome 21 between 21,001,558 bp and 22,000,294 bp which contains two inferred recombination events: a single cross-over event and a double cross-over event.

We find convincing evidence for a single cross-over event between 21,635,789 bp and 21,638,700 bp (Figure 5.18), where the posterior probability of the inheritance vector (00,11,10,00,11,10) switches from 0.98 to 1×10^{-29} and conversely the probability of inheritance vector (00,11,00,00,11,10) increases from 4×10^{-6} to 0.98. Hence we conclude that a cross-over event occurred in the chromosome transmitted from Pearl to Marlies.

Upstream, we find evidence for a double cross-over that spans 18,998 bp and has a posterior probability of over 99.99% for the inheritance vector (00,11,10,00,10,11) across 18 sites; we conclude that a double cross-over event occurred in the chromosome transmitted from Marlies to Dylan. I note that a contig junction occurs inside the double cross-over region (between BS000195.1 and BS000196.1), and that coverage is decreased at the interface between the two contigs. However the alignments in the double cross-over region are otherwise well behaved, and we see no obvious artefacts underlying the inference. Across all the recombination events the median fine-mapping resolution is 1,317 bp, with 30% of events being resolved to under 537 bp.

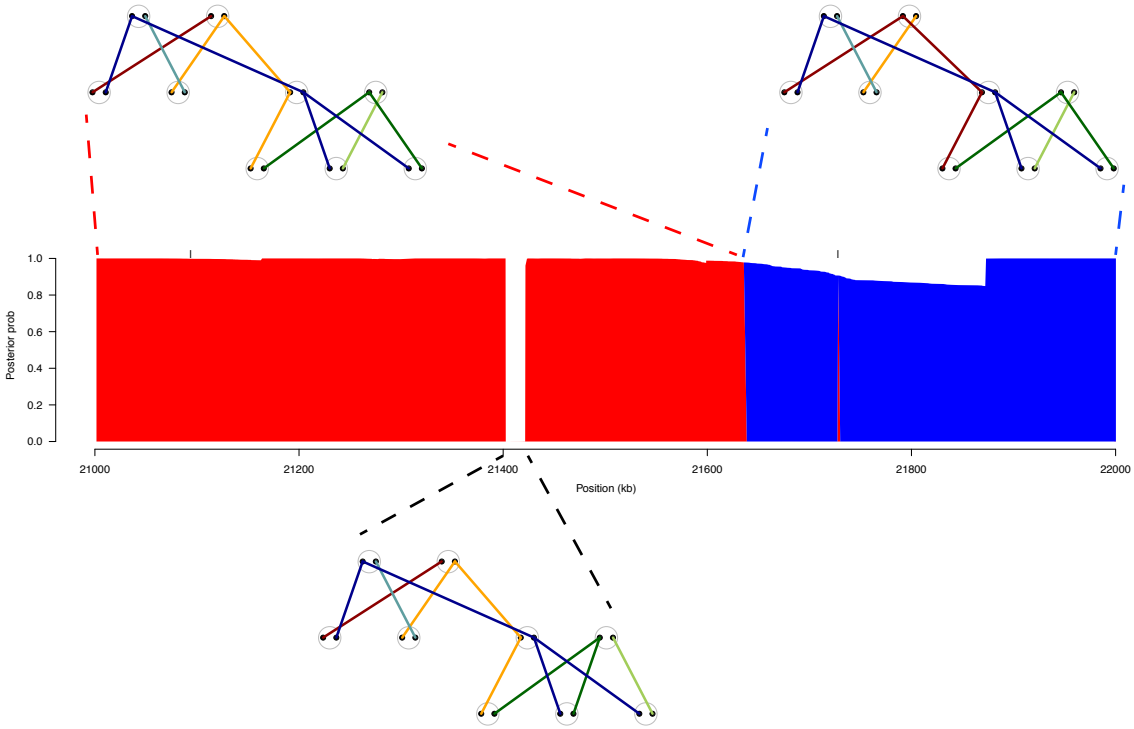


Figure 5.18: Posterior decoding of two recombination events. Posterior decoding over chromosome 21 between 21,001,558 bp and 22,000,294 bp. Inheritance vectors which maximise the posterior probability for the three regions are indicated. Black tick marks represent the hard called genotype vectors that are Mendel inconsistent.

5.4 Anatomy of Mendel inconsistent genotypes

Additionally, the inheritance information allows us to resolve non-standard events, and diagnose problems. Sites with MIC genotype vectors, reflect either *de novo* mutation events, transmission of heterozygous deletions, genotyping error, or false positives. Consequently, a major challenge is distinguishing true events from artefacts. Here I analyse MIC identified from the hard genotype calls.

Consider the same region on chromosome 21 (containing the double cross-over and single-cross over events), there are two MIC sites in this region.

First, we show an example of a how uncertainty successfully prevents a false MIC call. Consider the left-most hard called MIC candidate in Figure 5.18 at 21,093,998 bp on chromosome 21. The genotype likelihoods reflect high levels of uncertainty in the genotype generating the inconsistency (Figure 5.19a, Dirk1). This uncertainty arises because of low coverage, only five reads cover the variant site in Dirk1; four match the reference allele and one matches the alternative allele. When uncertainty is incorporated, as in the recombination inference, the false signal is removed, for instance the posterior probability profile at the site is consistent the local decoding. Hence, incorporating uncertainty successfully eliminates this false signal.

Second, we show an example of where genotyping error results in a cryptic MIC error. The right-most MIC candidate in Figure 5.18 at 21,728,271 bp on chromosome 21 has an inconsistent posterior probability relative to the local decoding. This arises because the genotype likelihood is incorrect (Figure 5.19, Marlies); of the 20 reads covering the site in Marlies, four contain the reference allele, the other 15 correspond to the alternative allele. Despite the evidence for a heterozygous genotype in Marlies, the relative likelihood of the data for the heterozygous call is estimated to be $10^{8.5}$ times less likely than the homozygous alternative model. Consequently, incorporating uncertainty does not solve the issue and the posterior decoding is distorted (Figure 5.18).

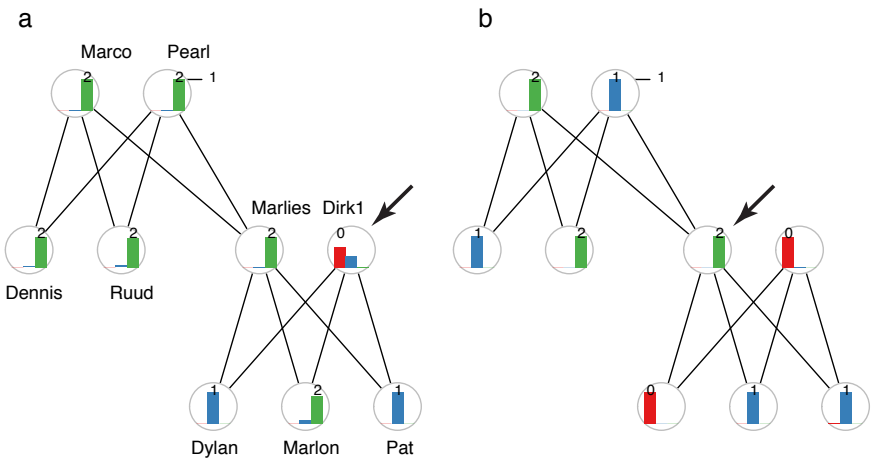


Figure 5.19: False Mendel inconsistent (MIC) signals generated through genotyping error. Hard-called genotypes represented by $\{0, 1, 2\}$ corresponding to homozygous reference, heterozygous, homozygous alternative genotypes respectively. Underneath the estimated posterior probability of the three genotype models. Arrows indicate errors. **(a)** Platypus called Mendel inconsistent genotype generated by genotyping error in Dirk1. Inspection of alignments indicate that Dirk1 is heterozygous at the site thereby resolving the Mendel inconsistency. **(b)** Platypus called Mendel inconsistent genotype generated by genotyping error in Marlies. Marlies is assigned a homozygous alternative genotype despite four of the 20 aligned reads containing the reference allele. This is an example of an poorly calibrated genotype likelihood. The sites correspond to MICs indicated in Figure 5.18

Most, but not all MICs are artefacts. For example, we found evidence for six MIC sites in a L1M interspersed repeat element on chromosome 21 between 497,732 bp and 506,610 bp (Figure 5.20). The inconsistency arises because Pearl is heterozygous for a $\sim$ 10kb deletion, which she transmits to her offspring Dennis and Marlies (Figure 5.21).

Analysis of MIC are a promising approach to prioritise of regions of interest, however, the focus of this Chapter is to investigate recombination detection in the pedigree. I therefore do not discuss methods to differentiate between sequencing error and *de novo* mutation, but I note that this is an active area of current research.

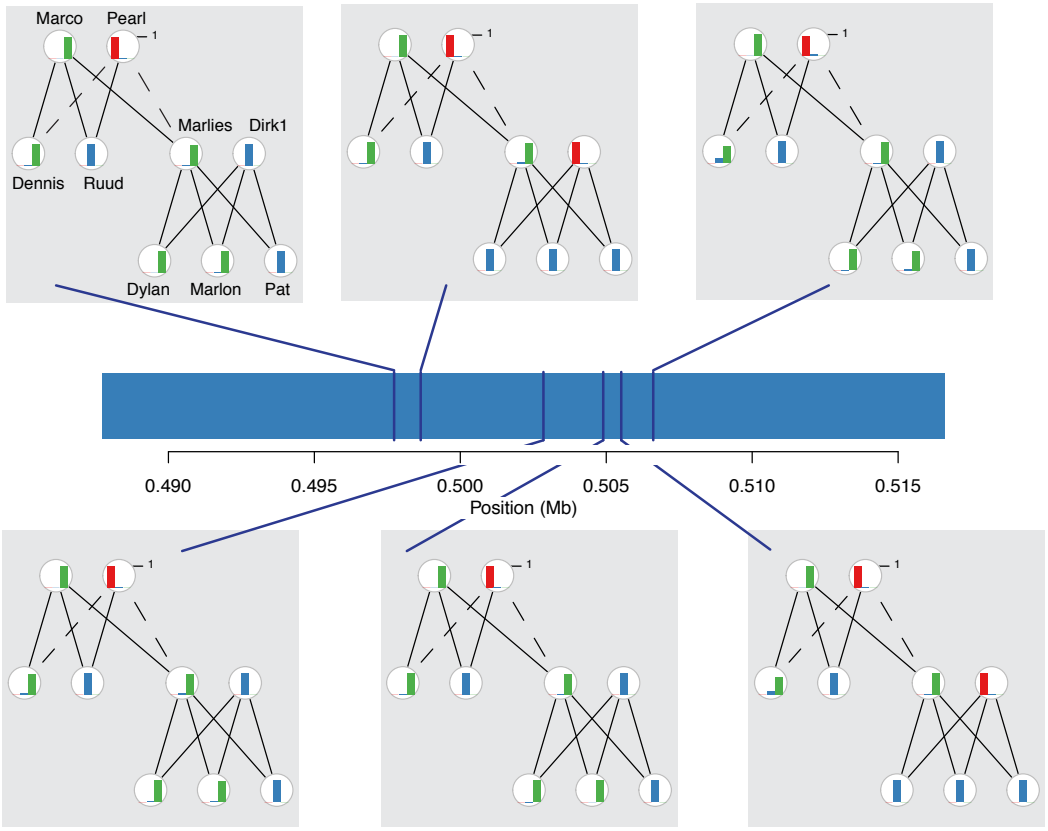


Figure 5.20: Genotype probabilities for sites in a segregating deletion on chromosome 21. Six grey panels show informative sites in the deletion. Red, blue, and green bars represent the posterior probability for homozygous reference, heterozygous, and homozygous alternative genotype models calculated under a uniform prior. Dashed lines indicate transmission of the deletion generating hemizygous alleles in Pearl, Dennis and Marlies. There is no evidence in the read data for Marlies transmitting the deletion to her offspring.

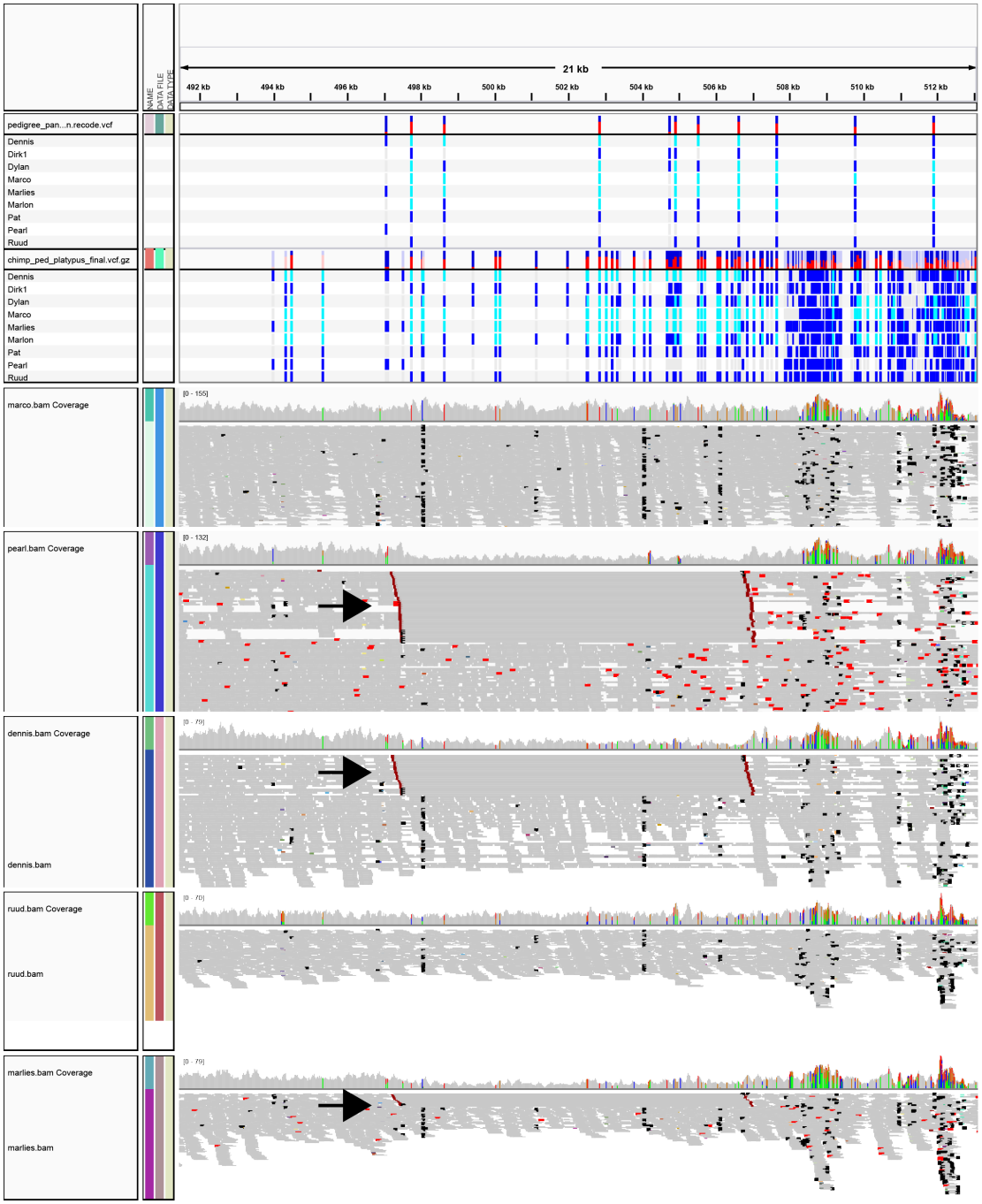


Figure 5.21: A segregating deletion. Alignments on chromosome 21 for first and second generation of the chimpanzee pedigree revealing a heterozygous deletion in Pearl that is transmitted to Dennis and Marlies generating seven Mendel inconsistent sites (see Figure 5.20). The top panel represents the genotypes for sites in the intersection set. The second panel represents Platypus calls. Grey sites represent homozygous reference, turquoise heterozygous, and blue homozygous alternative genotypes. Beneath are the alignments for Marco, Pearl, Dennis, Ruud, and Marlies. Horizontal grey rectangles represent reads, black rectangles indicate deletions, crimson rectangles represent inverted mate-pairs, and red rectangles represent reads with unmapped mate pairs. The heterozygous deletion is detected in Pearl, Dennis and Marlies (indicated by arrows).

5.5 Discussion

In this chapter I have quantified the sex-specific cross-over differences in Western chimpanzees and their total genetic distance. These results are concordant with observations in humans both in that males have ∼60% of the recombination of females, and similar levels of recombination are observed between the species. We do not see sex-specific differences in the distribution of cross-over events across chromosomes, though female rates seem to be more variable than males (Figure 5.15). I explain how uncertainty can be incorporated into analyses, and utilise inheritance vectors to both diagnose errors and identify cross over events across the chimpanzee genome. HTS provides the opportunity to detect a wide spectrum of genome changes and has enabled incredible advances in cataloguing SNPs, indels, and copy number changes. However, a major challenge in detecting these events is sifting signal from noise. Moreover, many changes that impact the genome have not yet been assessed using a HTS approach (microsatellites, minisatellites, gene conversion events, events around cross over and non cross over, or non allelic homologous recombination). These data provide the additional opportunity for accurate detection rare events in large genomes - a continuing research aim in this experiment.

Chapter 6

Discussion

This work is built upon, in principle, a modest HTS experiment. We estimated recombination rates from haplotypes inferred from 10 unrelated Western chimpanzee genomes sequenced to moderate coverage ($\sim 10\times$). In practice, the experiment was incredibly challenging and motivated methods to (1) incorporate statistical uncertainty arising from low coverage, (2) counteract sequencing error, and (3) overcome mapping artefacts arising from the deficiencies in the reference sequence. We demonstrated that the approach is not only feasible, it is also incredibly informative about biological processes. With current sequencing costs it would be possible to repeat the experiment at $15\times$ coverage with 15 individuals, which would reduce genotype error thereby improving rate estimates. Hence, the chimpanzee sequencing project could be considered a lower bound for the approach, and given its success, consequently opens the door for studying recombination rate variation at hotspot-level resolution in almost any species with a reference sequence.

Additionally, our experience in discovering and counteracting novel error modes in population sequence data can be translated into improved experimental design. Currently, frustratingly little of the potential information in HTS is utilised; predominantly limited by the higher error rates and statistical uncertainty in sequencing that

overwhelm true signal, especially for rare events. Improvements will come through reduced sequencing error, improved calibration of the measured uncertainty in assemblies and alignments, and more advanced variant detection models (of which *de novo* assembly of variant calls is particularly attractive).

A complementary and powerful approach to control the signal to noise ratio is to sequence pedigrees or population isolates, where descent can be used to establish robust catalogues of variation. Indeed, leveraging transmission potentially enables, with current technology, extensive and accurate study of a wide range of genome changes that impact large genomes. To this end, an on-going aim for the chimpanzee pedigree sequencing experiment is to develop methods to detect non-allelic homologous recombination, gene conversion, events around complex cross-overs, mircosatellites and minisatellites.

With respect to analyses of recombination, extended pedigree sequencing can augment LD-based analyses. LD-based analysis of population samples is an extremely powerful approach to estimate recombination rates at fine scales. However, there are limitations: (1) the inferred rate estimates are sex-averaged, and (2) only single cross-over events are modelled (though there has been attempts to incorporate gene-conversion; e.g. Hellenthal and Stephens, 2007). Through combining these approaches, the short comings in both analyses could be addressed, and would provide insight into how recombination operates over different historical timescales.

It has been extremely exciting to participate in a field invigorated by a recent discovery. PRDM9 has changed our thinking about the recombination process and, as the first example of genome-wide genome rearrangement modifier, has wider biological implications. This extends not only to differential hotspot usage, but likely differential chromosome instability (Myers et al., 2008) and therefore may provide insight into how genomes evolve structurally. One prediction would be different chromosome breakpoint syndromes in humans carrying different PRDM9 alleles, or more precisely

C-allele carriers would likely exhibit additional features to A-allele carriers.

I raised three questions in the Introduction Chapter regarding PRDM9. Given what we have learned about recombination in chimpanzees, what can we answer?

- What is the exact relationship between PRDM9 variants, their predicted motifs, and the resulting recombination activity?

There is convincing evidence in humans of PRDM9 motif specificity, both through differential localisation signals in human populations of different ancestry, and *in vitro* confirmation of binding predictions of human A- and I-alleles (Baudat et al., 2010). Indeed, the lack of overlap between human and chimpanzee hotspot locations is compatible with PRDM9 driving all hotspot locations. There are five additional, complementary strands of evidence that strongly suggest that PRDM9 specifies the location of all hotspots.

1. Sperm-typing at 17 hotspots indicated that loci, including those with no exact match to the 13 bp or 17 bp motif, were influenced by the *Prdm9* genotype (Berg et al., 2011, 2010).

2. Differential hotspot activity is observed both within species (Hinch et al., 2011), and between species (Auton et al., 2012).

3. Studies in humans and cattle show overwhelming signals of genome-wide association for hotspot usage at *Prdm9* in humans ($p < 10^{-200}$ Hinch et al., 2011; Kong et al., 2010), and differential regional cross-over intensity between alleles in cattle (Sandor et al., 2012).

4. ChIP-Seq experiments in mice against DMC1 foci in PRDM9 knockouts (Brick et al., 2012) suggest that PRDM9 determines the placement of almost all cross-over hotspots in mice apart from the PAR - the only locus in mammalian

genomes that undergoes recombination in every single viable meiosis (I will return to this observation later).

5. I estimated fine-scale recombination rates on chromosome 19 from the 20 wild mice (the HTS data are described in Section 2, note, however, that calls were filtered using VQSR with false positive rate ¡ 0.01%). The average fine-scale recombination rate profile around the Brick et al. (2012) DMC1 ChIP-Seq hotspots show strong, differential, local activity (Figure 6.1a). Given the evidence in human and chimpanzee, we can hypothesise that the differential signals arise from the PRDM9 alleles in the sample of 20 mice matching the 9R (B10.S-H2^{t4}/(9R)/J) and B6 (C57BL), and not the 13R (B10.F-H2^{pb1}/(13R)/J). Though the authors do not describe the ancestry of the donor strains used to derive 9R and 13R, which appear to be admixed, I found that the amino acid sequence for 9R exactly matches the *b*-haplotype reported by Baudat et al. (2010) from *M. m. domesticus*. In contrast 13R has six ZFs discordant with the *PRDM9* allele from the *b*-haplotype (ZF 4-9; Figure 6.1b). We note that there is no deviation in recombination rate above background in the 13R mice (red line), in contrast to the ∼6.4-fold increase in rate at DSB hotspots in 9R and B6. Lines 9R and 13R share almost identical genetic backgrounds but have different PRDM9 alleles (cf. Sup. Fig. 1; Brick et al., 2012), therefore we might speculate that the observations are compatible with all recombination hotspots in mice are driven by PRDM9 binding.

   Additionally, Brick et al. (2012) observe that the 13R determines most hotspot locations in F1 mice of 13R×9R reciprocal crosses (for the data displayed 86% (5,589) are 13R-derived hotspots and 12% (841) are 9R-derived). However, we observe that despite this dominance relationship, the 9R DSB hotspots in the F1 cross (purple line) show a 2.6-fold increase in rate, providing strong evidence that heterozygous PRDM9 allele combinations result in a superposition

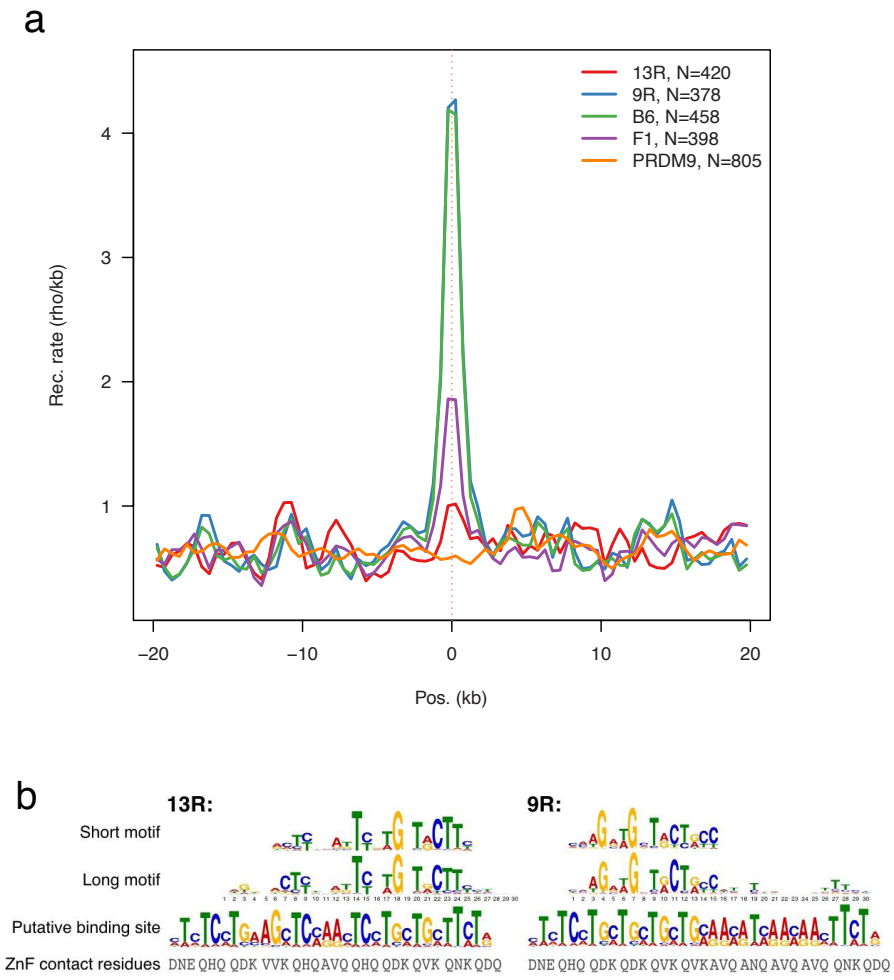


Figure 6.1: Average profile of fine-scale recombination rate variation estimated from the 20 wild mice at anti-DMC1 ChIP-Seq foci measured in different mouse strains by Brick et al. (2012). **(a)** Legend indicates the different mouse strains, and their associated ChIP-Seq peaks. Mice strains *9R* and *13R* share almost identical backgrounds but have different PRDM9 alleles. *B6* represents the C57Bl/6J strain. F1 represents the progeny of a 9R♂×13R♀ cross. *PRDM9* represents PRDM9 null mice B6 mice. Recombination rates estimated through LDhat applied to genotypes called in 20 wild mice. Variants were called and genotyped through the GATK using VSQR against SNP-chip data and additionally filtering variants to have read depth > 5 and genotype quality > 10. Genotype call set generated by Robbie Davies. **(b)** Is taken from Brick et al. (2012, Sup. Fig. 14) and shows the MEME-ChIP (Machanick and Bailey, 2011) identified motifs enriched in DSB hotspots (*short motif* and *long motif*) and the PRDM9-binging prediction from amino acid sequence.

of PRDM9-dependent maps. Note, however, that the localisation signal in F1 is 72% larger than would be predicted from the reported DSB allele-derived co-dominance (i.e. the percent 9R-derived hotspots in F1 times the localisation signal attributed to 9R plus background rate is 1.08 ρ/kb), implying a more complicated interaction between DSB, PRDM9 alleles and the resulting hotspots. PRDM9 dependence is further demonstrated at *Prmd9*$^{-/-}$ DSB hotspots that show no deviation from background rates (orange line). This indicates that with knockdown of PRDM9 DSB shifts to different locations not correlated with PRDM9 sites, and suggests that PRDM9 is necessary for hotspot positioning. Sequencing *Prdm9* in the 20 mice would make this conclusive.

In addition, Brick et al. (2012) also measured PRDM9-derived H3K4Me3 marks across the samples. Figure 6.2 indicates that PRDM9-specific H3K4Me3 marks are neither necessary nor sufficient for hotspot activity, but are associated with a subset of the resulting hotspots (less than 20% of the marks have a localisation signal above background rate).

These distinct lines of evidence show that value of comparative studies in identifying important components of the recombination process. The next question is motivated by the evidence for PRDM9 localising all hotspots, but the current low proportion of hotspots explained by PRDM9-bound motifs.

- How might PRDM9 influence hotspots without a clear match to the motif?

By aggregating fine-scale recombination rate variation profiles over different genome features in humans and chimpanzees, we observed that although hotspots evolve rapidly, the way in which they interact with genomic features can influence their placement. In particular CpG islands, the genomic features responsible for maintaining gene transcription activity (Ramirez-Carrozzi et al., 2009), are strong attractors in both species, but particularly in chimpanzees. Combined with the absence of any

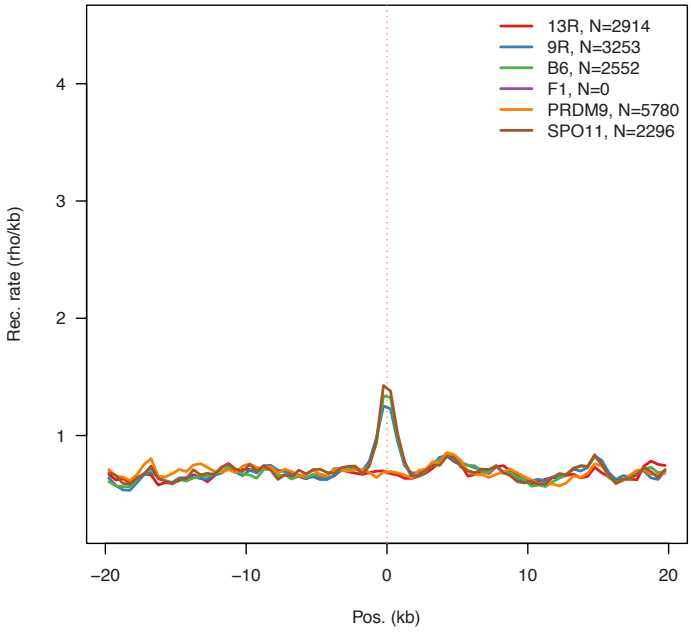


Figure 6.2: Average profile of fine-scale recombination rate variation estimated in 20 wild mice at anti-H3K4Me3 ChIP-Seq foci in different mouse strains. Legend indicates the different mouse strains from Brick et al. (2012), and their associated PRDM9-derived ChIP-Seq peaks. Mice annotation and recombination map as for Figure 6.1.

consistent evidence for localising activity at PRDM9 (sub)motifs in chimpanzees, even after accounting for shared motifs amongst PRDM9-binding predictions, this may indicate that chimpanzee hotspot placement is more reliant on chromatin state than primary sequence.

This hypothesis is not necessarily contradicted by observations in humans that point mutations in the 13 bp motif extinguish sperm-hotspots (Jeffreys et al., 2005; Jeffreys and Neumann, 2002; Jeffreys et al., 2001), which have been interpreted as evidence of strong PRDM9 motif specificity. However, the abundance of the 13 bp motif, and additionally the constellation of degenerate near-matches, the 'extinguished' hotspots could easily be explained by a shift in hotspot location to a nearby alternative match of marginally greater degeneracy, nested in the same region of accessible

DNA. Since sperm-typing has limited scope, even shifts of a hundred bases would not be detected. The 17 sperm typing results could therefore be interpreted as an indication of permissive sequence specificity in PRDM9. If motif binding is more permissive that currently thought, degenerate motifs could quite possibly underlie all recombination hotspots. To investigate this hypothesis I calculated the localisation signal at motifs with greater levels of degeneracy. Figure 6.4 indicates extended degeneracy, even for the allele with the strongest motif association - the A-allele. The degenerate 13 bp motif CCNCCNTNNCCNC results in a 43% increase in rate; introducing a one nucleotide change to the motif still results in a 17% increase in rate.

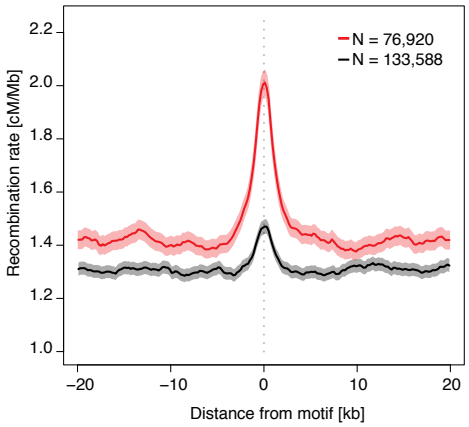


Figure 6.3: Fine-scale recombination rate variation in humans around matches to the degenerate 13 bp motif (red), and matches to motifs one nucleotide change away from the degenerate motif (black). The average recombination rate profile was estimated from the combined-population HapMap recombination map over 500bp intervals after thinning to remove motif clustering.

Indeed, Panizza et al. (2011) propose an *axis tethered DSB machinery* model for recombination initiation in yeast, where only one recombination-initiating DSB event occurs amongst a plethora of local candidate sites (Figure 1.2). Yeast provide an interesting reference point: no *Prdm9* ortholog has been found, nor a recombination motif, and it is thought that DSB hotspot locations are determined by factors such as chromatin state (Lichten and de Massy, 2011). Consequently, diverged *Sacchromyces*

species (more than 10× more divergent than humans and chimpanzees), show substantial hotspot overlap Tsai et al. (2010). Hence, at one end we have yeast and other species with no functioning PRDM9 whose hotspots are governed by chromatin state, and on the other we have humans in which hotspot placement is PRDM9 driven by motifs. These two examples could potentially mark out extrema on a plane of joint interaction between chromatin state and primary sequence affinity. In other words, for species with functional PRDM9, the protein may modulate hotspot locations through joint motif and chromatin accessibility.

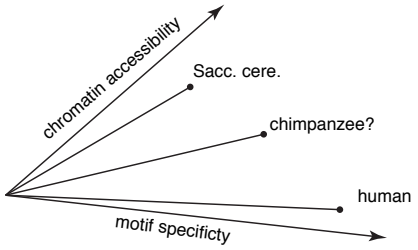


Figure 6.4: Space of interactions determining recombination hotspot placement.

To explore the impact of motif affinity on PRDM9-based hotspot placement, I consider a simplistic, enzyme kinetics based model where PRDM9 interacts reversibly with DNA through its zinc finger array, and then marks regions initiating hotspots through methylating Histone 3K. The reaction is written:

$$[\mathrm{E}] + [\mathrm{S}] \underset{\mathrm{k_r}}{\overset{\mathrm{k_f}}{\rightleftharpoons}} [\mathrm{ES}] \xrightarrow{\mathrm{k_{cat}}} [\mathrm{E}] + [\mathrm{P}]$$

where $[S]$ represents the concentration of unmarked motifs, $[E]$ the concentration of PRDM9, $[P]$ the concentration of H3K4Me3 marked motifs, k_f the forward binding constant, k_r the dissociation constant, and k_{cat} the turnover number. We assume that the concentration of the DNA-PRDM9 bound intermediate $[ES]$ does not change on the time-scale for histone methylation (i.e. we assume quasi-steady-state kinetics). Then the velocity ν of the reaction is described by the Michaelis-Menten kinetic equation:

$$\nu = \frac{V_{max}[S]}{K_m + [S]}, \tag{6.1}$$

where K_m and V_{max} are defined as

$$V_{max} = k_{cat}[E]_0 \tag{6.2}$$

$$K_m = \frac{k_{cat} + k_r}{k_f}, \tag{6.3}$$

and $[E]_0$ represents the initial concentration of PRDM9 in prophase I. We then assume that k_{cat} is less than both k_f and k_r. An informal motivation for this assumption can be made as follows: consider the activation energy of the two processes, the DNA interactions are electrostatic, whereas histone modification requires the formation of covalent bonds, therefore the rate of DNA interactions is more rapid than the histone modifications. We then consider two cases.

Case 1: $\mathbf{k_f \approx k_r}$. We assume that PRDM9 binding is very permissive; it binds and releases motif candidates at the same rate. Then the velocity of the reaction is:

$$\nu_1 = \frac{V_{max}[S]}{\cancel{K_m}^{1} + [S]} \tag{6.4}$$

$$= \frac{V_{max}[S]}{1 + [S]}. \tag{6.5}$$

Case 2: $\mathbf{k_f \gg k_r}$. We assume that PRDM9 binding preferential and the equilibrium is shifted far to the right, then the velocity of the reaction is:

$$\nu_2 = \frac{V_{max}[S]}{\alpha + [S]} \tag{6.6}$$

The ratio of the two cases quantifies the effect of motif specificity,

$$\frac{\nu_1}{\nu_2} = \frac{\alpha + [S]}{1 + [S]} \quad 0 \leq \alpha < 1 \tag{6.7}$$

Now let motif binding affinity increase,

$$\lim_{\alpha \to 0} \frac{\nu_1}{\nu_2} \to \frac{[S]}{1 + [S]} \approx [S], \tag{6.8}$$

since the motif frequency is $[S] \sim 10^{-3}$ motifs per 13-mer per genome. Hence the velocity of histone modification under the stated assumptions is governed by motif concentration (and consequently DNA accessibility), not the binding specificity of the zinc finger alleles. To test this hypothesis I propose the following experiment.

Experiment 1: Evaluating the relative importance of motif specificity and chromatin accessibility in chimpanzee PRDM9. A challenge in studying PRDM9 *in vivo* is that meiotic cells are hard to obtain, there are no commercial antibodies with high specificity against the protein, and it is hard to obtain chimpanzee material. Hence, we propose an experiment following the ChIP-Seq protocol of (Noor, Altemose, Myers et al., unpublished) to assess PRDM9-binding hotspots in engineered cell lines. Their experiment involves constitutive expression of the human A-allele in the HEK293 cell line. HEK293 cells are hypertriploid and are therefore potentially prone to artefacts arising from complex karyotype. I propose to knock-in each of the three known common chimpanzee alleles into a human primary fibroblast cell line or mononuclear blood cells (which have a titre of 1^6 cells per mL. This would allow the motif specificity of the chimpanzee alleles to be assessed relative to the HEK293 results. Importantly, comparison between the alleles would enable the roles of motif specificity and chromatin state to be quantified. Some research questions would be. How many regions are bound? Are these predominantly in promoters? How do the ChIP-Seq signals relate to recombination rates? Can we identify motifs from the ChIP-Seq peaks? An adventurous extension would be to manipulate chromatin pat-

terns through chemical treatment (for instance histone deacetylase inhibitors, such as VPA (Abbot Pharmaceuticals) are well tolerated), however, this experiment would be asymmetric - a negative result would be uninformative. Similarly, the observation of differential recombination activity at hyper- and hypo-methylated gene promoters could be tested.

- Why is a gene so important in recombination evolving so rapidly, and why is such a costly mechanism maintained over evolutionary time?

Myers et al. (2010) noted that the DNA sequence of the *Prdm9* ZF domain constitutes a coding minisatellite, implying a high intrinsic mutation rate resulting from repeat instability. Additionally, patterns of evolution at the ZFs, especially the coordinated and clustered evolution at the DNA contacting residues, point towards positive selection on binding specificity (Oliver et al., 2009), however the agent of selection is unknown. The agent of selection could possibly arise from the gradual erosion of motif targets by biased gene conversion, leading to loss of fitness through insufficient cross-over to support chromosome disjunction (Coop and Myers, 2007), or through the promotion of deleterious alleles at hotspots (Boulton et al., 1997; Coop and Myers, 2007).

A challenge would be to resolve how sufficient selective pressure is exerted through DNA binding given the abundance of motif near matches in the genome. For example, the ∼50,000 hotspots identified in human are overwhelmed by the number of potential binding sites (in the human genome there are 3.8 million matches one nucleotide change away from the degenerate 13 bp motif).

I discuss two potential candidates for restricting the number of bound motifs; either at the restricted PAR sequence, and/or across the entire genome through chromatin state:

1. The requirement for obligatory recombination at the PAR is a potential can-

didate for a source of restricted motifs. In humans, (a) a crossover in this region appears to be obligatory for completion of meiosis (Hale, 1994) and X-Y segregation (Hale, 1994; Hassold et al., 1991), (b) the PAR displays a remarkably high combination rate, estimated to be 1/53 cM/kb (Rappold, 1993), (c) measurements of the overall frequency of exchange in the PAR indicate that it occurs 3 - 12 times less frequently in females than males (Lichten and Goldman, 1995). The potential impact of the PAR-PRDM9 interaction could be detected through two signals.

First, we might expect to observe lineage specific divergence as a signal of evolved sequence to suppress genetic conflict (under this model the signals of recurrent mutation at PRDM9 ZFs would be interpreted as reflecting recurrent involvement of PRMD9 in suppressing genetic conflicts). Indeed there is evidence for differences at the PAR specific DXYS20 locus, a VNTR-like minisatellite (Nakamura et al., 1987), between populations of European and admixed African American ancestry (Table 6.1; Rappold et al., 1992). Rappold et al. (1992) note that, in contrast to other PAR-repeats, the DXYS20 sequence was not observed in other any other species tested, including primates. This mechanism would be somewhat analogous to the segregation distorter locus (SD) system in Drosophila, where SD/SD^{+} males almost exclusively produce *SD* bearing children. The target of the distortion is the *Responder* locus (Rsp), which corresponds to a A/T-rich minisatellite and has differential transmission depending on repeat size (Wu et al., 1988). This hypothesis could be tested through a paired sequencing experiment across human populations (1) high-throughput capture sequencing or RFLP sequencing of individuals' PAR regions and (2) the sequencing their PRDM9 alleles.

Second, we might expect to observe sex-bias in PAR transmission resulting from hybrid incompatibility. This could be assessed through sequencing the

Table 6.1: Haplotype frequencies across DXYS20 locus in individuals of European ancestry (N=108) and individuals of admixed African American ancestry (N=82). Three haplotypes were observed A, B and C. Given the high levels of diversity observed across the minisatellite, allele C was inferred to be African-specific. Data taken from (Rappold et al., 1992).

Ancestry	A	B	C
European	0.83	0.17	0
African American	0.28	0.24	0.48

sex-polarised F1 crosses of 13R and 9R mice and comparing the distribution of DSB locations across the PAR (Figure 6.5). Since males and females have characteristically different cross-over rates across the PAR, if distortion did occur it would be observed as differential DSB hotspot intensities between the two sex-polarised classes.

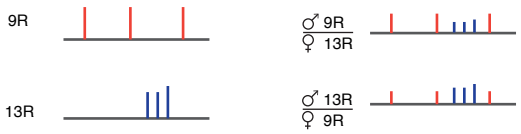


Figure 6.5: Hypothetical model of differential distributions of DSB locations between sex-polarised crosses of 13R and 9R mice. Horizontal grey lines represents the common, shared PAR sequence between the two strains. On the left are the 13R- and 9R-derived DSB hotspot locations (note that the distributions appear distinct between 9R and 13R (cf. Figure 3, Brick et al., 2012). On the right are hypothetical DSB profiles resulting from the sex-polarised F1 crosses.

2. Alternatively the motif restriction could occur genome-wide, where motif sites are restricted by limited DNA accessibility arising through chromatin state or epigenetic modification. Genome-wide methylation profiles are largely conserved between humans and chimpanzees, suggesting that chromatin state is conserved over evolutionary timescales (Molaro et al., 2011). These constant, genome-wide chromatin state profiles could act as an agent to restrict the available PRDM9-bound motifs. If this question was to be systematically assessed,

Experiment 1 would need to be extended, incorporating systematic RNAi knockdown of chromatin remodelling factors in a tagged-PRDM9 expressing cell line. Or alternatively, this could be assessed by attempting to rescue fertility in sterile mice hybrid crosses (e.g. PWD $\times$B6 (Mihola et al., 2009)) through either genetic or chemical manipulation of chromatin state.

In addition to PRDM9, Genome-wide association studies have identified several loci influencing genome-wide recombination rate in human. These include the *RNF212* gene harbouring variants with antagonistic effects in males and females (Kong et al., 2008), and the 17q21.31 inversion (Kong et al., 2010). Given the diverse interdependencies between factors that have been delineated in yeast and *C. elegans*, PRDM9 is likely to cooperate with other factors. Hence I propose an experiment to identify potential binding partners.

Experiment 2: Pull down of PRDM9 binding partners. A high throughput approach to identify potential PRDM9 trans-factors would be a ChIP-Seq experiment in a tagged-PRDM9 expressing cell line using an antibody against H3K4Me3 sites - a protein complex immunoprecipitation experiment (Co-IP). After the H3K4Me3 Co-IP, the precipitate is affinity selected using an antibody against the PRDM9 tag. Then the protein composition of both the flow through and the eluate would be assayed using Mass spectrometry or Western blot. The relative comparison of PRDM9-eluate and the flow through protein populations would improve the signal to noise ratio.

A third experiment to investigate the evolution of PRDM9 is to sequence ancient human DNA and reconstruct ancestral *Prdm9* sequences. Given the recent construction of a high coverage Denisovan genome ($> 30\times$, Meyer et al., 2012), it is potentially feasible to infer the ancestral PRDM9 sequence from ancient genomes given a reference panel of present day human and chimpanzee PRDM9 alleles.

The recent progress in the discovery of PRDM9 underlines the importance of multifaceted experimental, computational, and statistical approaches in contempo-

rary molecular biology. It is my intention to work at the interface of the three.

References

G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, 30(1): 97–101, Jan 2002. doi: 10.1038/ng786.

S. Anderssen, E. Sjøttem, G. Svineng, and T. Johansen. Comparative analyses of LTRs of the ERV-H family of primate-specific retrovirus-like elements isolated from marmoset, African green monkey, and man. *Virology*, 234(1):14–30, Jul 1997. doi: 10.1006/viro.1997.8590.

L. Aravind and E. V. Koonin. The HORMA domain: a common structural denominator in mitotic checkpoints, chromosome synapsis and DNA repair. *Trends Biochem Sci*, 23(8):284–6, Aug 1998.

A. Auton, A. Fledel-Alon, S. Pfeifer, O. Venn, L. Ségurel, T. Street, E. M. Leffler, R. Bowden, I. Aneas, J. Broxholme, P. Humburg, Z. Iqbal, G. Lunter, J. Maller, R. D. Hernandez, C. Melton, A. Venkat, M. A. Nobrega, R. Bontrop, S. Myers, P. Donnelly, M. Przeworski, and G. McVean. A fine-scale chimpanzee genetic map from population sequencing. *Science*, 336(6078):193–8, Apr 2012. doi: 10.1126/science.1216872.

E. Axelsson, M. T. Webster, A. Ratnakumar, L. Consortium, C. P. Ponting, and K. Lindblad-Toh. Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome Research*, pages 1–59, Oct 2011. doi: 10.1101/gr.124123.111.

J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler. Recent segmental duplications in the human genome. *Science*, 297(5583):1003–7, Aug 2002. doi: 10.1126/science.1072047.

N. H. Barton. Genetic hitchhiking. *Philos Trans R Soc Lond, B, Biol Sci*, 355(1403): 1553–62, Nov 2000. doi: 10.1098/rstb.2000.0716.

F. Baudat and B. de Massy. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome Res*, 15(5): 565–77, Jan 2007. doi: 10.1007/s10577-007-1140-3.

F. Baudat and A. Nicolas. Clustering of meiotic double-strand breaks on yeast chromosome III. *Proceedings of the National Academy of Sciences of the United States of America*, 94(10):5213–8, May 1997.

F. Baudat, J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B. de Massy. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, 327(5967):836–40, Feb 2010. doi: 10.1126/science.1183439.

A. R. Bellvé, J. C. Cavicchia, C. F. Millette, D. A. O'Brien, Y. M. Bhatnagar, and M. Dym. Spermatogenic cells of the prepuberal mouse. Isolation and morphological characterization. *J Cell Biol*, 74(1):68–85, Jul 1977.

I. L. Berg, R. Neumann, K.-W. G. Lam, S. Sarbajna, L. Odenthal-Hesse, C. A. May, and A. J. Jeffreys. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat Genet*, Sep 2010. doi: 10.1038/ng.658.

I. L. Berg, R. Neumann, S. Sarbajna, L. Odenthal-Hesse, N. J. Butler, and A. J. Jeffreys. Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(30):12378–83, Jul 2011. doi: 10.1073/pnas.1109531108.

A. Bergerat, B. de Massy, D. Gadelle, P. C. Varoutas, A. Nicolas, and P. Forterre. An atypical topoisomerase II from Archaea with implications for meiotic recombination. *Nature*, 386(6623):414–7, Mar 1997. doi: 10.1038/386414a0.

A. Bird, M. Taggart, M. Frommer, O. J. Miller, and D. Macleod. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, 40(1):91–9, Jan 1985.

N. P. Blackledge, H. K. Long, J. C. Zhou, S. Kriaucionis, R. Patient, and R. J. Klose. Bio-CAP: a versatile and highly sensitive technique to purify and characterise regions of non-methylated DNA. *Nucleic Acids Res*, 40(4):e32, Feb 2012. doi: 10.1093/nar/gkr1207.

Y. Blat, R. U. Protacio, N. Hunter, and N. Kleckner. Physical and functional interactions among basic chromosome organizational features govern early steps of meiotic chiasma formation. *Cell*, 111(6):791–802, Dec 2002.

V. Borde, N. Robine, W. Lin, S. Bonfils, V. Géli, and A. Nicolas. Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. *The EMBO Journal*, 28(2):99–111, Jan 2009. doi: 10.1038/emboj.2008.257.

R. H. Borts and J. E. Haber. Length and distribution of meiotic gene conversion tracts and crossovers in Saccharomyces cerevisiae. *Genetics*, 123(1):69–80, Sep 1989.

A. Boulton, R. S. Myers, and R. J. Redfield. The hotspot conversion paradox and the evolution of meiotic recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 94(15):8058–63, Jul 1997.

D. Bourc'his and T. H. Bestor. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*, 431(7004):96–9, Sep 2004. doi: 10.1038/nature02886.

Y. Brandvain and G. Coop. Scrambling eggs: meiotic drive and the evolution of female recombination rates. *Genetics*, 190(2):709–23, Feb 2012. doi: 10.1534/genetics.111.136721.

K. Brick, F. Smagulova, P. Khil, R. D. Camerini-Otero, and G. V. Petukhova. Genetic recombination is directed away from functional genomic elements in mice. *Nature*, 485(7400):642–5, May 2012. doi: 10.1038/nature11089.

K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, and J. L. Weber. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet*, 63(3):861–9, Sep 1998. doi: 10.1086/302011.

B. L. Browning and Z. Yu. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet*, 85(6):847–61, Dec 2009. doi: 10.1016/j.ajhg.2009.11.004.

S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*, 81(5):1084–97, Nov 2007. doi: 10.1086/521987.

J. Buard, P. Barthès, C. Grey, and B. de Massy. Distinct histone modifications define initiation and repair of meiotic recombination in the mouse. *The EMBO Journal*, 28(17):2616–24, Sep 2009. doi: 10.1038/emboj.2009.207.

F. Calafell, E. L. Grigorenko, A. A. Chikanian, and K. K. Kidd. Haplotype evolution and linkage disequilibrium: A simulation study. *Hum Hered*, 51(1-2):85–96, Jan 2001. doi: 22963.

J. L. Caswell, S. Mallick, D. J. Richter, J. Neubauer, C. Schirmer, S. Gnerre, and D. Reich. Analysis of chimpanzee history based on genome sequence alignments. *PLoS Genetics*, 4(4):e1000057, Apr 2008. doi: 10.1371/journal.pgen.1000057.

A. Chakravarti, K. H. Buetow, S. E. Antonarakis, P. G. Waber, C. D. Boehm, and H. H. Kazazian. Nonuniform recombination within the human beta-globin gene cluster. *Am J Hum Genet*, 36(6):1239–58, Nov 1984.

B. Charlesworth, M. Nordborg, and D. Charlesworth. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet Res*, 70(2):155–74, Oct 1997.

Y. Choo and A. Klug. Physical basis of a protein-DNA recognition code. *Curr Opin Struct Biol*, 7(1):117–25, Feb 1997.

D. M. Church, L. Goodstadt, L. W. Hillier, M. C. Zody, S. Goldstein, X. She, C. J. Bult, R. Agarwala, J. L. Cherry, M. DiCuccio, W. Hlavina, Y. Kapustin, P. Meric, D. Maglott, Z. Birtle, A. C. Marques, T. Graves, S. Zhou, B. Teague, K. Potamousis, C. Churas, M. Place, J. Herschleb, R. Runnheim, D. Forrest, J. Amos-Landgraf, D. C. Schwartz, Z. Cheng, K. Lindblad-Toh, E. E. Eichler, C. P. Ponting, and M. G. S. Consortium. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biology*, 7(5):e1000112, May 2009. doi: 10.1371/journal.pbio.1000112.

A. G. Clark, X. Wang, and T. Matise. Contrasting methods of quantifying fine structure of human recombination. *Annu Rev Genomics Hum Genet*, 11:45–64, Sep 2010. doi: 10.1146/annurev-genom-082908-150031.

V. Colot and J. L. Rossignol. Eukaryotic DNA methylation as an evolutionary device. *Bioessays*, 21(5):402–11, May 1999. doi: 10.1002/(SICI)1521-1878(199905)21:5⟨402::AID-BIES7⟩3.0.CO;2-B.

. G. P. Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, Oct 2010. doi: 10.1038/nature09534.

. G. P. Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, Nov 2012. doi: 10.1038/nature11632.

I. H. Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–61, Oct 2007. doi: 10.1038/nature06258.

I. H. G. S. Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, Oct 2004. doi: 10.1038/nature03001.

G. Coop and S. R. Myers. Live hot, die young: transmission distortion in recombination hotspots. *PLoS Genetics*, 3(3):e35, Mar 2007. doi: 10.1371/journal.pgen.0030035.

G. Coop, X. Wen, C. Ober, J. K. Pritchard, and M. Przeworski. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science*, 319(5868):1395–8, Mar 2008. doi: 10.1126/science.1151851.

H. B. Creighton and B. McClintock. A Correlation of Cytological and Genetical Crossing-Over in Zea Mays. *Proceedings of the National Academy of Sciences of the United States of America*, 17(8):492–7, Aug 1931.

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and . G. P. A. Group. The variant call format and VCFtools. *Bioinformatics*, 27(15): 2156–8, Aug 2011. doi: 10.1093/bioinformatics/btr330.

J. F. Degner, J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, and J. K. Pritchard. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–12, Dec 2009. doi: 10.1093/bioinformatics/btp579.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977.

R. C. Elston and J. Stewart. A general model for the genetic analysis of pedigree data. *Hum Hered*, 21(6):523–42, Jan 1971.

I. Eshel and M. W. Feldman. On the evolutionary effect of recombination. *Theor Popul Biol*, 1(1):88–100, May 1970.

B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8(3):186–94, Mar 1998.

L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 12(5):921–7, Sep 1995.

P. Fearnhead. Consistency of estimators of the population-scaled recombination rate. *Theor Popul Biol*, 64(1):67–79, Aug 2003.

P. Fearnhead and P. Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159(3):1299–318, Nov 2001.

P. Fearnhead and P. Donnelly. Approximate likelihood methods for estimating local recombination rates, Jan 2002.

M. W. Feldman, S. P. Otto, and F. B. Christiansen. Population genetic perspectives on the evolution of recombination. *Annu Rev Genet*, 30:261–95, Jan 1996. doi: 10.1146/annurev.genet.30.1.261.

L. Froenicke, L. K. Anderson, J. Wienberg, and T. Ashley. Male mouse recombination maps for each autosome identified by chromosome painting. *Am J Hum Genet*, 71 (6):1353–68, Dec 2002. doi: 10.1086/344714.

P. A. Fujita, B. Rhead, A. S. Zweig, A. S. Hinrichs, D. Karolchik, M. S. Cline, M. Goldman, G. P. Barber, H. Clawson, A. Coelho, M. Diekhans, T. R. Dreszer, B. M. Giardine, R. A. Harte, J. Hillman-Jackson, F. Hsu, V. Kirkup, R. M. Kuhn, K. Learned, C. H. Li, L. R. Meyer, A. Pohl, B. J. Raney, K. R. Rosenbloom, K. E. Smith, D. Haussler, and W. J. Kent. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*, 39(Database issue):D876–82, Jan 2011. doi: 10.1093/nar/gkq963.

M. Gardiner-Garden and M. Frommer. CpG islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2):261–82, Jul 1987.

J. Gay, S. Myers, and G. McVean. Estimating meiotic gene conversion rates from population genetic data. *Genetics*, 177(2):881, 2007.

J. L. Gerton, J. DeRisi, R. Shroff, M. Lichten, P. O. Brown, and T. D. Petes. Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11383–90, Oct 2000. doi: 10.1073/pnas.97.21.11383.

G. B. Golding. The sampling distribution of linkage disequilibrium. *Genetics*, 108 (1):257–74, Sep 1984.

N. L. Goodchild, D. A. Wilkinson, and D. L. Mager. Recent evolutionary expansion of a subfamily of RTVL-H human endogenous retrovirus-like elements. *Virology*, 196(2):778–88, Oct 1993. doi: 10.1006/viro.1993.1535.

C. Grey, F. Baudat, and B. de Massy. Genome-wide control of the distribution of meiotic recombination. *PLoS Biology*, 7(2):e35, Feb 2009. doi: 10.1371/journal.pbio.1000035.

C. Grey, P. Barthès, G. C.-L. Friec, F. Langa, F. Baudat, and B. de Massy. Mouse PRDM9 DNA-binding specificity determines sites of histone H3 lysine 4 trimethylation for initiation of meiotic recombination. *PLoS Biology*, 9(10):e1001176, Oct 2011. doi: 10.1371/journal.pbio.1001176.

R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *J Comput Biol*, 3(4):479–502, Jan 1996.

D. F. Gudbjartsson, K. Jonasson, M. L. Frigge, and A. Kong. Allegro, a new computer program for multipoint linkage analysis. *Nat Genet*, 25(1):12–3, May 2000. doi: 10.1038/75514.

H. Guillon and B. de Massy. An initiation site for meiotic crossing-over and gene conversion in the mouse. *Nat Genet*, 32(2):296–9, Oct 2002. doi: 10.1038/ng990.

D. W. Hale. Is X-Y recombination necessary for spermatocyte survival during mammalian spermatogenesis? *Cytogenet Cell Genet*, 65(4):278–82, Jan 1994.

T. J. Hassold, S. L. Sherman, D. Pettay, D. C. Page, and P. A. Jacobs. XY chromosome nondisjunction in man is associated with diminished recombination in the pseudoautosomal region. *Am J Hum Genet*, 49(2):253–60, Aug 1991.

K. Hayashi, K. Yoshida, and Y. Matsui. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. *Nature*, 438(7066):374–8, Nov 2005. doi: 10.1038/nature04112.

G. Hellenthal and M. Stephens. msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics*, 23(4):520–1, Feb 2007. doi: 10.1093/bioinformatics/btl622.

W. Hill and A. Robertson. Linkage disequilibrium in finite populations. *TAG Theoretical and Applied Genetics*, Jan 1968a.

W. G. Hill. Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, 33(2):229–39, Oct 1974.

W. G. Hill and A. Robertson. The effects of inbreeding at loci with heterozygote advantage. *Genetics*, 60(3):615–28, Nov 1968b.

A. J. Hilliker, G. Harauz, A. G. Reaume, M. Gray, S. H. Clark, and A. Chovnick. Meiotic gene conversion tract length distribution within the rosy locus of Drosophila melanogaster. *Genetics*, 137(4):1019–26, Aug 1994.

A. G. Hinch, A. Tandon, N. Patterson, Y. Song, N. Rohland, C. D. Palmer, G. K. Chen, K. Wang, S. G. Buxbaum, E. L. Akylbekova, M. C. Aldrich, C. B. Ambrosone, C. Amos, E. V. Bandera, S. I. Berndt, L. Bernstein, W. J. Blot, C. H. Bock, E. Boerwinkle, Q. Cai, N. Caporaso, G. Casey, L. A. Cupples, S. L. Deming, W. R. Diver, J. Divers, M. Fornage, E. M. Gillanders, J. Glessner, C. C. Harris, J. J. Hu, S. A. Ingles, W. Isaacs, E. M. John, W. H. L. Kao, B. Keating, R. A. Kittles, L. N. Kolonel, E. Larkin, L. L. Marchand, L. H. McNeill, R. C. Millikan, A. Murphy, S. Musani, C. Neslund-Dudas, S. Nyante, G. J. Papanicolaou, M. F. Press, B. M. Psaty, A. P. Reiner, S. S. Rich, J. L. Rodriguez-Gil, J. I. Rotter, B. A. Rybicki, A. G. Schwartz, L. B. Signorello, M. Spitz, S. S. Strom, M. J. Thun, M. A. Tucker, Z. Wang, J. K. Wiencke, J. S. Witte, M. Wrensch, X. Wu, Y. Yamamura, K. A. Zanetti, W. Zheng, R. G. Ziegler, X. Zhu, S. Redline, J. N. Hirschhorn, B. E. Henderson, H. A. Taylor, A. L. Price, H. Hakonarson, S. J. Chanock, C. A. Haiman, J. G. Wilson, D. Reich, and S. R. Myers. The landscape of recombination in African Americans. *Nature*, 476(7359):170–5, Aug 2011. doi: 10.1038/nature10336.

B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529, Jun 2009. doi: 10.1371/journal.pgen.1000529.

X. Huang and A. Madan. CAP3: A DNA sequence assembly program. *Genome Research*, 9(9):868–77, Sep 1999.

X. Huang, J. Wang, S. Aluru, S.-P. Yang, and L. Hillier. PCAP: a whole-genome assembly program. *Genome Research*, 13(9):2164–70, Sep 2003. doi: 10.1101/gr.1390403.

T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater,

J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. The Ensembl genome database project. *Nucleic Acids Res*, 30(1):38–41, Jan 2002.

R. Hubert, V. P. Stanton, H. Aburatani, J. Warren, H. Li, D. E. Housman, and N. Arnheim. Sperm typing allows accurate measurement of the recombination fraction between D3S2 and D3S3 on the short arm of human chromosome 3. *Genomics*, 12(4):683–7, Apr 1992.

R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*, 23(2):183–201, Apr 1983.

R. R. Hudson. Two-locus sampling distributions and their application. *Genetics*, 159 (4):1805–17, Dec 2001.

R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1):147–64, Sep 1985.

Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, and G. McVean. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet*, 44(2):226–32, Feb 2012. doi: 10.1038/ng.1028.

A. C. Jamieson, J. C. Miller, and C. O. Pabo. Drug discovery with engineered zinc-finger proteins. *Nat Rev Drug Discov*, 2(5):361–8, May 2003. doi: 10.1038/nrd1087.

A. J. Jeffreys and C. A. May. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet*, 36(2):151–6, Feb 2004. doi: 10.1038/ng1287.

A. J. Jeffreys and R. Neumann. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet*, 31(3):267–71, Jul 2002. doi: 10.1038/ng910.

A. J. Jeffreys and R. Neumann. The rise and fall of a human recombination hot spot. *Nat Genet*, 41(5):625–9, May 2009. doi: 10.1038/ng.346.

A. J. Jeffreys, A. MacLeod, K. Tamaki, D. L. Neil, and D. G. Monckton. Minisatellite repeat coding as a digital approach to DNA typing. *Nature*, 354(6350):204–9, Nov 1991. doi: 10.1038/354204a0.

A. J. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet*, 29(2): 217–22, Oct 2001. doi: 10.1038/ng1001-217.

A. J. Jeffreys, R. Neumann, M. Panayi, S. Myers, and P. Donnelly. Human recombination hot spots hidden in regions of strong marker association. *Nat Genet*, 37(6): 601–6, Jun 2005. doi: 10.1038/ng1565.

E. Jorgenson, H. Tang, M. Gadde, M. Province, M. Leppert, S. Kardia, N. Schork, R. Cooper, D. C. Rao, E. Boerwinkle, and N. Risch. Ethnicity and human genetic linkage maps. *Am J Hum Genet*, 76(2):276–90, Feb 2005. doi: 10.1086/427926.

L. Kauppi, M. Barchi, F. Baudat, P. J. Romanienko, S. Keeney, and M. Jasin. Distinct properties of the XY pseudoautosomal region crucial for male meiosis. *Science*, 331 (6019):916–20, Feb 2011. doi: 10.1126/science.1195774.

S. Keeney, C. N. Giroux, and N. Kleckner. Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell*, 88 (3):375–84, Feb 1997.

P. D. Keightley and S. P. Otto. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature*, 443(7107):89–92, Sep 2006. doi: 10.1038/nature05049.

W. J. Kent. BLAT–the BLAST-like alignment tool. *Genome Research*, 12(4):656–64, Apr 2002. doi: 10.1101/gr.229202.ArticlepublishedonlinebeforeMarch2002.

W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, Jun 2002. doi: 10.1101/gr.229102. ArticlepublishedonlinebeforeprintinMay2002.

S. Y. Kim, Y. Li, Y. Guo, R. Li, J. Holmkvist, T. Hansen, O. Pedersen, J. Wang, and R. Nielsen. Design of association studies with pooled or un-pooled next-generation sequencing data. *Genet Epidemiol*, 34(5):479–91, Jul 2010. doi: 10.1002/gepi. 20501.

N. Kleckner, A. Storlazzi, and D. Zickler. Coordinate variation in meiotic pachytene SC length and total crossover/chiasma frequency under conditions of constant DNA length. *Trends Genet*, 19(11):623–8, Nov 2003. doi: 10.1016/j.tig.2003.09.004.

A. Klug. The discovery of zinc fingers and their development for practical applications in gene regulation and genome manipulation. *Q Rev Biophys*, 43(1):1–21, Feb 2010. doi: 10.1017/S0033583510000089.

A. Kong, D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S. T. Palsson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, and K. Stefansson. A high-resolution recombination map of the human genome. *Nat Genet*, 31(3): 241–7, Jul 2002. doi: 10.1038/ng917.

A. Kong, G. Thorleifsson, H. Stefansson, G. Masson, A. Helgason, D. F. Gudbjartsson, G. M. Jonsdottir, S. A. Gudjonsson, S. Sverrisson, T. Thorlacius, A. Jonasdottir, G. A. Hardarson, S. T. Palsson, M. L. Frigge, J. R. Gulcher, U. Thorsteinsdottir, and K. Stefansson. Sequence variants in the RNF212 gene associate with genome-wide recombination rate. *Science*, 319(5868):1398–401, Mar 2008. doi: 10.1126/science.1152422.

A. Kong, G. Thorleifsson, D. F. Gudbjartsson, G. Masson, A. Sigurdsson, A. Jonasdottir, G. B. Walters, A. Jonasdottir, A. Gylfason, K. T. Kristinsson, S. A. Gudjonsson, M. L. Frigge, A. Helgason, U. Thorsteinsdottir, and K. Stefansson. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–103, Oct 2010. doi: 10.1038/nature09525.

L. Kruglyak, M. J. Daly, M. P. Reeve-Daly, and E. S. Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet*, 58(6):1347–63, Jun 1996.

M. K. Kuhner, P. Beerli, J. Yamato, and J. Felsenstein. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics*, 156(1):439–47, Sep 2000.

E. S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 84(8):2363–7, Apr 1987.

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la

Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and I. H. G. S. Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001. doi: 10.1038/35057062.

R. Lewontin. The interaction of selection and linkage. II. Optimum models. *Genetics*, 50:757–82, Oct 1964.

R. Lewontin. On measures of gametic disequilibrium. *Genetics*, 120(3):849–52, Nov 1988.

H. Li. Improving SNP discovery by base alignment quality. *Bioinformatics*, 27(8): 1157–8, Apr 2011. doi: 10.1093/bioinformatics/btr076.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–60, Jul 2009. doi: 10.1093/bioinformatics/btp324.

H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–8, Nov 2008. doi: 10.1101/gr.078212.108.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–9, Aug 2009. doi: 10.1093/bioinformatics/btp352.

N. Li and M. Stephens. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, pages 1–22, Dec 2003.

Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*, 34(8):816–34, Dec 2010. doi: 10.1002/gepi.20533.

M. Lichten and B. de Massy. The impressionistic landscape of meiotic recombination. *Cell*, 147(2):267–70, Oct 2011. doi: 10.1016/j.cell.2011.09.038.

M. Lichten and A. S. Goldman. Meiotic recombination hotspots. *Annu Rev Genet*, 29:423–44, Jan 1995. doi: 10.1146/annurev.ge.29.120195.002231.

G. Liti, D. M. Carter, A. M. Moses, J. Warringer, L. Parts, S. A. James, R. P. Davey, I. N. Roberts, A. Burt, V. Koufopanou, I. J. Tsai, C. M. Bergman, D. Bensasson, M. J. T. O'Kelly, A. van Oudenaarden, D. B. H. Barton, E. Bailes, A. N. Nguyen, M. Jones, M. A. Quail, I. Goodhead, S. Sims, F. Smith, A. Blomberg, R. Durbin, and E. J. Louis. Population genomics of domestic and wild yeasts. *Nature*, 458 (7236):337–41, Mar 2009. doi: 10.1038/nature07743.

G. Lunter and M. Goodson. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, Oct 2010. doi: 10.1101/gr.111120.110.

J. R. Lupski. Genomic rearrangements and sporadic disease. *Nat Genet*, 39(7 Suppl): S43–7, Jul 2007. doi: 10.1038/ng2084.

P. Machanick and T. L. Bailey. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, 27(12):1696–7, Jun 2011. doi: 10.1093/bioinformatics/btr189.

D. L. Mager and N. L. Goodchild. Homologous recombination between the LTRs of a human retrovirus-like element causes a 5-kb deletion in two siblings. *Am J Hum Genet*, 45(6):848–54, Dec 1989.

G. Marais and B. Charlesworth. Genome evolution: recombination speeds up adaptive evolution. *Current Biology*, 13(2):R68–70, Jan 2003.

E. R. Martin, D. D. Kinnamon, M. A. Schmidt, E. H. Powell, S. Zuchner, and R. W. Morris. SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics*, 26(22):2803–10, Nov 2010. doi: 10.1093/bioinformatics/btq526.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–303, Sep 2010. doi: 10.1101/gr.107524.110.

G. McVean. What drives recombination hotspots to repeat DNA in humans? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544):1213–8, Apr 2010. doi: 10.1098/rstb.2009.0299.

G. McVean, P. Awadalla, and P. Fearnhead. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160(3):1231–41, Mar 2002.

G. McVean, S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–4, Apr 2004. doi: 10.1126/science.1092500.

S. Meader, L. W. Hillier, D. Locke, C. P. Ponting, and G. Lunter. Genome assembly quality: assessment and improvement using the neutral indel model. *Genome Research*, 20(5):675–84, May 2010. doi: 10.1101/gr.096966.109.

M. Meistrich. Separation of spermatogenic cells and nuclei from rodent testes. *Methods in cell biology*, Jan 1977.

M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, and S. Pääbo. A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338(6104):222–6, Oct 2012. doi: 10.1126/science.1224344.

O. Mihola, Z. Trachtulec, C. Vlcek, J. C. Schimenti, and J. Forejt. A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science*, 323(5912):373–5, Jan 2009. doi: 10.1126/science.1163601.

J. Miller, A. D. McLachlan, and A. Klug. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *The EMBO Journal*, 4(6): 1609–14, Jun 1985.

E. P. Mimitou and L. S. Symington. Sae2, Exo1 and Sgs1 collaborate in DNA double-strand break processing. *Nature*, 455(7214):770–4, Oct 2008. doi: 10.1038/nature07312.

A. Molaro, E. Hodges, F. Fang, Q. Song, W. R. Mccombie, G. J. Hannon, and A. D. Smith. Sperm Methylation Profiles Reveal Features of Epigenetic Inheritance and Evolution in Primates. *Cell*, 146(6):1029–1041, Sep 2011. doi: 10.1016/j.cell.2011.08.016.

T. H. Morgan. Random segregation versus coupling in Mendelian inheritance. *Science*, 34(873):384, Sep 1911. doi: 10.1126/science.34.873.384.

S. Myers. The Detection of Recombination Events Using DNA Sequence Data. *University of Oxford*, Sep 2003.

S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746): 321–4, Oct 2005. doi: 10.1126/science.1117196.

S. Myers, C. Freeman, A. Auton, P. Donnelly, and G. McVean. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet*, 40(9):1124–9, Sep 2008. doi: 10.1038/ng.213.

S. Myers, R. Bowden, A. Tumian, R. E. Bontrop, C. Freeman, T. S. MacFie, G. McVean, and P. Donnelly. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, 327(5967):876–9, Feb 2010. doi: 10.1126/science.1182363.

S. R. Myers and R. C. Griffiths. Bounds on the minimum number of recombination events in a sample history. *Genetics*, 163(1):375–94, Jan 2003.

Y. Nakamura, M. Leppert, P. O'Connell, R. Wolff, T. Holm, M. Culver, C. Martin, E. Fujimoto, M. Hoff, and E. Kumlin. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science*, 235(4796):1616–22, Mar 1987.

R. Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154(2):931–42, Feb 2000.

R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–51, Jun 2011. doi: 10.1038/nrg2986.

M. Nordborg and S. Tavaré. Linkage disequilibrium: what history has to tell us. *Trends Genet*, 18(2):83–90, Feb 2002.

P. L. Oliver, L. Goodstadt, J. J. Bayes, Z. Birtle, K. C. Roach, N. Phadnis, S. A. Beatson, G. Lunter, H. S. Malik, and C. P. Ponting. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genetics*, 5(12): e1000753, Dec 2009. doi: 10.1371/journal.pgen.1000753.

S. P. Otto and T. Lenormand. Resolving the paradox of sex and recombination. *Nature Reviews Genetics*, 3(4):252–61, Apr 2002. doi: 10.1038/nrg761.

B. Padhukasahasram, J. D. Wall, P. Marjoram, and M. Nordborg. Estimating Recombination Rates From Single-Nucleotide Polymorphisms Using Summary Statistics. *Genetics*, 174(3):1517–1528, Sep 2006. doi: 10.1534/genetics.106.060723.

K. Paigen and P. Petkov. Mammalian recombination hot spots: properties, control and evolution. *Nature Reviews Genetics*, 11(3):221–33, Mar 2010. doi: 10.1038/nrg2712.

K. Paigen, J. P. Szatkiewicz, K. Sawyer, N. Leahy, E. D. Parvanov, S. H. S. Ng, J. H. Graber, K. W. Broman, and P. M. Petkov. The recombinational anatomy of a mouse chromosome. *PLoS Genetics*, 4(7):e1000119, Jul 2008. doi: 10.1371/journal.pgen.1000119.

K. Palin, H. Campbell, A. F. Wright, J. F. Wilson, and R. Durbin. Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genet Epidemiol*, 35(8):853–60, Dec 2011. doi: 10.1002/gepi.20635.

J. Pan, M. Sasaki, R. Kniewel, H. Murakami, H. G. Blitzblau, S. E. Tischfield, X. Zhu, M. J. Neale, M. Jasin, N. D. Socci, A. Hochwagen, and S. Keeney. A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell*, 144(5):719–31, Mar 2011. doi: 10.1016/j.cell.2011.02.009.

S. Panizza, M. A. Mendoza, M. Berlinger, L. Huang, A. Nicolas, K. Shirahige, and F. Klein. Spo11-accessory proteins link double-strand break sites to the chromosome axis in early meiotic recombination. *Cell*, 146(3):372–83, Aug 2011. doi: 10.1016/j.cell.2011.07.003.

P. J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–80, Oct 2009. doi: 10.1038/nrg2641.

E. D. Parvanov, P. M. Petkov, and K. Paigen. Prdm9 controls activation of mammalian recombination hotspots. *Science*, 327(5967):835, Feb 2010. doi: 10.1126/science.1181495.

A. V. Persikov, R. Osada, and M. Singh. Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics*, 25(1):22–9, Jan 2009. doi: 10.1093/bioinformatics/btn580.

T. D. Petes. Meiotic recombination hot spots and cold spots. *Nature Reviews Genetics*, 2(5):360–9, May 2001. doi: 10.1038/35072078.

A. L. Price, A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, 5(6):e1000519, Jun 2009. doi: 10.1371/journal.pgen.1000519.

J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, 69(1):1–14, Jul 2001. doi: 10.1086/321275.

S. Ptak, A. Roeder, M. Stephens, Y. Gilad, S. Paabo, and M. Przeworski. Absence of the TAP2 human recombination hotspot in chimpanzees. *PLoS Biology*, 2:849–855, 2004.

S. E. Ptak, D. A. Hinds, K. Koehler, B. Nickel, N. Patil, D. G. Ballinger, M. Przeworski, K. A. Frazer, and S. Pääbo. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet*, 37(4):429–34, Apr 2005. doi: 10.1038/ng1529.

V. R. Ramirez-Carrozzi, D. Braas, D. M. Bhatt, C. S. Cheng, C. Hong, K. R. Doty, J. C. Black, A. Hoffmann, M. Carey, and S. T. Smale. A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. *Cell*, 138(1):114–28, Jul 2009. doi: 10.1016/j.cell.2009.04.020.

G. A. Rappold. The pseudoautosomal regions of the human sex chromosomes. *Hum Genet*, 92(4):315–24, Oct 1993.

G. A. Rappold, A. Henke, M. Pohlschmid, and T. H. Huisman. Pseudoautosomal repeat displays higher variability in blacks than in Caucasians. *Hum Genet*, 88(5): 573–8, Mar 1992.

E. Revenkova, M. Eijpe, C. Heyting, C. A. Hodges, P. A. Hunt, B. Liebe, H. Scherthan, and R. Jessberger. Cohesin SMC1 beta is required for meiotic chromosome dynamics, sister chromatid cohesion and DNA recombination. *Nat Cell Biol*, 6(6):555–62, Jun 2004. doi: 10.1038/ncb1135.

J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nat Biotechnol*, 29(1):24–6, Jan 2011. doi: 10.1038/nbt.1754.

S. L. Salzberg and J. A. Yorke. Beware of mis-assembled genomes. *Bioinformatics*, 21(24):4320–1, Dec 2005. doi: 10.1093/bioinformatics/bti769.

C. Sandor, W. Li, W. Coppieters, T. Druet, C. Charlier, and M. Georges. Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. *PLoS Genetics*, 8(7):e1002854, Jul 2012. doi: 10.1371/journal.pgen.1002854.

J. P. Sanford, H. J. Clark, V. M. Chapman, and J. Rossant. Differences in DNA methylation during oogenesis and spermatogenesis and their persistence during early embryogenesis in the mouse. *Genes Dev*, 1(10):1039–46, Dec 1987.

M. Sasaki, J. Lange, and S. Keeney. Genome destabilization by homologous recombination in the germ line. *Nat Rev Mol Cell Biol*, 11(3):182–95, Mar 2010. doi: 10.1038/nrm2849.

S. Saxonov, P. Berg, and D. L. Brutlag. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 103(5):1412–7, Jan 2006. doi: 10.1073/pnas.0510310103.

T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–100, Oct 1990.

S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller. Human-mouse alignments with BLASTZ. *Genome Research*, 13(1): 103–7, Jan 2003. doi: 10.1101/gr.809403.

L. Ségurel, E. M. Leffler, and M. Przeworski. The case of the fickle fingers: how the PRDM9 zinc finger protein specifies meiotic recombination hotspots in humans. *PLoS Biology*, 9(12):e1001211, Dec 2011. doi: 10.1371/journal.pbio.1001211.

C. Sequencing and A. Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, Sep 2005. doi: 10.1038/nature04072.

F. Smagulova, I. V. Gregoretti, K. Brick, P. Khil, R. D. Camerini-Otero, and G. V. Petukhova. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature*, 472(7343):375–8, Apr 2011. doi: 10.1038/nature09869.

A. F. Smit. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev*, 9(6):657–63, Dec 1999.

C. Spencer, P. Deloukas, H. Sarah, J. Mullikin, and S. Myers. The influence of recombination on human genetic diversity. *PLoS Genet*, 2(9):e148, 2006.

H. Stefansson, A. Helgason, G. Thorleifsson, V. Steinthorsdottir, G. Masson, J. Barnard, A. Baker, A. Jonasdottir, A. Ingason, V. G. Gudnadottir, N. Desnica, A. Hicks, A. Gylfason, D. F. Gudbjartsson, G. M. Jonsdottir, J. Sainz, K. Agnarsson, B. Birgisdottir, S. Ghosh, A. Olafsdottir, J.-B. Cazier, K. Kristjansson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, A. Kong, and K. Stefansson. A common inversion under selection in Europeans. *Nat Genet*, 37(2):129–37, Feb 2005. doi: 10.1038/ng1508.

K. M. Steinberg, F. Antonacci, P. H. Sudmant, J. M. Kidd, C. D. Campbell, L. Vives, M. Malig, L. Scheinfeldt, W. Beggs, M. Ibrahim, G. Lema, T. B. Nyambo, S. A. Omar, J.-M. Bodo, A. Froment, M. P. Donnelly, K. K. Kidd, S. A. Tishkoff, and E. E. Eichler. Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat Genet*, 44(8):872–80, Jan 2012. doi: 10.1038/ng.2335.

M. Steinmetz, K. Minard, S. Horvath, J. McNicholas, J. Srelinger, C. Wake, E. Long, B. Mach, and L. Hood. A molecular map of the immune response region from the major histocompatibility complex of the mouse. *Nature*, 300(5887):35–42, Nov 1982.

M. Stephens and P. Donnelly. Inference in molecular population genetics. *... of the Royal Statistical Society: Series ...*, Jan 2002.

M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68(4):978–89, Apr 2001. doi: 10.1086/319501.

M. P. H. Stumpf and D. B. Goldstein. Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. *Current Biology*, 13(1):1–8, Jan 2003.

M. P. H. Stumpf and G. McVean. Estimating recombination rates from population-genetic data. *Nature Reviews Genetics*, 4(12):959–68, Dec 2003. doi: 10.1038/nrg1227.

J. W. Szostak, T. L. Orr-Weaver, R. J. Rothstein, and F. W. Stahl. The double-strand-break repair model for recombination. *Cell*, 33(1):25–35, May 1983.

E. Thompson. Statistical Inference from Genetic Data on Pedigrees. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 6:i–xiv+1–169, Jan 2000.

I. J. Tsai, A. Burt, and V. Koufopanou. Conservation of recombination hotspots in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 107(17):7847–52, Apr 2010. doi: 10.1073/pnas.0908774107.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory*, Jan 1967.

J. D. Wall, L. A. Frisse, R. R. Hudson, and A. D. Rienzo. Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. *Am J Hum Genet*, 73(6):1330–40, Dec 2003. doi: 10.1086/380311.

J. Wang, H. C. Fan, B. Behr, and S. R. Quake. Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. *Cell*, 150 (2):402–12, Jul 2012. doi: 10.1016/j.cell.2012.06.030.

H. Watanabe, A. Fujiyama, M. Hattori, T. D. Taylor, A. Toyoda, Y. Kuroki, H. Noguchi, A. BenKahla, H. Lehrach, R. Sudbrak, M. Kube, S. Taenzer, P. Galgoczy, M. Platzer, M. Scharfe, G. Nordsiek, H. Blöcker, I. Hellmann, P. Khaitovich, S. Pääbo, R. Reinhardt, H.-J. Zheng, X.-L. Zhang, G.-F. Zhu, B.-F. Wang, G. Fu, S.-X. Ren, G.-P. Zhao, Z. Chen, Y.-S. Lee, J.-E. Cheong, S.-H. Choi, K.-M. Wu, T.-T. Liu, K.-J. Hsiao, S.-F. Tsai, C.-G. Kim, S. OOta, T. Kitano, Y. Kohara, N. Saitou, H.-S. Park, S.-Y. Wang, M.-L. Yaspo, and Y. Sakaki. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature*, 429(6990):382–8, May 2004. doi: 10.1038/nature02564.

B. Weir. *Genetic data analysis II: methods for discrete population genetic data*. Apr 1997.

L. R. Weitkamp. Proceedings: Population differences in meiotic recombination frequency between loci on chromosome 1. *Cytogenet Cell Genet*, 13(1):179–82, Jan 1974.

W. Winckler, S. Myers, D. Richter, R. Onofrio, G. McDonald, R. Bontrop, G. McVean, S. Gabriel, D. Reich, and P. Donnelly. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, 308(5718):107, 2005.

C. Wiuf and J. Hein. Recombination as a point process along sequences. *Theor Popul Biol*, 55(3):248–59, Jun 1999. doi: 10.1006/tpbi.1998.1403.

C. Wiuf, T. Christensen, and J. Hein. A simulation study of the reliability of recombination detection methods. *Mol Biol Evol*, 18(10):1929–39, Oct 2001.

C. I. Wu, T. W. Lyttle, M. L. Wu, and G. F. Lin. Association between a satellite DNA sequence and the Responder of Segregation Distorter in D. melanogaster. *Cell*, 54(2):179–89, Jul 1988.

N. Yu, M. I. Jensen-Seaman, L. Chemnick, J. R. Kidd, A. S. Deinard, O. Ryder, K. K. Kidd, and W.-H. Li. Low nucleotide diversity in chimpanzees and bonobos. *Genetics*, 164(4):1511–8, Aug 2003.

P. A. Zweidler-Mckay, H. L. Grimes, M. M. Flubacher, and P. N. Tsichlis. Gfi-1 encodes a nuclear zinc finger protein that binds DNA and functions as a transcriptional repressor. *Mol Cell Biol*, 16(8):4024–34, Aug 1996.