Factors Influencing Genetic Variation in Wild Mice

Robert William Davies

Keble College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2015

Acknowledgements

I'd like to begin by thanking my supervisor Simon Myers. I am grateful for all I learned from you both statistically and scientifically, and hope to bring your passion, dedication, curiosity and rigour to my own scientific career. I'd like to thank collaborators and course directors Jonathan Flint, for always being there and always pushing me, and Richard Mott, for his insightful, thought provoking questions. Financially, I am grateful for the support provided to me by the Wellcome Trust.

I am thankful for having had such a wonderful set of friends and colleagues while at Oxford. I offer thanks to Hilary, Na, Nick and Winni for proof-reading various chapters of this thesis. Developmentally, I am grateful for the supportive environment of the Wellcome Trust Centre for Human Genetics, in particular Winni and Kiran, from whom I learnt much about programming and computers. Socially, I am thankful for having had such wonderful experiences with housemates, friends, and fellow Keble graduate students. Together, you enriched my time at Oxford immeasurably.

Personally, I would like to thank my family, in particular my parents. I offer thanks to my Dad, who got me into research at such as young age, and to my Mom, for always telling me what I needed to hear and always being able to offer the professors point of view.

Finally, I'd like to thank Roos. You've always been there for me throughout the DPhil, be it celebrating the exhilarating highs, or offering compassion during the crushing lows. I am eternally grateful.

Factors Influencing Genetic Variation in Wild Mice

Robert William Davies, Keble College, University of Oxford

Thesis submitted for the degree of Doctor of Philosophy

Michaelmas 2015

The mouse is a premier model organism for mammalian biological research. They have been instrumental in studies illuminating processes of fundemental importance to genetics including the control of recombination and the process of speciation. However, most of these studies are based on laboratory mice, which are an artificial population with a complex relationship to wild mice. Studies of mammalian biology would benefit from understanding how history has shaped the genome of the present day mouse.

In this thesis, I explore factors that contribute to patterns of genetic variation in wild mice, paying particular attention to the process of recombination. I use whole genome sequencing data from three wild populations: 20 French *M. m. domesticus*; 20 Taiwanese *M. m. castaneus*; and 10 Indian *M. m. castaneus*. In addition, I use 13 classical laboratory strains and 6 wild derived inbred strains.

These data show that the French and Taiwanese populations have been through recent, severe population bottlenecks, are recipients of considerable migration, and are partially inbred, with about 15-20% of the genome recently homozygous. All of these features are absent from the Indian mice. Signatures of selection reveal that the *Prl* gene in Indian mice was

the result of a full sweep originating from a highly diverged species of mouse. In terms of recombination, the linkage-disequilibrium based rate map of the French and Taiwanese mice is shown to be highly punctuate, while the Indian rate map is comparatively flat, and devoid of population level hotspots. The recombination landscape is shown to dominate nucleotide substitution, notably in the Indian mice, where it reverses a 2:1 strong (C/G) to weak (A/T) mutation bias in favour of a 2:1 weak to strong fixation rate. Dozens of long-extinct, and active, PRDM9 motifs are identified through a comparative genomic approach, which also reveals a highly localized picture of the distribution of GC-biased gene conversion. Separately, a low-coverage reference-panel free read-aware genotype imputation method named STITCH is presented, and shown to accurately impute genotypes in outbred laboratory mice and humans.

Contents

1 Introduction **1**

1.1 Thesis layout 2

1.2 Survey of knowledge on wild mouse genetics 4

1.3 Biological machinery of recombination 9

1.4 *Prdm9* diversity in wild mice 14

1.5 Statistical techniques used in this thesis 15

1.5.1 Expectation maximization 16

1.5.2 Hidden Markov models 19

2 Cataloguing variation in wild mice **24**

2.1 Methods 25

2.2 Results 35

2.2.1 Generation of high quality array sites 35

2.2.2 SNP genotyping and accuracy 36

2.3 Discussion 41

3 Factors that influence patterns of variation in wild mice **45**

3.1 Descriptive statistics 46

3.1.1 Estimating the callable genomes in the populations 46

3.1.2 Ancestral allele call 48

3.1.3 Comparison of variation in different populations 49

3.1.4 Discussion 55

3.2 Demography and population structure 56
3.2.1 Introduction 56
3.2.2 Methods 57
3.2.3 Results 62
3.2.4 Discussion 68
3.3 Inbreeding 71
3.3.1 Introduction 71
3.3.2 Methods 73
3.3.3 Results 81
3.3.4 Discussion 89
3.4 Selection 90
3.4.1 Introduction 90
3.4.2 Methods 92
3.4.3 Results 95
3.4.4 Top results 96
3.4.5 Discussion 113

4 The recent landscape of recombination in mice 118

4.1 Introduction 119
4.2 Methods 121
4.2.1 Rate map construction 121
4.2.2 Hotspot calling 122
4.2.3 Motif identification and refinement 122
4.3 Results 124
4.3.1 Construction and validation of maps 124
4.3.2 Hotspots 127
4.3.3 Effect of recombination on nucleotide substitution patterns . . 138
4.4 Discussion 142

4.4.1 Recombination rate variation and hotspots 142
4.4.2 Nucleotide substitution bias 146

5 The ancient landscape of recombination in mice 149

5.1 Introduction 150
5.2 Methods 152
5.2.1 Lineage definitions and SNP filtering 153
5.2.2 p-value calculations 155
5.2.3 *k*-mer clustering 156
5.2.4 Estimating parameters associated with gene conversion 159
5.3 Results 163
5.3.1 *de novo* motif identication 163
5.3.2 Estimating parameters associated with gene conversion 170
5.4 Discussion 176

6 Low coverage imputation in mice and humans 182

6.1 Introduction 182
6.2 Materials and Methods 184
6.2.1 Overview 184
6.2.2 Simulation under the model 185
6.2.3 Complete data probability under the model 187
6.2.4 Initialization 189
6.2.5 Maximization and parameter updating 189
6.2.6 Parameter bounding 191
6.2.7 Pseudo-haploid approximation 192
6.2.8 Useful variables 195
6.2.9 Parameter updating under haploid model 196
6.2.10 Parameter updating under pseudo-haploid model 199

6.2.11 Parameter updating under diploid model 201
6.2.12 Efficient calculation of forward backward variables 204
6.2.13 CFW sequencing 206
6.2.14 CONVERGE study 208
6.2.15 Beagle 208
6.3 Results 208
6.3.1 Model description 208
6.3.2 CFW outbred mice 210
6.3.3 CONVERGE study 212
6.3.4 Effect of subsampling reads and samples 216
6.3.5 Effect of variant filtration on imputation performance 217
6.4 Discussion 218
6.4.1 Algorithm performance 218
6.4.2 The effect of genetic history, sample size and sample coverage on performance 222
6.4.3 Implication for study design 224
6.4.4 Effect of variant filtration 228
6.4.5 Further considerations 229

7 Future directions and concluding remarks **231**

Bibliography **236**

List of Figures

1.1 Mouse evolutionary species tree 5

1.2 House mouse geographic range 6

1.3 House mouse migration routes 7

1.4 Model for repair of double strand breaks during meiosis 10

2.1 Sampling location of all mice 25

2.2 Sampling location of French mice 26

2.3 Sampling location of Taiwanese mice 27

2.4 Overview of variant calling pipeline 28

2.5 Sample sequencing depths 36

3.1 Ratio of sex chromosome coverage to autosomes 48

3.2 Shared polymorphic variation between species 51

3.3 Derived allele frequency spectra in the populations 53

3.4 Comparison of pairwise allele frequencies 54

3.5 PCA results for all mice 63

3.6 PCA results for wild mouse populations 64

3.7 PCA results for French and Taiwanese overlayed with sampling location 64

3.8 ADMIXTURE results for all mice 65

3.9 Neighbour joining tree 66

3.10 TreeMix on wild mice 66

3.11 PSMC for wild populations 67

3.12 Sensitivity and specificity for inbreeding in French mice 82
3.13 Sensitivity and specificity for inbreeding in Taiwanese mice 82
3.14 Example inbreeding assignment in wild mice 83
3.15 Homozygosity on chromosome 19 84
3.16 Distribution of homozygosity in the wild mice 85
3.17 Inactivating mutations in simulated wild mice 88
3.18 Genome-wide private fixed derived variants 97
3.19 Distribution of D and H genome-wide for the different populations . . 98
3.20 Distribution of singleton genotypes surrounding *Vkorc1* 99
3.21 Principal components analysis of genotypes surrounding *Vkorc1* . . . 100
3.22 Features surrounding *Brca2* in Taiwanese mice 103
3.23 Gene trees for region surrounding *Brca2* 104
3.24 Features surrounding *Prl* in Indian mice 109
3.25 Gene trees for region surrounding *Prl* 110
3.26 Sequence for *Prl* 111

4.1 Example recombination rate along genome for the three wild populations 125
4.2 Proportion of recombination in the maps explained by a given proportion of sequence 126
4.3 Correlation between wild recombination rate maps at varying scales . 127
4.4 Intensity of rate maps at previously published hotspots 128
4.5 Features of hotspots called from rate maps 129
4.6 Weak to strong substitutions around hotspots called from rate maps . 131
4.7 Best French motif from LD-hotspots 133
4.8 Example of bad French motif from LD-hotspots 133
4.9 Best Taiwanese motif from LD-hotspots 134
4.10 Best recombination motifs aligned to known motifs 135

4.11 Intersecting LD-hotspots stratified by posterior motif probability against known hotspots . . . 137
4.12 Nucleotide skew by derived allele frequency . . . 139
4.13 Association between recombination heat and nucleotide skew for Indian population . . . 141

5.1 SNP assignment to lineages for hotspot death . . . 154
5.2 Example crossovers . . . 161
5.3 QQ plots for hotspot death p-values . . . 165
5.4 Comparing WSB/EiJ p-values from lineage test to other lineages . . . 165
5.5 Comparing WSB/EiJ hotspot death lineage to AT to GC p-values . . . 166
5.6 Some top motifs across different lineages from hotspot death . . . 168
5.7 Top *Mus musculus* hotspot death motifs . . . 169
5.8 Motifs found without p-values from other test . . . 170
5.9 Best recombination and hotspot death motifs aligned to known motifs 171
5.10 Nucleotide substitutions around loss events from clusters . . . 172
5.11 All nucleotide substitutions around loss events from clusters . . . 173
5.12 Comparison of models for estimating parameters associated with GC-biased gene conversion . . . 175
5.13 Example of flexibility of four exponential hotspot death model . . . 178

6.1 Overview of STITCH . . . 209
6.2 Performance of method on CFW mice as assessed through external validation . . . 211
6.3 Performance of STITCH on CONVERGE humans as assessed through external validation . . . 215
6.4 Performance of method on reduced coverage CFW and CONVERGE datasets . . . 217

List of Tables

2.1 VCF annotation definitions 32

2.2 Quality metrics for callsets 37

3.1 Callable genome percent for different populations 47

3.2 Summary of variation in different populations 49

3.3 Comparison of pairwise variation 52

3.4 Number of stop-gains per wild mouse genome 87

3.5 Number of inactivated genes per wild mouse genome 87

3.6 Genotypes inside *Brca2* gene 102

3.7 Branches for *Brca2* gene 106

3.8 Genotypes inside *Prl* gene 108

4.1 Estimate of effective population size using genetic maps 122

4.2 Intersecting known-hotspots with LD-hotspots 136

5.1 2×2 table for lineage test 155

5.2 2×2 table for AT to GC test 156

5.3 Number of SNPs per lineage for hotspot death analyses 164

5.4 Number of significant k-mers and PWMs per lineage 166

5.5 Estimation of gene conversion parameters 176

6.1 Definitions for commonly used variables 186

6.2 Performance of CFW study under different programs and options . . 210

6.3 Performance of CONVERGE study under different programs and options with no reference panel 213
6.4 Performance of CONVERGE study under different programs and options with a reference panel 214
6.5 Performance of STITCH on CONVERGE study original imputation . 216
6.6 Effect of filtering on imputation performance 218

Chapter 1

Introduction

The mouse is a premier model organism, routinely used both for biomedical research as well as to investigate topics of broad biological interest. For example, much has been learned about the process of recombination, and its relationship with speciation, by studying the actions of the recombination-directing gene *Prdm9* in strains of laboratory mice. However, the mice typically studied in laboratories have a complex historical relationship to their wild counterparts, due to strong founding bottlenecks, limited genetic diversity, and introgression between subspecies [1]. Furthermore, factors that influence patterns of genetic variation in wild populations, such as selection, are either partially or fully disabled in these mice, through housing in highly artificial and controlled environments, and directed breeding. Studying processes such as recombination in laboratory mice therefore cannot fully capture the breadth and diversity of these processes in their wild counterparts. Studies of variation in wild mice will enable better interpretation of results from laboratory mice and will enhance our understanding of mouse biology more generally. Studies of variation will also aid future phenotype studies by enhancing study design, for example by identifying untapped pools of genetic variability in populations of mice, and aid in understanding phenotypic comparisons with humans, for example by understanding the prevalence of inbreeding in mice and its effect on protein coding variation.

Therefore, the primary goal of this thesis is to explore the influence of recombina-

tion and other processes on wild mouse genetic variation. To do this, I made use of whole-genome sequencing data from three populations of wild mice, as well as classic laboratory strains of mice and wild-derived inbred strains. By applying both existing and novel population genetic approaches to these data, I am able to draw conclusions on the past and recent organization of *Prdm9* usage in mice and its influence on nucleotide substitution, as well as demographic processes, inbreeding and selection.

In this thesis, I begin with a brief introductory chapter, which begins with a layout of the thesis, and follows with material and concepts which are relevant to multiple chapters. As an example, knowledge of mouse sub-speciation history and world-wide dispersal informs all aspects of downstream interpretation of genetic variation. Subsequent chapters will be largely self contained, with a preamble on chapter contents, further introductory text, methods as appropriate, results, and discussion.

The rest of this chapter will be organized as follows. First, I provide an overview of the layout of the thesis. Next, I briefly survey the known history of wild mice, followed by the biological machinery of recombination, and *Prdm9* diversity in wild mice. Finally, I review statistical topics relevant for the analyses performed in this thesis, such as hidden Markov models and expectation maximization.

1.1 Thesis layout

In this thesis, I explore factors that contribute to patterns of genetic variation in wild mice, paying particular attention to the process of recombination. I use whole-genome sequencing data from three mouse populations: a population of French *M. m. domesticus*, a population of Taiwanese *M. m. castaneus*, and a population of Indian *M. m. castaneus*. The use of these three populations across two subspecies allows for two important comparisons: between subspecies *M. m. domesticus* and *M. m. castaneus*; and between the ancestral Indian population with the derived French and Taiwanese populations. I also compare the wild mice with commonly used laboratory

strains, along with 6 wild-derived inbred strains, which are fully homozygous mice derived from wild mice. For this study, 3 outgroup strains are available, representing *M. caroli*, *M. famulus*, and *M. spretus*, as well as one representative each from *M. m. domesticus*, *M. m. musculus* and *M. m. castaneus*.

I begin by generating catalogues of single nucleotide polymorphism (SNP) variation in the mice from whole-genome sequencing data. A number of sensitivity and error estimates are used to show the data are of high quality. I present summaries on the numbers of SNPs in the different populations, and compare how variation is shared between them. A variety of standard population genetic approaches are used to analyze the relatedness and histories of the populations. I next investigate patterns of inbreeding in the populations using a novel hidden Markov model, which reveals substantial variation both between and within populations, and analyze how it relates to observed differences in protein coding variation among the studied mice. I also look for evidence of selection in the populations, finding several clear examples of adaptation, including adaptive introgression.

I next analyze patterns of recombination in the different wild populations using linkage-disequilibrium based rate maps. The French and Taiwanese maps are highly punctuate, unlike the qualitatively smooth Indian map, which is unlike any previously studied mammalian rate map. I validate these maps using previously published currently active hotspots (narrow regions with large rate increases), and I call French and Taiwanese hotspots from the rate maps. I am able to identify motifs in the French and Taiwanese hotspots, compare them to previously published motifs, and comment on the past use of motifs and hotspots in the different subspecies. I also look at the effect of recombination on nucleotide substitution, and show that guanine and cytosine (GC)-biased gene conversion has a dominant role in nucleotide evolution in all populations, but in particular in the Indian population, where due to the smoothness of the recombination rate map, it operates nearly genome-wide.

I next analyze the past history of recombination in mice using a technique modelled on hotspot drive. This re-identifies the previously found motifs for *M. m. domesticus* and *M. m. castaneus*. Using base-pair resolution identification of ancient motifs, I plot GC-biased gene conversion at very fine resolution, and use it to estimate parameters of interest in gene conversion, which may be interpreted as relating to crossovers and noncrossovers.

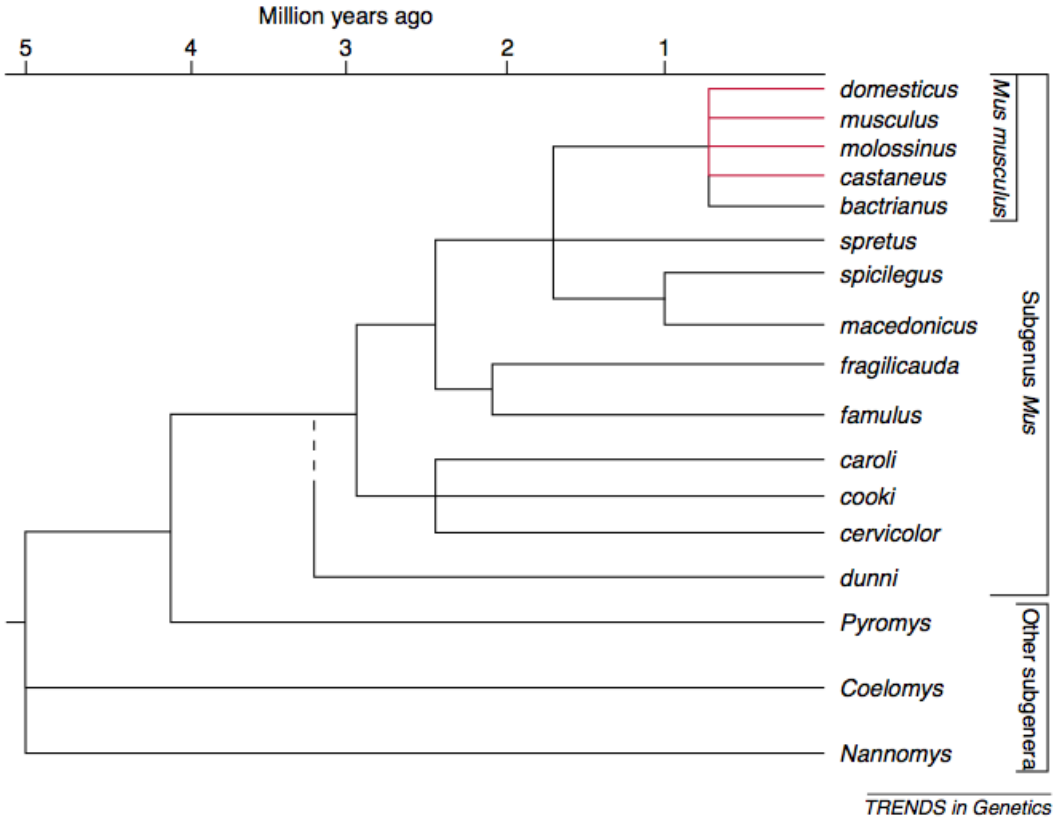
Finally, I present STITCH (sequencing to imputation through constructing haplotypes), which is an algorithm for low-coverage, read-aware, reference-panel free imputation. Although STITCH is unrelated to the analysis of variation in wild mice, its development builds upon hidden Markov models used earlier in this thesis. In presenting STITCH, I develop the methodology, and show its applicability on two populations: a laboratory population of outbred mice; and a human female Han Chinese population. On the mouse population, this method is the only known method which successfully imputes in this ultra-low coverage setting, while in the human population, this method performs as well as a method that uses a reference panel, and does so 6 times faster. I then discuss its applicability more broadly.

1.2 Survey of knowledge on wild mouse genetics

Mice, as well as rats, are members of the order *Rodentia*. The terms mice and rats do not correspond to a specific taxonomic grouping, but generally refer to small or large rodents with general physiological traits. In this thesis, I am principally interested in studying the house mouse *Mus musculus*, and its relations with recent ancestors, due to the role of *Mus musculus* in biomedical research. Throughout this thesis, I will use the term mouse when speaking of the house mouse and other closely related species in the genus *Mus*.

The relationship between the house mouse, *Mus musculus*, and several closely related species of interest are highlighted in Figure 1.1. These include the western

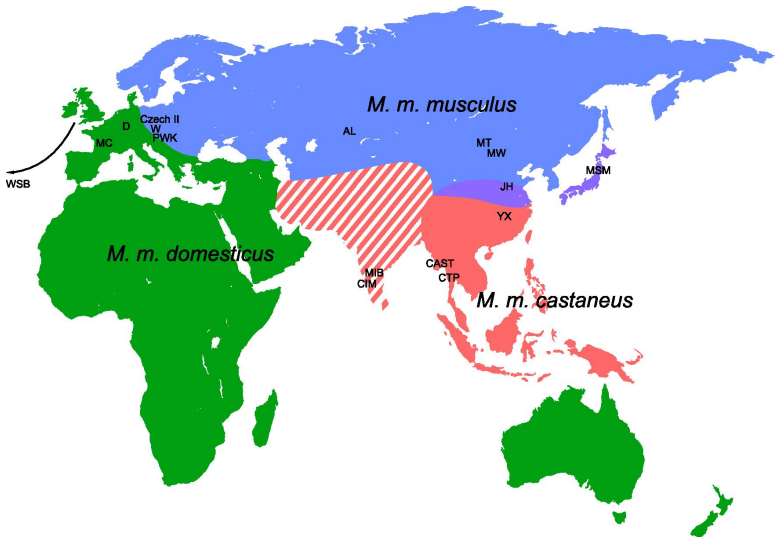
Figure 1.1. Mouse evolutionary species tree Figure from Guenet *et al.* [2]. Timings are calibrated using rat divergence.



Mediterranean mouse *Mus spretus*, the Indian *Mus famulus* and the Asian *Mus caroli* [2]. More distant relationships outside the subgenus *Mus* include karyotypically distinct species such as *Pyromys* and *Coelomys*. The divergence between mice and rats, assessed using *Mus musculus* and *Rattus norvegicus*, is thought to have occurred 12 million years ago, with the two species having about 16-18% pairwise diversity [3,4].

Within *Mus musculus*, classification of subspecies has been a notoriously difficult problem, due to subtle morphological and coat-colour variation, combined with widespread geographical range. More recent detailed genetic and biological studies have typically focused on 3 major well-characterized subspecies with wide geographic ranges, along with additional subspecies that have more narrow geographic ranges or unclear levels of support for subspecies designation [2, 5]. The present-day range

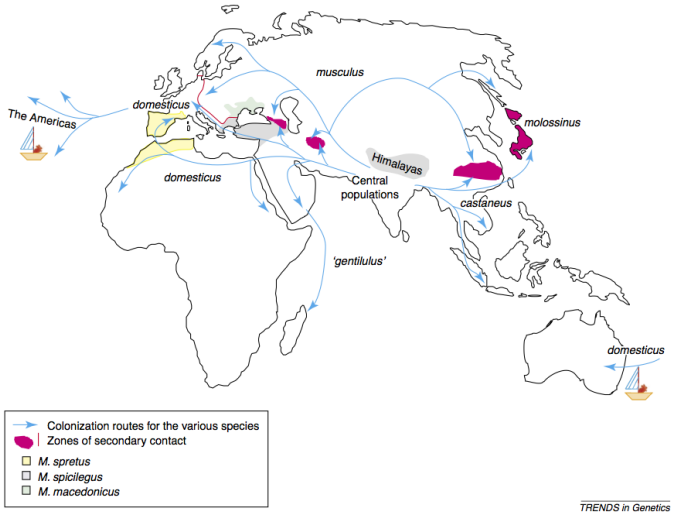
Figure 1.2. House mouse geographic range Figure from Lilue *et al.* [5]. Barred red refers to approximate region of ancestral range of *Mus musculus*, which is currently principally inhabited by *M. m. castaneus*.



of the 3 major subspecies is shown in Figure 1.2, and include: *domesticus*, found in Western Europe, North America and Australia; *musculus*, found in Eastern Europe and Central Asia; and *castaneus*, found in Eastern Asia. Additional subspecies with varying levels of support include *M. m. molossinus*, found in Japan and composed of a recent admixture between *musculus* and *castaneus* [6]; *bactrianus*, the name given to a southwestern Asian population thought to be closely related to the source population for house mice; *gentilus*, found in the Arabian peninsula [7, 8]; and an independent and currently unnamed population sampled in Nepal [8]. In this thesis, I will overwhelmingly refer to the three best-characterized subspecies, *domesticus*, *musculus* and *castaneus*.

Evidence suggests that *Mus musculus* traces its origins to southwestern Asia, from which it dispersed mirroring human expansion, as shown in Figure 1.2 [9]. There have been many efforts to categorize *Mus musculus* dispersal history over the years, with a particularly in depth recent approach coming from Suzuki *et al.* [8], that I will briefly summarize. In their work, where they sample *Mus musculus* mitochondrial DNA across a broad geographical range, they reach a number of conclusions. First, mi-

Figure 1.3. House mouse migration routes Figure from Guenet *et al.* [2].



tochondrial haplogroups within *Mus castaneus* differentiates into 4 sub-groups, with 3 of these are confined to southwestern Asia and only one of them making up the majority of *castaneus* samples found throughout Asia. Second, haplogroups within *musculus* are split into two subgroups, with the first principally confined to eastern and central Asia, and the second spanning from eastern Europe through Asia, with further subdivision agreeing with geographic coordinates. Third, the *domesticus* haplogroups are equally related and show no obvious subdivision. Finally, the ancestral source for *Mus musculus* likely ranges over southwestern Asia, from Iraq to the west to northwestern India to the east. As an aside, it is worth noting that of the *castaneus* haplogroups, their Taiwanese sample, as well as a Vietnemese sample, are the only two castaneus samples from outside southwestern Asia not to fall within the haplogroup which has expanded generally throughout Asia. It is worth noting that this might be due simply to a recent migratory event or stochastic deep coalescence, as an alternative to a larger older migratory event.

Subspecies split time estimates from the Suzuki *et al.* paper give a range of 0.37-0.47 million years ago, with perhaps *musculus* and *castaneus* having split more recently from each other than with *domesticus*, and a more recent split of 0.17 to

0.27 million years ago for splits within the *castaneus* subspecies and about 0.14 to 0.16 million years ago for a split within the *musculus* subspecies [8]. They note recent *castaneus* range expansion, which they date to 3,800 to 7,600 years ago, and is consistent with timings pertaining to regional human agricultural development, as in the expansion of *domesticus* dating to approximately 12,000 years ago [8].

It should be noted that incomplete lineage sorting is extensive in *Mus musculus* [8, 10–14]. Therefore, one should view mitochondrial studies as a good start, with the possibilities for more refined histories to come from larger studies of nuclear DNA. More recent work based on microsatellite nuclear data from Hardouin *et al.* finds generally similar results, and singles out the Iranian plateau as a more specific geographic home to many diverse *Mus musculus*, and perhaps represents an ancestral home for the species [15]. Additional estimates of population histories and sizes come from Geraldes *et al.*, and suggest a population size of 200,000 - 400,000 for *M. m. castaneus*, 100,000 - 200,000 for *M. m. domesticus* and 60,000 - 120,000 for *M. m. musculus*, with a split time between all subspecies of about 500,000 years ago [11].

As for cross-subspecies fertility, crosses between various subspecies occur at different efficiencies in the laboratory. Famously, male offspring between *M. m. domesticus* and *M. m. musculus* are often infertile [16], a well-studied phenomenon that helped identify the first known mammalian speciation gene [17]. Despite this, introgression between species has been documented many times, for example between *M. m. domesticus* and *M. m. musculus* in European populations of house mice [18]. More distant introgression from *M. spretus* into *domesticus* has also been noted, in the case of resistance to the rodenticide warfarin [19], as well as more generally genome-wide [20].

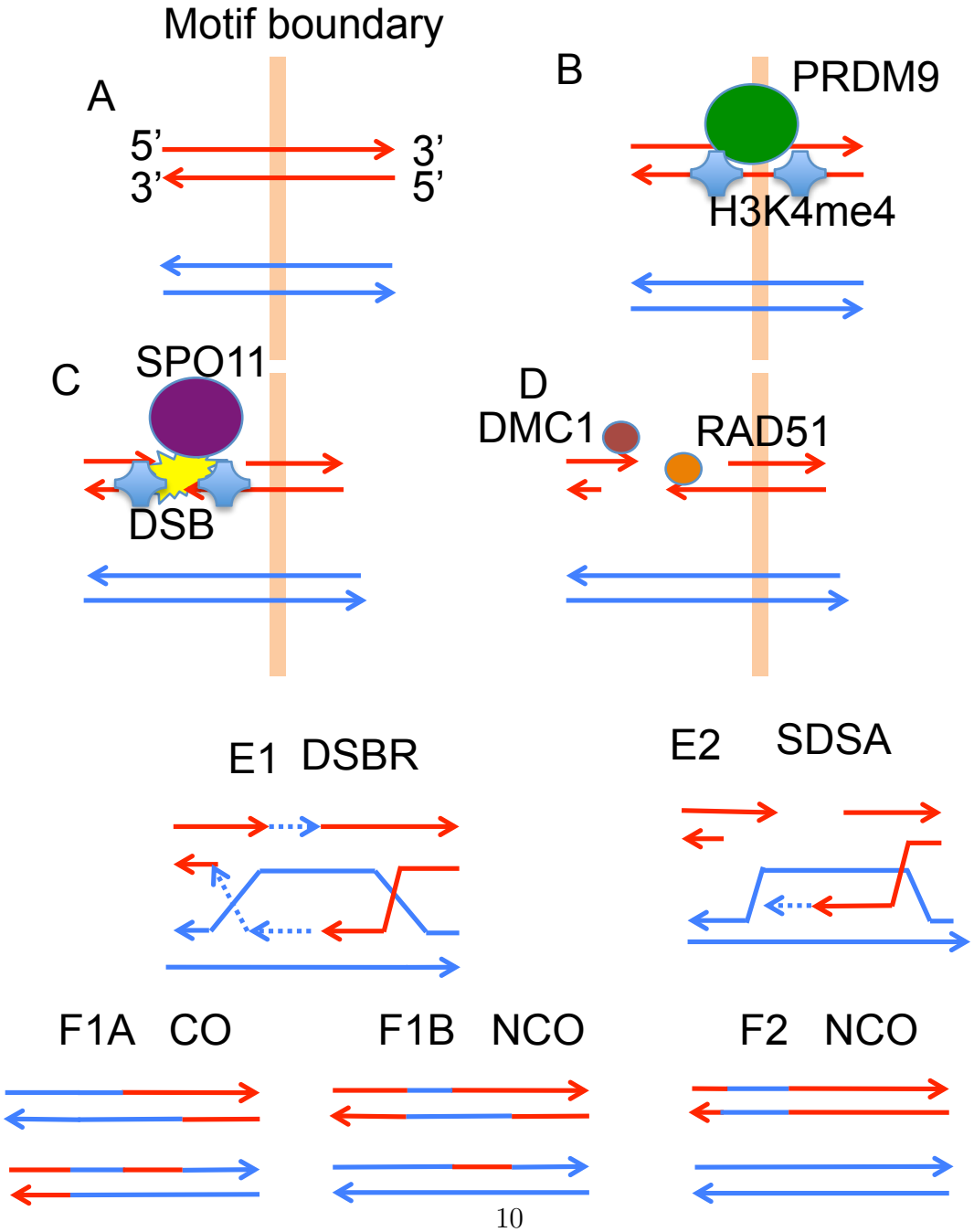
1.3 Biological machinery of recombination

Full reviews of the steps and proteins involved in mammalian meiotic recombination are available elsewhere, *e.g.* [21,22]. Here I seek to introduce the genes, their relationships and roles in mammalian meiotic recombination as necessary for interpretation of the results of this thesis. I pay particular attention to the role of PRDM9, given its role in localizing recombination at fine scales, and therefore its ability to shape the recombination landscape in different species and populations. I focus specifically on recombination in the mouse, and occasionally human, and avoid species with inactive forms of PRDM9, such as dogs or birds [23].

Meiosis in a diploid eukaryotic meiotic cell begins with a round of DNA replication to produce two copies, or chromatids, of each homologous chromosome, leaving four double-stranded copies of DNA in each chromosome [22]. This meiotic cell then divides in two during meiosis I, with segregation of homologous chromosomes to different cells, before the sister chromatids of each homologue separate in meiosis II [22]. In meiosis I, after DNA replication, during prophase I, double strand breaks (DSBs) are localized, induced and repaired, with repair occurring through two pathways (more below) and resolution leading to either crossovers or noncrossovers (more below), leading to the proper alignment (synapsis) of homologous chromosomes. One crossover per chromosome is required for proper segregation of homologues during meiosis; without such a crossover, the resulting germ cells may be aneuploid (have non-standard chromosome counts), with considerable reductions in fertility [22].

During prophase I of meiosis I, DSBs are induced and resolved, as shown in Figure 1.4. These DSBs do not occur at random, but at "hotspots" defined by the protein PRDM9, which promotes DSB formation through promoting trimethlyation of histone 3 at lysine 4 (H3K4me3) (more below) [25–27]. DSBs are catalyzed by the topoisomerase-like protein SPO11 [28]. Following DSB formation, broken DNA is resected by an exonuclease in a 5' to 3' orientation, leaving single-stranded overhangs

Figure 1.4. Model for repair of double strand breaks during meiosis One sister chromatid of double-stranded DNA is shown for two homologous chromosomes, in red and blue, with a PRDM9 binding motif region highlighted vertically in red/pink. During meiosis, homologous chromosomes (A) are bound by PRDM9 (B), which leads to H3K4me3 deposition. Double-strand breaks are then induced by SPO11 (C), followed by 5' to 3' strand resection, and subsequent protein loading, including DMC1 and RAD51 (D). Subsequently, repair occurs through either the DSBR (E1), with either crossovers (F1A) or noncrossovers (F1B), or through the SDSA pathway (E2), with resolution as a noncrossover (F2). Mismatches within heteroduplex DNA (dsDNA with different colours in F) are subsequently resolved (not shown). Modelled on [21, 24]



that are rapidly bound by RAD51 and DMC1 [28]. The single stranded DNA then invades the homologous chromosome, and leads to resolution as either a crossover, where one chromatid from each pair of homologous chromosomes are broken and rejoined, or a noncrossover, with no such reciprocal exchange among the chromatids. Studies from yeast have shown that DSB resolution occurs through one of two pathways, the double strand break repair (DSBR) pathway, or the synthesis-dependent strand annealing (SDSA) pathway [24]. Both crossovers and noncrossovers are associated with gene conversion, the unequal exchange of DNA between sister chromatids as part of DSB resolution [21]. These two repair pathways, as well as crossovers and noncrossovers, are highlighted as part of Figure 1.4.

The protein PRDM9 is a zinc finger protein, composed of several conserved domains, and a variable number of zinc fingers, that contact DNA and determine DNA binding specificity [23]. As a result, different PRDM9 alleles with different zinc fingers contact DNA with different affinities, leading to different binding motifs. In mice, PRDM9 has been shown to nearly fully determine the fine scale localization of nearly all hotspots during recombination [29]. Once bound to DNA, the SET domain in PRDM9 catalyzes histone H3 Lys4 trimethylation (H3K4me3), which leads, by unknown mechanisms, to eventual DSB formation by SPO11 [28].

It has been shown that if an individual is heterozygous for a hotspot-disrupting mutation, and hence has one hotter (higher affinity) and one colder (lower affinity) allele of DNA at that hotspot, then during gene conversion, the colder allele is likely to be overtransmitted, referred to as biased gene conversion [30, 31]. This overtransmission of alleles from a heterozygote during meiosis is referred to more generally as meiotic drive [32], has similar population genetic dynamics as selection, and will promote the fixation of the cold allele in the population over time [33]. This erosion of hotspots over time was originally viewed as a sort of paradox, in that hotspots should lead to their own erosion, yet they were clearly still present in humans and other

species [30, 33, 34]. Resolution of this paradox has largely been achieved through empirical determination of the role and structure of PRDM9 in hotspot determination, and how rapid mutagenesis of the zinc finger array due to its minisatellite structure [35] gives rise to new PRDM9 alleles with a corresponding complete shift to new uneroded binding locations and hence hotspots genome-wide.

During gene conversion, separate from biased gene conversion, mismatches that arise during repair of SNPs on homologous chromatids tends to lead to conversion of the weak (adenine or thymine) base in favour of the strong (cytosine or guanine) base [32]. This second, separate form of transmission distortion during meiosis, which may be the result of heteroduplex DNA repair choice, is called GC-biased gene conversion, and it promotes the fixation of weak to strong mutations near PRDM9 binding motifs over time [32]. Experimental evidence in favour of GC-biased gene conversion has been shown, for example in humans using sperm typing [36]. On a broad scale, GC-biased gene conversion is a possible mechanism to explain the correlation between GC content and recombination rate in the genomes of various species [37].

Recently, several studies have further increased or solidified our knowledge of mammalian recombination, including several in the mouse. I will highlight a few of these key papers here.

In 2014, Cole *et al.* provided a detailed analysis of individual tetrads (two pairs of sister chromatids from homologous chromosomes in meiosis) in either mouse oocytes or spermatocytes, and from this observed several key features of crossover and non-crossover behaviour in mice [38]. First, each of the crossovers they observed was associated with a unidirectional gene conversion tract from the unbroken to the broken homologue, with an average width of about 500-600 bp, consistent with double Holliday junction resolution in the DSBR pathway. Each noncrossover was also associated with a unidirectional gene conversion, at a shorter but more variable length than the crossover gene conversion. The use of unidirectional gene conversion, versus

bi-directional gene conversion, suggests that in mice, noncrossovers all but exclusively arise through the SDSA pathway. They also noted that despite tight DSB localization near PRDM9 binding motifs, many crossover gene conversion tracts do not include the PRDM9 motif itself.

In 2014, Pratto *et al.* made individual recombination maps of the human genome by performing chromatin immunoprecipitation followed by single-stranded DNA sequencing on DMC1 [39]. Among other analyses, they looked at the presence of rare and common SNP variation surrounding their ∼30,000 hotspots, stratified by ancestral and derived base, from which they make two conclusions. First, they observe an increase in rare variants that is strand-specific and not symmetric about the *Prdm9* motif, particularly for transitions, which is consistent with mutation of the exposed strand during DSBR. Second, they observed a proportionally larger increase in high frequency variants, that is symmetric and strand unaware, particularly for weak to strong substitutions, consistent with GC-biased repair of heteroduplexes in SDSA. Echoing the finding that low frequency mutations are enriched around DSB sites, in 2015, Arbeithuber *et al.* provided more direct evidence of the increased mutagenicity of crossover sites by sequencing 5,796 crossovers and showing an increased mutation rate among crossover versus non-recombining control DNA [40]. More precise comments were difficult in their work due to low sample sizes, with only 17 mutations on CO background versus 3 in non-recombining control regions.

In 2014, Odenthal-Hesse *et al.* used strand-specific PCR amplification of pools of sperm to detect crossovers and noncrossovers in six individual human hotspots [36]. They found transmission distortion from weak to strong bases only for noncrossover events very close to the PRDM9 binding site, and not at sites further away, although they had weak power for noncrossovers away from the PRDM9 binding motifs, due to a scarcity of detected events.

To recap, PRDM9 is a zinc finger protein that facilitates deposition of H3K4me3 at

(some of) its binding motifs during meiosis, leading to double strand breaks by SPO11. Once resected at the DSB, DNA is bound by the proteins DMC1 and RAD51, among others. In mice, repair of the double strand breaks occurs through either DSBR, leading to crossover with a 500-600 base pair gene conversion, or through SDSA, which leads to a noncrossover with a much smaller gene conversion thought to be less than 100 bp. Hotspots with heterozygous mutations leading to different PRDM9 binding affinities lead to an increased risk of passing on the hotspot-inactivating allele, or biased gene conversion. Similarly, repair of mismatches in heteroduplex DNA often favours the strong base over the weak base, known as GC-biased gene conversion. The universality of GC-biased gene conversion to noncrossovers and crossovers is not yet fully clear, given the limited hotspot-specific nature of analyses conducted so far. Finally, there is evidence that crossovers are mutagenic for strong to weak transitions.

1.4 *Prdm9* diversity in wild mice

Two recent papers, by Kono *et al.* [14] and by Buard *et al.* [12], are highly informative about the distribution of allelic forms of *Prdm9* in subspecies of *Mus musculus*. I review these papers briefly here.

In their work, Kono *et al.* sequenced various components of *Prdm9* and other regions of nuclear DNA from mice including widespread geographical sampling from the three subspecies [14]. Examining variation across the whole length of *Prmd9* in 12 mice, they found that variation is heavily concentrated within the zinc finger array (ZFA), although various non-synonymous mutations are observed outside the ZFA, including in the KRAB domain, but none in the PR/SET domain. They sequenced the *Prdm9* ZFA from 79 mice, observing 36 unique zinc fingers (ZFs), with considerable variation at the DNA-contacting residues at amino acid positions -1, 3 and 6 within the α-helix domain of each ZF. From these 36 ZFs, they observed 57 ZFAs, ranging in size from 9 to 16 ZFs, which, after phylogenetic tree construction, were

clustered into 5 groups: a *M. m. musculus* group, a predominantly *M. m. domesticus* group, and three predominantly *M. m. castaneus* groups. Except for the *musculus* groups, no group exclusively contains a single subspecies. Somewhat surprisingly, the multiple *castaneus* groups do not show a similar trend to the mitochondrial results from Suzuki *et al.* [8], with the Suzuki results implying *castaneus* mice spread out across southwestern Asia are all more closely related to each other than other *castaneus* mice, while the *Prdm9* results of Kono *et al.* show a broad geographical distribution for all 3 *castaneus* groups. It is also noteworthy that of the Japanese mice analyzed in this study, representing a recently founded island population, 12 of 13 were homozygous for the same ZFA, suggesting that despite (seemingly) high levels of diversity, founding bottlenecks likely highly restrict *Prdm9* diversity but do not otherwise obviously encumber population survival.

In their work, Buard *et al.* similarly sequenced 93 wild mice from a wide geographic sampling, observing 78 ZFAs and 118 ZFs [12]. Similar to Kono *et al.*, the DNA-contacting residues -1, 3 and 6 of the α-helix were by far the most variable. Constructing a phylogenetic tree, the mice grouped into three major subtrees, with a predominantly *M. m. domesticus* subtree, a predominantly *M. m. castaneus* subtree, and a mixed *M. m. musculus* and *M. m. castaneus* tree. Buard *et al.* conclude that the diversity of *Prdm9* they observe is consistent with previous reports on genomic variability in mice, for example from [41], and therefore that variation in sequence of *Prdm9* in these populations is not due to selective forces.

1.5 Statistical techniques used in this thesis

In this thesis, I make use of a variety of software packages for both data processing and statistical inference. However, there were instances where tools were unavailable to perform the desired inference. In those cases, I developed methods, based on hidden Markov models and expectation maximization, to perform the desired in-

ference, such as in Chapter 3, for homozygosity determination, or in Chapter 6, for low-coverage imputation. Here, I lay out the necessary framework required to understand the statistical underpinnings and motivation for both hidden Markov models and expectation maximization. Other statistical techniques developed in this thesis, such as parameter estimation in Chapter 5, which do not span multiple chapters, are explained as they are used.

1.5.1 Expectation maximization

Parameter estimation and model fitting are often made challenging by having to deal with unobserved hidden, or latent, variables. In such a case, directly calculating the likelihood (the probability of the parameters of the model given the observations) by summing over the space of hidden variables is often computationally infeasible. In such situations, the expectation maximization (EM) algorithm is a method for finding maximum likelihood estimates. It works by alternating between estimating the unobserved variables, and then determining a new set of parameters by maximizing the expected log likelihood with respect to the hidden states determined by the previous parameter set.

Now for a more formal treatment. We are interested in maximizing the log likelihood of λ, the set of parameters describing the model, with respect to the observed data O,

$$l(\lambda|O) = \log(P(O|\lambda)) \tag{1.1}$$

However, to do so, we need to integrate over the hidden variables $Q \in \mathcal{Q}$

$$l(\lambda|O) = \log(P(O|\lambda)) = \log\left(\sum_{Q\in\mathcal{Q}} P(O,Q|\lambda)P(Q|\lambda)\right) \tag{1.2}$$

This is often infeasible, given the space of $\mathcal{Q}$. Instead, consider the full likelihood

$$l(\lambda|O,Q) = \log(P(O,Q|\lambda)) \tag{1.3}$$

In the expectation step, we calculate the function that is the expectation of the log-likelihood given the hidden states Q calculated using current parameters λ' and O

$$U(\lambda, \lambda') = \mathbb{E}[l(\lambda|O, Q)|O, \lambda'] = \sum_{Q \in \mathcal{Q}} P(Q|O, \lambda') \log(P(O, Q|\lambda)) \tag{1.4}$$

In the maximization step, we maximize $U(\lambda, \lambda')$ with respect to λ', and use the λ as the new parameter estimates. The EM algorithm is therefore composed of first initializing λ, and then alternating between the expectation step, where $U(\lambda, \lambda')$ is calculated based on the parameter estimates λ' from the previous iteration, and the maximization step, where new parameters λ are obtained so as to maximize $U(\lambda, \lambda')$. Now, it is not necessarily clear that the above should work. A new set of parameters λ is generated from λ' such that $\mathbb{E}[l(\lambda|O, Q)|O, \lambda']$ is increased, but it is not clear that $l(O|\lambda) \geq l(O|\lambda')$. To see why this is the case, consider the following. For any Q with non-zero probability, $P(Q|O, \lambda)$, then

$$\begin{aligned} P(O|\lambda) &= \frac{P(O, Q|\lambda)}{P(Q|O, \lambda)} \\ \implies \log(P(O|\lambda)) &= \log(P(O, Q|\lambda)) - \log(P(Q|O, \lambda)) \end{aligned} \tag{1.5}$$

Taking the expectation of both sides over Q for a fixed O and λ' gives

$$\begin{aligned} \sum_{Q \in \mathcal{Q}} P(Q|O, \lambda') \log(P(O|\lambda)) &= \sum_{Q \in \mathcal{Q}} P(Q|O, \lambda') \log(P(O, Q|\lambda)) - \sum_{Q \in \mathcal{Q}} P(Q|O, \lambda') \log(P(Q|O, \lambda)) \\ &= U(\lambda, \lambda') + V(\lambda, \lambda') \end{aligned} \tag{1.6}$$

where $V(\lambda, \lambda') = -\sum_{Q \in \mathcal{Q}} P(Q|O, \lambda') \log(P(Q|O, \lambda))$. By Gibbs' Inequality (Theorem 1, supported by Lemma 1), $V(\lambda, \lambda') \geq V(\lambda', \lambda') \, \forall \lambda$, and since $U(\lambda, \lambda') \geq U(\lambda', \lambda')$ by the choice of λ, then

$$\begin{aligned} \log(P(O|\lambda)) - \log(P(O|\lambda')) =& U(\lambda, \lambda') - U(\lambda', \lambda') + \\ & V(\lambda, \lambda') - V(\lambda', \lambda') \\ \geq& 0 \end{aligned}$$

and therefore, by the choice of λ, the EM algorithm increases the likelihood with each iteration.

Lemma 1.1. $\forall x > 0$, $\log(x) \leq x - 1$

Proof. The Taylor series of e^x is $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$. Consider x and e^{x-1}. The

$$\begin{aligned} e^{x-1} &= 1 + (x-1) + \frac{(x-1)^2}{2!} + \frac{(x-1)^3}{3!} + ... \\ &= x + \frac{(x-1)^2}{2!} + \frac{(x-1)^3}{3!} + \frac{(x-1)^4}{4!} + \frac{(x-1)^5}{5!} + ... \\ &= x + \frac{(x-1)^2}{2!}\left(1 + \frac{x-1}{3}\right) + \frac{(x-1)^4}{4!}\left(1 + \frac{x-1}{5}\right) + ... \\ &\geq x* \end{aligned}$$

$$\implies \log(e^{x-1}) \geq \log(x)$$

$$\implies x - 1 \geq \log(x)$$

where * holds due to the terms $(1 + \frac{x-1}{n}) > 0 \quad \forall n > 0, n \in \mathbb{N}$ with $x > 0$

□

Theorem 1.1 (Gibbs's Inequality)**.** *Let* $\sum_{i=1}^n p_i = 1$, $\sum_{i=1}^n q_i = 1$, $0 \leq p_i \leq 1$, $0 \leq q_i \leq 1$. *Then* $-\sum_{i=1}^n p_i \log(p_i) \leq -\sum_{i=1}^n p_i \log(q_i)$

Proof. Let $I = \{i | p_i > 0\}$. Then

$$\begin{aligned} -\sum_{i=1}^n p_i \log(q_i) + \sum_{i=1}^n p_i \log(p_i) &= -(\sum_{i=1}^n p_i \log(q_i) - \sum_{i=1}^n p_i \log(p_i)) \\ &= -(\sum_{i \in I} p_i \log(q_i) - \sum_{i \in I} p_i \log(p_i)) \\ &= -\sum_{i \in I} p_i \log(q_i/p_i) \\ &\geq -\sum_{i \in I} p_i (\frac{q_i}{p_i} - 1)* \\ &= -\sum_{i \in I} q_i + \sum_{i \in I} p_i \\ &= -\sum_{i \in I} q_i + 1 \\ &\geq -1 + 1 = 0 \end{aligned}$$

where * holds due to the previous lemma. Therefore, $-\sum_{i=1}^{n} p_i \log(p_i) \leq -\sum_{i=1} p_i \log(q_i)$, as required

□

1.5.2 Hidden Markov models

The Markov property is where the probability of a future state depends only on the current state and not on previously occurring states. A Markov model is a model where the states satisfy the Markov property. A hidden Markov model (HMM) is one where the states are "hidden" and satisfy the Markov property, and there are separate observations, which are visible and are emitted from the hidden states in a probabilistic fashion. In Genetics, HMMs have wide appeal, due to their ability to plausibly model biological processes, their computational simplicity and their flexibility. One such example from population genetics is HapMix, which uses a hidden Markov model where the hidden states are ancestry along a chromosome, and the observations are genotypes [42].

Here, I first outline a Markov model in more detail, and then expand this to a hidden Markov model. I omit an example, but these can be found elsewhere, for example in a tutorial by Rabiner [43]. I then explain two techniques, the Viterbi algorithm, and the forward backward algorithm, for calculating hidden state probabilities given observations. Together, the outline presented here, as well as detail in later chapters, should be sufficiently detailed to enable implementation of the methods presented in this thesis (except for computational implementation issues, such as scaling).

Markov model

Consider a system with K discrete states, numbered $1, ..., K$. Consider a set of states q_1, q_2, ..., ordered in some way (consider here by time) with $q \in \{1, ..., K\}$. Then the Markov property states that

$$P(q_{t+1} = k_{t+1} | q_1 = k_1, ..., q_t = k_t) = P(q_{t+1} = k_{t+1} | q_t = k_t) \tag{1.7}$$

To fully specify this model, two things are needed. First, initial state probabilities, denoted π_k, with $\pi_k = P(q_1 = k)$. Second, probabilities to describe transitions between states. In a homogeneous system, one where transitions between states are invariant with respect to time, then $A_{i,j} = P(q_{t+1} = j|q_t = i)$, with $0 \leq A_{i,j} \leq 1 \quad \forall i, j$. In heterogeneous systems, we can write $A^t_{i,j}$, to reflect that these probabilities change over time.

Hidden Markov model

In an HMM, states are observed indirectly, through observations O_t at time t. The observations O_t in an HMM satisfy the condition

$$P(O_t|q_1 = k_1, ..., q_t = k_t) = P(O_t|q_t = k_t) \tag{1.8}$$

To fully specify this model requires the initial state probabilities and transition probabilities from the Markov model, as well as emission probabilities defined by $b_k(O_t) = P(O_t|q_t = k)$, where $b_k(O_t)$ represent the emission probabilities for state k over the range of O_t. More complicated emission probabilities are possible, as are seen in later chapters of this thesis.

Let $\lambda = (A, b, \pi)$ be the parameters of the model. Then the probability of a set of observations $O = \{O_1, ..., O_T\}$ is calculated as

$$P(O_1, ..., O_T|\lambda) = \sum_{Q \in \mathcal{Q}} P(Q, O_1, ..., O_T|\lambda) \tag{1.9}$$

$$= \sum_{Q=(k_1,k_2,...,k_T)} P(q_1 = k_1) \prod_{t=1}^{T-1} P(q_{t+1} = k_{t+1}|q_t = k_t) \prod_{t=1}^{T} P(O_t|q_t = k_t) \tag{1.10}$$

However, directly performing this calculation is computationally infeasible in most real applications, as there are K^T different possibilities for Q. HMMs allow for such calculations to be performed with relative ease.

When using an HMM, people, such as Rabiner [43], often describe there being three specific challenges associated with their implementation:

1. Given the set of observations $O = \{O_1, ..., O_T\}$, calculate $P(O|\lambda)$

2. Given O, calculate the Q that offers the "best" explanation (discussed below)

3. Optimize λ given O

The solution to the first problem is a handy by-product of a way of calculating the second problem, while the third problem is dealt with in this thesis by expectation maximization, which was already discussed. Here, I focus on efficient calculation of the second problem.

There are two commonly used ways to try and solve the second problem. The first, the Viterbi algorithm, calculates the most likely state sequence, call it $q*$, *i.e.* $q* = \arg\max_q P(Q = q|O, \lambda)$. The second, using the forward backward algorithm, allows for the calculation of the marginal probability of each state at every time, or $P(q_t = k|O, \lambda) \quad \forall k, t$, and hence allows one to calculate the most likely set of marginal states. The forward backward algorithm is often further used for updating model parameters in EM, as will be seen in later chapters.

Viterbi algorithm

The Viterbi algorithm proceeds by recursion to find the state path that maximizes the likelihood, denoted q^*, given O and λ. To use this recursive method, one first initializes at $t = 1$, then proceeds through $t = 2, ..., T$. This yields the most likely final state (q_t^*), but not the full path of how to get there ($\{q_u^*|1 \leq u < t\}$). However, if one keeps track of the best state at time t to yield each state at time $t + 1$, one can backtrack and calculate the single most likely state sequence.

More formally, let $\delta_t(k_t)$ be the maximum probability over the states and the observations up to time t, given that the state at t is k_t, or

$$\delta_t(k_t) = \max_{k_1,...,k_{t-1}} P(q_1 = k_1, ..., q_{t-1} = k_{t-1}, q_t = k_t, O_1, .., O_t|\lambda) \tag{1.11}$$

$$= \max_j \left(\delta_{t-1}(j) A_{j,k}\right) b_k(O_t) \tag{1.12}$$

Similarly, let $\psi_t(k)$ be the state at time $t-1$ most likely to have yielded state k at time t

$$\psi_t(k) = \arg\max_j(\delta_{t-1}(j)A_{j,k}) \tag{1.13}$$

To initialize δ, *i.e.* $\delta_1(k)$, we use the priors and the emission probabilities. After recursion, the probability of q^* given O and λ is $P(q^*|O,\lambda) = \max_k(\delta_T(k))$, and the state at the final time is $q_T^* = \arg\max_k(\delta_T(k))$. Calculation of the remaining states, *i.e.* the backtracking for $t = T-1, ..., t = 1$, is done by recursively calculating $q_t^* = \psi_{t+1}(q_{t+1}^*)$.

Forward and backward algorithms

Much like the Viterbi algorithm, the forward and backward algorithms are recursions. Consider first the forward algorithm. We are interested in the joint probability of being in state k at time t along with the set of observations up to time t

$$\alpha_t(k) = P(O_1, ..., O_t, q_t = k|\lambda) \tag{1.14}$$

To calculate $\alpha_{t+1}(k)$ from $\alpha_t(k)$, notice that all that changes is the inclusion of the observation O_{t+1} and the change in state from t to $t+1$. Therefore, we can easily calculate $\alpha_{t+1}(k)$ from $\alpha_t(k)$ by summing over the previously calculated $\alpha_t(j)$ multiplied by the transition probabilities from j to k and the observations at time $t+1$

$$\alpha_{t+1}(k) = \left[\sum_{i=1}^{K} \alpha_t(i)A_{i,k}\right] b_k(O_t) \tag{1.15}$$

Initialization of the forward algorithm is done with the priors and the emission probabilities to calculate $\alpha_1(k)$. Note that as a useful by-product, we get the full probability of the observations given the parameters as

$$P(O|\lambda) = \sum_{k=1}^{K} \alpha_T(k) \tag{1.16}$$

The backward algorithm uses a similar idea. We are interested in the joint probability of the remaining observations conditional on being in state k at time t,

$$\beta_t(k) = P(O_{t+1}, O_{t+2}, ..., O_T | q_t = k, \lambda) \tag{1.17}$$

$\beta_t(k)$ is calculated given $\beta_{t+1}(j)$ by considering all $\beta_{t+1}(j)$ for states j at time $t+1$, adding in the emission for that state, and then transitioning back to state k at time t

$$\beta_t(k) = \sum_{i=1}^{K} \beta_{t+1}(i) b_i(O_{t+1}) A_{k,i} \tag{1.18}$$

To initialize, $\beta_T(k) = 1 \quad \forall k$. Recursion is then done from $t = T-1$ to $t = 1$, $\forall k$.

Marginal state probabilities can then be calculated using the probabilities from the forward and backward probabilities with

$$P(q_t = k_t | O, \lambda) = \frac{\alpha_t(k)\beta_t(k)}{P(O|\lambda)} \tag{1.19}$$

Finally, as an aside, note that, crucially, in the implementation given here, the forward and backward algorithms have computational complexity proportional to K^2T. Efficient consideration of symmetries in transition matrices used commonly in statistical genetics can be used to reduce this to KT, as noted in, for example, in the fastPHASE algorithm [44]. This is considered in Chapter 6.

Chapter 2

Cataloguing variation in wild mice

The major aim of this thesis is to characterize variation in wild mice. The purpose of this chapter is to describe the acquisition of samples and data, and the generation of SNP callsets. Newly acquired samples include 20 French *M. m. domesticus* mice and 20 Taiwanese *M. m. castaneus* mice. Previously sequenced and published samples include 10 Indian *M. m. castaneus* mice and 1 *M. famulus* sample from Halligan *et al.* [3], 13 classical laboratory strains and 4 wild derived inbred strains from Keane *et al.* (known as the "17 Strains" dataset) [13], and an *M. caroli* sample from Stefflova *et al.* [45]. The subsequent analysis of variant callsets from either the complete set of 69 mice or population-specific callsets forms the majority of the remainder of this thesis.

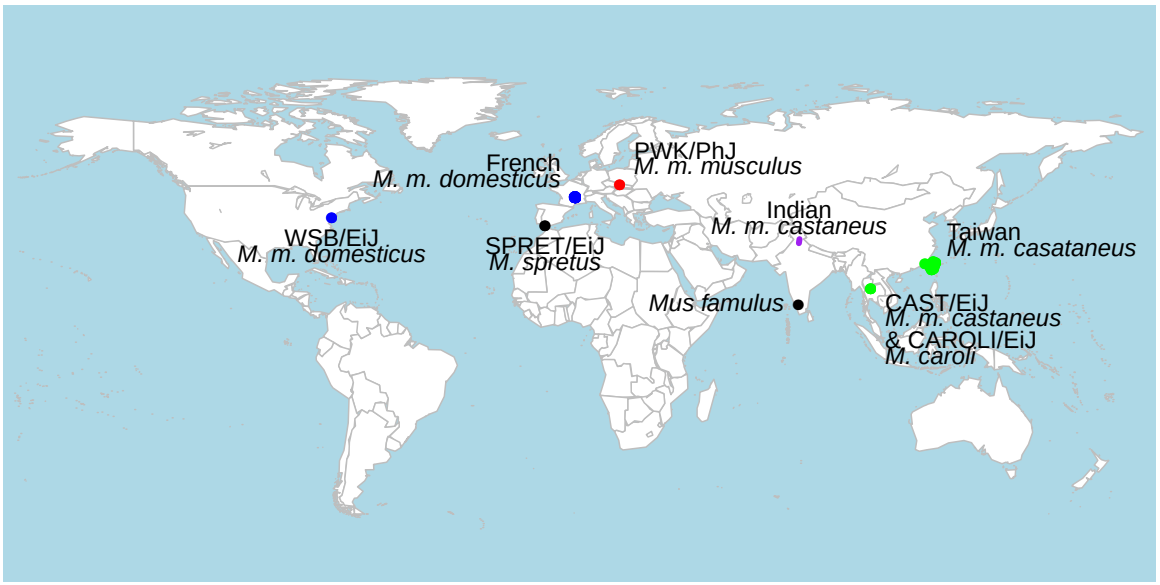
The rest of this chapter will be organized as follows. I begin with a Methods section, where I present sampling information for the mice used in this thesis, outline the pipeline used to call variants, and give detailed methodological information. In the Results, I describe the sampling and sequencing of the mice, present several SNP callsets for each of the populations, calculate quality metrics for them, and then outline a choice of a combined as well as population-specific callsets. Finally, in the Discussion, I review these results in light of previously published works, and talk about how they might be improved with recent methodological innovations.

2.1 Methods

Sampling and data acquisition

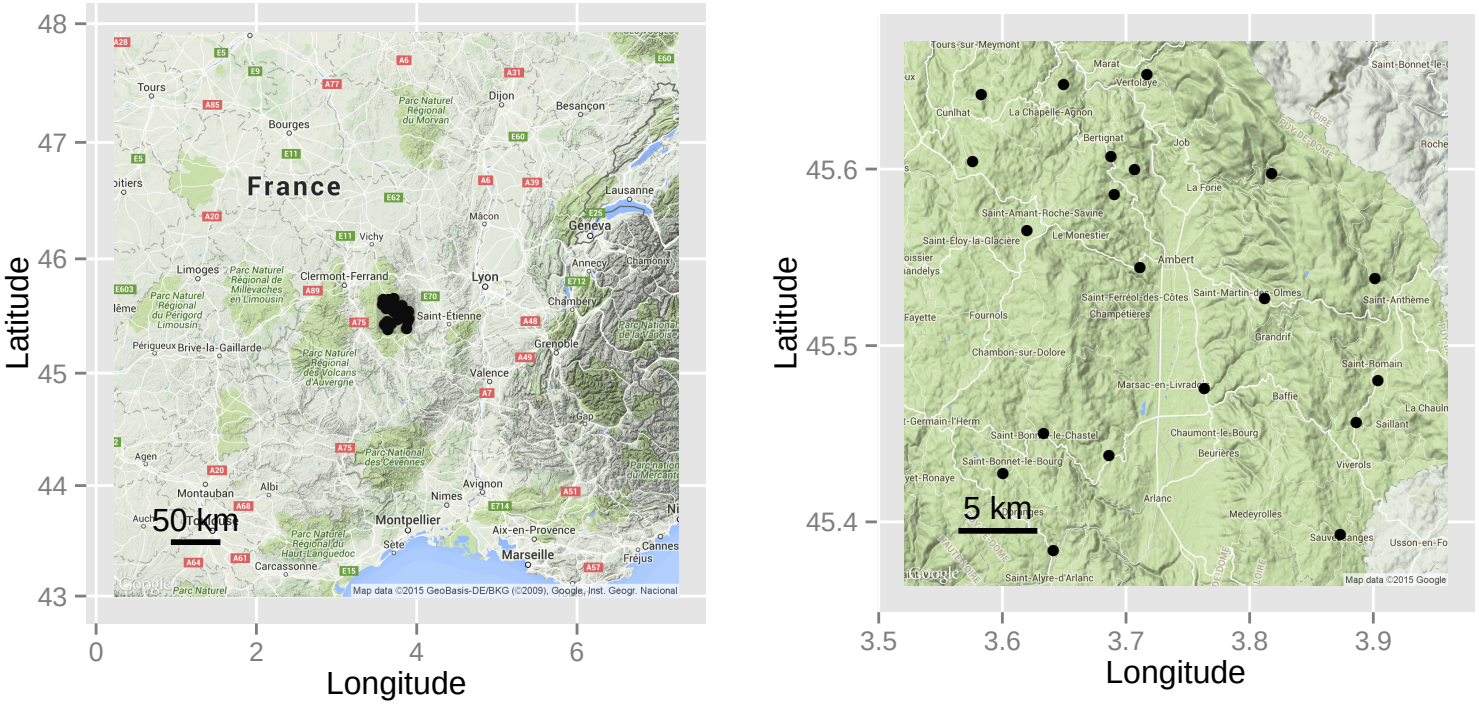
Figure 2.1 shows approximate geographical sampling of all mice considered in this thesis. This includes mice sequenced for the first time as part of work relating to this thesis, as well as previous sequencing efforts for which data was downloaded. More detailed sampling information is now given.

Figure 2.1. Sampling location of all mice Shown are the approximate sampling locations for all mice used in this thesis, as well as the subspecies or species designation in italics. Sampling locations for the classical laboratory strains are not given, as they are of more mixed origin. The French and Taiwanese mice sampling information is not previously published, while the Indian sampling locations are from [46], and the other mice are from [47]. CAST/EiJ and CAROLI/EiJ sampling locations are given as Thailand so are overlapping in this plot.



We generated data for 20 French *M. m. domesticus* and 20 Taiwanese *M. m. castaneus* wild mice. The French mice were collected by Amelie Baud and Binnaz Yalcin over an approximately 50km region with each mouse sampled at least 1.7 km apart, with specific locations shown in Figure 2.2. The Taiwanese mice were collected by members of the lab of Alex Yu from around the periphery of Taiwan, with specific

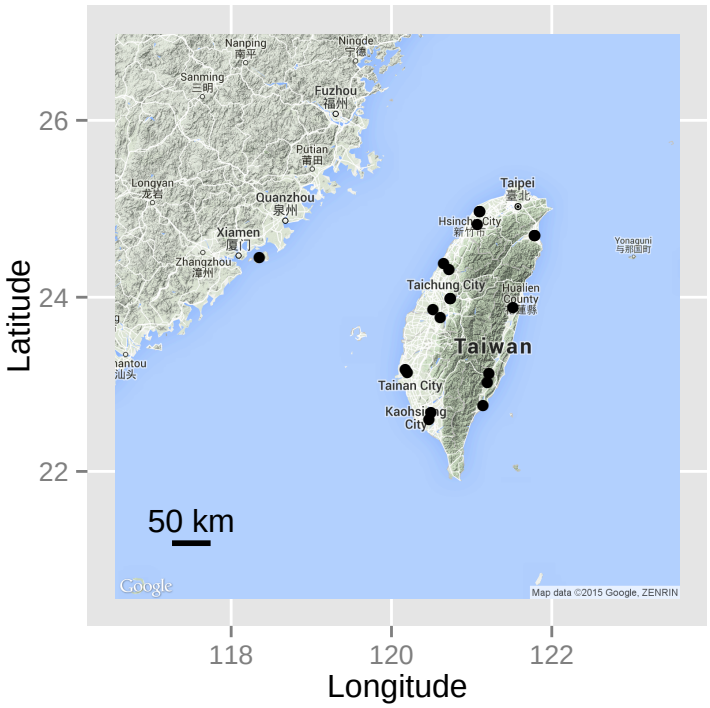
Figure 2.2. Sampling location of French mice Plots show sampling ranges for wild French mice. Left plot is zoomed out, right plot is zoomed in



locations shown in Figure 2.3. DNA extraction methods are not provided but followed standard lab protocols. Samples were sequenced from short insert paired end libraries to 10X on Illumina HiSeq machines. One French sample had an additional 30X of sequence generated for quality control purposes, to determine the influence of 10X versus 40X sequencing depth on genotype error.

I also downloaded raw (FASTQ) data on 10 wild Indian mice and 1 wild derived inbred strain *M. famulus* sample from Halligan *et al.* [3]. The *M. famulus* sample traces its origins to Kotagiri, Tamil Nadu, India, by way of the Montpellier Wild Mice Genetic Repository [46]. The Indian mice were collected in 4 localities over a 130 km transect through India, as shown on Figure 2.1, in 2003, with each mouse collected at least 500 m apart [46]. These mice are part of a subset of 15 of 38 Indian mice genotyped at 60 microsatellite markers and showing local clustering, based on

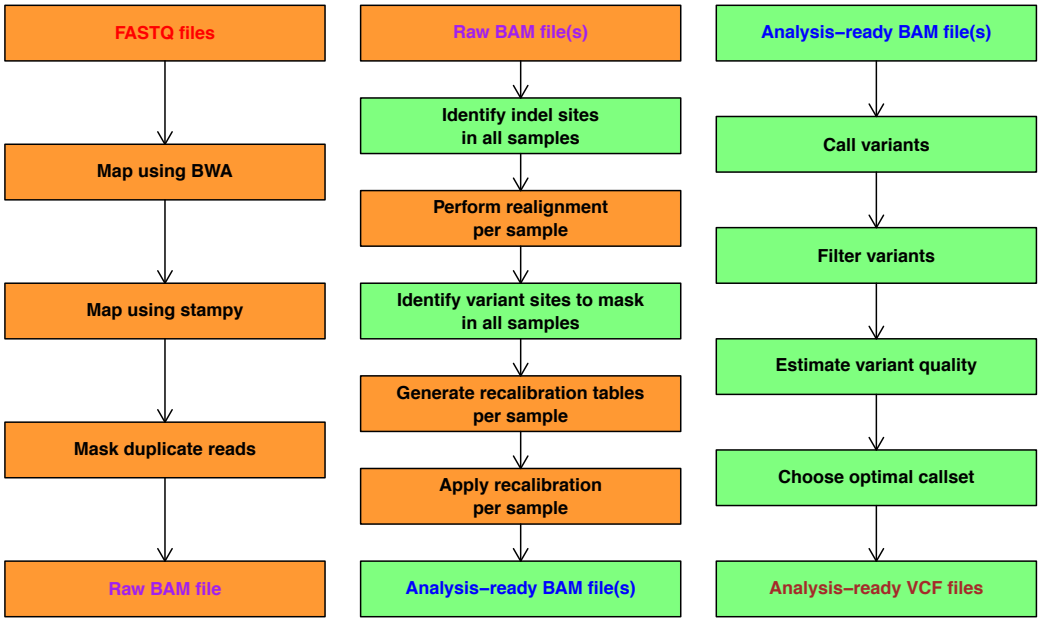
Figure 2.3. Sampling location of Taiwanese mice Plot show sampling ranges for wild Taiwanese mice. Note that at this resolution, some sampling locations appear overlapping



a STRUCTURE analysis [3]. Of note, one of the 15 mice showed a substantially higher proportion of homozygosity based on the microsatellilte investigation and was excluded from sequencing for that reason. Sequencing was reported, and used, from 10 of these 14 mice.

Finally, BAMs were available for 13 classical laboratory strains, primarily of *M. m. domesticus* origin [1], and 4 wild derived inbred strains, one each of *M. m. domesticus* (WSB/EiJ), *M. m. musculus* (PWK/PhJ), *M. m. castaneus* (CAST/EiJ) and *M. spretus* (SPRET/EiJ), from Keane *et al.* [13]. Finally, a single fully inbred *M. caroli* sample from Stefflova *et al.* [45] was obtained as raw FASTQ files.

Figure 2.4. Overview of variant calling pipeline The creation of analysis-ready VCFs from FASTQ files is broken conceptually into three parts, represented by three columns. The leftmost column, the making of raw BAM files, is run per-sample. The central column, the making of analysis-ready BAM files, is done using all samples from a population at once, and uses a mixture of per-sample and across sample analyses. The right column, the creation of analysis ready VCFs, is done using all samples from a population. Individual steps are broken into per-sample (orange) or across sample (green) steps.



Variant calling pipeline overview

In generating filtered SNP variant call format (VCF) files from raw read information in FASTQ files, I followed the pre-processing pipeline recommended by the Genome Analysis ToolKit (GATK) team as laid out in [48]. An overview of the approach taken in this thesis is given in Figure 2.4. Conceptually, this can be broken into three stages. The first stage is run per-sample, starts with raw reads (FASTQ files), maps them, removes or marks PCR duplicates, and creates one raw BAM file per sample. The second stage involves processing the BAM files to improve their quality

by realigning reads near indels and empirically re-estimating read base qualities. The third stage involves calling and filtering variants, as well as estimating their error rates and choosing a filtered callset with optimal quality criteria. To filter SNPs, I used the Variant Quality Score Recalibrator (VQSR), which builds a Gaussian mixture model (GMM) at known true positive sites using quality control metrics from the raw VCF, and then filters out novel variants which are probabilistically in the tail of the distribution defined by the GMM on the known variants, based on a user-set sensitivity threshold. I made use of three sets of true positive variants, one from the 17 Strains paper [13], and two using Affymetrix DNA genotyping microarray data [49]. All of the French mice were genotyped on the Affymetrix array, so the two Affymetrix training sets include one where all array sites were used without filtration, and a second for only the French mice where only those SNPs which showed high concordance between array and sequencing genotypes for the French mice were used. Finally, to evaluate accuracy of the SNP callsets, I make use of several metrics, some of which are customized and described in this Methods section.

Except for the 17 Strains data, which was downloaded as analysis ready BAMs, pre-processing was done using the steps in the left column of Figure 2.4. Per-population BAM refinement was done according to the central column of Figure 2.4, separately for the 3 wild populations, the *M. famulus*, *M. caroli*, and 1 high coverage French mouse. Variant calling was performed either across all mice, or for the 3 wild populations individually.

Generating BAMs from FASTQs

Mapping of sample reads was done against mm9 using bwa aln with flag -q10 [50], merged using bwa sampe, re-aligned using stampy with –bamkeepgoodreads [51], and converted back to BAM format using samtools [52]. BAMs were merged into library level using Picard MergeSamFiles.jar with SORT_ORDER=coordinate (Picard URL `http://broadinstitute.github.io/picard/`). PCR duplicates were either marked

or removed using Picard MarkDuplicates, and the resulting BAMs merged into a sample level BAM, again using Picard.

BAM pre-processing

Version 2.1-11-g13c0244 of the Genome Analysis ToolKit (GATK) [53] was used unless otherwise noted. I did not consider two regions of the genome by using the -XL flag, one on chromosome 9 between 3,000,000 and 3,084,000, and the other on chromosome 2 between 98,500,000 and 98,510,000, as these regions contained a large number of poorly mapped reads and were problematic when running the GATK in parallel mode.

Local indel realignment for a given set of mice was conducted in two stages. First, for that set of mice, the GATK RealignerTargetCreator was run on the full set of BAM files to create a list of intervals needing realignment. Next, the GATK IndelRealigner was run one mouse at a time using the interval list from RealignerTargetCreator.

Base Quality Score Recalibration was done similarly but in three stages. First, for a given population, sets of putatively polymorphic sites were generated by running the GATK UnifiedGenotyper on the full population under default conditions. Second, these sites were used for each mouse in a given population as a list of sites to ignore when generating base quality score recalibration tables for each mouse separately using the GATK BaseRecalibrator. Finally, recalibrated BAM files were printed one at a time using the GATK PrintReads using the recalibration tables from the GATK BaseRecalibrator.

Affymetrix Mouse Diversity Genotyping array

All of the 20 French mice were genotyped on the Affymetrix Mouse Diversity Genotyping Array as part of a larger set of 65 genotyped samples processed in Oxford. Genotype calling of the 65 samples was done along with 351 samples downloaded from the Jax website `http://cgd.jax.org/datasets/diversityarray/CELfiles.shtml` [54], for a total of 416 samples called in one batch.

Genotypes were called for the Affymetrix Mouse Diversity Genotyping Array (MDGA) using the BRLMM-P algorithm of Affymetrix Power Tools (apt-1.15.0) with default options. Calling with BRLMM-P was done after SNPs previously identified as poorly performing by Affymetrix were removed (584,729 SNPs were called). With respect to the Affymetrix annotation file na32, I removed sites where the annotation file contained unmapped sites and retained only a single instance of duplicate probes. I used the reference sequence to remove sites where neither the A or B allele matched the reference sequence. I also used the reference sequence to convert Affymetrix allele A and B calls to reference and alternate. I further removed 167 SNPs where the reference allele I determined was different from the reference allele used by the MouseDivGeno software, an alternative genotype caller [54]. Calling of array sites using the GATK for wild French mice was done using the UnifiedGenotyper under –genotyping_mode GENOTYPE_GIVEN_ALLELES –output_mode EMIT_ALL_SITES –stand_call_conf 0 –stand_emit_conf 0.

High quality array sites were defined as those which included the following in comparison to sequencing genotypes: at least one heterozygous call or at least one homozygous reference and one homozygous alternate call; no disagreements between array and sequencing results; no more than 1 missing sequencing call; and no more than 3 missing genotyped calls.

Generation of 17 Strains VQSR training sites

The 2011-11-02 callset from the 17 Strains sequencing paper was used [13]. To derive training sets for *M. m. domesticus* and *M. m. castaneus* populations, I selected sites which had alternate genotype calls for the respective wild derived inbred strain (WSB/EiJ and CAST/EiJ) and were of high quality (Above Threshold Genotype of 0 or 1 [13]). A prospective population was then genotyped at those sites using the GATK UnifiedGenotyper using flags –genotyping_mode GENOTYPE_GIVEN_ALLELES –output_mode EMIT_ALL_SITES -stand_call_conf 0

Table 2.1. VCF annotation definitions Descriptions are taken verbatim from VCF headers and can be found at the GATK website
`https://www.broadinstitute.org/gatk/`

Symbol	Definition
DP	Approximate read depth; some reads may have been filtered
Dels	Fraction of Reads Containing Spanning Deletions
HaplotypeScore	Consistency of the site with at most two segregating haplotypes
HRun	Largest Contiguous Homopolymer Run of Variant Allele In Either Direction
MQ	RMS Mapping Quality
MQ0	Total Mapping Quality Zero Reads
MQRankSum	Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities
ReadPosRankSum	Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias
QD	Variant Confidence/Quality by Depth
SB	Strand Bias

-stand_emit_conf 0, and then fixed sites were removed using the GATK SelectVariants with flag -env.

Generating SNP callsets

SNP discovery and genotyping was done using the GATK UnifiedGenotyper, using additional flags –stand_call_conf 0 –stand_emit_conf 0 -A HardyWeinberg –A HomopolymerRun, under SNP mode.

Variant filtration using SelectVariants

SNPs were filtered using the GATK SelectVariants. Annotations used are described in Table 2.1. The choice of parameter settings described here reflects a mixture of recommendations for filtering variant callsets in a paper using 10 *Pan troglodytes* samples sequenced to 10X [55], default GATK recommendations for performing cutoff based variant filtration, as well as manual interrogation of BAM files. SNPs were filtered out if any of HRun > 4, MQ0 > 40, $((\text{MQ0}/(1.0 * \text{DP}))) > 0.40$, QD < 1.0,

MQ < 25.0, Dels > 0.0, HaplotypeScore > 40.0 or SB > 1.0 were true. For the French and Taiwanese mice, SNPs were further filtered if they had DP < 60 or DP > 340 on the autosomes, DP < 45 or DP > 255 for chromosome X, and DP < 15 or DP > 85 for chromosome Y. Similar cutoffs adjusted for the number and depth of coverage of the Indian mice led to DP lower bounds of 90, 67, and 22, while upper bounds were 510, 383, and 128, for the autosomes, X and Y chromosome, respectively.

Variant filtration using the VQSR

Building a model for the VQSR was done using the different training sets and for a given raw callset using the GATK VariantRecalibrator using annotations -an QD -an HaplotypeScore -an MQRankSum -an ReadPosRankSum -an MQ -an DP -an SB -an InbreedingCoeff -an HRun. Annotations are described in Table 2.1. These annotations were chosen based on GATK best practices guidelines when the filtering was performed. Callsets were ultimately generated for a given sensitivity using the GATK ApplyRecalibration. Filtering for the combined dataset did not use the InbreedingCoeff annotation.

Genotype masking

A filtered callset was further filtered at the genotype level after site filtration by setting genotypes to missing if the read depth (RD) was less than 5 or the genotype quality (QC) was less than 10. For male samples on chromosome X, this was relaxed to 2 for read depth. All fixed SNPs were subsequently removed using the GATK SelectVariants with the -env and -ef flags.

Sensitivity estimates using *M. Spretus* derived singletons near *Vkorc1*

This metric of sensitivity is based on the presence of an *M. Spretus* introgressed haplotype in a French mouse. More details of this can be found in Section 3.4.4, but briefly, one of the French *M. m. domesticus* mice (Mouse 34) contained an

introgressed segment from approximately 134 Mbp to 138 Mbp on chromosome 7. Further examination confirmed that both the majority of these singleton calls were also observed in the *M. Spretus* sample and that the mutations found in *Vkorc1* were the same as those seen in Warfarin resistant mice throughout Western Europe [19]. I therefore utilized the recovery of these singleton SNPs in French mouse 34 as a quality control check on the ability to recover singleton heterozygote calls, a particularly challenging type of variant to call.

Specifically, I made use of the 4 wild derived inbred strains from the 17 Strains paper, SPRET/EiJ (*M. spretus*), CAST/EiJ (*M. m. castaneus*), WSB/EiJ (*M. m. domesticus*), and PWK/PhJ (*M. m. musculus*). I intersected SPRET/EiJ specific SNPs from the 17 Strains sequencing paper (bi-allelic with SPRET/EiJ = Alt, CAST/EiJ = Ref, WSB/EiJ=Ref, PWK/PhJ=Ref) with raw calls from the combined callset for the region of chromosome 7, yielding 37,511 SNPs specific to SPRET/EiJ. Of those, again from the combined raw callset, I found 28,731 heterozygote genotype calls for Mouse 34. I defined as a quality check the proportion of those 28,731 SNPs for which Mouse 34 had a heterozygote Ref/Alt genotype call in a given callset.

Comparison against high coverage French mouse

For a callset of interest containing French mice, I intersected their results with the high coverage unfiltered results, removing sites at which the alleles disagreed. I then evaluated concordance in the callset between the mouse sequenced to low and high coverage as the proportion of sites including a non-reference call (Ref/Alt or Alt/Alt) which agreed between low and high coverage, divided by the number of calls in which both the low and high coverage mouse had calls. It should be noted that this procedure likely suffers in that it overestimates confidence in that sites which are called heterozygous or homozygous alternate in the low coverage data cannot be correctly determined if they are homozygous reference in the high coverage data, as such sites are not called in the high coverage sample. This approach does benefit from simplic-

ity, reproducibility between callsets, and not requiring filtering of the high coverage callset, which would be difficult given the VQSR strategy employed.

2.2 Results

2.2.1 Generation of high quality array sites

To determine a high quality set of SNPs polymorphic in both the sequencing and array results, to use as both a quality control metric and to obtain filtered SNPs for use with the VQSR, I used the GATK to genotype the French sequencing data at array sites. Genotypes in this dataset were not filtered for read depth or genotype quality in the comparisons reported here. Call rates for the array at sites genotyped by both sequencing and array were generally good for the 20 French mice (90.56% to 96.77%) except for one sample (Mouse 54, 80.24%), which was subsequently removed from this phase of analysis. When a mouse was genotyped on more than one array, I selected the array with the highest call rate for that mouse. For assessing performance between array and sequencing results, 15 mice were selected to be a training set from which to derive high quality sites, and the rest used to determine accuracy at those sites. I selected test set mice with the 1st, 7th, 13th and 19th highest array call rates, as genotype call rate is indicative of array quality. The mouse sequenced to 40X coverage (Mouse 311) happened to have the highest array call rate (96.77%).

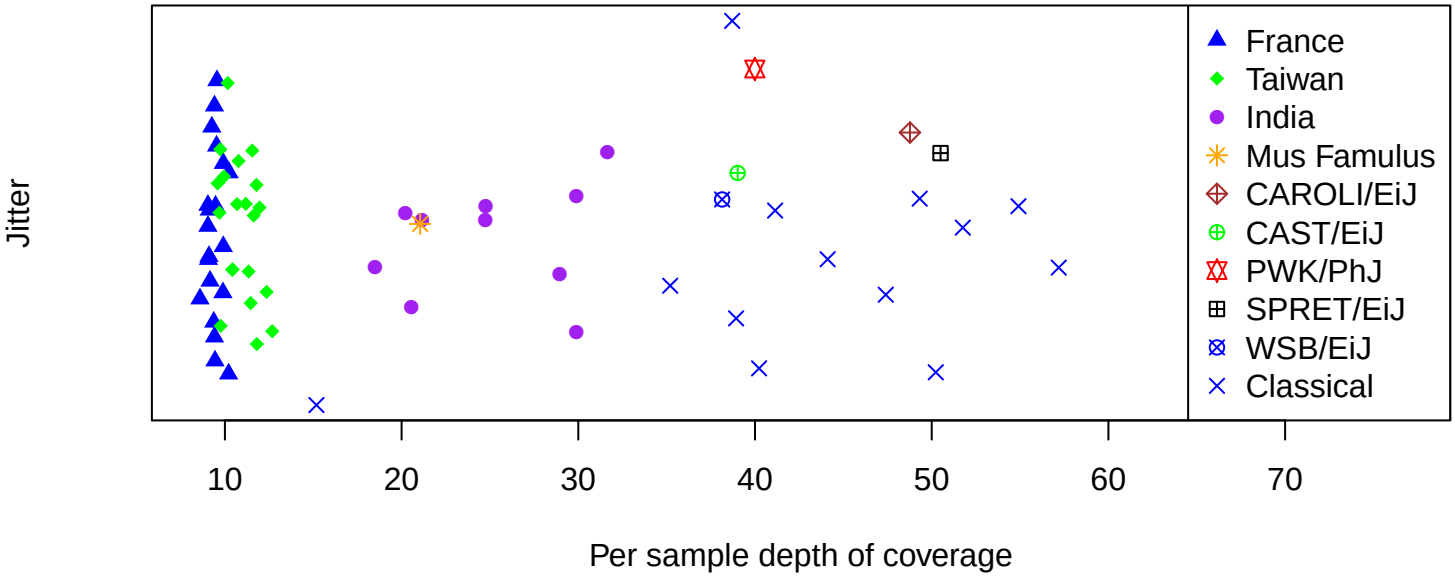
Intersecting the genotypes between the array and sequencing, and keeping only high quality SNPs, yielded 146,106 SNPs. For the 4 test samples, accuracy on these SNPs between sequencing and array was 98.6%, 98.8%, 98.2% and 99.0%, in order of increasing array call rate. For the 183,067 SNPs which failed a QC requirement but were polymorphic, accuracy on the test set was 86.5%, 85.0%, 86.4% and 90.2%, again ordered by array call rate, smallest to largest.

I subsequently employed this list of 146,106 SNPs as a set of high quality filtered sites for the Variant Quality Score Recalibrator for the French *M. m. domesticus*

mice, or used the full set of 621,272 SNPs for all populations of mice.

2.2.2 SNP genotyping and accuracy

Figure 2.5. Sample sequencing depths Plot of sequencing depth for each of the 69 mice in this study



Depth of whole genome sequencing coverage for each of the mice is shown in Figure 2.5. Two sets of variants were called using this data; a single combined callset involving all 69 mice for the comparative analyses, and 3 population-specific callsets for the creation of recombination rate maps. These different callsets were created to allow for the tailoring of tradeoff between accuracy and sensitivity, and to take advantage of population-specific distributions of error metrics for the VQSR in the population-specific callsets, which should allow for more accurate variant filtration.

For each of these raw callsets, I generated a series of filtered datasets using either the GATK SelectVariants, or the GATK Variant Quality Score Recalibrator (VQSR)

Population	VQSR Type	VQSR Sensitivity	Error hets in hom	Error chromosome X	Error against 40X	Error against array	Sensitivity vkorc1	Sensitivity array	Number of SNPs	Ti/Tv
France	Array F	95	0.64	1.97	0.47	0.92	43.33	94.02	12,957,830	2.20
France	Array F	97	0.72	2.28	0.50	0.93	51.32	96.01	14,606,149	2.19
France	**Array F**	**99**	**1.12**	**3.62**	**0.69**	**0.94**	**66.69**	**98.09**	**17,353,264**	**2.16**
France	Array NF	95	2.06	5.82	1.03	0.93	73.82	96.58	18,071,593	2.14
France	Array NF	97	2.97	8.24	1.42	0.93	79.75	98.01	19,369,816	2.10
France	Array NF	99	6.11	15.73	2.50	0.94	86.38	99.20	22,008,978	2.01
France	17 Strains	95	1.29	3.89	0.77	0.93	57.86	93.49	16,805,717	2.14
France	17 Strains	97	2.20	6.52	1.15	0.93	68.94	96.49	18,547,713	2.11
France	17 Strains	99	4.19	11.63	1.90	0.94	81.84	98.62	20,843,679	2.04
France	Hard Filt.		5.36	16.37	1.63	0.91	75.27	96.96	19,805,592	2.06
Taiwan	**Array NF**	**95**	**2.05**	**11.20**					**36,344,063**	**2.12**
Taiwan	Array NF	97	2.87	14.67					39,183,932	2.10
Taiwan	Array NF	99	6.34	25.57					42,864,322	2.05
Taiwan	17 Strains	95	1.83	10.32					29,748,456	2.11
Taiwan	17 Strains	97	2.16	11.20					34,112,325	2.11
Taiwan	17 Strains	99	3.66	15.80					39,549,666	2.08
Taiwan	Hard Filt.		6.11	19.44					33,692,857	2.04
India	**Array NF**	**95**	**1.11**	**1.80**					**66,190,390**	**2.18**
India	Array NF	97	1.59	2.57					71,134,757	2.16
India	Array NF	99	3.70	5.56					78,220,348	2.11
India	17 Strains	95	0.67	1.16					57,674,209	2.18
India	17 Strains	97	1.09	1.63					65,981,654	2.17
India	17 Strains	99	2.63	3.31					75,103,886	2.13
India	Hard Filt.		5.41	72.61					78,487,616	2.10
All	**Array NF**	**95**	**1.90**	**8.95**	**1.11**	**0.93**	**86.16**	**96.74**	**140,765,771**	**2.04**
All	Array NF	97	2.38	13.99	1.35	0.93	89.84	98.20	160,347,803	2.03
All	Array NF	99	4.52	22.73	2.36	0.94	93.28	99.36	184,763,078	1.99

Table 2.2. Quality metrics for callsets Error metrics are described in the main text. For VQSR types, Array F means filtered, NF means not filtered, Hard Filt. means Hard Filtering using the GATK SelectVariants. Error metrics for the combined callset using all populations is only for the French subset. Callsets that are deemed most optimal for their population or the combined set, and were subsequently used in the thesis, are bolded.

using various training sets and sensitivities. For the VQSR-based procedures, two training sets were utilized. The first was based on either the Affymetrix array as just described, either all sites with no filtration (621K SNPs) for any population or the high quality filtered sites (146K) for the French population only. The second training set was obtained for a population by first genotyping SNPs previously defined as variant in the appropriate wild derived inbred strain (WSB/EiJ for French, CAST/EiJ for Taiwan and India), and then retaining those also polymorphic in the population of interest (5.4M for French, 16.6M for Taiwan, 15.9M for India). Results for the VQSR were reported using each of three sensitivities of 99%, 97% and 95%.

Therefore, 10 callsets were generated for the French population, 7 callsets for the Taiwanese population, 7 callsets for the Indian population and 3 callsets for the combined dataset. Table 2.2 summarizes these callsets and looks at their performance assessed using a variety of metrics. I used four different metrics to assess error rate, and two to assess sensitivity, as well as looking at the transition-transversion ratio. For the error metrics, two of these are defined for any population, while two are specific to the French population. Both sensitivity measures are specific to the French population. I now go over these error metrics in more detail, and go on to choose the callsets based on their performance on the error metrics.

The first error metric, "Error hets in hom", leverages the fact that the wild mice contain multi-megabase regions with minimal heterozygote genotype calls, which reflect recent inbreeding. In Section 3.3, I describe this phenomenon in more detail, and outline a hidden Markov model used to partition the genome into inbred and non-inbred regions. The error metric used here is the proportion of genotypes (erroneously) called heterozygous in inbred regions of the genome, divided by the percentage which are heterozygous in outbred regions of the genome. This normalization allows us to infer what percent of heterozygote genotypes in outbred regions of the genomes are falsely called, and as such represents a heterozygote genotyping error

rate estimate. The second error metric, "Error chromosome X", calculates a similar metric, but this time uses only chromosome X, and calculates an error rate for heterozygote genotypes by dividing the proportion of genotype calls for males on chromosome X which are heterozygous by the proportion of genotype calls for females on chromosome X which are heterozygous. The third, "Error against 40X", compares the genotypes of the French mouse sequenced to 40X versus its 10X sequence from the French population. The fourth, "Error against array", compares the sequencing genotypes against the array genotypes at high quality array sites, for the 4 French samples which were not used to generate the list of high quality array sites.

Two metrics were used to assess sensitivity for the callsets. The first, "Sensitivity *Vkorc1*", is as described in the Methods, and is based on recovering private singleton SNPs in a French mouse on an *M. spretus* derived haplotype containing a warfarin resistant form of *Vkorc1*. The second measure of sensitivity, "Sensitivity array", determines the sensitivity against the 146,106 high quality array SNPs.

Values obtained on the metrics as shown in Table 2.2 were generally quite good, and as would be expected, qualities tended to get better as the number of novel SNPs decreased, and vice-versa. The "Error chromosome X" measure, despite being very similar to the "Error hets in hom" measure, was uniformly higher, most likely due to the combined issues of the poorer quality of its assembly and lower sequence coverage. The "Error against 40X" measure was low at no more than a few percent, suggesting that the 10X genotypes were not much worse than 40X genotypes would be, although the 10X sample had much higher levels of missing genotype calls due to masking of genotypes with low read depth or genotype quality (See Table 3.1 for missing data proportions in the combined callset for each population after filtration). Finally, "Error against array" was extremely similar across all filtration strategies, as was the sensitivity against the array as compared to the VQSR sensitivity, most likely as the array sites are in areas of the genome which are highly amenable to high

quality sequencing and are easily recovered with all the filtering strategies. Note that for the French population, the sensitivity for array sites without filtering is not a perfect match for the VQSR sensitivity due to genotype filtering post-VQSR.

Within all populations, as well as across populations, I did see large variation in the metrics calculated. In general, the hard filtering using the GATK VariantFiltration did not perform as well as the VQSR approaches. As well, while there was overlap in performance, in general the 17 Strains based VQSR did not perform as well as the array based VQSR; for example, for the French population, the 95% sensitivity 17 strains based VQSR performed worse than the 99% sensitivity array filtered set, despite the later having slightly more SNPs.

In general, when analyzing the same VQSR training set across populations, I observed that the Indian population gave the best results, and the combined callset the worst. For instance, using the array unfiltered at 95% sensitivity, the Indian Ti/Tv was 2.18 vs 2.14 for the French, 2.12 for the Taiwanese and 2.04 for the combined callset. It should be noted that the lower quality of the combined callset is not spread evenly across all types of variants; for instance, many quality metrics are very similar between the French mice called in the combined callset and the French mice when called on their own; as such, I suspect much of the reduction in the Ti/Tv in the combined callset is due to poor quality calling of the more diverged strains, in particular the *M. famulus* and *M. caroli* samples.

For the recombination rate map construction, low error rates were prioritized. For the French *M. m. domesticus* population, I chose to use the high quality array sites filtered dataset at a sensitivity of 99%. For the Taiwanese *M. m. castaneus* and Indian *M. m. castaneus* callsets, I chose to use the array unfiltered sites at a sensitivity of 95%. For the combined callset, I chose to use the array unfiltered sites at a sensitivity of 95%.

2.3 Discussion

In this Chapter, I began by presenting the sampling and acquisition of mice used in this thesis. I then presented the analysis pipeline in detail, the SNP callsets made, estimates of error and sensitivity on them, and the final choice of a combined callset with 140M SNPs, and 3 population specific callsets with fewer SNPs.

A more detailed analysis of these callsets will follow, but it is worth noting here that this represents an extremely dense and accurate SNP callset. The combined callset, when it was made, integrated all known published whole genome sequencing datasets representing wild and laboratory mice, and includes multiple wild populations, laboratory mice, and wild derived inbred strains. Comparing it to existing datasets in other mammals, it is comparable in both number of variants, number of samples, and sampling diversity to a recent survey of variation in Great Apes, which catalogued 89M SNPs in 79 great apes across 5 species and multiple subspecies [56], as well as variation in wild and domesticated pigs, including a catalogue of 41M SNPs from 111 pigs, either domesticated from China or Europe, or wild [57].

In terms of quality, the metrics generated in Table 2.2 show that the callset is highly accurate, with low false heterozygote error rates of 1-2%, low error against the high quality array genotypes of about 1%, and higher but still manageable error on chromosome X. As for sensitivity, there is excellent sensitivity against array sites, but much weaker for the difficult to call *M. spretus* derived private singletons, although these are better in the combined callset, which includes the SPRET/EiJ sample. In general, it was felt that subsequent analyses, such as recombination rate map construction, would benefit more from a lower error rate than higher sensitivity, so in general the choice was made to use sensitivities for VQSR on the lower end of the range considered here. The choice of 10X sequencing for the French and Taiwanese populations also seems reasonable, as there is only a modest amount of error in comparing the single French mouse at 10X versus 40X, and Ti/Tv ratios between the

Indian and Taiwanese mice showed only a modest decline, yet this meant about 3 times as many mice could be sequenced for the same budget.

Although good, and certainly sufficient for this thesis, the callsets presented here would likely be further improved by recent methodological updates, outlined below, much like they were improved compared to earlier versions thanks to methodological updates. The large number of quality and sensitivity metrics shown in Table 2.2 is a reflection of earlier troubles to generate accurate genotypes from these samples. Originally, callsets using these mice were generated using the hard filtering cutoffs, and downstream work, such as the recombination rate maps, was highly suspect. The implementation of the VQSR led to substantial decreases in error, from a heterozygote error rate of 5-6% down to a more manageable 1-2%, as shown in Table 2.2. Similar further improvements would likely come from using the newer reference genome mm10 versus mm9, which would likely help reduce error rate further by reducing the number of mapping artefact induced false heterozygote genotype calls. Furthermore, SNP calling using the UnifiedGenotyper has generally been superseded by local *de novo* re-assembly methods, including the GATK HaplotypeCaller [48] or Platypus [58]. It is also worth noting that the quality of indel genotypers was not felt to be sufficiently high at the time this thesis was initiated to warrant substantial analysis, although software now routinely used, such as Platypus [58] or a full *de novo* assembly approach like Cortex [59], would likely have made them sufficiently accurate to be analyzed alongside the SNP variation. The same can be said about larger scale structural variants [60, 61].

It is also worth noting that this represents the best attempt to evaluate the error rate of the Affymetrix Mouse Diversity Genotyping array on wild samples. Although generally the array has been used for laboratory mice [49], as it was originally intended, it has been used to genotype wild mice, notably by Staubach *et al.* [18], which used the array to genotype wild mice and search for signals of selection and

introgression, which will be discussed in Chapter 3. These results suggest that naively using the array on wild mice, even following the recommended procedure of genotyping a large number of samples simultaneously to reduce error rates, would lead to an error rate of about 5-10% in *M. m. domesticus* mice, due to off-target variation. Although methodological improvements exist, notably the MouseDivGeno software [54], it is difficult to imagine this being particularly successful for more modest samples sizes. It is worth noting that the French mice were originally genotyped on the arrays not for quality control purposes, but as a fore-runner to whole genome sequencing, as it was thought that reliable recombination rate maps could be generated from the arrays. However, much like VCF filtering using cutoffs, the array data proved too error-prone to generate reliable maps. Using the array to genotype on more diverse mice, such as the Indian mice, would be even more difficult, given the probe based nature of the array, and the higher density of variants in that population. Although the Affymetrix MDGA is not commonly used anymore, alternative succesors such as the MUGA and MegaMUGA arrays (Neogen, Lincoln, NE, USA), would likely suffer the same problems. Although this work suggests that removal of probe sites near known polymorphic variants would lead to lower array error rates, a simple alternative to arrays has emerged in the form of genotype-by-sequencing (GBS) [62]. In this approach, DNA is first digested with restriction enzymes which cleave the genome into long fragments, and then only the ends of those fragments are sequenced. This lead to high coverage of a fraction of the genome. Although coverage is often highly variable, a method like the low coverage imputation algorithm described in Chapter 6 could be used to fill in the blanks. To summarize, this work suggests how bad the Affymetrix MGDA is for genotyping wild mice, casts doubt on previous results on wild mice using the array which don't account for such a high error rate, and suggest how it might be improved, although this likely will not be necessary as subsequent studies of variation in wild mice which do not require genome wide variation would

likely be better served by using GBS as opposed to arrays.

In conclusion, these callsets hint at extraordinary diversity in wild mice, discussed in Chapter 3, and are sufficiently sensitive and accurate to form the basis for the comparisons which will follow in the remainder of the thesis.

Chapter 3

Factors that influence patterns of variation in wild mice

In Chapter 2, I presented a set of 69 wild and laboratory mice that have been whole-genome sequenced, and described SNP callsets either on all mice or specific to a population. Here, I use the callsets to compare patterns of variation both within and between populations, subspecies, and species. The major aims of this chapter are to learn about factors that influence patterns of variation in these mice, and to take advantage of whole-genome sequencing to perform these analyses with greater resolution than previously possible.

The rest of this chapter will be organized as follows. First, the variant calls from Chapter 2 will be examined in more depth to assess how they are distributed both within and between populations. Second, the population histories of these mice will be investigated using a variety of methods, and related to previously estimated genetic histories. Third, I present a novel method for identifying homozygosity due to inbreeding using a hidden Markov model, validate it, describe results in the different wild populations, and analyze how this relates to loss-of-function mutational burden. Finally, I look at the distribution of some commonly used statistics to infer selection in the wild mouse populations, and identify several new examples of strong selective events, including cross-species mammalian introgression.

3.1 Descriptive statistics

3.1.1 Estimating the callable genomes in the populations

While the filtered SNP callsets give us reasonable sensitivity and error rates, they do not give a complete picture about the distribution of these metrics genome-wide. In particular, sensitivity and error can be affected locally by many population-specific features, such as copy number variants private to one population, and a failure to take this into account can lead to false inferences, for example, a population specific reduction in diversity which is actually the result of a copy number variant. Therefore, given the interest in comparative genomics between different mice and populations throughout this thesis, I sought to identify the "callable genome" in the data.

The GATK VQSR, which I used for SNP filtration, employs a two stage filtration procedure. First, using a raw SNP callset and a training set of true positive variation, the model builds a Gaussian mixture model (GMM) which partitions the space spanned by the annotation variables into a "pass" region and a "fail" region, based on the desired sensitivity to retain training sites. Second, using this model, the algorithm labels each SNP in the raw callset as passing or failing at a given sensitivity as to whether it is within the pass region.

Theoretically, depending on the choice of annotations, it should be possible to run all sites genome-wide through the GMM built by the VQSR. However, this is not a provided option, and extracting the constructed GMM for manual reimplementation is not obvious. Furthermore, running all sites is computationally impractical. Therefore, I utilized an approximating approach to estimate callability genome-wide.

Similar to the approach taken in the PSMC paper [63], I chose to assess callability using deviations from expected depth of coverage. I used the GATK DepthOfCoverage to calculate coverage against the autosome, chromosome X and chromosome Y non-N regions for each mouse. Coverage for the samples generated for this study was shown previously in Figure 2.5. I then labeled regions as uncallable where, for a given mouse,

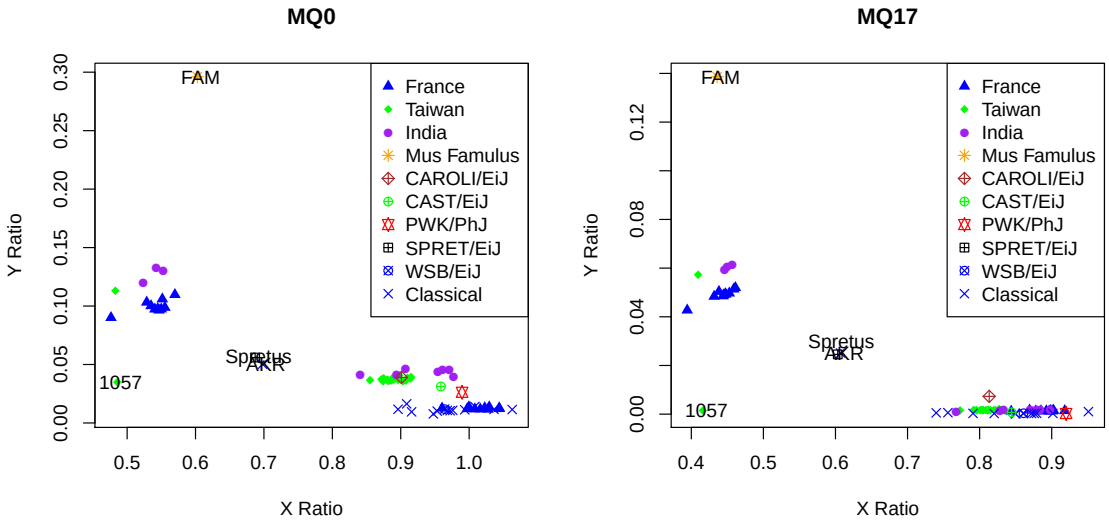
	% callable	% miss pre	% miss post
French	83.39	7.23	12.41
Taiwanese	79.58	5.26	12.58
Indian	84.75	0.63	3.29
13 Classical Strains	87.89	0.61	2.83
WSB/EiJ	90.97	0.39	1.26
CAST/EiJ	88.17	0.87	2.10
PWK/PhJ	88.29	0.94	2.02
SPRET/EiJ	85.14	2.03	3.26
CAROLI/EiJ	79.54	4.16	6.94
M. famulus	76.84	3.83	10.31

Table 3.1. Callable genome percent for different populations Results are for the combined dataset of 140.8 M SNPs. % missing refers to the percent of genotypes missing on average in each mouse in that group either before (pre) or after (post) implementation of the callable genome mask.

the coverage was less than one half or more than three times this value. For males for chromosome X, I divided the cutoffs in half. For a population, I classified bases as uncallable if fewer than 80% of samples were callable at that base. For these analyses, I considered 10 different groups - French, Taiwanese, Indian, classical (13 classical laboratory strains from the 17 Strains set), and the 6 wild inbred lines (WSB/EiJ, CAST/EiJ, PWK/PhJ, SPRET/EiJ, the *M. famulus* sample, and CAROLI/EiJ). The fraction of the genome of each population that is callable, as well as subsequent per-mouse missing data rates, are shown in Table 3.1 for each of the 10 groups. Depending on the analysis, non-missing genotypes in uncallable regions in a given population were masked out.

Following coverage calculations, I compared the average coverage on chromosomes X and Y to the autosomal coverage to evaluate gender assignment. Results are shown in Figure 3.1. Mouse 1057 from Taiwan has a ratio of X-chromosome reads that mimics males in the sample but has no signal from chromosome Y, making it XO and phenotypically female. Two samples (SPRET/EiJ (Spretus) and AKR) from the 17 Strains dataset showed values inconsistent with classification as either a male or female sample. Breaking the results down by sequencing library confirmed a

Figure 3.1. Ratio of sex chromosome coverage to autosomes Axes show average of sex chromosome coverage to autosomal coverage, with horizontal representing chromosome X and vertical chromosome Y. Left panel shows using all reads with mapping quality of 0 or above, right panel is for reads with mapping quality of 17 or above.



clear separation between those samples that were male in origin and those that were female (data not shown). Conversations with Thomas Keane, the lead author of the 17 Strains paper, indicated that additional data from these two strains was generated at a later date, and it is likely that sample mixup occurred at that point. Thankfully, as these mice are completely inbred, the impact is very minor genome-wide.

3.1.2 Ancestral allele call

Many analyses require knowledge of the ancestral allele. To do this, I made use of the wild-derived inbred strains within the dataset. As we saw in the Introduction, previous analyses indicate that among the mice seen here, the three *M. musculus* subspecies are the most closely related, with *M. Spretus* the next most related, followed by *M. famulus* and then *M. caroli* [2, 64]. This was confirmed in our data, as shown for example in Figure 3.9 or Figure 3.10.

I therefore chose to implement a fairly simple ancestral allele call based on the *M. caroli* and *M. famulus* samples. I did not additionally use *M. spretus* as DNA of *M. spretus* origin is known to segregate within the French *M. m. domesticus* population; for example, a recent multi-megabase introgressed segment of *M. spretus* DNA as shown previously by Song *et al.* [19], and as will be shown in Section 3.4. For the ancestral allele call, I first set *M. caroli* and *M. famulus* to missing if they had a heterozygote genotype, then took their genotype when they agreed, and had no ancestral call if they disagreed. If one sample had a missing genotype, I used the call from the other.

3.1.3 Comparison of variation in different populations

	Number	Average Coverage	% of Genome Callable	Number of Polymorphic SNPs	Number of Fixed Derived SNPs	$\hat{N}_e$	$\hat{\theta}_k$ per Kb	$\hat{\theta}_\pi$ per Kb
French	20	9.4	83.4	13,560,355	7,688,747	99,174	1.5	1.6
Taiwanese	20	10.9	79.6	24,527,196	3,762,291	187,970	2.8	3.2
Indian	10	25.0	84.7	51,682,589	929,828	449,893	6.8	5.8
Classical	13	43.4	87.9	10,060,567	9,731,978	78,162	1.2	1.4

Table 3.2. Summary of variation in different populations Descriptive statistics for the 3 wild population and lab strains on the combined callset with 140.8M SNPs. The estimate of $\hat{N}_e$ using Watterson's estimator is adjusted for the callable genome, as is $\hat{\theta}_k$ and $\hat{\theta}_\pi$. Note that these values are calculated without requiring an ancestral allele.

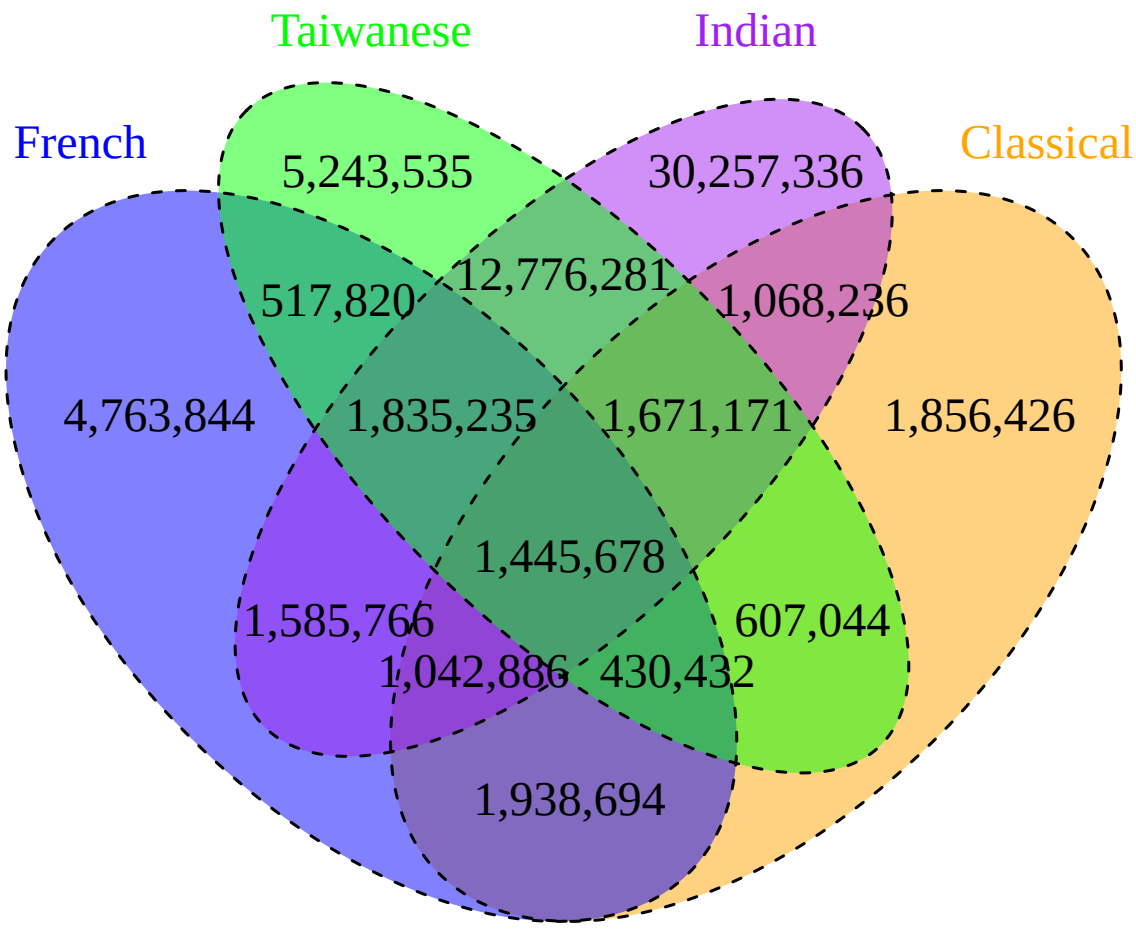
Table 3.2 shows a basic summary of the variation found in each of the 3 wild populations, as well as the classical laboratory strains. In terms of raw numbers of

SNPs, in decreasing order, the Indian mice are the most variable, followed by the Taiwanese, French and finally the classical laboratory strains. The converse is seen for the number of fixed derived variants. I also made a preliminary estimate of N_e using Watterson's estimator, where I estimate the population mutation rate $\theta = 4N_e\mu$ using $\hat{\theta_k} = \frac{s}{\sum_{i=1}^{n} \frac{1}{i}}$, where n is the number of haplotypes sequenced and s is the number of polymorphic variants, and I assume $\mu = 3.79 \times 10^{-9}$/gen/bp from [3], based on rat-mouse divergence. Using this estimate, I see a 4.5 fold variation in size from the smallest population, the French mice, to the largest one, the Indian mice, whose N_e is estimated at $450,000$ by this method. I see qualitatively similar results with $\hat{\theta_\pi}$, defined in Equation 3.14. In terms of the estimates of θ, $\hat{\theta_k}$ and $\hat{\theta_\pi}$, given that differences between these are the numerator for Tajima's D, I can look at the sign for basic evidence of departures for uniform histories, and see a slightly positive value for the Indian mice, while both the Taiwanese mice and French mice have a slightly negative difference, consistent with historic bottlenecks having affected these groups. I will return to Tajima's D in Section 3.4, and a more detailed examination of the population history and relatedness in these mice in Section 3.2.

Table 3.3 shows how variation is shared pairwise between the populations. The classical mice, which are predominantly *M. m. domesticus* in origin, share only about half of their segregating variation with the French mice, although a larger proportion of their fixed derived variants are shared between the two populations. Comparing the Taiwanese and Indian mice, most of the Taiwanese variation is also segregating in the Indian mice, while most Taiwanese fixed derived variants are also still segregating in the Indian mice. In general, across all population comparisons, the Indian population shares quite a lot of segregating variation with all other populations, for example sharing more polymorphic sites with the classical and French groups, than they do with each other.

To more explicitly compare levels of segregating variation between populations, for

Figure 3.2. Shared polymorphic variation between species Given a variant is segregating in at least one population and is not missing in any population (68.2 M SNPs), given are the counts of variants segregating in each possible combination of populations. Areas are not proportional to counts.



	Missing	Poly	FixedA	FixedD	Unknown	Sum
Missing	955,825	1,212,692	3,735,080	1,066,679	569,140	7,539,416
Poly	170,157	4,857,690	5,362,039	2,178,679	991,790	13,560,355
FixedA	279,277	2,401,937	32,465,237	67,074	0	35,213,525
FixedD	82,009	1,155,753	31,439	6,419,546	0	7,688,747
Unknown	43,478	432,495	0	0	3,793,690	4,269,663
Sum	1,530,746	10,060,567	41,593,795	9,731,978	5,354,620	68,271,706

(a) French & Classical

	Missing	Poly	FixedA	FixedD	Unknown	Sum
Missing	783,093	5,623,825	903,432	87,979	141,087	7,539,416
Poly	450,722	5,909,565	6,012,595	412,601	774,872	13,560,355
FixedA	316,201	30,184,384	4,579,675	133,265	0	35,213,525
FixedD	169,181	6,441,557	782,026	295,983	0	7,688,747
Unknown	101,033	3,523,258	0	0	645,372	4,269,663
Sum	1,820,230	51,682,589	12,277,728	929,828	1,561,331	68,271,706

(b) French & Indian

	Missing	Poly	FixedA	FixedD	Unknown	Sum
Missing	2,339,321	2,235,071	2,267,994	385,818	311,212	7,539,416
Poly	1,392,711	4,229,165	6,232,385	830,218	875,876	13,560,355
FixedA	2,661,971	12,303,305	19,655,211	593,038	0	35,213,525
FixedD	690,073	3,962,110	1,083,347	1,953,217	0	7,688,747
Unknown	386,819	1,797,545	0	0	2,085,299	4,269,663
Sum	7,470,895	24,527,196	29,238,937	3,762,291	3,272,387	68,271,706

(c) French & Taiwanese

	Missing	Poly	FixedA	FixedD	Unknown	Sum
Missing	941,729	1,285,351	3,653,312	1,027,531	562,972	7,470,895
Poly	284,179	4,154,325	13,375,936	4,661,394	2,051,362	24,527,19
FixedA	224,411	3,549,170	23,860,150	1,605,206	0	29,238,93
FixedD	44,756	575,291	704,397	2,437,847	0	3,762,291
Unknown	35,671	496,430	0	0	2,740,286	3,272,387
Sum	1,530,746	10,060,567	41,593,795	9,731,978	5,354,620	68,271,70

(d) Taiwanese & Classical

	Missing	Poly	FixedA	FixedD	Unknown	Sum
Missing	763,166	5,665,754	843,778	71,278	126,919	7,470,895
Poly	791,296	17,728,365	4,982,037	419,194	606,304	24,527,196
FixedA	160,010	22,647,842	6,428,918	2,167	0	29,238,937
FixedD	56,132	3,245,975	22,995	437,189	0	3,762,291
Unknown	49,626	2,394,653	0	0	828,108	3,272,387
Sum	1,820,230	51,682,589	12,277,728	929,828	1,561,331	68,271,706

(e) Taiwanese & Indian

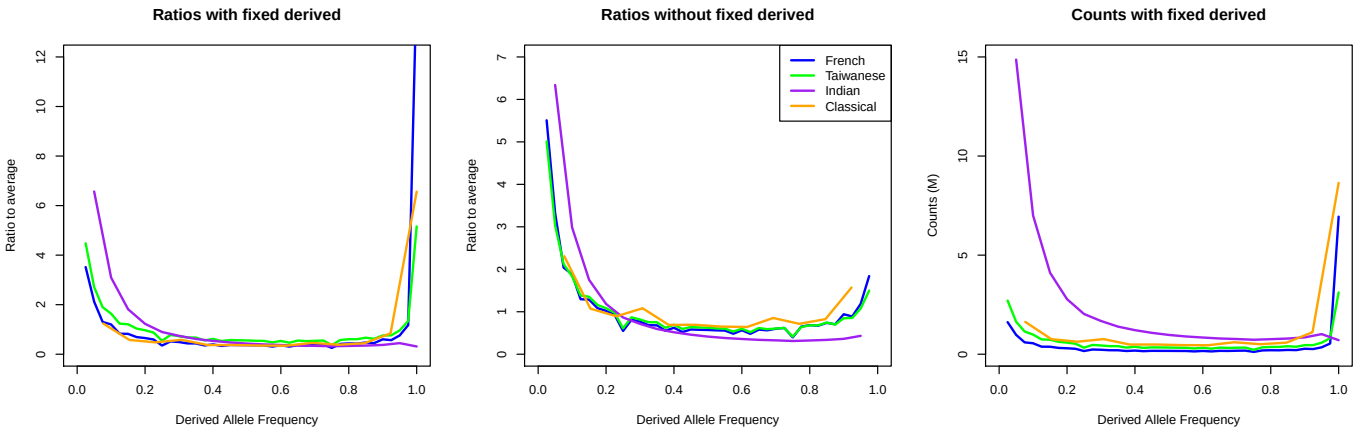
	Missing	Poly	FixedA	FixedD	Unknown	Sum
Missing	462,637	484,712	468,569	252,633	151,679	1,820,230
Poly	915,763	5,227,971	33,760,753	7,668,284	4,109,818	51,682,589
FixedA	120,415	3,645,229	7,194,374	1,317,710	0	12,277,728
FixedD	12,313	254,065	170,099	493,351	0	929,828
Unknown	19,618	448,590	0	0	1,093,123	1,561,331
Sum	1,530,746	10,060,567	41,593,795	9,731,978	5,354,620	68,271,706

(f) Indian & Classical

Table 3.3. Comparison of pairwise variation Subtables give pairwise comparison of variation between populations from the combined SNP callset. The population listed first is associated with the row names, while the population listed second is associated with the column names. Only SNPs where there exist some difference between the populations, not including missing data, are included. Columns are: Missing, at least 25% missing data; Poly, polymorphic; FixedA, fixed for the ancestral allele; FixedD, fixed for the derived allele; Unknown, fixed but unknown ancestral allele call; Sum, sum across row or column.

SNPs for which an allele frequency was available for all populations, *i.e.* in a callable region with a sufficient number of mice with genotypes, one can study a Venn diagram (Figure 3.2). Somewhat interestingly, all possible combinations are represented by at least a few hundred thousand variants, with the least common combinations involving sharing between groups from excluding India, or between distinct subspecies, such as the Taiwanese and either French or Classical. In terms of private variation, the Indian population has the most, at 30.2 M out of the total 68.2 M SNPs, while the French and Taiwanese possess similar numbers, though France is a much larger proportion of its total, and the Classical population the least.

Figure 3.3. Derived allele frequency spectra in the populations For a given population, I binned SNP frequencies relative to the number of sampled haplotypes, and took the relative count for each frequency, either including (left plot) or excluding them (center plot). The right plot shows the counts for the different populations.



The distributions of allele frequency spectra across the populations, making use of the ancestral allele call, are shown in Figure 3.3. The classical, French and Taiwanese frequency spectra are dominated by fixed derived variants, while as we saw earlier in the tables, the Indian mice have few. Once the fixed derived variants are removed, the frequency spectra become more similar between the populations. However, the French, Taiwanese and classical populations all show strong increases in the number of variants at high frequency compared to more moderate frequencies. This is consistent with migration into those populations.

Next, for SNPs which are segregating in at least one population and have an ancestral allele call, I compared joint allele frequencies between populations by contrasting the observed sharing against what would be expected based on joint allele frequencies being a product of marginal allele frequencies (Figure 3.4). We see that low frequency variants in the French are less likely to be seen in Taiwan, implying diversity which has arisen in the French mice since the two source populations split. Similarly, low-frequency variation in the French and Indian mice is shared less between populations. The Taiwanese and Indian mice show a stronger correlation of

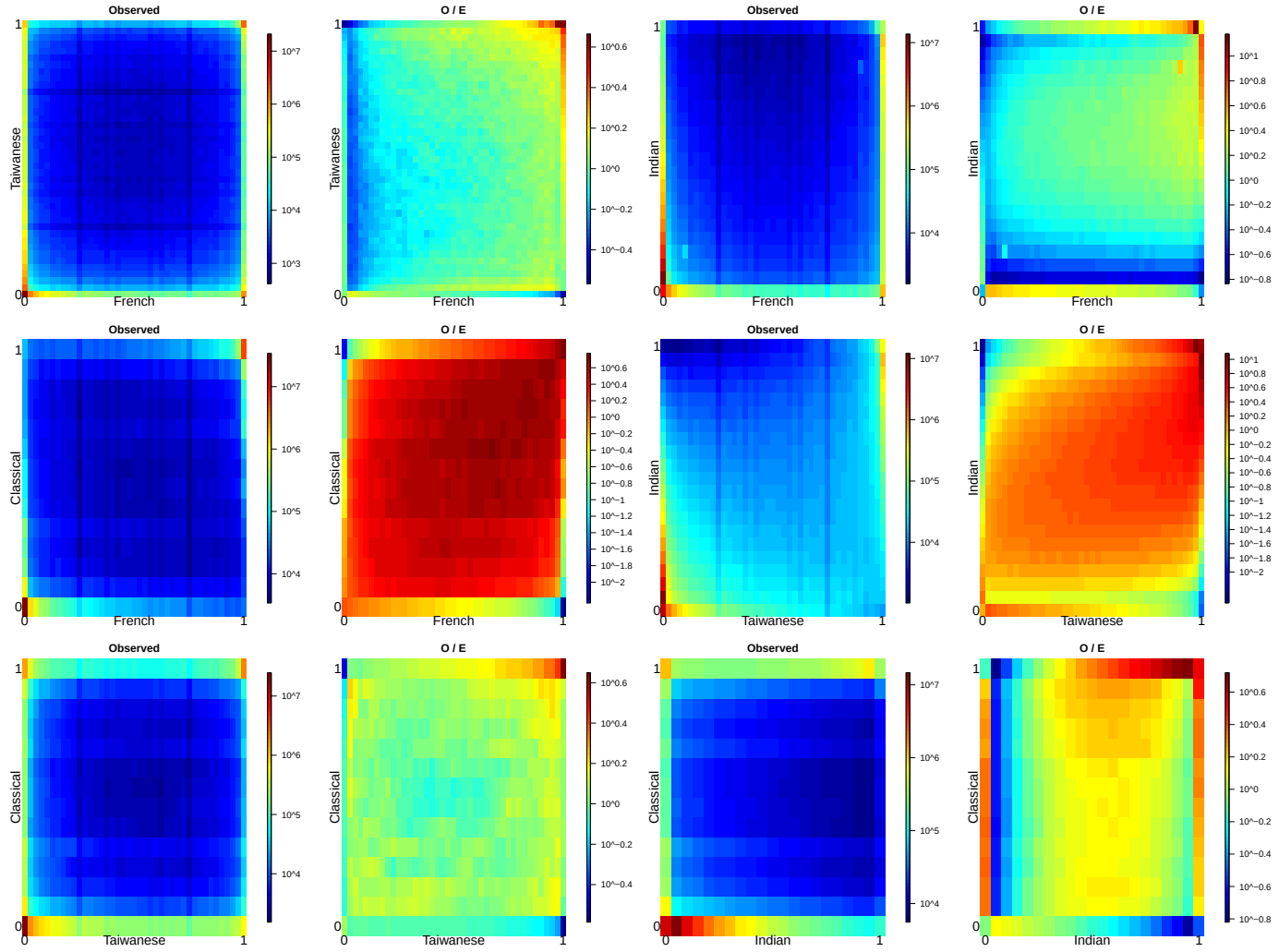


Figure 3.4. Comparison of pairwise allele frequencies Each pair of populations is represented by two plots: Observed, with the actual counts; and O/E, where the observed counts are divided by the product of the two marginal frequencies of the two populations. The populations are indicated by the values in the X and Y axes. Each axis shows the derived allele frequency spectra for variants in that population. Plots are heat maps, with the legend on the right of each subfigure corresponding to the number of SNPs or the ratio in each cell. Note that bars around 25% and 75% for French and Taiwanese correspond to artefacts due to rounding and high missingness.

allele frequencies, typical of a less drifted pair of populations. As for the Classical population, they share most variation with French *M. m. domesticus* mice. However, as expected, they share almost no relationship between frequencies of segregating variants with the other populations, consistent with a strong founding bottleneck, except that rare (and hence recent) Indian variation is more likely to be absent in the Classical strains.

3.1.4 Discussion

In the initial examination of variation in the different populations, some features are clear. The Indian mice have by far the most variation, consistent with the region they were sampled from representing a source population for *Mus musculus* [8, 9]. Using simple estimates, they appear to have a population size in the hundreds of thousands, an order of magnitude larger than human population sizes up until the recent past [63].

These results also begin to highlight the benefits of comparative genomics, with the comparison of multiple populations each with unique aspects to its history. The derived allele frequency spectra of the Indian mice shown in Figure 3.3 is unlike that of the French, Taiwanese and classical populations. Since, under neutrality and panmixia, the number of variants at a given frequency is a decreasing function of their frequency, then the observation that there are more variants at high compared to moderate frequencies in these populations implies the action of other forces, such as selection or migration. Although the French and Taiwanese results could in theory be due to ancestral allele miscalling, the consistently decreasing Indian allele frequency curve for the same SNPs allows this to be rejected as a major factor. As for the classical population, their even flatter curve is consistent with strong bottlenecking and drift, which is not surprising given the artificial structure of this population.

3.2 Demography and population structure

3.2.1 Introduction

The size of a population is rarely static through time. Populations expand and contract, depending on their relationships with other species, the local environment, and their ability to adapt and evolve. Populations are also often structured, for either geographical or sociological reasons, and there may be migration between different populations. Each of these factors will influence patterns of genetic variation.

Characterizing the history of a population serves two purposes. First, especially for humans, it is a topic for study in its own right. A major focus of genetic analysis is to understand our own history as a species. Examples of these types of studies include those that shed light on human migratory events out of Africa [65], as well as the more recent colonization of the Americas [66]. Aside from understanding the events of history, a second reason to learn about population histories is to help frame expectations of genetic variation under neutrality. For example, knowledge of population sizes and diversity sheds light on whether burdens of putatively functional mutations carried in different populations might be in line with expectation [67], while knowledge of the variance in genetic diversity genome-wide could help identify selective events through identification of patterns of genetic variation outside the variance expected through the known history of a population [18].

In other species, efforts to study demographic processes have been no less fruitful. In mice, targeted sequencing of many mice over small genomic regions revealed a rich history, as was shown in the Introduction. Recently, several species have seen their histories updated through the sequencing of multiple wild animals from diverse geographic locations [68–70].

Methods for analyzing the history of populations have evolved as the techniques available to study them have changed. Sometimes, non-genetic avenues are available for study, including written and oral histories, censuses, and fossil records. However,

these sources of data are not always available, for instance with mice, and have limited range and utility.

I therefore analyzed the history of mice through detailed analyses of their patterns of genetic variation. What follows is a brief description of methods commonly used to analyze population histories, as well as a brief overview of the methods used here. I then give results for applying these methods to the data, discuss them, relate them to previous works, and discuss what future studies might seek to focus on.

3.2.2 Methods

One important consideration when considering programs to infer history is to distinguish between those that use genotype or haplotype data. Many recent methods to investigate population structure or history do work specifically on haplotypes and not genotypes. Chromopainter and fineSTRUCTURE are two programs that work together, where the former seeks to create a matrix of relatedness between individuals based on how often they copy haplotypic segments from each other, while the later uses input from the former to infer population structure [71]. Results from this method are highly informative, and are more sensitive than those based on non-haplotype based genome-wide summaries [72]. Other methods also use haplotypes, for instance to estimate recent population sizes [73].

However, haplotypes can be difficult to infer. Early projects such as HapMap were specifically designed around trio data to facilitate the generation of extremely accurate haplotypes from genotype data [74]. Modern approaches to phasing genotype data generate accurate results, but do so using thousands of samples at a time [75]. They can be improved with the inclusion of related samples, even if the relationship is not explicitly laid out [76]. Despite the small sample size and lack of related samples, I briefly investigated whether I could accurately phase the mice using ShapeIt2 with read-aware phasing [77]. Informal evaluations of the ability to phase accurately were

performed, by making "pseudo-diploid" haplotypes using mice from the French and Taiwanese populations. To do this, I used the inbreeding results from Section 3.3, and made new BAM files by combining half the reads from pairs of different mice in the same population with long overlapping stretches of homozygosity, calling variants, filtering them to the known callset, and phasing them using ShapeIt2. I then compared phasing results for the pseudo-diploid mouse against the known haplotypes. This gave estimates of a switch error rate on the order of 25 kbp in the data, clearly too short for phase based inference.

Therefore, I focused analysis on methods which did not require phased haplotypes, but could instead use the diploid genotype data. In this section, I apply a variety of genome-wide methods to investigate the histories of the mice sequenced here. These include: principal components analysis (PCA), a commonly used model-free approach (see, for example, [78]); ADMIXTURE , a model based but label-free method [79]; PHYLIP, another label-free method to create a neighbour joining tree [80]; TreeMix, a method to assess drift between labeled populations with migration [81]; and the pairwise sequentially Markovian coalescent (PSMC), a method to estimate population size changes using individual genome sequences [63].

PCA

Principal components analysis is a commonly used statistical data reduction technique. Given some set of data with L variables and N subjects, the goal of principal components analysis is to find a transformation of the data such that one obtains new representations of the variables as principal components, which successively explain as much of the remaining variation in the data as possible. Consider a genotype matrix G with entries $G_{j,i} \in \{0, 1, 2\}$ for bi-allelic SNP j for sample i. One first normalizes the matrix G using $f_j = \frac{1}{2N} \sum_{l=1}^{N} G_{j,l}$ to get $X_{j,i} = G_{j,i} - f_j$. Further normalization can be done by dividing by $f_j(1-f_j)$, discussed in [78] (I did not do this for the mice). One then calculates a matrix $M = \frac{1}{L} X^{\mathrm{T}} X$ and perform resulting eigenanalysis on the

matrix M. The resulting eigenvectors of the matrix M are the principal components, and often capture aspects of existing population structure [82], because structure has a genome-wide impact on variability. In some cases, results have a direct connection to the underlying genealogy [78]. I ran PCA using R, either using all samples at once or separately for the different wild populations.

ADMIXTURE

ADMIXTURE is a model based procedure to estimate genome-wide admixture proportions [79]. This model makes a strong assumption: that present day samples are admixed from K ancestral populations, with observable data (the genotypes), hidden data (sample admixture proportions) and parameters (allele frequencies for the SNPs in each ancestral population). More specifically, individual i in the model would be represented by a hidden vector of ancestry $q_{i,k}$, for each of the K populations, with $\sum_{k=1}^{K} q_{i,k} = 1$ and $0 \leq q_{i,k} \; \forall \; k$. Further, the probability of the genotype at each SNP j for sample i is calculated based on the frequencies of SNP j, $f_{k,j}$, in each of the k ancestral populations. The model is fit to the observed data for a given value of K using expectation maximization, with admixture proportions for each sample being the desired output. This relatively simple formulation of the problem belies substantial complexity in actually finding an optimal solution using EM, given the large parameter space involved, and the interrelatedness of the variables in the likelihood.

For the implementation of ADMIXTURE, I thinned the data to lessen the impact of SNPs in linkage disequilibrium, a procedure recommended by the authors. Finally, I ran ADMIXTURE for K from 2 to 7, and plotted the resulting ancestry proportions for each sample.

Neighbour joining

The neighbour joining method constructs trees to parsimoniously relate samples based on a distance matrix. Here, I use an allele-sharing distance, where for sample i at

SNP l with $g_{l,i} \in \{0, 1, 2\}$, the distance metric between samples i and j is $d(i, j) = \sum_{i=1}^{L} |g_{l,i} - g_{l,j}|$. This is further normalized by missingness in the two samples. Now, for a brief conceptual overview of how trees are built using this method, consider a tree as a connected acyclic graph, with one vertex for each sample, and a series of internal vertices. Consider the simplest tree, one with a single internal vertex X, and an edge leading to all sample vertices s_i for $i = 1, ..., n$. The length of the edge from a sample vertex to the internal vertex is half the average distance between that sample and all other samples. The total length of the graph would be $S_o = \sum_{i<j} d(i, j) = \sum_{i=1}^{N} L_{i,X}$, where $L_{i,X}$ is the length of the edge from i to X. If one considers adding another internal node, with some samples s_x connecting to internal vertex X and some samples s_Y to internal vertex Y, with internal vertices X and Y connected by a new edge, then this edge gets length L_{XY} equal to the average distance between vertices s_x and s_y, subtracting out the distance from s_x to X and s_y to Y. In this fashion, adding internal nodes shrinks the total branch length of the constructed tree. Tree construction continues in a fashion, outlined in the paper [80], so as to minimize the sum of edge lengths on the tree, and continues until the graph has no vertices of degree greater than 3.

In constructing the neighbour joining trees, I used all autosomal SNPs, using code written in R, combined with Rphylip [83], an R wrapper to PHYLIP [80], a method which implements the construction of neighbour joining trees.

TreeMix

TreeMix is a method designed to infer tree structure and identify migratory events using SNP data on pre-identified populations [81]. In this method, given a population B that recently descended from population A, a SNP at frequency x_A in population A, and some measure of drift between population A and B as $c_{A,B}$, then the frequency in population B is modeled using $X_B \sim x_A + \epsilon_B$, where $\epsilon_B \sim N(0, c_{A,B} x_A (1 - x_A))$. Allele frequencies in successive populations can be modeled in a similar fashion, and

the relationships between populations can be modeled using an acyclic graph. Migrations between populations are added in by adding additional edges to the graph, such that modeling allele frequency in a population is driven by $1-w$ times its source population plus w times the migratory population. In this fashion, one could sample allele frequencies for a given graph structure and drift weights. The observed allele frequencies are therefore the observed data, and one can construct a likelihood for a graph structure encompassing migration and a set of weights, and optimize accordingly. For a fit model, the observed sample covariances can be compared to the expected sample covariances, and from these residuals, standard errors formed. These are particularly useful for inferring whether additional migratory events might be warranted.

I ran TreeMix [81] using autosomal SNPs with between 0 and 10 migratory events allowed, using a block size of 1000, and the predefined population labels.

PSMC

The pairwise sequentially Markovian coalescent (PSMC) by Li and Durbin is a method to reconstruct ancestral population sizes using only sequence information from a single diploid sample [63]. The method models present day genotypes of a diploid sample (the observed data) based on the underlying time to most recent common ancestor (TMRCA) of the sample's two haplotypes (hidden data), as well as the ancestral population size, the population mutation rate and the population recombination rate (parameters). To simplify calculations and increase robustness, the observed data is a list of windows across the genome of a given width (default 100bp), and whether there is at least one site difference between the two haplotypes, or if the region is uncallable. Under default conditions, windows are marked uncallable if at least 90 bases are uncallable in a window, with individual bases declared uncallable if the average sequencing depth at that base is greater than twice or less than one third the genome-wide average. PSMC uses EM to learn the parameters of the model,

and uses user supplied values for the per-generation per-base pair mutation rate, as well as the generation time in years, to allow for the interpretation of the ancestral population sizes in terms of individuals and years. This makes the output readily interpretable, although distinct demographic processes, such as increased population size or migration, are not distinguishable under the model [63].

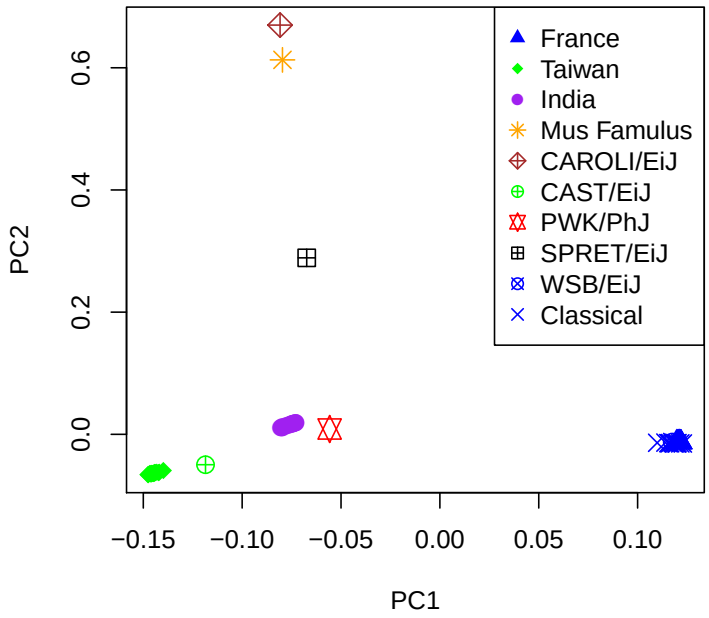
I ran PSMC for each of the wild mice, separately, using the GATK generated genotypes, with filtering to remove uncallable regions in each mouse, to reflect PSMC user guidelines. I used a generation time of 0.5 years and 3.79×10^{-9}/bp/gen, as used by Staubach *et al.* and originally estimated using rat-mouse divergence [3].

3.2.3 Results

The first interrogation of the wild mice was using principal components analysis on the entire dataset, with results shown in Figure 3.5. These results clearly show that the principal components generally cluster mice by ascertained subspecies and species, and no mice were obviously erroneously classified. The first principal component separates the predominantly *M. m. domesticus* mice, composed of the French mice, the classical strains, and WSB/EiJ, from an *M. m. castaneus* grouping of the Taiwanese mice and CAST/EiJ, while the Indian mice cluster separately in between these two groups. The second principal component separates out the outgroups from the *Mus musculus* samples.

To look for evidence of further structure within the populations, I repeated PCA on each of the wild populations separately (Figure 3.6). Repeating the analysis by using either even or odd chromosome numbers qualitatively preserved relationships among mice, implying non-random structure (not shown). To gain further insight, I overlayed the PCA results for the French and Taiwanese mice with sampling location, as previously shown in Figure 2.2 and Figure 2.3. Unfortunately sampling locations for the Indian mice were not available. Results are shown in Figure 3.7. For the

Figure 3.5. PCA results for all mice



French mice, there is a clear, if moderate, association: the first principal component is broadly associated with a north-west to south-east sampling gradient, while the second principal component is associated with a north-east to south-west gradient. A similar pattern is seem in the Taiwanese mice, with mice sampled more closely together in physical space being more closely related along the first two principal components.

I next ran ADMIXTURE on the full 69 mice (Figure 3.8). Results here, as for the PCA, generally lead to clustering along the originally ascertained species and subspecies designation. At higher values of K, while the French, Taiwanese, Indian and classical populations are all represented by their own ancestral population, there is no clear evidence of admixture, and no K value identifies admixture in either France or Taiwan.

Figure 3.6. PCA results for wild mouse populations

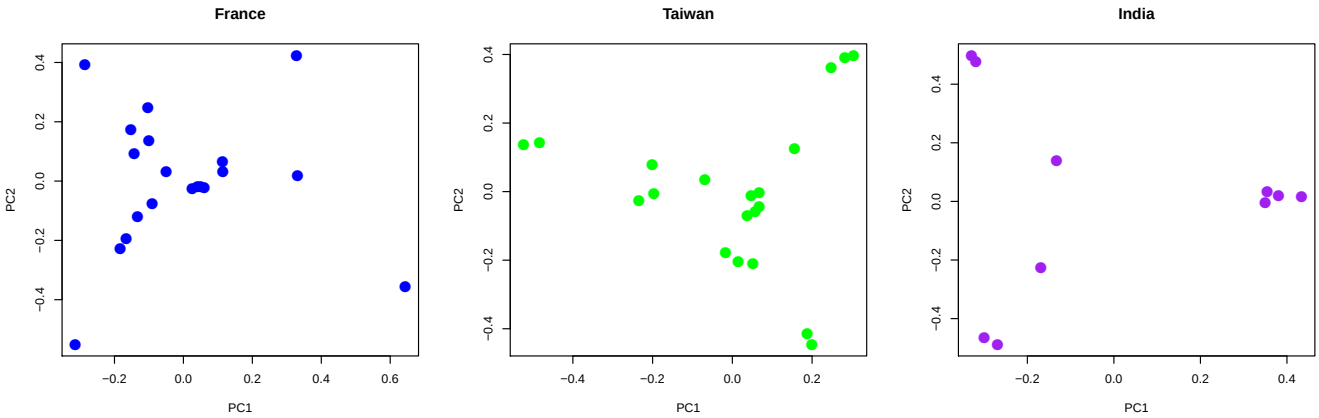


Figure 3.7. PCA results for French and Taiwanese overlayed with sampling location Left plot shows French mice, while right plot shows Taiwanese mice. Principal component vector values for the mice are shown as either colours (PC1) or sizes (PC2) for each sampled mouse.

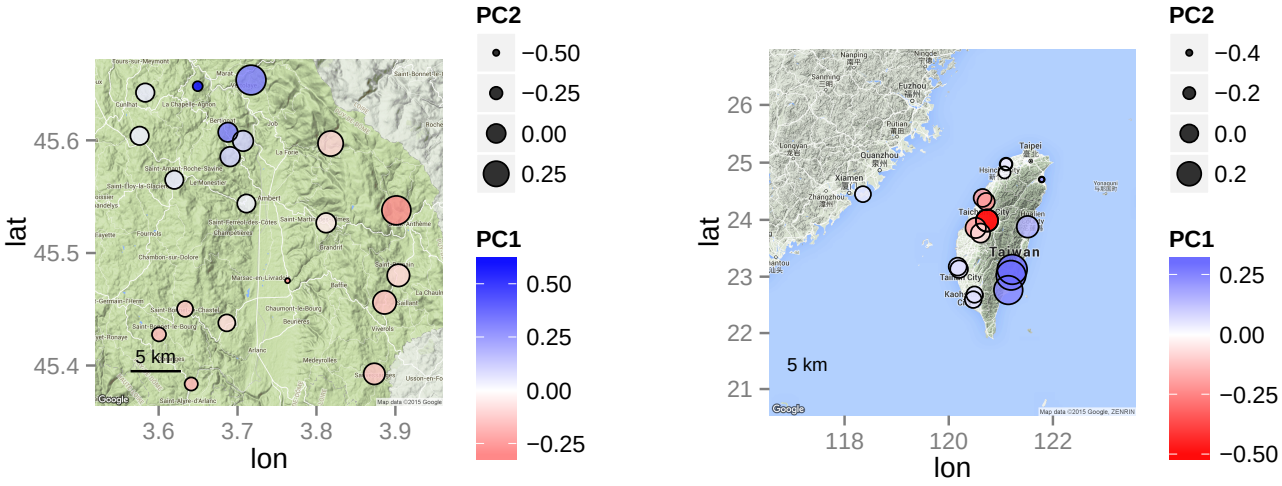
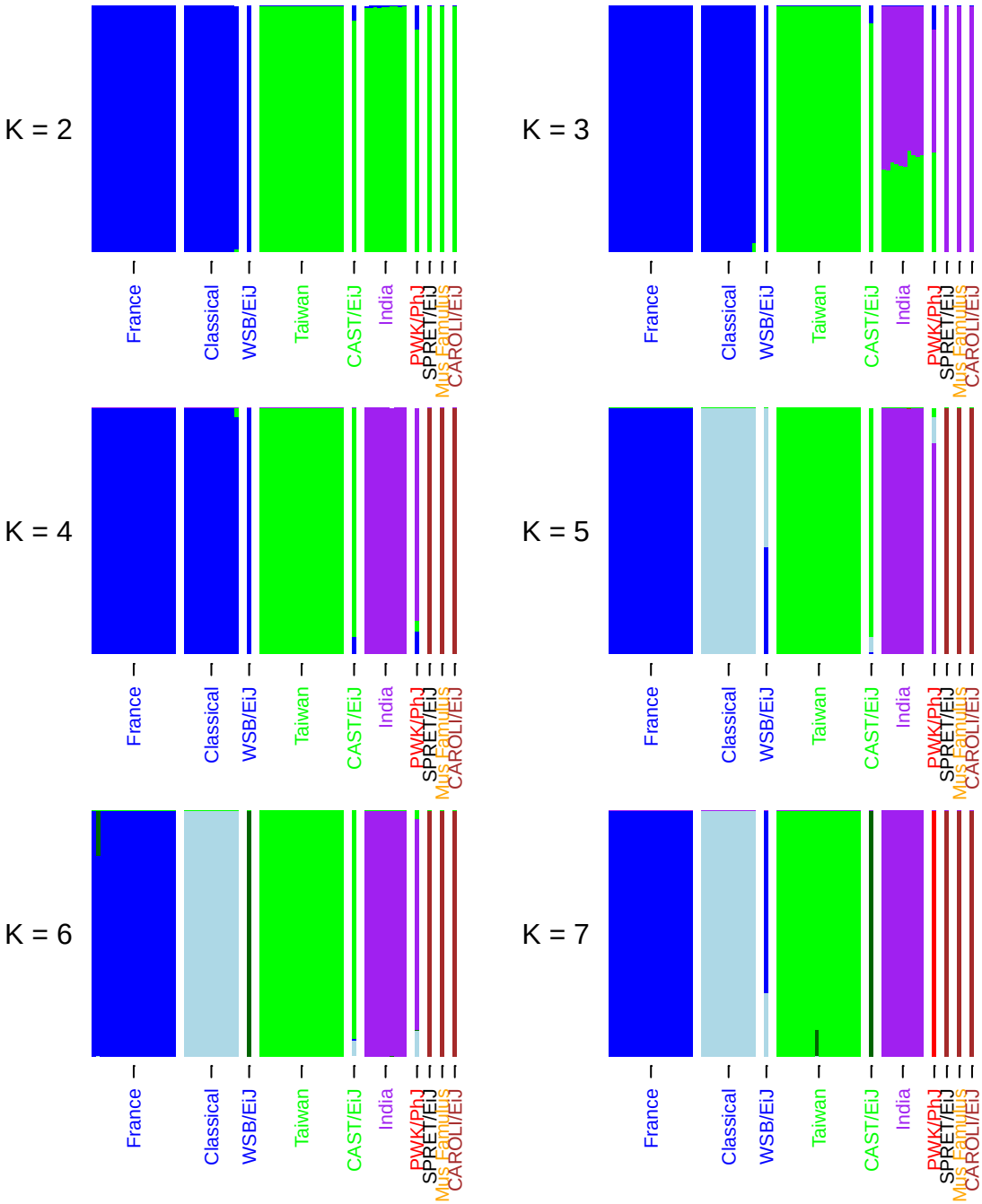


Figure 3.8. ADMIXTURE results for all mice Results are shown using LD-thinned input on genome-wide data for a range of K. Labels underneath represent conventional population designation used in this text, with colours representing different subspecies or species, with *M. m. domesticus* in blue, and *M. m. castaneus* in green but India labelled specifically with purple. Colours for different values of K have been chosen to try and replicate this colour scheme as much as possible.



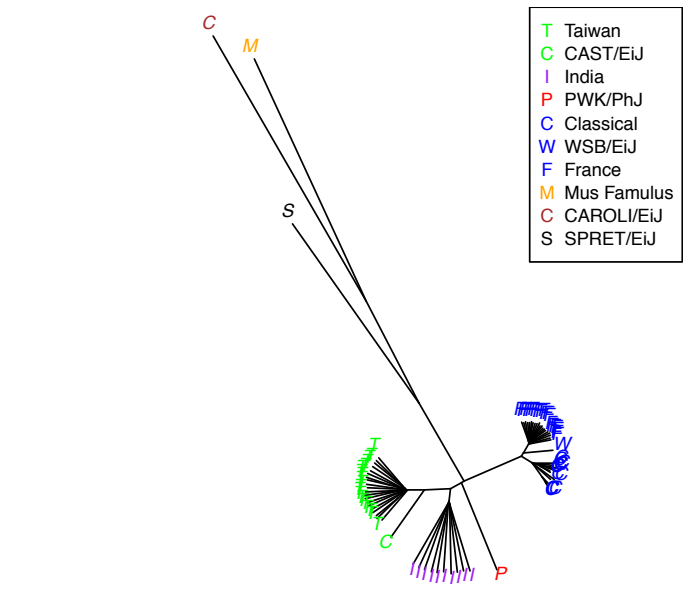


Figure 3.9. Neighbour joining tree Shown are all samples, with labels given in the legend.

Figure 3.10. TreeMix on wild mice TreeMix is run on labelled populations, with 5 added migration events, with the strength of migration given by the colour legend.

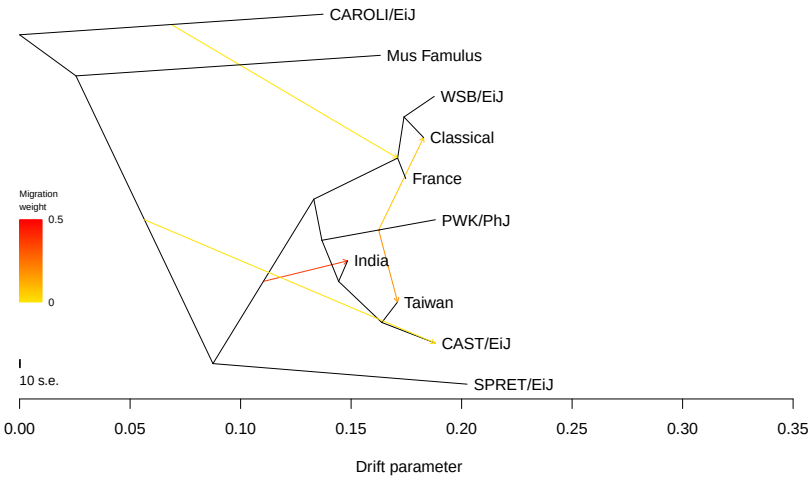
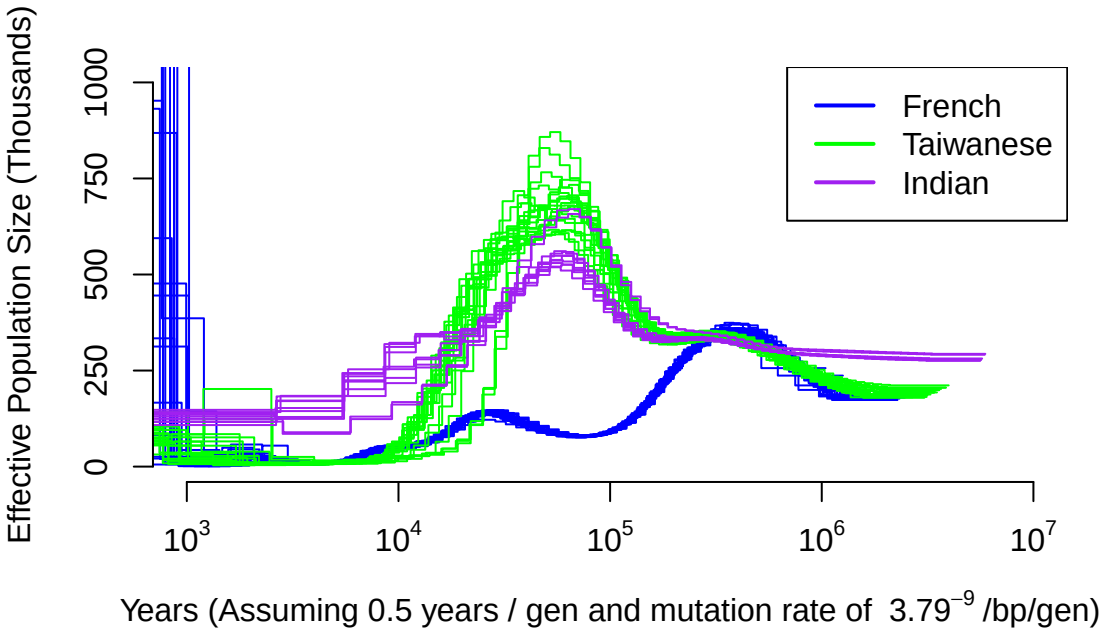


Figure 3.11. PSMC for wild populations



A neighbour-joining tree reveals expected relationships, with long separations, clustering with subspecies, classical strains closely clustering with French mice, and branch lengths for each wild sample consistent with the Ne estimates (Figure 3.9). TreeMix results (Figure 3.10), grouping the mice by labeled population, are shown for 5 migration events. This is a subjective choice (no formal guidelines are given in [81]) that achieves a reasonable reduction in the maximum standard error of the model (from 62.7 with 0 migratory events, to 3.9), while still allowing for ease of interpretability. Overall tree topology is consistent with earlier results, for example the neighbour joining results. For migration, the strongest migration signal in seen from the ancestral *M. musculus* into the Indian population. Next, the PWK/PhJ (*M. m. musculus*) sample has migration into both the Taiwanese and classical laboratory strains. Finally, I see events from the ancestor of *M. spretus* and *M. musculus* into CAST/EiJ, as well as, surprisingly, *M. caroli* into the *M. m. domesticus* mice.

Finally, to more explicitly interrogate population size changes, I ran PSMC on each of the wild mice. Results are shown in Figure 3.11. Here, I see stark departures from a constant population size in all populations. The Indian mice are the most uniform, with population sizes fluctuating between approximately 125,000 and 600,000, and generally lying near to the 450,000 from Watterson's estimator in Table 3.2. By contrast, the Taiwanese mice show an extremely strong bottleneck, with the average minimum population size seen in the mice dropping from $> 500,000$ to about 7,000 about 5,000 years ago, while the French mice see a reduction to about 5,000 over the same time period.

3.2.4 Discussion

The methods presented here give a fairly consistent picture of relatedness among the sampled mice. The label-free methods of PCA, ADMIXTURE and neighbour-joining all give similar results, with PCA being slightly more informative in that it identified within population structure. TreeMix, but not ADMIXTURE, identified admixture events between the species, notably into Taiwan, and PSMC revealed that the two commensal populations sampled here experienced extreme bottlenecks in their recent histories.

These results are in broad agreement with what is already known, as was seen in the Introduction, and confirm that nothing unusual occurred with the sampling of the mice. The most novel result is likely the strength of the bottlenecks in the derived populations. Due to the inability of PSMC to resolve sharp changes in demography [63], as would occur during extreme founding scenarios, the reduction in population size in the French and Taiwanese populations may be more extreme than was presented here. There remain, however, at least two areas in which either further analysis of these mice, or newly acquired data, would shed light on unresolved issues: a joint model of the split times, migration levels, and population sizes of the three

populations; and a more thorough understanding of the extent of introgression in these subspecies.

Methods exist to jointly model multiple demographic parameters of a population, of which the dadi approach of Gutenkunst *et al.* bears mention [65]. In using this method, one first constructs a likely demographic scenario with unknown parameters for one of more populations, and then uses the joint allele frequency spectra between those populations to determine the parameters that best fit the data. For example, one could model two populations A and B as having split some time T_1 in the past, with population A experiencing a bottleneck at time T_2. Using dadi one would obtain estimates for present day population sizes, T_1, T_2 and the extent of the bottleneck. One could further compare different demographic scenarios by comparing the fit of the different models. I tried to use this approach on the three wild populations, but was unable to differentiate between different classes of scenarios with qualitatively different fits: between longer split times and larger ancestral effective population sizes with little migration; versus shorter split times and smaller ancestral effective population sizes with more migration. While dadi is able is able to resolve such questions on more recently separated populations, such as humans [65], new methodology would likely need to be developed to perform the same inference on the wild mouse populations studied here.

As for introgression, despite several suggestive results, its full extent remains to be characterized in these populations. As under neutrality under the coalescent, the number of variants decreases as the allele frequency increases, the sharply elevated increases in high frequency derived variants seen in Figure 3.3 in the French and Taiwanese populations suggest high levels of migration impacting those populations. Some known examples exist, for instance large multi-megabase introgression jointly between *M. m. domesticus* and *M. m. musculus* [18], and I see some evidence of it in the data from TreeMix, as well as the *M. spretus* into *M. m. domesticus* example

which follows in Section 3.4. I spent considerable time early in my DPhil trying to develop a hidden Markov model in the spirit of HapMix [42] to detect and characterize introgression in the wild populations using the four wild derived inbred strains as references; however, given the large ancient N_e of the populations of mice seen here, I was unable to confidently distinguish all but the largest introgressive episodes from incomplete lineage sorting.

To better characterize migration and introgression in mice, I would suggest more diverse sampling is needed. Depending on price considerations, one could use the Indian population to devise an array, or for a less biased approach, one could use genotype by sequencing (GBS) approach [62] to sequence a fraction of the genome at modest coverage for a reasonable cost, for example 5% of the genome at 10X for approximately 100 GBP a sample. One could sample a large number of populations, say 50 populations with 20 mice each, and run something like HapMix [42] or a successor algorithm to identify and characterize recent introgression. Having such a large pool of mice would also make identifying sources of introgression easier, something I struggled to do with the smaller diversity of mice sampled here.

In conclusion, I saw broad agreement with the histories and relatedness of the mice with what was previously known based on more limited sampling information, although PSMC more clearly resolves what Watterson's estimator cannot. I also see novel fine-scale population structure in mice, and previously uncharacterized and distinct introgression into each mouse subspecies, but given the large historical population sizes involved, am unable to more rigorously categorize it. Future studies with more sampling of mice may be informative here.

3.3 Inbreeding

3.3.1 Introduction

In most species, inbreeding is assumed to be a generally deleterious phenomenon, in that more inbred organisms will be less fit than similarly outbred organisms. Two ways inbreeding affects the fitness of an organism are through the exposure of recessive mutations with negative fitness consequences, or alternatively, through increased homozygosity for alleles at loci where heterozygosity is an advantage [84]. Understanding how inbreeding is distributed in different species and populations may shed light on the importance of these different phenomena on fitness.

In humans, inbreeding can have devastating phenotypic consequences through the exposure of recessive lethal mutations, famously including royal families such as the Habsburgs of Austria [85]. Segments of homozygosity from older inbreeding or from lower population diversity have more subtle effects on human phenotypes, including on height and cognition [86]. In contrast to humans, many other species, in particular domesticated animals, are tolerant of much higher levels of inbreeding, perhaps because most deleterious mutations have already been selected out by a long period of managed breeding by humans. For example, on average 23% of the genome of a domesticated pig is homozygous, with similar levels of homozygosity observed within wild animals [87]. Studies of mice in the wild have found similarly high levels of inbreeding, for instance a study of *M. m. domesticus* mice from Arizona yielded an inbreeding coefficient in that population of about 0.20 [88]. In mice, tolerance of high levels of inbreeding is quite obvious, as many fully inbred wild derived inbred strains have been bred [47], and most classical laboratory strains are nearly or fully inbred [13].

However, certain questions remain. The population of Arizona mice was genotyped on an array with a few thousand markers [88]. Other work in mice has similarly sampled only a handful of microsatellite or SNP genotypes [41]. In such cases, it can

be difficult or impossible to discern reductions in heterozygosity due to long-term founding bottlenecks, and past versus recent inbreeding, due to a sparsity of markers. In addition, more precise knowledge of homozygosity segment lengths could inform our understanding of the dynamics of mouse breeding structure and how homozygosity arises in the wild. Finally, and most importantly, our whole-genome sequencing allows us to study the relationship between loss-of-function gene variants, which are exposed in homozygous stretches, with levels of inbreeding or population diversity in mouse populations, for the first time.

I therefore analyzed the genomes of the wild mice for the presence of inbreeding. To do so, I developed a hidden Markov model, that classifies the genomes of each mouse into regions of homozygosity and heterozygosity. In designing such a method, I was principally interested in longer regions of homozygosity which are unambiguously due to recent mating, as opposed to shorter segments, which might relate either to local inbreeding or population structure (In addition to the motivating factors previously interested, robust knowledge of where the mice are homozygous allows for the creation of quality control metrics to evaluate SNP callsets, as was seen in Chapter 2). Furthermore, we were interested in knowing where the mice were homozygous in constructing recombination rate maps, since recombination events captured on homozygous haplotypes are effectively counted twice. This will be explained fully in Chapter 4, but briefly, we merged SNP genotypes from mice with overlapping homozygous segments to create pseudo-heterozygous mice, when constructing recombination rate maps.

In this section, I first lay out a novel method to detect homozygous segments. Next, I validate the method by assessing its sensitivity and specificity. I then describe the homozygosity results in the different populations, compare them with loss-of-function mutations, and discuss the results and their implications.

3.3.2 Methods

Here, I start by borrowing terminology from the locus level and refer to a segment of DNA as being homozygous if it comprised of two haplotypes that have not recombined over their length since their most recent common ancestor, and heterozygous otherwise. One method to detect such homozygous regions, implemented in the program PLINK [89], is to use a sliding window approach, in which one counts either the number or proportion of heterozygous versus homozygous genotype calls over a fixed number of SNPs or base pairs, and calls the region homozygous if there are sufficiently few heterozygote genotypes [86]. Such a method is certainly appealingly simple.

I considered such a method but decided against using it, for the following reasons. Studies that typically use such methods are based on array genotypes, which, particularly for humans, are incredibly accurate, with error rates well below 1%. This is because array data typically interrogates only "well-behaved" regions of the genome, free from common copy number variable or duplicated regions. Arrays are also usually run on a large number of individuals, and poorly performing array sites can be evaluated through metrics such as deviation from Hardy-Weinberg equilibrium and removed.

By contrast, especially in the lower coverage French and Taiwanese mice, I estimate a heterozygote genotype error of about 1-2%, and notably, this is not distributed evenly genome-wide, likely due to mapping and other such artefacts. As such, due to the higher error rate, and suspected variation in error rate genome-wide, I chose to employ a hidden Markov model for the determination of homozygous or heterozygous regions, with parameter updating using expectation maximization. A hidden Markov model was an appropriate choice of model due to the physical linkage properties of DNA, *i.e.* that information about whether a given window of DNA is homozygous or not depends only on the information in that window and whether the previous

window was homozygous.

For reasons of robustness and practicality, I consider homozygosity and heterozygosity to take place over intervals of a fixed size, for example $500,000$ base pairs. An observation is defined to be the number of heterozygote genotypes observed in that interval. The underlying premise is that many heterozygous genotypes should be observed in the heterozygous state, while the homozygous state should see only a very small number of heterozygote genotypes (errors or *de novo* mutations). One should thus be able to define regions of homozygosity as extended stretches with low number of heterozygote genotypes.

Here, I outline the model by describing how one would simulate data under the model given knowledge of the parameters. In the following subsection, I more formally lay out how I generate hidden state probabilities. Finally, I describe how the parameters would be updated.

Sampling under the model

The hidden Markov model (HMM) has $K = 2$ states, and let $k = 1$ refer to where the mouse is homozygous and $k = 2$ refer to where the mouse is heterozygous. Consider an individual chromosome, and let it be broken up into T windows of equal physical width. At the broad scale required, recombination rates in mice vary at most modestly (see for example [90], and Chapter 4), so transition rates are modelled homogeneously genome-wide.

To start, consider modelling homozygous and heterozygous chunks on a chromosome. To do so, start by choosing which of the two states the chromosome starts in, governed by initial state probabilities π_k. Next, for each window chromosome wide, the probability of transitioning between states k_t in window t to k_{t+1} in window $t + 1$ is given by probability $A_{k_t,k_{t+1}}$. Finally, for each window along the genome, conditional on whether the two chromosomes are homozygous or not, the number of heterozygous genotypes observed O_t is modelled by a negative binomial distribution

with parameters p_k and r_k. I did not choose a Poisson distribution for the observations, seemingly the more natural choice based on an infinite sites distribution, as variation in histories in the different windows genome-wide were so great that there would need to be a different Poisson rate for each window. The negative binomial distribution, which can be considered a Poisson distribution with random Gamma distributed mean, accomplishes this requirement, and offered a much better fit to the observed data.

HMM formulation

Now for a more explicit set of definitions. I build on the notational framework introduced in Section 1.5.2 for dealing with HMMs.

Recall that there are two hidden states, where $k = 1$ refers to the homozygous state and $k = 2$ the heterozygous state. Let the chromosome be broken up into T windows of fixed physical width, and let $Q = (q_1, q_2, ..., q_T)$ be a path of hidden states through the windows of the chromosome. Let O_t be the number of heterozygote genotypes in window t.

As mentioned above, I use parameters for initial state probabilities of π_k, where $\pi_1 + \pi_2 = 1$ and $0 \leq \pi_k \leq 1$ for $k = 1$, $k = 2$ so that

$$P(q_1 = k) = \pi_k \tag{3.1}$$

For transitions, I use a homogeneous rate matrix, $A_{k_t,k_{t+1}}$, with $\sum_{j=1}^{2} A_{k,j} = 1$ and $0 \leq A_{k,j} \; \forall \; k, j$, such that the transition probability from state k_t at window t to state k_{t+1} at window $t + 1$ is

$$P(q_{t+1} = k_{t+1} | q_t = k_t) = A_{k_t,k_{t+1}} \tag{3.2}$$

For the emissions, I use parameters r_k and p_k to parameterize the negative binomial distribution of each state. Let $b_k(O_t)$ be the probability mass function for the negative

binomial distribution for the two k, and I get that

$$P(O_t|q_t = k) = b_k(O_t) = \binom{O_t + r_k - 1}{O_t}(1 - p_k)^{r_k} p_k^{O_t} \tag{3.3}$$

Let $O = \{O_t|t = 1, ..., T\}$ be the set of observations for a given sample for a chromosome, let $Q = (q_1, q_2, ..., q_T)$ be a hidden path, and let $\lambda = \{\pi, A, p, r\}$ be the parameters of the model. This gives the following joint probability for the observations and hidden path

$$P(O, Q|\lambda) = \pi_{q_1} \prod_{t=1}^{T-1} A_{q_t, q_{t+1}} \prod_{t=1}^{T} b_k(O_t) \tag{3.4}$$

Taken together, the equations in this section define the probabilities necessary for me to run the forward and backward algorithms, as well as calculate Viterbi paths, as laid out in Section 1.5.2.

Parameter initialization

Manual inspection of the data for the French and Taiwanese suggested substantially higher proportions of the genome as being inbred than for the Indian mice. I therefore initialized $\pi_1 = 0.2$ for the French and Taiwanese, and $\pi_1 = 0.02$ for the Indian mice. For the transition matrix, I initialized the probability of a transition either from homozygous to heterozygous or vice-versa as $A_{1,2} = A_{2,1} = 0.01$. For the emissions, I took the observations genome-wide (*i.e.* O_t across all chromosomes and subjects), and sorted them. I then took the smallest $\pi_k \times 100\%$ of the observations, obtained maximum likelihood parameter estimates for a negative binomial model, and set this to be p_1 and r_1. I similarly used the last $(1 - \pi_k) \times 100\%$ of the observations to estimate parameters for the heterozygous portion of the genome.

Likelihood and updating

As was introduced in Section 1.5.1, expectation maximization is used to determine the parameters of the model that maximize the likelihood. Recall that a "complete

dataset" D includes the observed data (O, the heterozygote genotype counts), as well as the hidden parameters (Q, the hidden states). Given a set of parameters λ, the log-likelihood of the complete data is $L(\lambda) = \log(l(\lambda|D)) = \log(l(\lambda|D = (O, Q)))$. Given a current set of parameters λ^i, a new set of parameters λ^{i+1} is generated to maximize the expectation of $l(\lambda^{i+1})$ with respect to the distribution of hidden parameters obtained by λ^i, $U(\lambda^{i+1}, \lambda^i) = \sum_Q P(Q|O, \lambda^i) \log(P(O, Q|\lambda^{i+1}))$. The purpose of the EM algorithm is that by choosing λ^{i+1} that maximize $U(\lambda^{i+1}, \lambda^i)$, the likelihood of λ^{i+1} over the observed data, $l(\lambda^{i+1}|O)$, also increases. A formal proof of this was given in Section 1.5.1.

In applying the EM algorithm, initialization is first done with a set of parameters λ^0, as described above. For each subsequent iteration $i = 1, 2, ...$, I alternated performing an "Expectation" phase, where $U(\lambda^{i+1}, \lambda^i)$ is calculated, and a "Maximization" phase, where λ^{i+1} is chosen to maximize $U(\lambda^{i+1}, \lambda^i)$ by taking derivatives of $U(\lambda^{i+1}, \lambda^i)$. In the Expectation phase, the crucial component is calculating the state probabilities $P(Q|O, \lambda^i)$ - these are calculated using the forward and backward algorithms. To calculate the derivatives and updates in the Maximization stage, let us first consider the following.

Suppose we have a set of hidden states across all samples. Then we can count the number of times a sample is found in a given state. Let n_k^1 be the number of samples that start in state k, $n_{k_t,k_{t+1}}^t$ be the number that switch between states k_t and k_{t+1} between windows t and $t+1$, and let $m_{k,b}^t$ be the number of samples in state k that emit b heterogyzous SNPs in window t. Let b range from 1 to B. We can further calculate $n_{k_t,k_{t+1}} = \sum_{t=1}^{T-1} n_{k_t,k_{t+1}}^t$ and $m_{k,b} = \sum_{t=1}^{T} m_{k,b}^t$. Furthermore, consider expanded notation to reflect sample labelling. Let $q_{n,t}$ be the hidden state for sample n in window t, let $O_{n,t}$ be the number of observed heterozygous genotypes for sample n in window t, let O_n be the vector of observations for sample n, and let O be the

matrix of observations across all samples. Then the complete data log-likelihood is

$$\begin{aligned} l(\lambda) &= \log(P(O, Q|\lambda)) \\ &= \sum_{k=1}^{K} n_k^1 \log(\pi_k) \\ &+ \sum_{i=1}^{K}\sum_{j=1}^{K} n_{i,j} \log(A_{i,j}) \\ &+ \sum_{k=1}^{K}\sum_{b=1}^{B} m_{k,b} \log(b_k(b)) \end{aligned} \tag{3.5}$$

Calculating updates for a parameter is done by taking the derivative of $U(\lambda^{i+1}, \lambda^i)$ with respect to that parameter, setting it equal to 0 and solving. For the state probabilities, this yields

$$\pi_k^{i+1} = \frac{\mathbb{E}_{\lambda^i}[n_k^1]}{N} \tag{3.6}$$

where $\mathbb{E}_{\lambda^i}[n_k^1] = \mathbb{E}_{\lambda^i}[n_k^1|O, \lambda^i] = \sum_{n=1}^{N} P(q_{n,1} = k|O_n, \lambda^i)$. Next, for the transition matrices,

$$A_{l,j}^{i+1} = \frac{n_{l,j}}{n_{l,1} + n_{l,2}} \tag{3.7}$$

where $n_{l,j} = \sum_{t=1}^{T-1}\sum_{n=1}^{N} P(q_{n,t+1} = j|q_{n,t} = l, O, \lambda^i)$.

Finally, the negative binomial parameters are more complicated. Recall the negative binomial formulation is $b_k(b) = \binom{b+r_k-1}{b}(1-p_k)^{r_k}p_k^b$, as previously shown in Equation 3.3. Then the final summation of Equation 3.5 can be re-written to give

$$\sum_{k=1}^{K}\sum_{b=1} m_{k,b}\left[\log(\binom{b+r_k-1}{b}(1-p_k)^{r_k}p_k^b)\right] \tag{3.8}$$

$$\begin{aligned} &= \sum_{k=1}^{K}\sum_{b=1} m_{k,b}\Big[\log(\Gamma(b+r_k)) - \log(b!) - \log(\Gamma(r_k)) \\ &+ r_k\log(1-p_k) + b\log(p_k)\Big] \end{aligned} \tag{3.9}$$

Then solving for p_k by setting the derivative to 0 gives

$$\begin{aligned} 0 = \frac{dl}{dp_k} &= \sum_{b=1} m_{b,k}\left[\frac{r_k}{1-p_k}(-1) + \frac{b}{p_k}\right] \\ \implies p_k &= \frac{\sum_b b m_{k,b}}{\sum_b b m_{k,b} + r_k\sum_b b} \end{aligned} \tag{3.10}$$

with $m_{k,b} = \sum_{n=1}^{N} \sum_{t=1}^{T} P(q_{n,t} = k | O_{n,1}, ..., O_{n,t} = b, ..., O_{n,T})$. Doing the same thing for r_k gives

$$0 = \frac{dl}{dr_k} = \sum_b m_{k,b} \Big[\psi(b + r_k) - \psi(r_k) + \log(1 - p_k) \Big] \tag{3.11}$$

where $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$. I used a numerical optimization technique to calculate r_k (note that this meant r_k need not be an integer, which works as the binomial coefficient can be re-written using the gamma function for non-integer r_k).

Finally, after a set number of iterations (10), which in practice was sufficient to ensure convergence as measured by a lack of increase in the log likelihood of the data, the parameters from the model were taken as optimized, and the most likely hidden state for each sample was taken as the Viterbi path for each sample. Note that although the notation used here is for a single chromosome in many samples, we ran this on all chromosomes (19 of different length) for a population of mice at once.

Filtering

For uninformative regions of the genome with fewer than 500 SNPs in a window, separately per population, the probability of an emission was set to 1 for each of the $K = 2$ states. As such, a state call is still made, based on information from surrounding regions.

Sensitivity and specificity

I used simulations to assess the ability to identify both heterozygous and homozygous regions. Simply put, as input for the simulations, I made use of very long 10+ Mbp inbred regions that overlapped between different mice within the same population. I used this to determine if I would call these regions outbred, as intended, or would call them inbred. Similarly, I would make pseudo-outbred regions but leave in regions which were homozygous, and determine the ability to recover these regions.

I first ran the full algorithm, including both the EM algorithm to determine parameters, and a Viterbi algorithm to call the most likely state for each mouse, for a window size of 1 Mbp, and created a list of overlaps between pairs of mice of at least 10 Mbp. I then took overlaps between mice in which one overlap could be contained within the overlap of the other mouse or were exactly the same size, so that I could "spike" the larger homozygous region into the smaller one and create a continuous heterozygous run in the mouse with the smaller overlap, whereby outside the overlap region there should exist heterozygosity on either side of the pseudo-heterozygosity in the overlap region.

To assess a measure akin to specificity using this procedure, I re-ran the Viterbi algorithm (per chromosome) using the parameters from the final iteration, where the overlap was completely pseudo-heterozygous, and obtained the number and distribution of homozygous calls made within the overlap region. Then, for a range of insert sizes (500 kbp, 1 Mbp, 2 Mbp, 3 Mbp, 4 Mbp and 5 Mbp), I determined the number of observed calls (real and false) made of that size (running the algorithm on the entire dataset) versus those made in the pseudo-heterozygous overlap regions (false only), correcting for the difference in size of genome interrogated. I then took, as the measure, a ratio of the difference between the observed and the pseudo-heterozygous calls divided by the observed calls, for each of the different insert sizes. Note that this was felt to be the best measure of specificity, since both formal truth data is missing, and the assignments of homozygosity and heterozygosity are not strictly binary due to the desire to stratify them by different lengths.

To assess sensitivity, for a range of insert sizes, for each overlap, I created a pseudo-homozygous region within the pseudo-heterozygous region by randomly choosing a subset of the pseudo-heterozygous region to revert to the original homozygous genotype of that mouse. Successful recovery of a homozygous segment, and hence sensitivity, was defined by whether or not there existed a homozygous call within the

homozygous insert region and whether it overlapped with the insert region by 50% or more. For example, consider a window size of 1 Mbp and a homozygous insert size of 5 Mbp. For an overlap region on a chromosome between two mice from physical positions 10 Mbp to 20 Mbp, then a homozygous insert might be simulated between physical positions 12,300,000 and 17,300,000. If the HMM calls homozygous between 14 Mbp and 18 Mbp and heterozygous elsewhere, then overlap would be defined as 57.9 % =(17,300,000 - 14,000,000) / (18,000,000-12,300,000), which would be defined as successful recovery of that segment.

3.3.3 Results

Sensitivity and specificity

Results for the performance estimations for three window sizes (250 kbp, 500 kbp and 1 Mbp) for the French *M. m. domesticus* mice are shown in Figure 3.12, while results for the Taiwanese mice are shown in Figure 3.13. The Indian population had an insufficient amount of overlapping homozygosity to make such an assessment. For sensitivity, larger window sizes have poorer sensitivity for detecting homozygous segments smaller than 3 Mbp in size, although at 3 Mbp and beyond, power is very similar. For specificity, for the French mice, results for the larger window size led to substantially fewer small false positive calls, while the Taiwanese mice saw few real calls at that size. Note that the dip in the estimated proportion of calls real for the Taiwanese sample for a region of length 5 Mbp is due to a single homozygous call between two mice on chromosome 7 spanning 2 Mb of poor quality, unmapped genome (Figure 3.13, right panel, solid line). There were no false positive calls for either mice of greater than 5 Mb. I chose to proceed with a 1 Mbp window size, confident in the ability of the method to be both sensitive and specific for most recently inbred segments of more than 3 Mbp.

Figure 3.12. Sensitivity and specificity for inbreeding in French mice Left panel assesses sensitivity, where the horizontal axis gives the length of simulated homozygous regions, and the vertical axis gives the power, or proportion of times the region was correctly called homozygous. Right panel assesses a measure akin to specificity, where horizontal axis is homozygous calls of a given length, and the vertical axis is the estimated proportion of the homozygous calls that are truly homozygous. For both panels, results for different window sizes are indicated with different dashed lines.

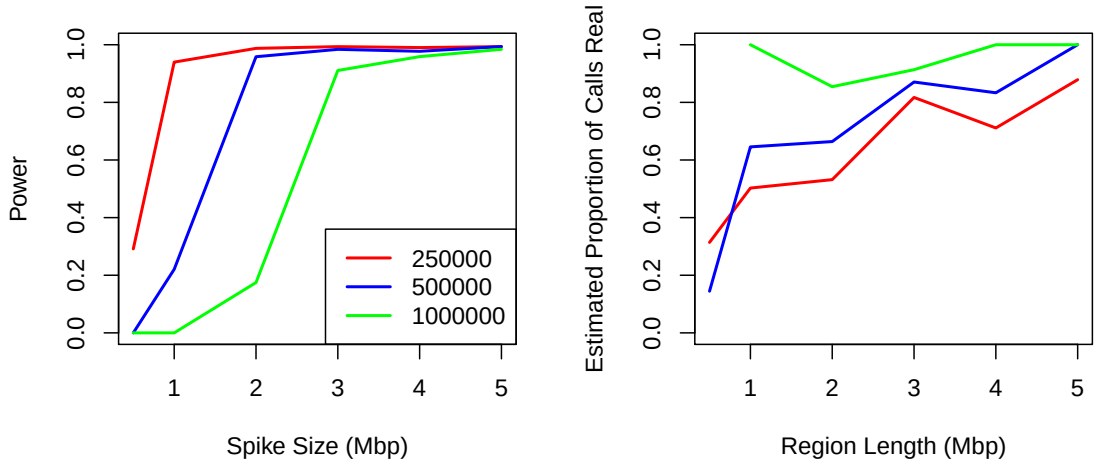


Figure 3.13. Sensitivity and specificity for inbreeding in Taiwanese mice Panels as in Figure 3.12 for the Taiwanese mice

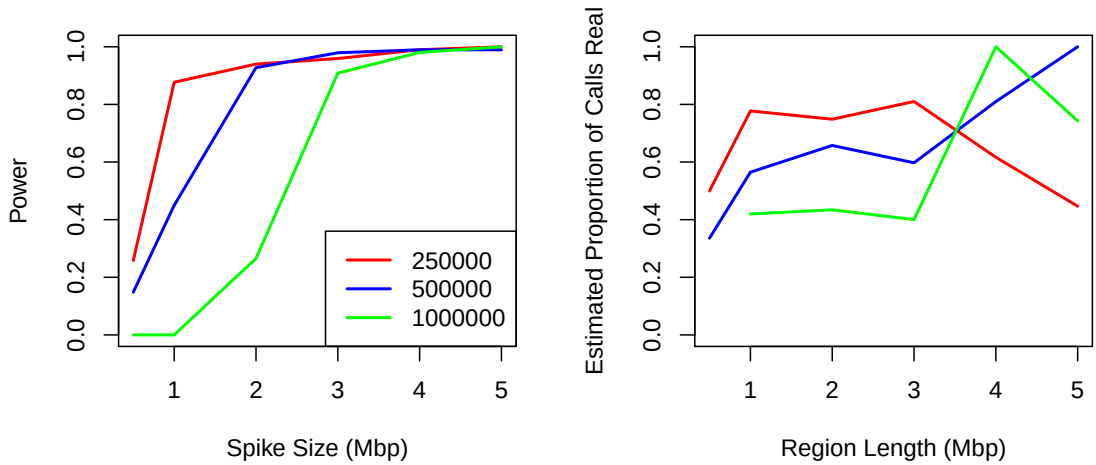


Figure 3.14. Example inbreeding assignment in wild mice Shown are two mice, with yellow regions referring to regions called homozygous by the algorithm, green regions called heterozygous, and red regions masked out due to insufficient genotypes.

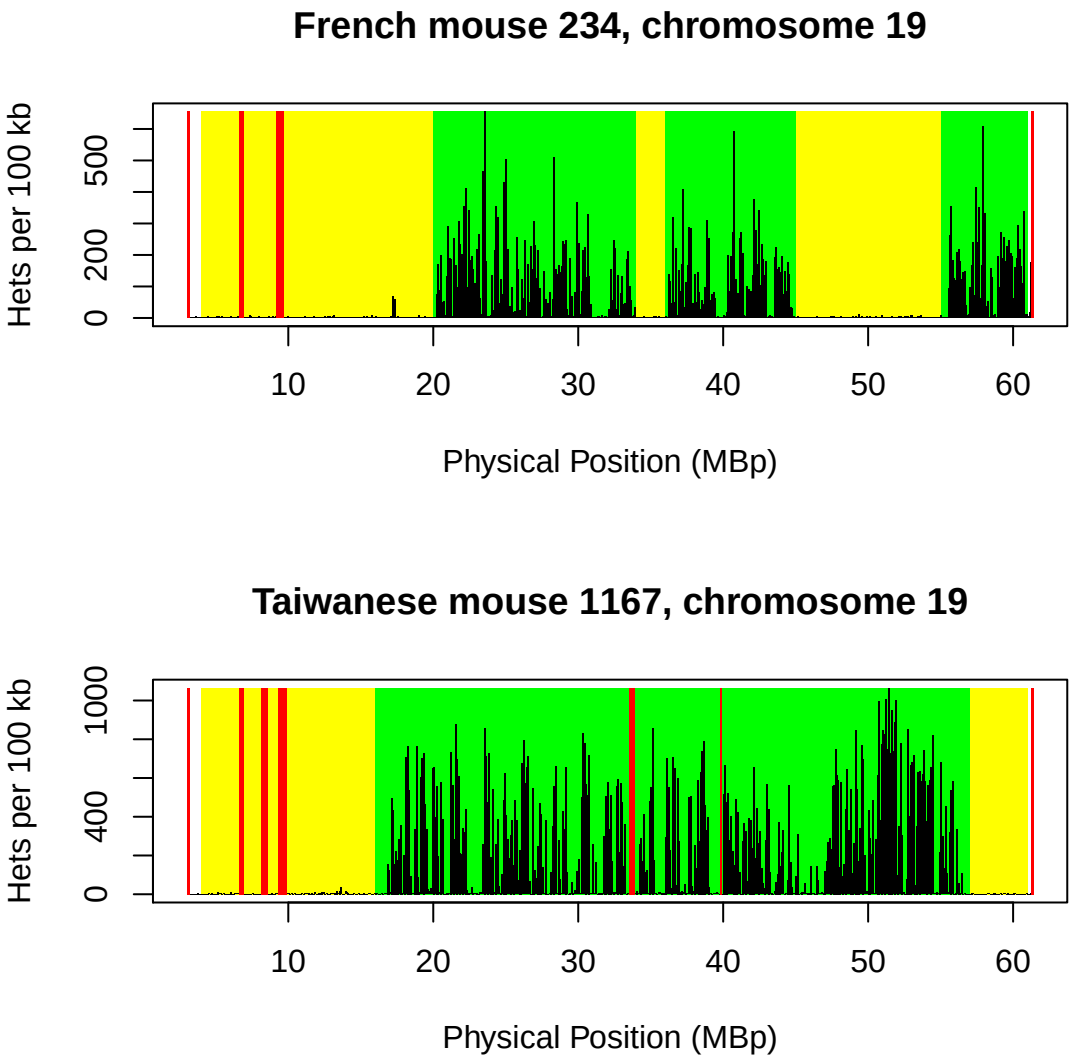
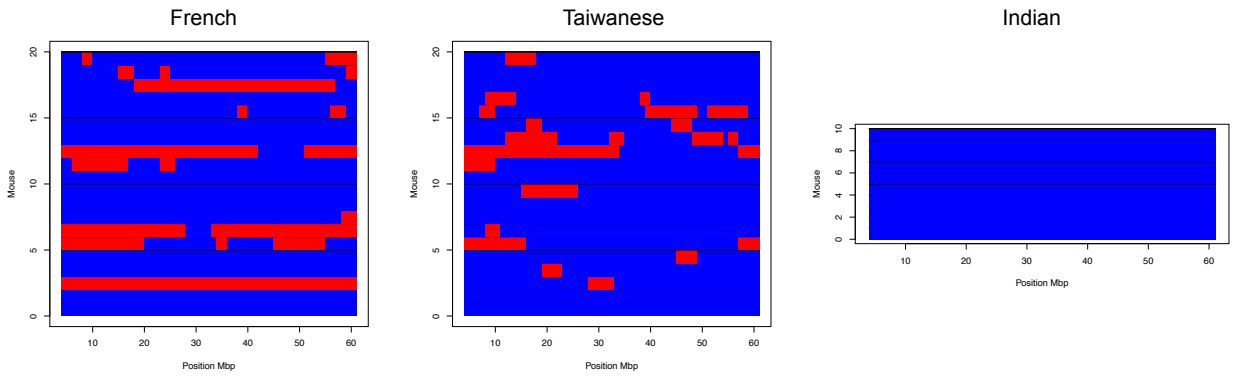


Figure 3.15. Homozygosity on chromosome 19 Rows give mice, colours represent heterozygous (blue) or homozygous (red). The Indian population has no homozygous segments on this chromosome.

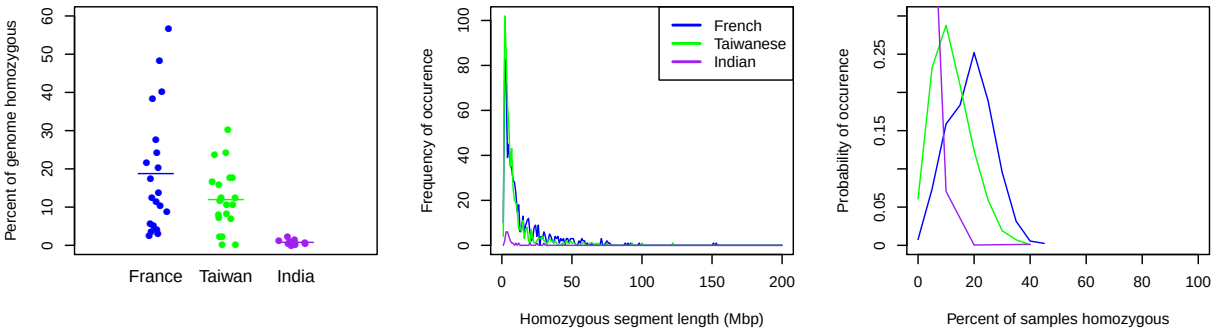


Distribution in wild mice

An example call of homozygous and heterozygous regions is overlayed on top of heterozygous genotype calls for one of the French and Taiwanese mice in Figure 3.14. These results indicate the HMM is working broadly as anticipated, and shows that while the outbred regions of the genome have large range in heterozygote genotype density, the inbred regions are starkly near zero. Example results for chromosome 19 are shown in Figure 3.15, showing homozygous regions in the three populations.

Running the HMM on all three populations of mice, we see that the average genome of a French mouse is 19 % homozygous, while the average in a Taiwanese mouse is 14 % and the average for an Indian mouse is 0.4 %, as shown in Figure 3.16. There is considerable variance in estimates between mice, with one French mouse having in excess of 50% of their genome homozygous, and others having less than 5%. In the Indian mice, very little homozygosity is observed at all, with no mouse having more than 5%. The distribution of segment lengths was heavily skewed away from full chromosome length homozygosity to much shorter segments in all populations, although the French contained the most long segments. Looking at at clustering of homozygous regions in the mice suggested an approximately normal distribution of

Figure 3.16. Distribution of homozygosity in the wild mice Left plot shows what percent of each of the genomes of the wild mice are homozygous. Centre plot shows a histogram of homozygous segment size. Right plot shows distribution of the number of mice that are homozygous in any given region of the genome.



the number of animals homozygous in any given region, and did not suggest any particular region was more likely to be homozygous.

Loss-of-function mutations

To investigate the distribution of loss-of-function mutations in the different populations, and how it relates to inbreeding, I identified loss-of-function mutations in the different populations and compared them. For this analysis I used the ancestral allele call from *M. caroli* and *M. famulus*, along with snpEff [91], to find derived SNP stop-gain mutations *i.e.* a SNP which changed an amino acid triplet to a stop-codon.

Because loss-of-function mutations are notoriously difficult to call accurately [92], I performed an initial check on performance by looking at the locations of SNP stop-gain mutations in the reference strain C57BL/6J. Encouragingly, virtually all (70/71) stop-gain mutations identified in this mouse occurred at precisely the end of annotated genes, giving confidence that this class of loss-of-function mutations were accurately called. Other potential loss-of-function mutation categories were also considered, such as splice site variants, but were not further pursued as they performed more poorly

on this measure. Since the reference genome C57BL/6J is more closely related to the French mice as opposed to the other populations, and functional annotations are based upon C57BL/6J, a concern is that genes independently inactivated in *M. m. domesticus* and *M. m. castaneus* would not be properly counted in *M. m. castaneus* when the mutation occurred later in the gene than in *M. m. domesticus*; as such, to help mitigate this risk, I considered only the first SNP stop-gain mutation per gene. This can, for example, occur in C57BL/6J if the annotated stop codon is in fact a mutational changed gained by C57BL/6J.

Summaries of the number of stop-gain mutations found in an average mouse in each population are shown in Table 3.4. For the total counts, the Indian mice contain a larger count than the Taiwanese mice ($p = 2 \times 10^{-12}$, Student's t-test) or the French mice ($p = 2 \times 10^{-15}$). However, by contrast, the Indian mice carry fewer total inactivated genes than the French ($p = 5 \times 10^{-11}$) or the Taiwanese ($p = 3 \times 10^{-6}$), as shown in Table 3.5.

The functional impact of stop-gain mutations depends on how likely they are to be found in a homozygous state in a population and lead to loss of gene function. This is influenced both by their frequency, and the degree of inbreeding found in the populations. To better examine our results in the different populations, I attempted to simulate results for each population for entirely inbred and outbred mice. To do this, I first determined the frequency of each stop-gain mutation. Next, I simulated single haploid genomes with inactivating mutations according to the frequencies of those mutations in a given population. From this I could estimate the number of genes inactivated in a fully inbred mouse, while for a fully outbred mouse, I combined two different haploid genomes and counted genes inactivated on both haplotypes. Results are given in Figure 3.17, and are consistent with observed results and the results in the Tables, revealing no evidence of a departure from expectations given the observed inbreeding levels. This implies that patterns observed in real mice are driven by

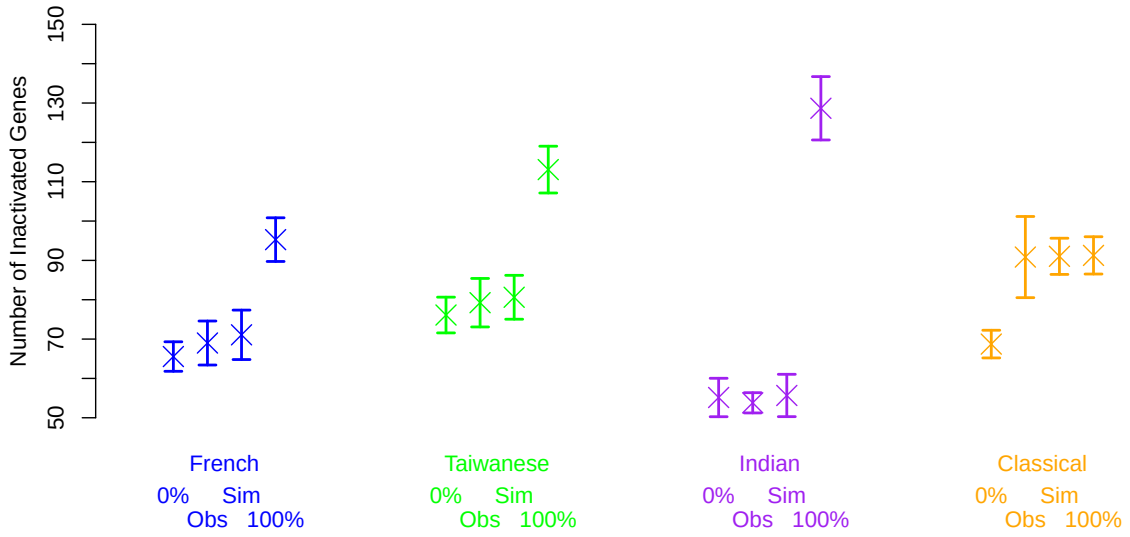
Table 3.4. Number of stop gains per wild mouse genome For each mouse, for each class of stop-gain mutation, the number of derived stop-gain mutations per diploid genome was counted, with heterozygous stop-gain mutations counted once and homozygous stop-gain mutations counted twice. Shown are the mean and standard deviation per mouse within each class and population. Unique classes of stop-gain mutations are: "Rare" means less than 5% frequency; "Polymorphic" means greater than 5% frequency; "Fixed" means fixed derived (or missing) in all mice in that population. "Total" means any one of the previous classes. Note that due to missing genotypes, there is variation in the "Fixed" count within a population.

	France		Taiwan		India		Classical	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Rare	9	3	10	4	71	6	16	10
Polymorphic	113	10	173	9	172	8	93	20
Fixed	62	3	37	2	15	1	74	0
Total	184	12	220	10	259	9	183	21

Table 3.5. Number of inactivated genes per wild mouse genome For each mouse, for each class of inactivated gene, the number of inactivated genes (*i.e.* homozygous derived stop-gain) per diploid genome was counted. Shown are the mean and standard deviation per mouse within each class and population. Unique classes of inactivated genes are: "Rare" means knocked out in only one mouse, "Polymorphic" means segregating and not rare, "Fixed" means homozygous derived (or missing) in all mice in that population. "Total" means any one of the previous classes. Note that due to missing genotypes, there is variation in the "Fixed" count within a population.

	France		Taiwan		India		Classical	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Rare	3	2	3	2	8	2	4	3
Polymorphic	35	4	58	6	38	3	50	10
Fixed	31	1	18	1	8	0	37	0
Total	69	6	79	6	54	3	91	10

Figure 3.17. Inactivating mutations in simulated wild mice The plot shows the mean number of loss-of-function mutations plus or minus one standard deviation, for either simulated outbred (0% homozygous), the observed results (Obs), simulated observed (Sim), or fully simulated inbred (100% homozygous) mice. Haplotypes are simulated from the derived allele frequency distribution for loss-of-function mutations in each population. Fully inbred animals are simulated by double sampling a single haplotype, while fully outbred animals are simulated by independently sampling two haplotypes. Simulated observed means mice are simulated with inbreeding the same inbreeding levels as observed in the wild mice.



differences in the frequencies of stop-gain mutations in the various populations. This also implies the "burden" of inbreeding is biggest in India, and smallest in the French mice, matching the observed levels of inbreeding in those populations. Finally, as could be expected, the Classical observed results are consistent with fully inbreeding French mice.

3.3.4 Discussion

In this section, I presented a method to classify regions of wild mouse genomes as being either homozygous or heterozygous. I showed that for a 1 Mbp window size, the method has good power for segments of homozygosity of greater than 3 Mbp, and low rates of false positives. This method will be used subsequently in Chapter 4 in the construction of recombination rate maps to mitigate the influence of mating structure on recombination rate estimation, and was used previously in Chapter 2 as a quality control metric for evaluating SNP callsets. As for results in the different populations, we see that the French mice contain on average the most and the longest homozygosity, followed by the Taiwanese, with the Indian mice containing very little. Locations of homozygosity are not obviously unusually distributed genome-wide. Given the relatively uniform recombination rate across the chromosome in mice compared to other mammals [90, 93], it is unlikely we would expect such an enrichment, unless there were large, ongoing selective sweeps. If we look at loss-of-function mutations, we see that Indian haplotypes contain the largest burden, but have the smallest number of genes knocked out on average.

The loss-of-function results also allow for an interesting comparison of the number of derived stop-gain mutations carried in the different populations. Recall that considered here are sites where variation exists in any of the wild mice populations or the classical laboratory strains. As such, one can consider the 100% inbred results to be the same as would be observed along a single haplotype, and so the larger number

of loss-of-function mutations in the Indian mice leads to the conclusion that selection is more effective at removing them in the French and Taiwanese populations, presumably due to the high levels of inbreeding in those populations allowing (or requiring) these groups to quickly purge deleterious recessive mutations. However, due to allele frequency distributions in the populations, the Indian mice carry the lowest burden on average in a completely outbred mouse, perhaps reflecting more effective selection in this group due to their large historical population size (Figure 3.11). Somewhat peculiarly, the Taiwanese mice carry the largest number of inactivated genes in a given mouse, despite carrying an intermediate level on a single haplotype. Some caution should be used in the interpretation of these results: for example, this argument suggests that many of the observed "loss-of-function" mutations, particularly in more inbred groups, may act on inconsequential genes and have little or no fitness consequence.

Future improvements would come from a PSMC style classifier [63] of homozygosity, so that a more formal TMRCA could be calculated along the genome, and homozygous versus heterozygous stretches could be considered more explicitly. As for interpretation, it should be possible to use the distribution of homozygous region lengths observed here to infer parameters associated with mouse colony structure, for instance size and transience, by comparing either analytically derived or simulated length distributions to observed distributions, and selecting model parameters that offer the best fit to the observed data. Note that at a minimum, the many short segment lengths inferred here suggests that long-term colonies are likely common.

3.4 Selection

3.4.1 Introduction

To search for and to characterize regions under selection is informative both generally, for instance to understand the effect of selection on biodiversity, as well as specifically,

to find particular genes contributing to the changing phenotype of a population. Such analyses are therefore a topic of considerable interest in humans and other species. Large-scale surveys of human genetic variation, such as the HapMap Project and the 1000 Genomes Project, have included companion studies to better characterize and subsequently understand recent selective events in the human genome [74, 94]. An often cited example of a gene under selection in the human population is the Lactase gene in European populations, where an allele rose to high frequency in European populations over the last 5,000 to 10,000 years by providing a selective advantage to its carriers, most likely by conferring the ability to consume dairy products into adulthood [95]. One of the highlighted reasons for sequencing the Neanderthal genome was in its ability to help identify regions of recent selection in the human genome - to try and find regions responsible for making us human [96].

These have been similar efforts to find selected regions in other species. For example, surveys of selection have been performed in honeybees [70], wild pigs [97], and wild mice [18]. Work has also focused on understanding selection in domesticated animals (cattle [98]), or on the possibility of polygenic selection and adaptation (*e.g.* rabbits [99]). Some of these involved considerable follow-up, and helped to shape knowledge of what processes are under selection in wild or domesticated animals. In dogs, sequencing of pooled dog versus wolf genomes revealed multiple regions of the genome likely under selective pressure, and pointed to the importance of a switch from a starch-poor to a starch-rich diet during the domestication of dogs [100].

I therefore sought out to test for signals of selection in the wild mice, for example to identify noteworthy examples of strong selective events which could shed light on biological features of particular importance to the different populations. An important concern is that we expect limited power due to the small sample sizes available here, particularly given our as-of-yet incomplete knowledge of mouse population genetics and history.

3.4.2 Methods

Many popular methods used to search for regions of the genome with non-neutral histories are based on one of three overlapping signals that selection has taken place or is ongoing [101]: unusually long and related haplotypes; deviations from the expected allele frequency spectra; and unusual patterns of differentiation between populations. I first discuss the former and its potential use in this dataset, and then move on to the later two, focusing in particular on deviations from the expected allele frequency spectra.

One common technique to evaluate whether an allele is under selection in a population is to look at the length of haplotypes which contain that allele, versus those that do not. The theory is that haplotypes containing an allele under selection increase in frequency so rapidly that their linkage with neighbouring variants decays more slowly than for haplotypes that do not carry that allele. Therefore, one can measure the decay in haplotype length for an allele, and deduce that alleles associated with unusually long haplotypes may be under selective pressure. One popular implementation of this is the extended haplotype homozygosity (EHH) approach of Sabeti *et al.* [102]. Simply, the method works by considering a region of SNPs, then asking for all members with an allele (SNP), whether the probability that haplotypes that contain that allele are more frequently identical (homozygous) then those that do not. A follow up method, the cross population extended haplotype homozygosity (XP-EHH) [103], does something similar, except it compares a set of results surrounding a SNP within a population to those outside of it. Other approaches, such as the Integrated Haplotype Score of Voight *et al.* [104], use similar ideas. A recent publication [105] includes the composite of multiple statistics (CMS) approach [101] and combines multiple measures of haplotype lengths, including XP-EHH, and iHS, to find evidence of regions of the genome under selection.

However, to use such methods, one must be able to phase genotype data into

haplotypes. As was shown in Section 3.2.2, we are not able to generate haplotypes with a sufficient degree of accuracy to detect unusual decay properties of haplotypes. I therefore was unable to pursue haplotype-based methods of detecting selection.

As mentioned earlier, an alternative to using haplotypes is to look for deviations in the allele frequency spectra, or summary measures of it, to find regions of the genome under selection. Commonly used statistics include Tajima's D [106] or Fay and Wu's H [107]. I pursued analysis of the wild mouse sequencing data using these statistics. I additionally looked for genes with a large number of private, fixed non-synonymous changes as evidence for recent, full selective sweeps. Finally, in subsequent follow up analysis, I examined measures of population differentiation, such as F_{st}, to help characterize the results.

Tajima's D

In 1989, Tajima proposed comparing the difference between different estimators of the population mutation rate $\theta = 4N_e\mu$ as evidence of departures from neutral genetic variation, where one such cause of departure could be the effect of selection [106]. The first estimator, θ_k, is based on the fact that under the assumption of constant population size, neutrality, random mating, and infinite sites, the expected number of segregating sites S has expectation $\mathbb{E}[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i}$, and so we can estimate θ with

$$\hat{\theta}_k = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}} \tag{3.12}$$

where n is the number of haplotypes sampled. Similarly, the average pairwise differences between samples, π, has expectation $\mathbb{E}[\pi] = \theta$, as so we get the estimator

$$\hat{\theta}_\pi = \frac{\sum_{j=1}^{n-1} \sum_{i=j+1}^{n} k_{i,j}}{\binom{n}{2}} \tag{3.13}$$

where $k_{i,j}$ the number of nucleotide differences between samples i and j. This is equivalent to

$$\hat{\theta}_\pi = \frac{2n}{n-1} \sum_{i=1}^{S} p_i(1-p_i) \tag{3.14}$$

where p_i is the derived allele frequency of SNP i. The difference between these two estimates, $d = \hat{\theta}_\pi - \hat{\theta}_k$, has expected value 0, and the variance $\hat{V}(d)$ can be calculated analytically (omitted here but given in the original 1989 paper [106]), to yield the statistic

$$D = \frac{d}{\sqrt{\hat{V}(d)}} \tag{3.15}$$

Under the assumptions listed above, D has mean 0 and variance 1. Upon deviation from these assumptions, D will be either positive or negative. If we consider for a given number of sites that $\hat{\theta}_k$ is invariant, then under changes which make the average allele frequency higher, such as population bottlenecks or balacing selection, D will be positive. Conversely, changes which make the average allele frequency lower, such as population growth, or a selective sweep, will make D negative.

Fay and Wu's H

Fay and Wu's H is also a test statistic designed to determine deviations from neutrality, and is also the difference between two estimates of θ, but involves a different estimator of θ. The expected number of sites at which i samples contained the ancestral allele and $n-i$ sites contain the derived allele, ξ_i, has expectation $\mathbb{E}[\xi_i] = \frac{\theta}{i}$ [108]. From this fact, in 1997 Fu proposed a family of estimators of θ [109], each with expectation θ, that took the form

$$L(r) = \frac{\sum_{i=1}^{n-1} i^{1-r}\xi_i}{\sum_{i=1}^{n-1} i^{-r}} \tag{3.16}$$

For instance, Watterson's estimator is just $L(1)$. In 2000, Fay and Wu [107] proposed using $L(-1)$ as an estimator of θ, which becomes

$$\hat{\theta}_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i \tag{3.17}$$

This can be re-written as

$$\hat{\theta}_H = \frac{2n}{n-1} \sum_{i=1}^{S} p_i^2 \tag{3.18}$$

Taking the difference between θ_π and this new estimator, we get Fay and Wu's H [107]

$$H = \hat{\theta}_\pi - \hat{\theta}_H \tag{3.19}$$

Note that H from Equation 3.19 is not normalized, unlike Tajima's D from Equation 3.15. Under the assumptions given before, of constant population size, neutrality, random mating, and infinite sites mutation, H will have expectation 0 and some variance (not calculated analytically). Upon deviation from neutrality, H will be either positive or negative. If we consider H as the difference between Equation 3.14 ($\frac{2n}{n-1} \sum_{i=1}^{S} p_i(1-p_i)$) and Equation 3.18 ($\frac{2n}{n-1} \sum_{i=1}^{S} p_i^2$), then we see that, generally speaking, this will be negative when there are more high frequency variants then intermediate frequency variants, as would be expected either at or just nearby to a locus which has recently undergone a sweep (variants which have "hitch-hiked" to high frequency). Conversely, when there are more intermediate frequency variants then expected, H will be positive. This could be the result of balancing selection, whereby population fitness is maximized by intermediate frequencies of multiple alleles, or purifying selection, where there are fewer low frequency mutations then expected. Note that demographic processes, such as bottlenecks and migrations, will systematically affect Fay and Wu's H, just like Tajima's D.

3.4.3 Results

To look for the effects of selection, I divided the genome into non-overlapping 1 Mbp intervals, and looked at the number of fixed derived variants private to a particular population, as well as the number of private fixed derived non-synonymous variants. Windows with insufficient callability (defined here as less than 50%) were not analyzed. Results are shown in Figure 3.18. Two results stood out - in the Taiwanese

population, the gene *Brca2*, and in the Indian population, the gene *Prl*. These genes are discussed in more detail below.

I next looked at the distribution of D and H genome-wide over the same 1 Mbp intervals. Again, regions with insufficient callability were removed, while the estimators in H were normalized to the callability of the region. Results were summarized into histograms and scatter plots and are shown in Figure 3.19. The distribution of D observed here shows that the French population had slightly positive D (mean=0.69, sd=1.40), the Taiwanese population also saw slightly positive D (mean=1.19, sd=1.99), while the Indian population had slightly negative D on average (mean=-0.73, sd=0.41). These results are in agreement with our earlier work showing the French and Taiwanese populations were subject to strong bottlenecks, while the Indian population is perhaps expanding more than expected. For H, all 3 populations had an average below 0, with the French and Taiwawnese results showing greater variances.

In terms of individual regions which are unusually different from the majority of observations, the most notable result is in the French population with a set of low D and H coming from a region of chromosome 17 between approximately 134 Mbp and 138 Mbp, including the gene *Vkorc1* (more on this below). The *Prl* region of India also contains low values of D and H, nearly the lowest genome-wide.

I therefore chose to focus on these top results for further investigation - *Vkorc1* in the French population, the region surrounding the *Brca2* gene in the Taiwanese population and the region surrounding the *Prl* gene in the Indian population.

3.4.4 Top results

Vkorc1 in the French population

To further explore this region, I examined genotypes from the 20 French mice for unusual variation. One of the mice seemed to have an unusual amount of private

Figure 3.18. Genome-wide private fixed derived variants Values were calculated over 1 Mbp non-overlapping windows, with insufficiently callable regions removed. The x-axis denotes physical position in ascending order genome-wide, with red and blue colours denoting alternating chromosomes. Private fixed derived means the other two populations are fixed ancestral while the population of interest is fixed derived. Extreme results are highlighted for Taiwan *Brca2* and India Prl.

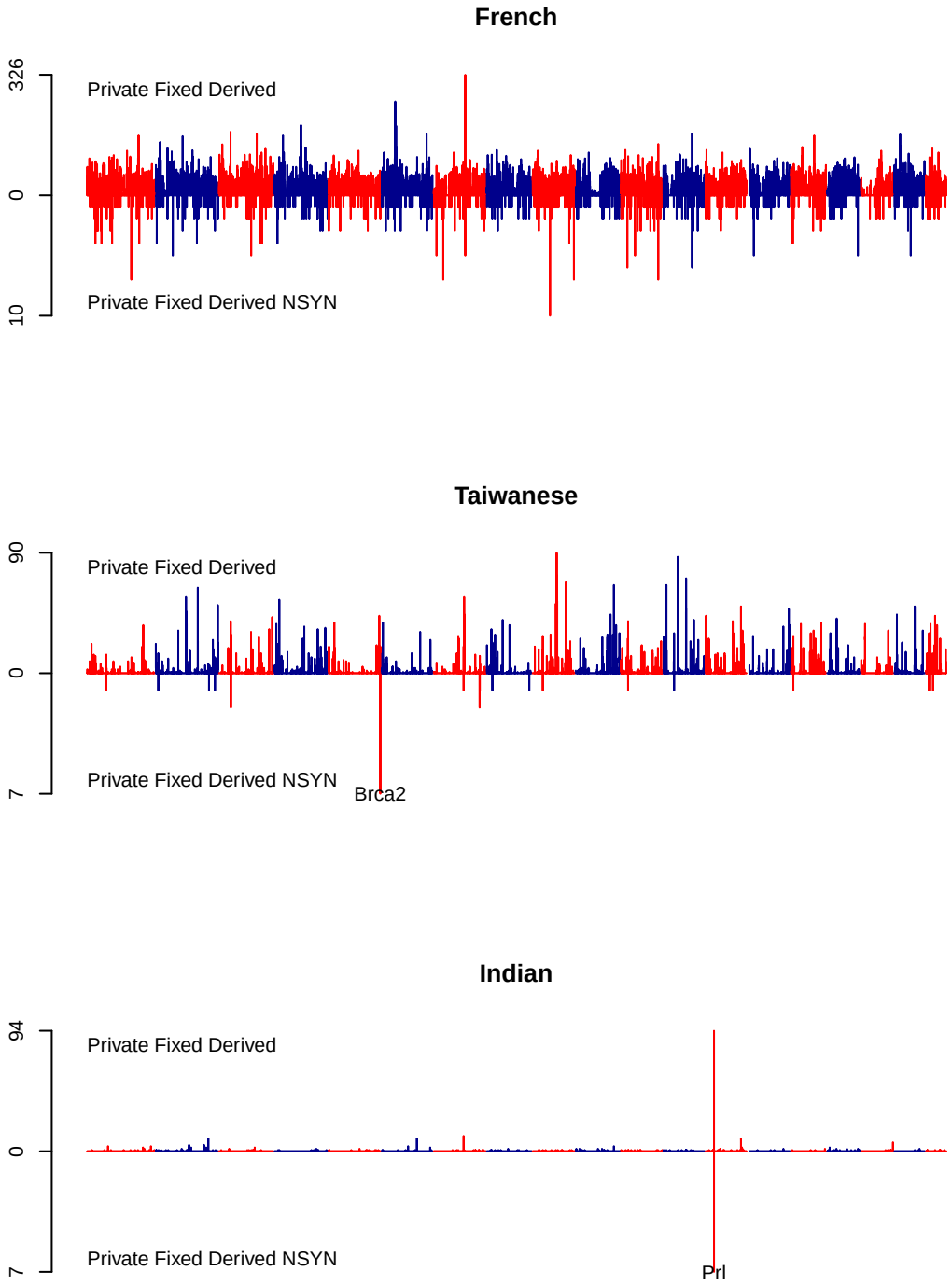


Figure 3.19. Distribution of D and H genome-wide for the different populations Values were calculated over 1 Mbp non-overlapping windows, with insufficiently callable regions removed. Red lines in the scatterplot denote average values, while black lines are through 0. Results surrounding *Vkorc1* and *Prl* are highlighted in their populations of interest

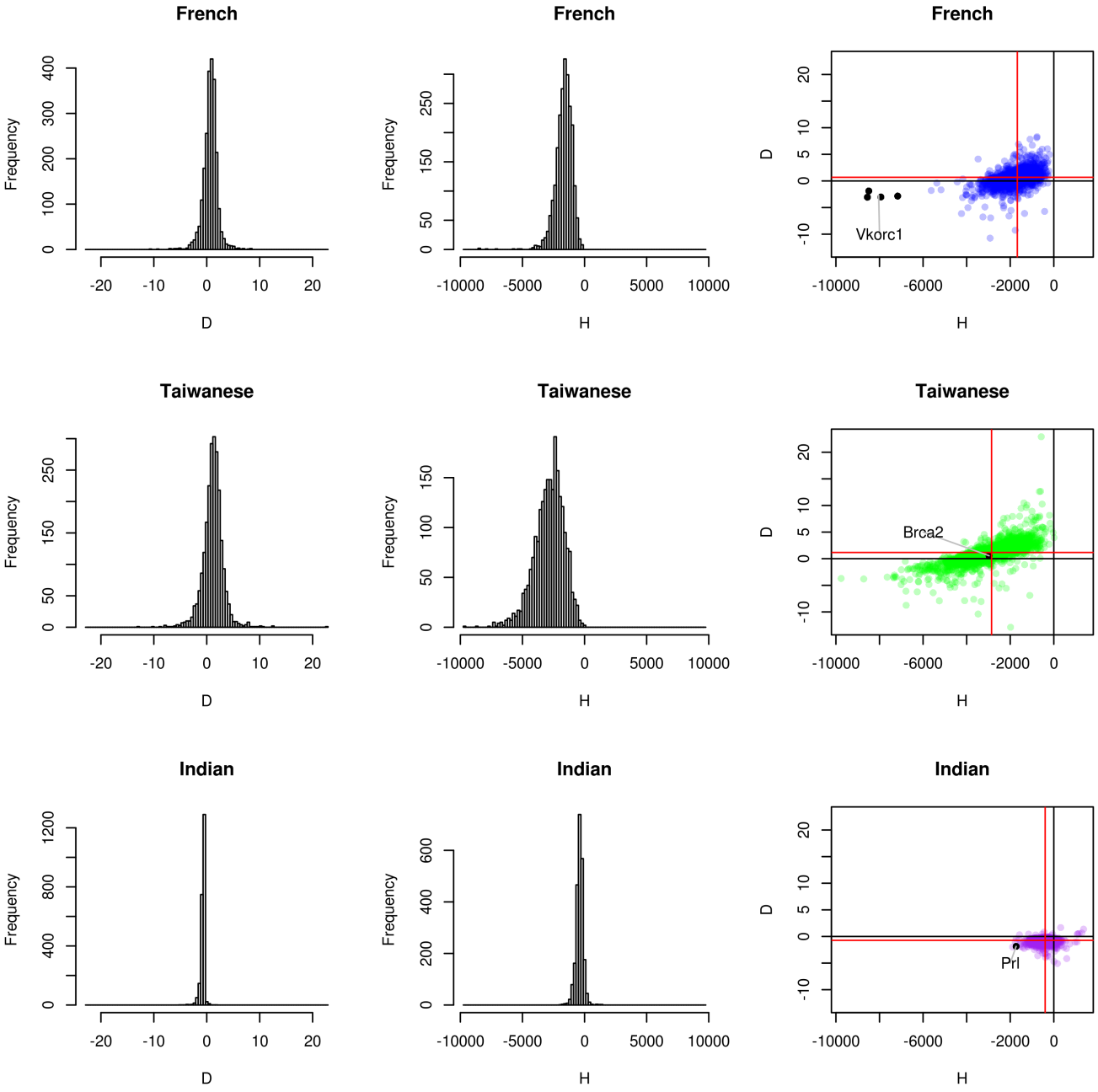


Figure 3.20. Distribution of singleton genotypes surrounding *Vkorc1*
Top panel shows per-mouse cumulative number of singleton (private heterozygous) genotypes in each of the 20 French mice. The position of *Vkorc1* is higlighted with a vertical bar. An outlier mouse is highlighted in red. Bottom plots are zoomed in at the start and end of the region.

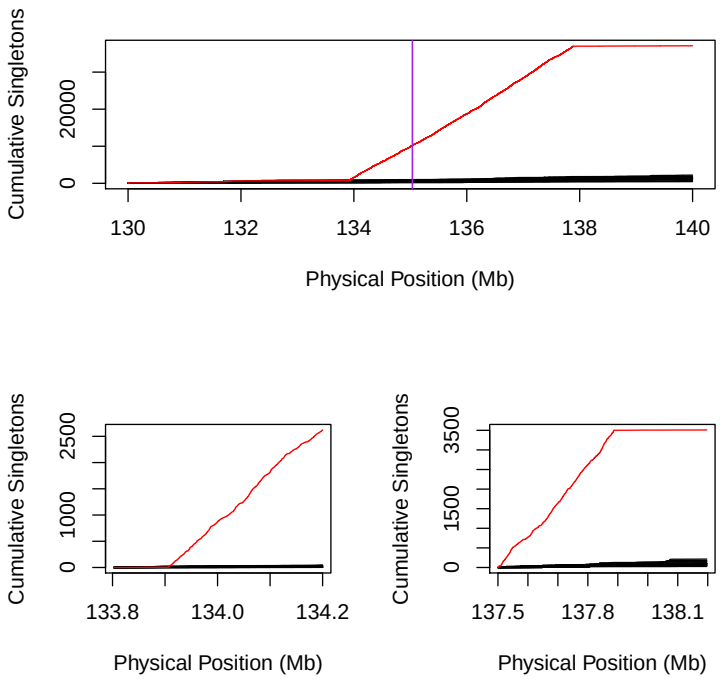
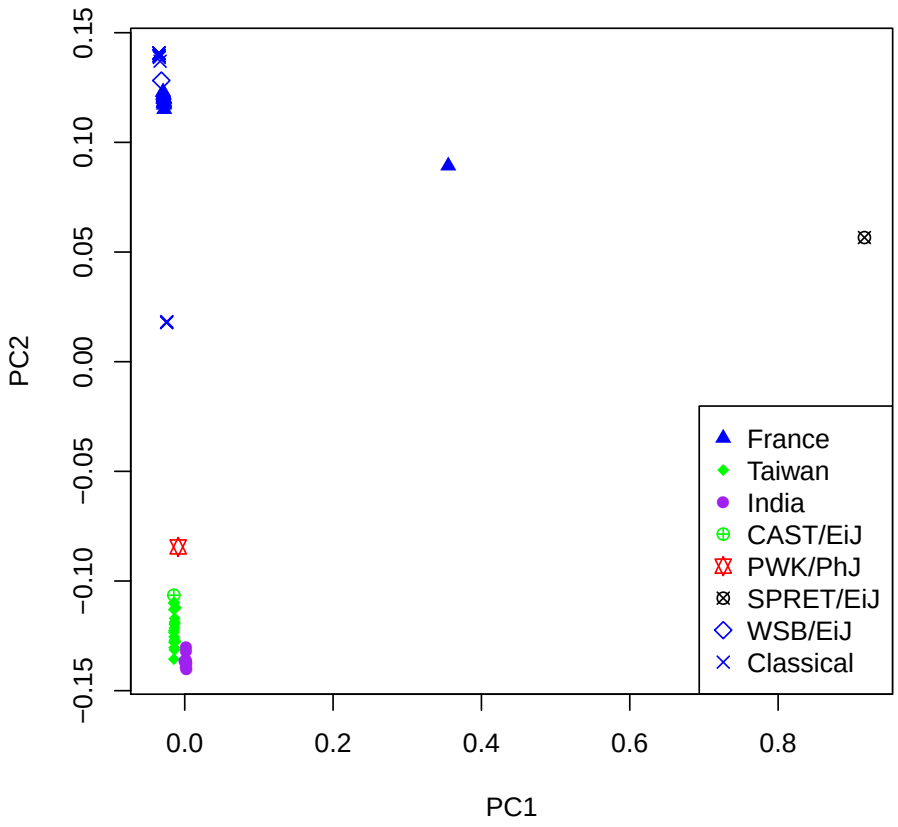


Figure 3.21. Principal components analysis of genotypes surrounding *Vkorc1* Results were calculated from chromosome 7 between 133.9 and 137.9 Mbp. The French mouse containing the introgression, mouse 34, is clustered midway between the SPRET/EiJ sample and the *M. m. domesticus* cluster. The three classical strains highlighted,129S1, 129S5 and LPJ, cluster midway between the other *M. m. domesticus* samples and the *M. m. castaneus* and *M. m. musculus* samples, due to part of this region being introgressed into these *M. m. domesticus* mice from either *M. m. castaneus* or *M. m. musculus*.



variation, as shown in Figure 3.20. I ran a PCA on the genotypes, and this mouse seemed to cluster midway between *M. m. domesticus* and *M. spretus*, as shown in Figure 3.21.

Investigation of the literature revealed that the region of interest for the French mouse was a known example of intersubspecific introgression [19]. *M. spretus* mice from north Africa and southern Spain acquired resistance to the rodenticide Warfarin, and this introgressed into *M. m. domesticus* and spread throughout Europe [19]. I searched for the 4 protein coding mutations from [19] in our data; indeed, the French mouse possessed all 4 known non-synonymous mutations that confer warfarin resistance (R12W, A26S, A48T, R61L), and these were absent from all other French mice. Therefore, it is highly likely that the unusual result in the French mice is driven by introgression from *M. Spretus* into European *M. m. domesticus*. The highly negative values of D and H in this region were caused by a substantial increase in low and high frequency variation caused by the introgressed segment.

Brca2 in the Taiwanese population

Results shown in Figure 3.18 implicated *Brca2* as unusual in the Taiwanese population for containing 7 private, fixed derived non-synonymous variants. A fine-scale survey of private Taiwanese fixed derived variants compared to the French and Indian mice, regardless of whether they were protein coding, revealed they were highly concentrated in the region spanning from approximately 151.339 to 151.360 Mbp. I surveyed other genes nearby, and found no other genes with private fixed non-synonymous substitutions in the Taiwanese mice.

A close inspection of the protein coding genotypes in *Brca2* is shown in Table 3.6. This revealed 3 additional non-synonymous changes of interest - two in which *M. famulus* and *M. caroli* agreed with the Taiwanese genotype, possibly an ancestral allele misassignment, and one in which *M. famulus* and *M. caroli* disagreed, and so an ancestral call was not made previously. The positions of these mutations, spread

Position	Ref	Alt	AA change	Effect	F	T	I	WSB	CAST	PWK	Spret	FAM	Car	Alignment
151,341,340	A	C	I665L	N-SYN	0	2	0	0	2	0	0	0	0	IIIIIT=
151,341,761	C	A	S805Y	N-SYN	0	2	0	0	2	0	0	0	0	SSYDSSLS
151,341,967	T	A	C874S	N-SYN	0	2	0	0	2	0	2	2	2	CSRRSRKG
151,342,027	C	A	Q894K	N-SYN	0	2	0	0	2	0	2	2	2	QKNNREEE
151,342,082	C	T	A912V	N-SYN	0	2	0	0	2	0	0	2	0	AEKKKKKE
151,342,266	A	G	S973	SYN	0	2	0	0	2	0	0	2	2	
151,342,479	T	G	F1044L	N-SYN	0	2	0	0	2	0	0	0	0	FF–LLWS
151,342,922	A	G	N1192S	N-SYN	0	2	0	0	2	0	0		0	NKNNNNSF
151,344,447	T	A	S1700	SYN	0	2	0	0	2	0	0	0	0	
151,344,584	C	T	P1746L	N-SYN	0	2	0	0	2	0	0		0	PQEEEE-V
151,345,011	A	C	Q1888H	N-SYN	0	2	0	0	2	0	0	0	0	QQQQQQK-
151,345,640	A	G	N2098S	N-SYN	0	2	0	0	2	0	0	0	0	NNNNNNS
151,347,471	A	C	G2229	SYN		2	0	0	2	0	0	0	0	

Table 3.6. Genotypes inside *Brca2* gene Shown are genotypes at sites fixed in *Brca2* for the Taiwanese mice as compared to the Indian and French mice. Sites are not polarized by ancestral allele call but by the reference mm9 genome - 0 = homozygous reference, 2 = homozygous alternate. Only synonymous (SYN) and non-synonymous (N-SYN) changes are shown. Genotypes are F=French, T=Taiwanese, I=Indian, W=WSB/EiJ, C=CAST/EiJ, P=PWK/PhJ, S=SPRET/EiJ, F=*M. famulus*, Car=*M. caroli*. Alignment is from UCSC Genome Browser and is amino acid sequence for, in order, mouse, rat, human, orangutan, dog, horse, oppossum and chicken. In alignment, - means gap in alignment, while = means no alignment available for that species.

Figure 3.22. Features surrounding *Brca2* in Taiwanese mice All plots are aligned vertically with respect to physical position, listed on the bottom. Top plot is frequencies for the Taiwanese population, anchored by *M. spretus* results in the region, *i.e.* *M. spretus* genotypes would all be 0 for this plot. Colours are shown in the legend on the right, while uncoloured (black) is for points where none of the other criteria are met, for example due to another population having missing data. Private fixed derived Taiwanese variants are denoted with triangles, which are given a vertical scatter for ease of viewing. Middle plot shows genes locally, with bars representing transcribed regions, with *Brca2* highlighted in red. Below are empirical p (*i.e.* quantile) for H, D and F_{st} against other populations, in 100 kbp windows, where F=France, T=Taiwan, I=India.

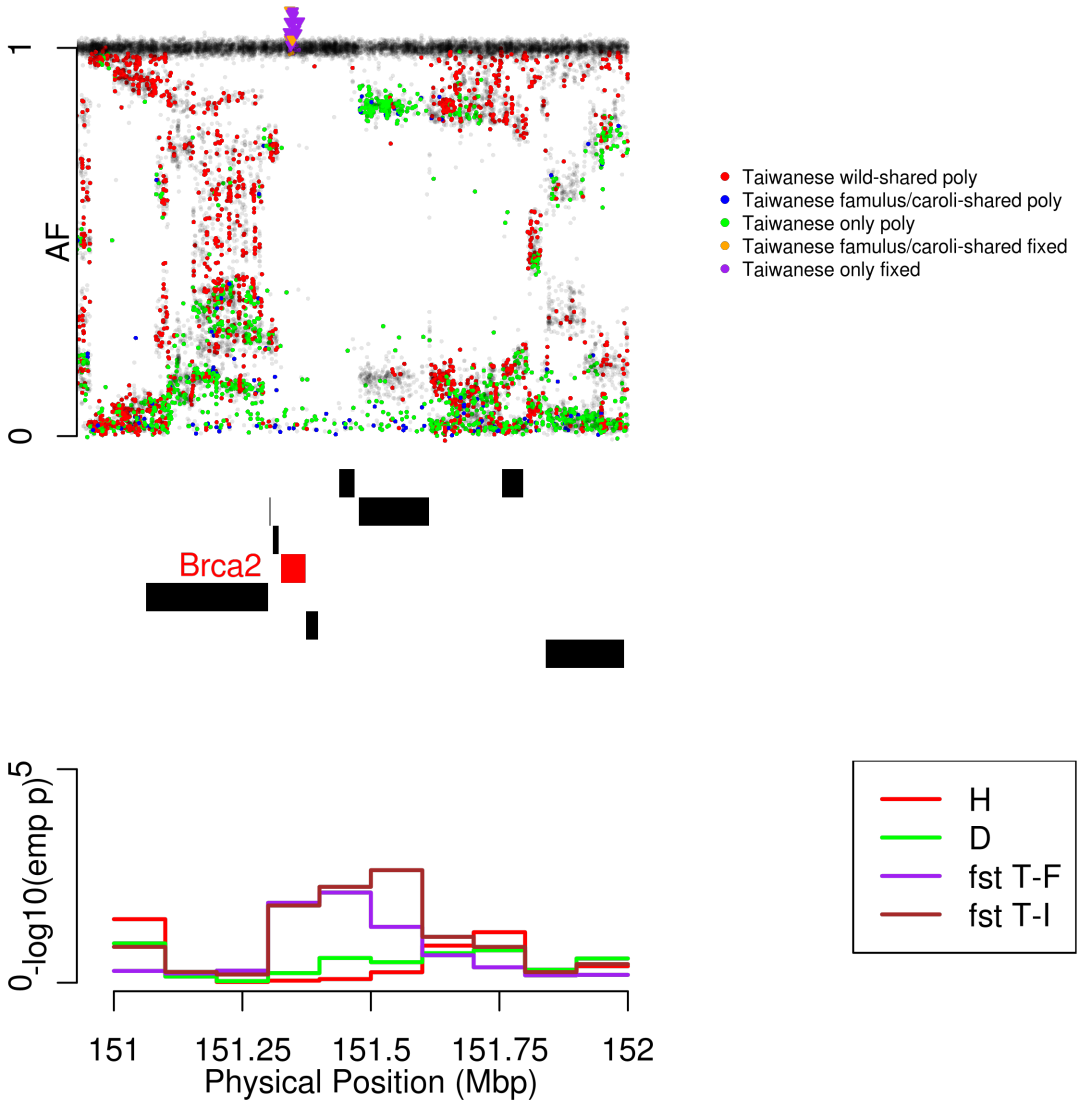
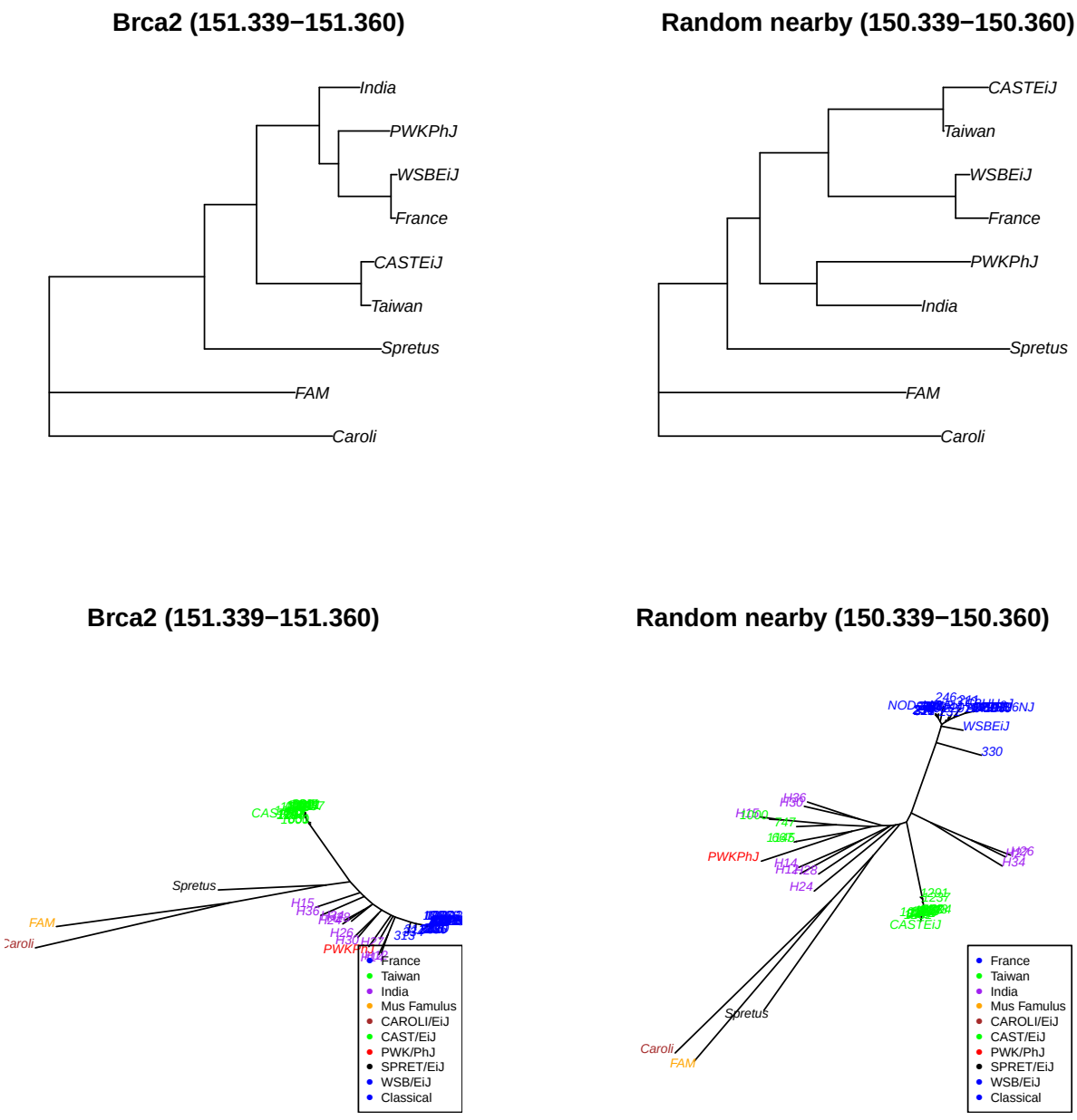


Figure 3.23. Gene trees for region surrounding *Brca2* Results are shown for region surrounding *Brca2* as well as a random region of the same size 1 Mbp away. The region surrounding *Brca2* was chosen based on the extent of private Taiwanese fixed derived variants, as explained in the text. The "random" region nearby was taken with a physical position 1 megabase earlier. The top plot is using MEGA, while the bottom plot is a neighbour-joining tree on all samples, with sample colourings explained in the legends.



out over the 5,000 bp, as well as 2 of the 7 changes being away from GC, suggest against their being the result of GC-biased gene conversion.

To investigate how well conserved the non-synonymous changes in Table 3.6 were, I investigated the default UCSC Genome Browser multiz alignment, looking at protein conservation across several species. Particularly well conserved mutations (to horse) were seen for I665L, Q1888H, and N2098S. Of these, isoleucine to leucine at position 665 is a modest substitution, given those amino acids are similarly sized with hydrophobic side chains. Similarly, asparagine (N, polar uncharged) to serine (S, polar uncharged) at position 2098 is similarly modest. By contrast, glutamine (Q, polar uncharged) to histidine (H, positively charged) at position 1888 represents a change predicted to induce a larger conformational change in the structure of the protein.

Further investigation of metrics for selection are shown in Figure 3.22. There is a dearth of intermediate frequency variation in the Taiwanese population in the immediate vicinity of *Brca2*. However, perhaps due to the recent extreme population changes in the Taiwanese population, this does not appear unusual per-se, as evidenced by the fact that D and H are not particularly noteworthy on a genome-wide scale. Although D and H should be relatively unremarkable on the sweep region itself, immediately outside the sweep region, there should be signatures of unusual D and H, as is shown later for Prl for the Indian population. F_{st} versus the Indian and French mice was more unusual (minimum $p = 0.0023$ versus Indian, minimum $p = 0.007$ versus French).

I looked more closely at the region surrounding *Brca2* by building simple trees in two ways in the immediate vicinity of the gene, as shown in Figure 3.23. First, I collapsed all population results to a single sample by taking the most common genotype in the population, then built maximum likelihood trees with MEGA-cc/MEGA6 [110, 111]. Second, I built neighbour-joining trees using PHYLIP and Rphylip using

allele-sharing matrices over the region of interest [83], similar to approaches taken by others to survey genome-wide relatedness [18, 70, 97], as I did in Section 3.2. Both of these results confirm that the CAST/EiJ sample is very similar to the Taiwanese samples, no other sample looks especially close to these samples, and the Taiwanese samples are outliers to all other *Mus musculus* mice, but not *M. spretus*.

I next investigated how unusual seeing 7 fixed non-synonymous mutations was in one gene, given there were 26 such changes genome-wide across all genes, and given that the gene, at 3,330 amino acids long, is the 172$^{\text{nd}}$ largest out of 24,322 genes. Given the large variability in differentiation in the genome, I first calculated, for each gene, how callable that gene was (number of callable bases in that population), and Fst between the Taiwanese and either the French or Indian mice, for a region of either 50,000 bp, centered at the middle of the gene, or the length of the gene, whichever was largest. Next, for all callable genes ($>$ 50% callable), I took the 10% of genes with the nearest Fst difference to *Brca2*. Including *Brca2*, this represented 17 of the 26 amino acids, out of a set of genes with 1,767,894 amino acids. Taken together, the *Brca2* results (7 out of 3,330 mutated versus 10 out of 1,764,564) was highly significant ($p = 1.6 \times 10^{-15}$), offering evidence that the clustering of non-synonymous substitutions in *Brca2* result was highly significant.

	N-SYN	Not N-SYN	Fraction N-SYN
Private CAST/EiJ	39	87	0.31
Private PWK/PhJ	12	23	0.34
Private WSB/EiJ	12	27	0.31

Table 3.7. Branches for *Brca2* gene Values are for between 151,339,000 and 151,360,000, the approximate extent of fixed variants for the Taiwanese population. 7,242 of 21,000 bases are coding in this region, approximately 34.4%. No ancestral allele alignment is done here, hence why variants are reported as being private or specific to a particular branch.

Finally, to further evaluate whether *Brca2* contains an unusually large number of non-synonymous differences, I focused on the CAST/EiJ sample, as it is quite

similar to the Taiwanese samples, and compared it to the WSB/EiJ and PWK/PhJ samples, given they are all haplotypes. Results are summarized in Table 3.7, which show that between 151,339,000 and 151,360,000, the number of non-synonymous differences private to the CAST/EiJ sample is approximately proportional to the length of the branch, as compared to what is seen in PWK/PhJ or WSB/EiJ. This result suggests that the amino acid differences found in the Taiwanese mice for *Brca2* are not obviously unusual.

To recap, *Brca2* contains 7 private fixed non-synonymous substitutions in the Taiwanese mice. Several of these are well conserved, and 1 in particular is well conserved and represents substitutions involving two amino acids with quite different chemical properties. Although *Brca2* is a large gene, it contains 7 of 26 such differences genome-wide, far more than the next most mutated gene at 2, and far more than would be expected by chance for genes with similar levels of divergence between Taiwanese and either Indian or French mice. However, when comparing CAST/EiJ with WSB/EiJ and PWK/PhJ, the fraction of sites along each branch that are non-synonymous is fairly constant. Further, signals of selection such as D and H are not particularly unusual, either at the sweep site or nearby, although the large variability of these measures in the Taiwanese populations makes them not particularly informative for selection anyway. Taken together, these pieces of evidence are suggestive, but not obviously conclusive, of some or all of the non-synonymous mutations having functional consequences, and their being fixed in the Taiwanese mice having been brought about by selection.

Prl in the Indian population

In the Indian mice, *Prl* was unusual for two reasons: first, it was the only gene with private, fixed derived non-synonymous variants in the Indian mice compared to the other populations, as seen in Figure 3.18; and second, it had one of the lowest values of H genome-wide, as seen in Figure 3.19. Further examination of the region showed

Position	Ref	Alt	AA change	Effect	F	T	I	WSB	CAST	PWK	Spret	FAM	Car	Alignment
27,151,365	G	A	R47H	N-SYN	0	0	2	0	0	0	0	0	0	RRRRRRRR
27,151,375	A	G	I50M	N-SYN	0	0	2	0	0	0	0	2	2	IMVVIIMK
27,151,391	C	T	H56Y	N-SYN	0	0	2	0	0	0	0	0	0	HHHHHHHH
27,151,413	T	C	F63S	N-SYN	0	0	2	0	0	0	0	0	0	FFFFFFFF
27,151,415	A	C	I64L	N-SYN	0	0	2	0	0	0	0	0	0	IISSNNNN
27,153,378	G	A	V71M	N-SYN	0	0	2	0	0	0	0	0	0	VVTTAAAA
27,153,400	T	C	V78A	N-SYN	0	0	2	0	0	0	0	2	2	VATTTTTT
27,153,415	A	T	D83V	N-SYN	0	0	2	0	0	0	0	0	0	DDSSSSSG
27,156,218	G	T	G130V	N-SYN	0	0	2	0	0	0	0	0	0	GGGGGGSR
27,156,237	A	T	E136D	N-SYN	0	0	2	0	0	0	0	0	2	EDEEDEDD
27,156,281	A	G	Q151R	N-SYN	0	0	2	0	0	0	0	2	2	QRRRRRRR
27,156,306	A	T	I159	SYN	0	0	2	0	0	0	0	0	0	
27,156,787	A	G	L182	SYN	0	0	2	0	0	0	0	2	2	
27,156,826	C	T	R195	SYN	0	0	2	0	0	0	0	2	2	

Table 3.8. Genotypes inside *Prl* gene Shown are genotypes at sites fixed in *Prl* for the Indian mice as compared to the French and Taiwanese mice. Sites are not polarized by ancestral allele call but by the reference mm9 genome - 0 = homozygous reference, 2 = homozygous alternate. Only synonymous (SYN) and non-synonymous (N-SYN) changes are shown. Genotypes are F=French, T=Taiwanese, I=Indian, W=WSB/EiJ, C=CAST/EiJ, P=PWK/PhJ, S=SPRET/EiJ, F=*M. famulus*, Car=*M. caroli*. Alignment is from UCSC Genome Browser and is amino acid sequence for, in order, mouse, rat, human, orangutan, dog, horse, oppossum and chicken. In alignment, - means gap in alignment, while = means no alignment available for that species.

Figure 3.24. Features surrounding *Prl* in Indian mice All plots are aligned vertically with respect to physical position, listed on the bottom. Top plot is frequencies for the Indian population, anchored by *M. spretus* results in the region, *i.e.* *M. spretus* genotypes would all be 0 for this plot. Colours are shown in the legend on the right, while uncoloured (black) is for points where none of the other criteria are met, for example due to another population having missing data. Private fixed derived Indian variants are denoted with triangles, which are given a vertical scatter for ease of viewing. Middle plot shows genes locally, with bars representing transcribed regions, with *Prl* highlighted in red, and *Prl* isoforms shown in blue. Below are empirical p (*i.e.* quantile) for H, D and F_{st} against other populations, in 100 kbp windows, where F=France, T=Taiwan, I=India.

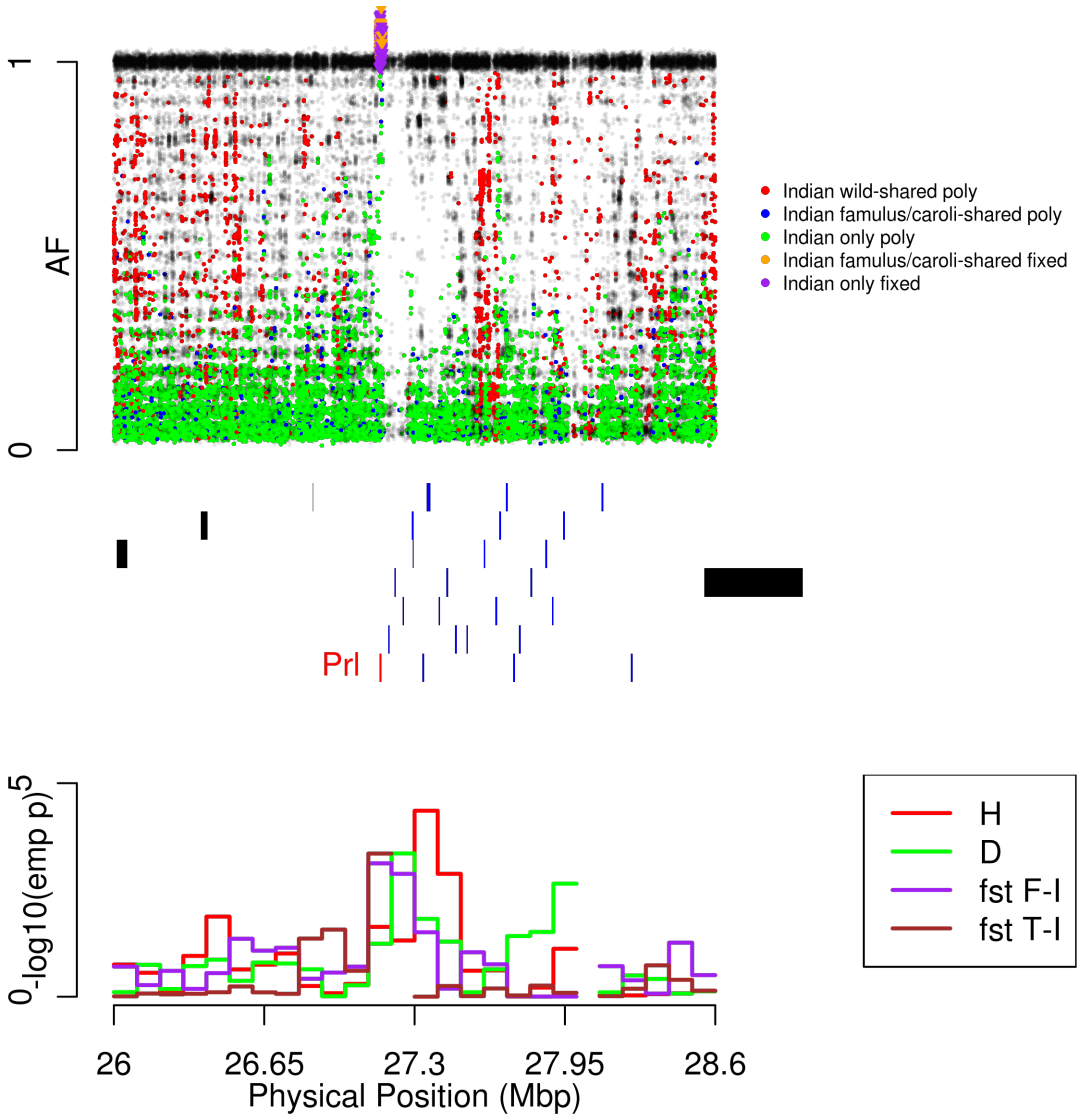


Figure 3.25. Gene trees for region surrounding *Prl* Results are shown for region surrounding *Prl* as well as a random region of the same size 1 Mbp away. The region surrounding *Prl* was chosen based on the extent of private Indian fixed derived variants, as explained in the text. The "random" region nearby was taken with a physical position 1 megabase earlier. The top plot is using MEGA, while the bottom plot is a neighbour joining tree on all samples, with sample colourings explained in the legends.

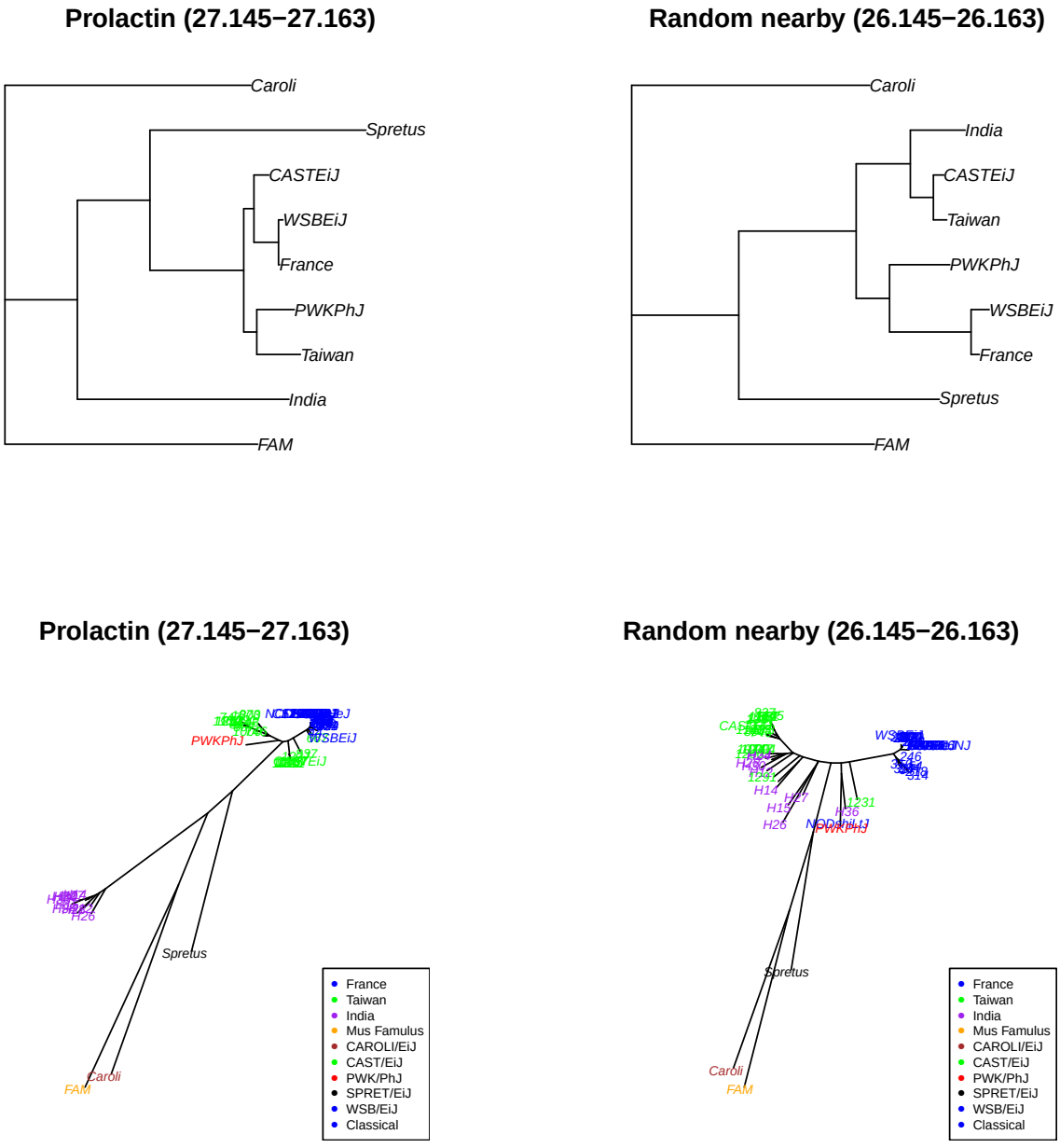


Figure 3.26. Sequence for *Prl* Sequence for *Prl* for the different mice, rat and human. Sequences were aligned using clustalW [112]. There were no polymorphic amino acid changes in Indian mice. D = domesticus = WSB/EiJ and C57/BL6, M=PWKPhJ, C=CASTEiJ, S=SPRET/EiJ, F=Mus Famulus, A=CAROLI/EiJ, I=Indian mice, R=Rat, H=Human.

```
     1
D :  MNSQGSAQKG-TLLLLLISNLLFCQNVQPLPICSAGD--CQTSLRELFDRVVILSHYIHT
M :  ..........-..........................--..................
C :  ..........-.......................T..--..................
S :  ..........-..........................--..................
F :  ..........-..........................--.............M.......
A :  ..........-..........................--.............M.......
I :  ..........-..........................--..........H..M.....Y.
R :  ....V..R.AG.....MM..........T..V..G..--...P.P.......M.......
H :  ..IK..PW..-S.....V....L..S.A.....PG.AAR..VT..D....A.V......N

     61
D :  LYTDMFIEFDKQYVQDREFMVKVINDCPTSSLATPEDKEQALKVPPEVLLNLILSLVQSS
M :  ............................................................
C :  ............................................................
S :  .....................................D......................
F :  ......................A.A...................................
A :  ......................IA....................................
I :  .....SL......M......A....V.................................
R :  ......................IA.A..................Q..............H.W
H :  .SSE..S....R.THG.G.IT.A..S.H.............QQMNQKDF.S..V.ILR.W

     121
D :  SDPLFQLITGVGGIQEAPEYILSRAKEIEEQNKQLLEGVEKIISQAYPEAKGNGIYFVWS
M :  ............................................................
C :  ............................................................
S :  ............................................................
F :  ....................................R...........................
A :  ...................D..............R...........................
I :  ............V.....D..............R...........................
R :  N.........L...H...DA.I...........R....I..............E..L...
H :  NE..YH.V.E.R.M.....A...K.V.....T.R....M.L.V..VH..T.E.E..P...

     181
D :  QLPSLQGVDEESKILSLRNTIRCLRRDSHKVDNFLKVLRCQIAHQNNC
M :  ...................................................
C :  ...................................................
S :  ...................................................
F :  ...................................................
A :  ...................................................
I :  ...................................................
R :  ..............D.AFY.N.............Y..F.....V.K...
H :  G.....MA....RLSAYY.LLH........I..Y..L.K.R.I.N...
```

no other genes with similar fixed mutations.

A closer inspection of the genotypes in the various mice for *Prl* is shown in Table 3.8. Aside from the 7 non-synonymous variants previously reported, 3 variants found in the Indian mice were also seen in both of *M. caroli* and *M. famulus*, while one was also seen in *M. caroli*. Given that GC-biased gene conversion can cause increased rate of fixation of strong (guanine (G) or cytosine (C); 3 intrastrand base interactions) as opposed to weak (adenine (A) or thymine (T); 2 intrastrand base interactions) nucleotides, I examined the position and bases of these variants. The 7 non-synonymous variants are spread out over three exons over a region of 5,000 bp, although locally close within exons, with 4 from strong to weak, 2 weak to strong, and 1 neutral (weak to weak), and argue against GC-biased gene conversion as their cause.

Further inspection of the non-synonymous mutations from Table 3.6 revealed several substitutions at highly conserved sites, including R47H, H56Y, F63S, G130V. Of these, arginine (R) to histidine (H) at position 47 is a more modest change, as both have positively charged side chains, while histidine (H, positively charged) to tyrosine (Y, hydrophobic) at position 56, phenylalanine (F, hydrophobic) to serine (S, polar uncharged) at position 63, glycine (small) to valine (hydrophobic) at position 130, and glutamine (Q, polar uncharged) to arginine (R, positively charged), have the potential to cause more conformational change. Interestingly, Q151R represents a site where the Indian mice carry the ancestral form, and the other *Mus musculus* mice as well as *M. spretus* carry a derived form at an otherwise highly conserved site.

I next examined several features of the region surrounding *Prl* more closely, as shown in Figure 3.24. *Prl* is located nearby to a large number of *Prl* paralogs. The region immediately containing *Prl* had relatively little segregating variation in the Indian mice, even high frequency derived mutation. Nearby, recovery of variation leads to the strong signals for D and H which are unusual genome-wide. The location of the fixed changes between the Indian and other populations are all highly

concentrated between 27.145 and 27.163 Mbp on chromosome 13.

I looked into the relatedness among the mice for this region as was done for *Brca2* earlier, with results in Figure 3.25. Results show that the sequence contained by the Indian mice, which is about 1.3% diverged from the French mice, would appear to fit into the species tree somewhere between *M. spretus* and the two outgroup species *M. caroli* and *M. famulus*. Finally, I examined amino acid sequences of *Prl* in the subspecies, as well as from rat and human. Results are shown in Figure 3.26.

Taken together, these results suggest that the mutations seen in *Prl* are the result of a recent, strong selective sweep, acting on a region of DNA introgressed from a species of mouse which is yet unknown but is more divergent from the Indian mice than *M. spretus* but less than the outgroups *M. caroli* or *M. famulus*.

3.4.5 Discussion

The results presented here offer a snapshot of some of the strongest examples of selection at work in the genomes of these wild mice. The introgressed region surrounding *Vkorc1* is about 4 Mbp long, clearly *M. spretus* in origin, and contains amino acid substitutions known to confer resistance to the rodenticide Warfarin [19]. The results at *Prl* in the Indian mice are also demonstrative of adaptive introgression, albeit much older. The source of this DNA is unclear, except that it is very divergent to the other *Mus musculus* mice. While the precise function of these mutations to *Prl* in the Indian mice are unclear, Prolactin itself is a major hormone regulator responsible for regulating the activity of hundreds of biological processes, including the stimulation of mammary tissue proliferation and the secretion of milk proteins [113, 114]. Therefore it is not unreasonable to assume these changes to *Prl* are likely to relay to the control of pregnancy and child rearing in the Indian population. Further study, to delienate the specific effects, as well as the extent of the mutations, is warranted. In the Taiwanese mice, *Brca2* contains a large number of private, fixed derived mutations,

several of which are highly conserved. It contains more such mutations than any other gene in the genome, and more than would be expected by chance. Although selection statistics are not suggestive of selection, they are not well powered given recent demographic extremes in Taiwan. In terms of its origin, this haplotype appears to come from within *Mus musculus*, although it is somewhat of an outlier to traditional *M. m. castaneus* DNA, and its exact origin and age is certain. *Brca2* acts to help repair double stranded DNA breaks, including during homologous recombination through interactions with *Rad51* [115]. In humans, inactivating mutations in *BRCA2*, as well as the related protein *BRCA1*, are well known to substantially increase the risk of certain cancers, notably breast and ovarian caners [116]. While the exact function of the private fixed non-synonymous mutations in *Brca2* seen in Taiwanese mice is unclear, it is entirely possible that these mutations have been driven to fixation through the actions of selection in the Taiwanese mice. A further study of the history of these mutations, or better yet, detailed functional studies, would help reveal what, if any, functional consequences these mutations have in Taiwanese mice.

There have been no attempts to survey the landscape of selective events previously in wild mice using whole-genome sequencing data. The only other recent, major attempt to survey selection in wild mice came in 2012 by Staubach *et al* [18]. In their work, using two populations of *M. m. domesticus* (France, N=11; Germany, N=11) and two populations of *M. m. musculus* (Czech Republic, N=11; Kazakhstan, N=11), they found a large number of regions in the genome subject to apparent selective sweeps, that introgression was common, and that the frequencies associated with introgressed events are not in agreement with neutral expectations, suggesting they are frequently targets of selection. Knowing what we know now, and reading their paper carefully, suggests that while the existence of large introgressed tracts is likely true, the large number of selected regions, their conclusions about the large number of small introgressions and possibly the fact that introgressed segments are under

non-neutral selective pressure, might not be accurate. First, their data came from the Affymetrix Mouse Diversity Array, with genotypes being called using BRLMM-P, which is similar to the data we had for the French mice and on which we estimated an error rate of about 5%. Although in Staubach *et al.* they do recognize the higher error rate of the array on wild versus laboratory mice, and do attempt to correct the genotype calls to mitigate the effect of off-target variation, it is unclear what the real error rate of their genotypes is. Second, marker density on the array is low in their study, with about 264 K useable polymorphic SNPs in the *M. m. domesticus* populations and 167 K in the *M. m. musculus* populations, meaning not many SNPs would be available in each of the 100 kbp bins they looked at for their selection statistics, and could leave them quite vulnerable to ascertainment biases. Third, they declared regions as significant based on comparisons against neutral simulations which were over a larger range (1Mbp) than the regions they tested (100 kbp), therefore underestimating the variance of their statistics and underestimating the cutoff required, while also using neutral simulations without migration and with population size estimates which are probably too generous based on what we know now. Taken together, these suggest many of the selected regions might be false positive - tellingly, very few of the regions identified by different methods overlap in their work, and they report only 9 with more confidence in their Table 2. As for introgression, the average reported lengths are short (German 189 kbp, French 150 kbp, Czech 347 kbp, Kazakh 461 kbp) given the number of SNPs in the implicated regions, the HapMix miscopying parameter they chose to use of 0.0005 too agressive (akin to thinking 0.5% of your genotypes are errors), the number of samples used too small, and the method of phasing for the donor populations (fastPHASE [44]) too unreliable, that these results do not seem plausible. However, the much longer tracts are not suspectible to many of these problems, and therefore seem more plausible, especially given the results we see here, while the non-neutrality of introgressed segments is again questionable given

the potential diversity of demographic factors at play which we even now do not fully understand.

Taken together, these results and the results from Staubach *et al.* do suggest a few key points which will inform future efforts to survey selection in wild mice. First, it is very difficult, although not impossible, to find selected regions in derived populations of wild mice like the French and Taiwanese populations. Given the extreme bottlenecks involved, the variance in genetic histories at neutrally evolving loci is so large on a genome-wide scale that it makes delineation of selected regions all but impossible when analyzing results for deviations from genome-wide patterns, as we have here. A more fruitful approach might be to analyze multiple more closely related populations, such as French and German populations, and look for population differentiation to identify selected regions which have been selected more recently and since the bottlenecking which introduced the variance in genetic history. It would also be beneficial to gain a better handle on the nature and extent of introgression in multiple populations world wide through the sequencing of a large number of single or pairs of mice.

Second, the results from the Indian population suggest this is a population that is much more amenable to study. Although haplotype-based methods will be all but impossible, given the large effective population size and number of samples that would be required to phase accurately, I showed here that summary statistic based methods can be quite a powerful approach, and given the milder genome-wide distribution of D and H for the Indian mice, sequencing a larger number of samples should allow for more regions to be found and greater confidence to be placed in classifying regions of the genome as either neutrally or non-neutrally evolving.

Third, searching for signals of selection, even in the Indian mice, would be greatly facilitated if a clearer picture of the extent of introgression between closely and distantly related mice were available. Both the *Vkorc1* and *Prl* results seen here are due

to introgression, while the derived allele frequency plots seen in the French and Taiwanese plots (Figure 3.3) suggests extensive migration into those populations. This is echoed by the finding of long introgression tracts in Staubach *et al.*, and informally in our own data, where we see many long (5-10 Mbp) tracts of introgressed DNA, particularly in the Taiwanese mice and to a lesser degree in the French mice, as evidenced through either PCA or counting singleton genotype methods (data not shown). A more thorough understanding of the frequency and intensity of introgression would shed light on our understanding of natural selection in these populations.

In conclusion, we found a new result of selection at *Prl* in the Indian mice, a suggestive signal at *Brca2* in the Taiwanese mice, and a known example of adaptive introgression at *Vkorc1* in the French mice. It will be difficult to find signals of selection in derived highly bottlenecked populations like the French and Taiwanese, while sequencing additional mice from the Indian population would likely identify more regions under selection and could yield insights into the relationship between selection and other genomic features, such as recombination and diversity. Finally, sequencing a larger number of single or pairs of mice from a large geographic span would likely help better understand signals of introgression seen here and in general, and help better characterize the histories of signals found.

Chapter 4

The recent landscape of recombination in mice

In Chapter 2, I presented whole genome sequencing data on 3 populations of wild mice, as well as laboratory and wild derived inbred strains of mice. I used this data in Chapter 3 to learn about how these mice were related, as well as their inbreeding, population histories, and selection. Here, the whole genome sequencing data was used by Oliver Venn to construct linkage-disequilibrium based recombination rate maps. I subsequently use the rate maps to learn how recombination is distributed at fine and broad scales between and within populations, as well as compare rates between populations and against known ChIP-seq based results.

The rest of this chapter will be organized as follows. First, I briefly re-introduce and expand upon known features of recombination in mice. Second, I present methods for constructing, analyzing and interpreting the rate maps. Third, I give results, including comparing the rate maps at fine and broad scales, calling hotspots, comparing those hotspots to known hotspots, examining nucleotide substitutions at hotspots, and comparing recombination rate variation at broad scales to patterns of genetic variation. Finally, I discuss these results in light of known biology, and suggest future avenues of research.

4.1 Introduction

As was reviewed in Chapter 1, recombination in mice, much like in many mammals, occurs in narrow regions of the genome referred to as hotspots [29, 93, 117, 118]. In humans, mice, and other species, this fine-scale localization is due to the zinc finger protein PRDM9 [25–27], which directs the recombination machinery to its binding motifs during meiosis. Individuals with different PRDM9 motifs can therefore possess completely different fine scale recombination hotspots [29].

Recombination in mice is very well studied. PRDM9 was first identified as a speciation gene in mice in 2009 [17]. In 2010, the role of PRDM9 in controlling fine scale recombination localization was found simultaneously in three studies, of which two relied heavily on analyses in mice: Parvanov *et al.* fine mapped a trans-acting recombination rate modifier locus using a cross between CAST/EiJ and C57BL/6J, which turned out to be PRDM9 [26], while Baudat *et al.* used, among other lines of evidence, polymorphism in PRDM9 in different strains of mice with different hotspots [27]. Further studies of mice recombination relied on extensive analysis of chromatin immunoprecipitation sequencing (ChIP-seq). In 2011, Brick *et al.* found that recombination localizes at testes-specific H3K4me3 marks. In 2012, Brick *et al.* showed that PRMD9 controls the localization of nearly all hotspots genome-wide, except in the pseudoautosomal region. They also showed that testes-specific H3K4me3 are induced by PRDM9, as opposed to through DSB repair, and that, at least in mice, PRDM9 effectively localizes recombination away from functional genomic elements (*e.g.* genes). In 2014, Baker *et al.* showed that PRDM9 leads to nucleosome reorganization, with PRDM9 binding leading to about 150 bp surrounding the PRDM9 motif not being bound by a nucleosome. In 2015, Baker *et al.* looked at haplotype specific initiation of recombination, revealing some history of the CAST/EiJ and C57BL/6J PRDM9 alleles [119]. They showed that the CAST/EiJ PRDM9 allele arose in *Mus musculus* before the split of *M. m. domesticus* and *M. m. castaneus*, while the C57BL/6J *M.*

m. domesticus allele arose afterwards or concurrently with the split.

Yet, despite these and additional works, questions remain about recombination in mice, particularly as it pertains to usage in different species, subspecies and populations of mice, and its consequences on patterns of linkage-disequilibrium and the direct effects of recombination on the genome. It has been shown before, and we saw in Chapter 3, that effective population sizes in mice are much larger [11] than previously studied mammals, such as humans. As individual PRDM9 alleles are typically studied in ChIP-seq, and subspecies of mice have large, somewhat overlapping PRDM9 repertoires [12, 14], the extent to which the recombination landscape are shared between subspecies is unclear. Further, it is unknown to what extent recombination is still localized into population level hotspots within the subspecies. Also, while Baker *et al.* [119] recently looked at ancestral hotspot sharing between subspecies, these are using hotspots defined by ChIP-seq for only single PRDM9 alleles and may not fully reflect PRDM9 usage within populations. Finally, in all populations, but in particular the Indian population, the direct effect of recombination on the genome, through the actions of GC-biased gene conversion on nucleotide substitution, and hence genome evolution, is unclear [120].

To address these, and further questions, recombination rate maps were therefore constructed and analyzed from the SNP callsets for the French, Taiwanese and Indian populations. Linkage-disequilibrium based recombination rate (LD-rate) map construction and initial analysis was performed by Oliver Venn. I followed up further analysis of the recombination rate map distributions within and between populations, and how they compared to ChIP-seq results from Brick *et al.* (2012, [29]) and Baker *et al* (2014, [121]). I next called hotspots in the LD-rate maps (LD-hotspots), found PRDM9 motifs, and compared both to previously published hotspots and motifs. Finally, I also sought to examine, at broader scales, the relationship between nucleotide substitutions and recombination in the different populations.

4.2 Methods

4.2.1 Rate map construction

Rate maps were constructed by Oliver Venn using the LDhat interval algorithm [122]. Briefly, before Oliver ran LDhat, I first filtered population specific callsets to remove singleton and multi-allelic SNPs. To minimize the impact of recent inbreeding, as described in Section 3.3, which could lead to distorted rate estimation in homozygous segments of the genome by spuriously increasing the influence of recombinations that took place on homozygous chromosomes while decreasing rate estimation by creating long identical haplotypes that then represent strong LD, I identified and collapsed overlapping homozygous chromosomes from different mice into surrogate "diploid" chromosomes at a 1 Mbp scale. For example, if two homozygous mice have genotypes $g_{i,1}, g_{i,2} \in \{0, 2\}$, where 0 is homozygous reference and 2 is homozygous alternate, then the new "diploid" chromosome would have $g_i^* = \frac{g_{i,1}+g_{i,2}}{2}$, with heterozygous or missing genotypes in either mouse leading to missing genotypes in the pseudo-diploid mouse. Note that this results in slight variations in the number of mice used to construct the map in each 1 Mbp interval. I then divided each 1 Mbp bin into bins of no more than 4,000 segregating sites each, overlapping by 100 segregating sites. Oliver then estimated recombination rates using LDhat for each bin using $\theta = 0.001$, a block penalty of 5, 60×10^6 iterations, a sampling period of 40×10^3 iterations, and 500 burn in sampling iterations. The individual bin rates were then combined by assigning the mean rate for overlapping sites between bins.

As in previous studies that estimated recombination rates from whole genome sequencing, for example work by Auton *et al.* [55], Oliver identified regions with artefactual, elevated recombination rate in the three populations. To correct these artefacts, Oliver found intervals with physical length > 50 kb and recombination rate in the top 0.001% of all intervals, and assigned them and the surrounding 50 SNPs the average recombination rate for that chromosome.

Population	N_e	r^2
French	20431	0.994
Taiwanese	17697	0.989
Indian	314063	0.997

Table 4.1. Estimate of effective population size using genetic maps From Oliver Venn. Estimates are from linear regression of chromosome genetic distance in each population against Cox *et al.* F1-estimated genetic distances [90].

To convert the estimated recombination rates from $\rho = (4 \times N_e \times r)$ to units of centi-Morgans (where N_e is the effective population size and r is the recombination rate), Oliver estimated the effective population size for each population through linear regression on the estimated chromosome genetic distance against the Cox *et al.* map [90]. This assumes that, at the chromosome-scale, wild mouse and the Cox *et al.* populations share the same genetic distance, *i.e.* the per-generation number of recombinations was the same in the two groups. Consequently, the slope coefficient estimates $4 \times N_e$.

4.2.2 Hotspot calling

Hotspots are loosely defined as regions of the genome with substantial elevation in recombination rate compared to background. I called hotspots for a given population as follows. First, I calculated smoothed recombination rates over 1 kbp. For example, if at position i in the original map there is rate r_i, I calculated the smoothed rate $r_i^s = \frac{1}{1001} \sum_{j=i-500}^{j=i+500} r_j$. Next, I looked for local maxima with smoothed rates greater than 10 cM/Mb. Finally, for a given local maxima, I took the boundaries of the hotspot to be the positions where the rate dropped to less than 50% of the peak smoothed hotspot rate.

4.2.3 Motif identification and refinement

I attempted to identify motifs in the French and Taiwanese populations using as of yet unpublished motif finding code written by Simon Myers. For a given set of

DNA sequences, the motif finder attempts to model the following: a set of (PRDM9) binding motifs, represented by position weight matrices (PWMs); the probability a given sequence of DNA contains either zero or one of the motifs; and where in the sequence the motif is found. Directly estimating these jointly is impractical, so a Gibbs sampler is used, which iterates by sampling through the joint probability density as follows

1. for each position of each DNA sequence, calculate the posterior probability that each motif is found. By extension, for each DNA sequence, this gives the posterior probability it contains each motif, as well as the probability the sequence contains no motif. Then sample either no motif being found, or a motif found at a specific position

2. for each motif, with its set of DNA sequences, estimate a new PWM based on the consensus overlap from DNA sequences containing that motif

3. update model parameters, such as the prior probability of motif location within sequence, and the prior probabilities of a sequence of DNA containing a motif (and by extension, no motif)

The method was used over other approaches, notably MEME [123], as it can handle a large number of wide sequences. The method also allows for joint modeling of multiple PWMs at the same time, although this feature was eventually not used.

Seeding the motif finder was done as follows. For a given set of hotspots, I defined the central portion (central 400bp) and the edges (rest) of the hotspots. Then, letting a k-mer be a string of DNA of length k, using $k = 8$, I generated counts of the number of times every k-mer was observed in hotspots centers, edges, and the rest of the genome. I then generated two sets of p-values. For the first, I compared the hotspot centers against the edges, by making a 2×2 chi-square table for each k-mer, where the first row is the counts in the centers and edges, and the second row is the same

count summed across all other k-mers with the same GC content and number of CpG dinucleotides. I then repeated this by generating p-values for the hotspots centers against the rest of the genome. From the k-mers with an odds ratio greater than one for hotspot centers against edges, I chose as a seed the k-mer with the smallest p-value for enrichment in hotspot centers versus the rest of the genome. As such, the chosen seed k-mers were enriched both centrally within hotspots and genome-wide. I next went on to generate a seed motif (PWM) from the seed k-mers by clustering together k-mers with similar central and genome-wide enrichment which were close in physical space, *i.e.* similar k-mers, which is fully described in Section 5.2.3.

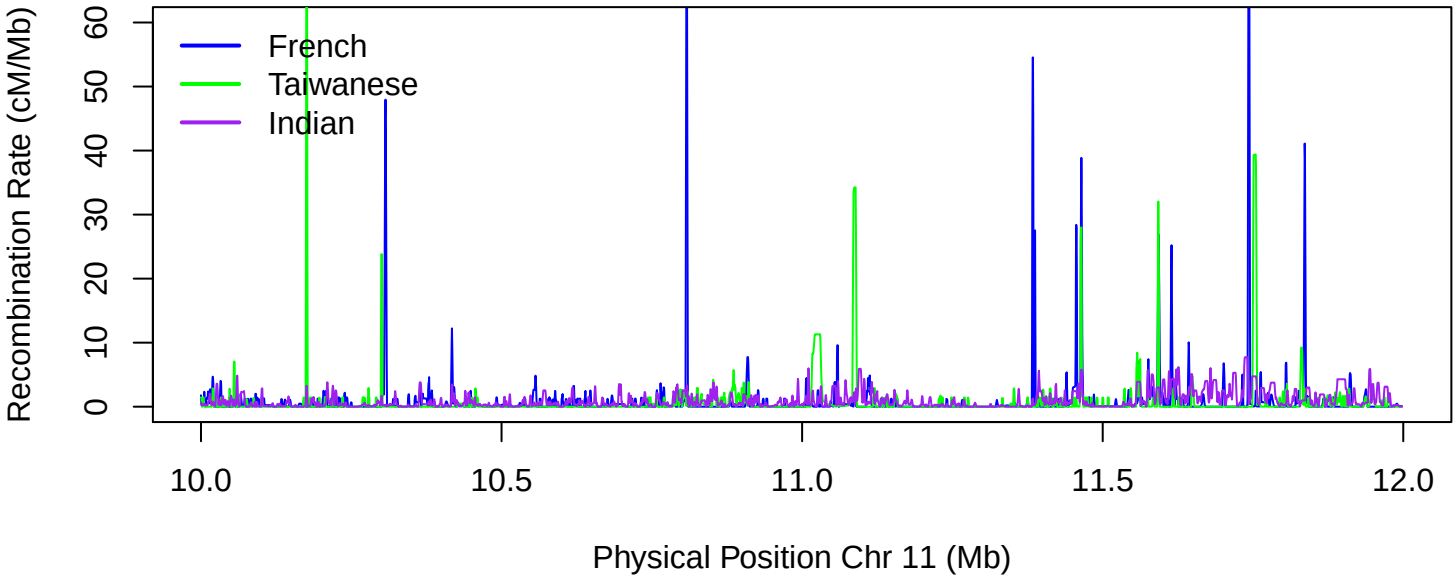
To find and refine motifs in the hotspots for both the French and Taiwanese populations separately, I began by taking the set of hotspots, getting their sequences, and masking out repeats. I then iterated PWM discovery as follows. First, I took the available hotspot set, found a seed k-mer, and generated a seed PWM. Next, I used this to initialize the motif refining code (Gibbs sampler) and ran it for 30 iterations. Finally, I kept the retained PWM, and removed hotspots which were likely to contain this motif. I iterated this procedure 4 times and found 4 PWMs. Since only one plausible PWM was found per subspecies (see Results), for each subspecies, I ran one final iteration for the plausible PWM on the entire unmasked set of hotspots without repeat masking, to discover sequences containing masked motif matches.

4.3 Results

4.3.1 Construction and validation of maps

An example of recombination rates on a 2 Mbp region of chromosome 11 for the three populations is shown in Figure 4.1. The maps for the French and Taiwanese populations showed strong clustering of recombination rates into hotspots, similar to what has been observed before in other mammals. By contrast, the Indian rate map was qualitatively flatter. This is seen in more detail by summarizing results

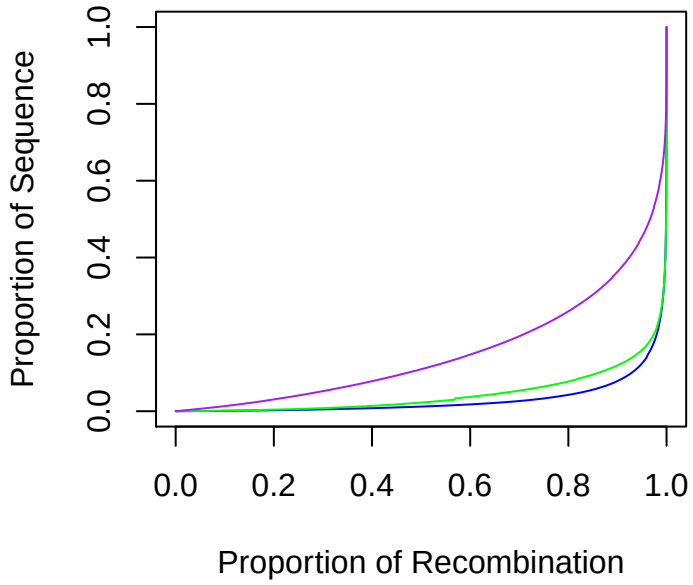
Figure 4.1. Example recombination rate along genome for the three wild populations



genome-wide and looking at what proportion of sequence explains a given proportion of recombination (Figure 4.2), with 95% of recombination occurring in the hottest 12.4% of the French genome, 15.9% of the Taiwanese genome, and 45.4% of the Indian genome. Conversely, for cold regions, defined as regions of the map where the rate was less than 10% of the genome-wide average, 69.7% of the French genome, 69.6% of the Taiwanese genome, and 33.0% of the Indian genome was cold.

If we consider the Indian rate map as a superposition of many PRDM9 alleles, which is not implausible given the larger population size and known PRDM9 diversity, one might expect cold regions in the Indian map to reflect some genomic regions with particular resistance to the recombination machinery. I therefore looked at what percentage of cold regions of the genome in Indian mice were similarly cold in the French and Taiwanese population, as defined in the previous paragraph. While elevated, at

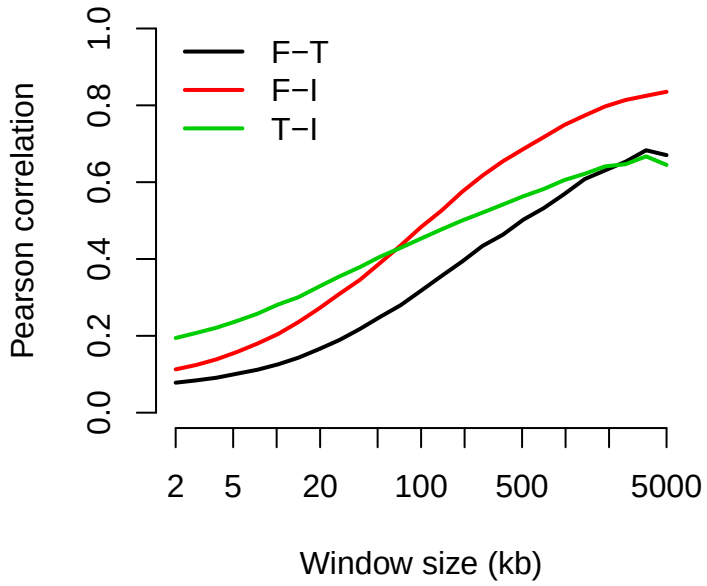
Figure 4.2. Proportion of recombination in the maps explained by a given proportion of sequence Colouring as in Figure 4.1. Interpretation for a given horizontal (x) and vertical (y) position is that the hottest fraction x of the map is distributed over the fraction y of the genome.



73.5% (versus 69.7%) and 73.6% (versus 69.6%) for the French and Taiwanese populations, respectively, it is only a modest enrichment compared to the genome-wide French and Taiwanese values.

I also compared maps to each other. Correlations between maps are weak at fine scales (< 10 kbp), and high at broad scales (> 1 Mbp) (Figure 4.3). At fine scales, the Taiwanese and Indian maps show the highest (albeit weak) correlation in rates, but this is not mirrored at the highest scales, where the correlation is highest between the French and Indian maps. The correlations seen here are different to similar landscapes observed among human groups, and instead mirror inter-species differences seen between humans and chimpanzees [55] and imply little overlap in recent ancestral *M. m. domesticus* (French) and *M. m. castaneus* (Taiwanese) PRDM9 alleles.

Figure 4.3. Correlation between wild recombination rate maps at varying scales Correlation (r) is shown for a variety of window sizes. F = French, T = Taiwanese, I = Indian

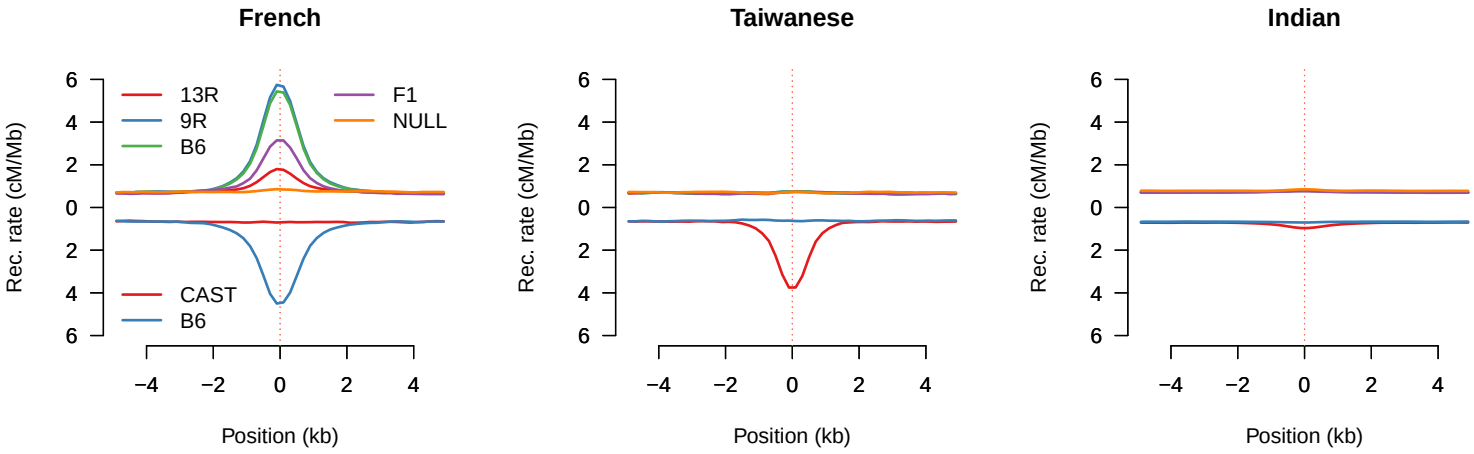


4.3.2 Hotspots

To look at hotspot usage, and characterize differences between subspecies, I began by looking at previously defined hotspots based either on DMC1 ChIP-seq (Brick *et al.* 2012 [29]) or meiosis specific H3K4me3 ChIP-seq (Baker *et al.* 2014 [121]). Recall from Chapter 1 that DMC1 is a protein that assembles at the sites of double stranded breaks during meiosis, while H3K4me3 deposition by PRDM9 in meiosis is a precursor of SPO11 double strand break induction. I will generally refer to these previously published sets of hotspots as "known"-hotspots, to distinguish the ones I later generate from the linkage-disequilibrium (LD) based recombination rate maps, which I call LD-hotspots.

At the known-hotspots, in the French recombination rate map, there is a clear increase in rate at previously identified *M. m. domesticus* hotspots, while we see no

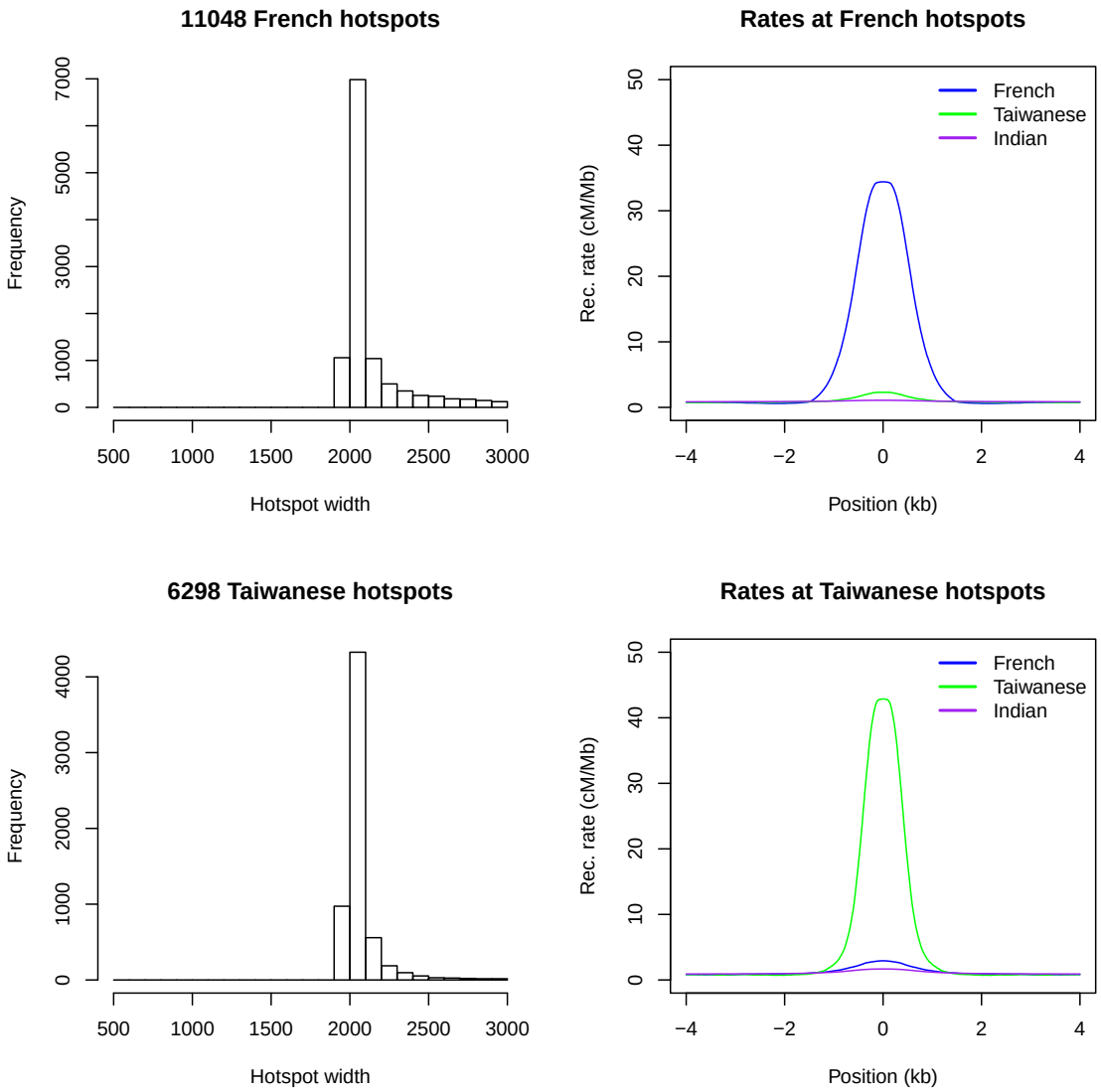
Figure 4.4. Intensity of rate maps at previously published hotspots
Upper; tests DMC1 ChIP-seq [29], F1=9R/13R, NULL=PRDM9 double knockout: lower; testes PRDM9 specific B6 and CAST H3K4me3 marks [121]



rate increase in the French map at *M. m. castaneus* defined hotspots (Figure 4.4). There is a similar result for the Taiwanese rate map, with elevations at previously defined *M. m. castaneus* hotspots but no such increase at *M. m. domesticus* hotspots. These results corroborate the previous results that the hotspot landscape of the French and Taiwanese mice is largely distinct (Figure 4.3). The strength of these hotspots in general is comparatively weak, compared to the example Figure 4.1. For the Indian population, there is a barely perceptible increase in recombination rate at the *M. m. castaneus* defined known-hotspots, similar to the slightly higher correlation at fine scales seen between Taiwan and India (Figure 4.3).

I next called LD-hotspots based on the linkage disequilibrium based recombination rate maps (Figure 4.5). With the definition of hotspot adopted in this thesis, there are 11,048 French LD-hotspots, 6,298 Taiwanese LD-hotspots, and 0 Indian LD-hotspots. The widths of these hotspots are distributed tightly at about 2,000 bp. The average Taiwanese rate is slightly higher in their LD-hotspots, at just over 40 cM/Mb, compared to the French rate in their LD-hotspots, at about 35 cM/Mb. Overlap

Figure 4.5. Features of hotspots called from rate maps Histogram show hotspot widths binned per 100 bp, while rate plots show rates in the three populations at the hotspots of a given population. Summary rates on the right are not otherwise smoothed.

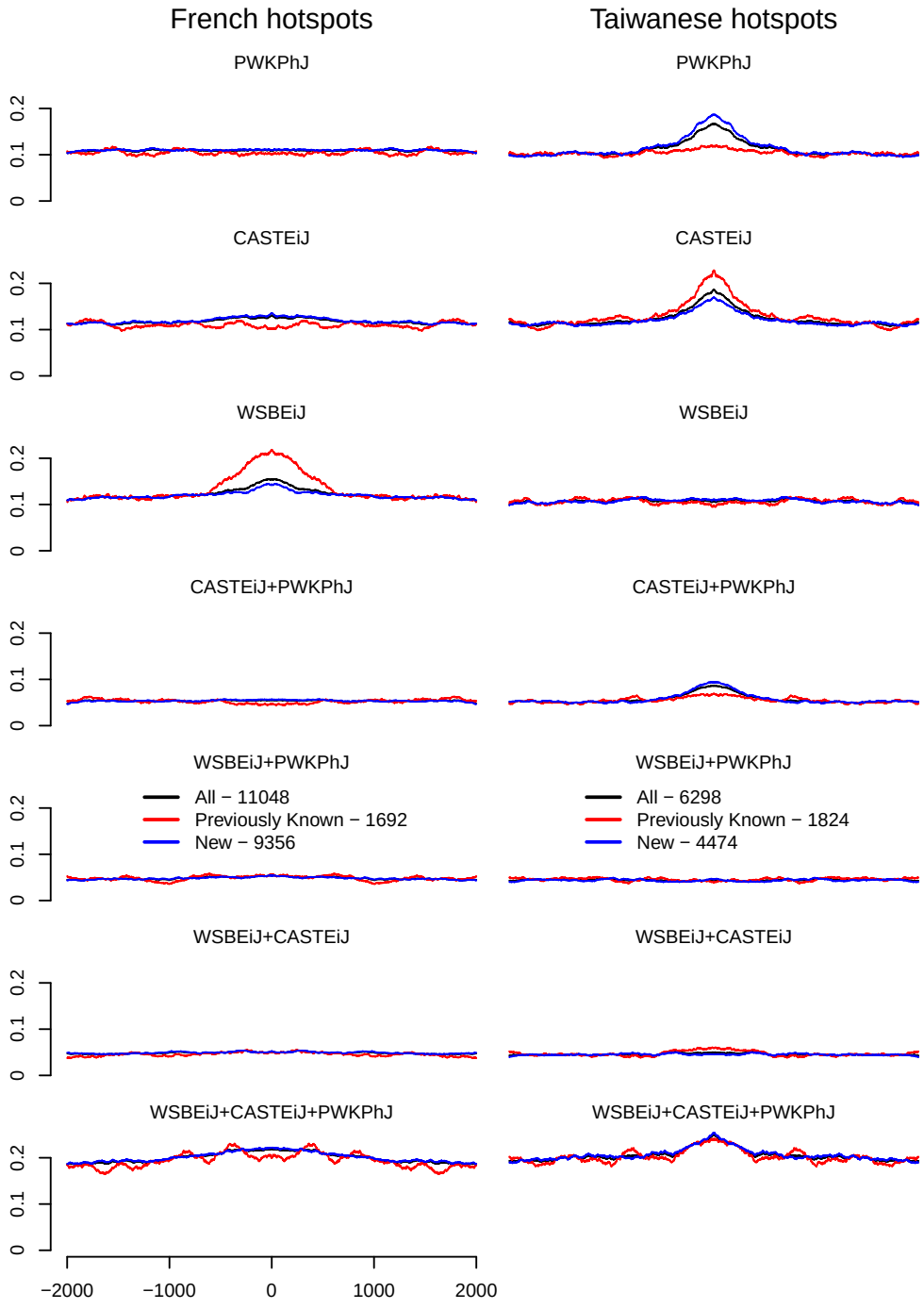


between LD-hotspots from the two populations is low, with 4.5% of French LD-hotspots overlapping Taiwanese hotspots and 7.8% of Taiwanese hotspots overlapping French hotspots. Recombination rates at the LD-hotspots were the same whether or not they were conditioned on also being a known hotspot (not shown). Note that the lack of hotspots in the Indian population is not due to bounding cutoffs used when generating the rate maps with LDhat, as the recombination rate bound in cM/Mb in the Indian population was the same as the French and Taiwanese populations.

To validate the LD-hotspots, I looked for local deviations in nucleotide substitution from weak (adenine or thymine) to strong (guanine or cytosine) bases in the hotspots. The motivation for such validation is that GC-biased gene conversion operates to promote fixation during meiosis at PRDM9 binding motifs [120], and so one would expect to see increases in weak to strong substitution rates in the LD-hotspots. I looked for fixed substitutions down combinations of one of the three representative wild derived inbred strains, WSB/EiJ from *M. m. domesticus*, CAST/EiJ from *M. m. castaneus* and PWK/PhJ from *M. m. musculus*, as compared to an ancestral sequence using *M. spretus*, *M. caroli* and *M. famulus*. I then classified variants uniquely with derived or ancestral status, and compared rates of increase of derived weak to strong substitutions at the hotspot sets. I further stratified the LD-hotspots into whether they intersected with the known hotspots, or were novel.

Results are shown in Figure 4.6, and yield several interesting findings. First, French LD-hotspots show local increases in weak to strong fixation in WSB/EiJ, as do the Taiwanese mice for CAST/EiJ, results that are in broad agreement with expectations, and validate the LD-hotspots. Second, at the LD-hotspots, the strength of fixation is stronger in the subset of these that are also in the known-hotspots, than for those that are novel. Multiple factors could explain this, for instance false positives hotspots or a large repertoire of rarer PRDM9 motifs with consequentially weaker GC-biased gene conversion, due to the more infrequent binding, and hence

Figure 4.6. Weak to strong substitutions around hotspots called from rate maps Shown are the fixation of weak to strong substitutions in the three *Mus musculus* wild derived inbred strains, with WSB/EiJ representing *M. m. domesticus*, PWK/PhJ representing *M. m. musculus* and CAST/EiJ *M. m. castaneus*. Variants were classified uniquely according to the 7 categories shown here. Values are smoothed 100 bp either side. Intersection with previously known is against strain specific H3K4me3 marks from Baker *et al* [121].



GC-bias, at these hotspots during meiosis. Third, there is an elevated rate of weak to strong substitutions in PWK/PhJ at the novel, but not the known, Taiwanese LD-hotspots. These results are consistent with sharing of some of the Taiwanese hotspots, and hence PRDM9 motifs, between *M. m. castaneus* and *M. m. musculus*, although the lack of increase at known LD-hotspots indicates the particular allele from CAST/EiJ marking the known-hotspots is inactive in *M. m. musculus*. The results further confirm that the PRDM9 motifs active in *M. m. domesticus* and *M. m. castaneus* have been inactive in the reciprocal subspecies since their split. Fourth, an increase in the rate of fixation of variants present across all 3 inbred strains of mice in the Taiwanese hotspots suggests this allele was active in the *Mus musculus* ancestor, and the lack of signal down the WSB/EiJ *M. m. domesticus* line suggests its loss down that subspecies lineage. Finally, it is worth noting that the signal at the Taiwanese LD-hotspots is tighter in spread than the French LD-hotspots, which may reflect differences in PRDM9 biology, but alternatively could just reflect that the Taiwanese rate maps, and hence LD-hotspots, are more precisely defined due to the greater SNP density in that population.

I next tried to find motifs enriched in the French and Taiwanese LD-hotspots. To recap from the Methods, the idea is to iteratively find motifs on successive subsets of the LD-hotspots. To start, I would take the available LD-hotspots, mask out repeats, use the seeder code to get a starting position weight matrix (PWM), and then use the Gibbs sampler to refine them. After each iteration, I would set aside the finished motif, as well as remove from consideration any hotspot sequence that had a high posterior probability of containing the motif. I would then continue with the remaining set of LD-hotspots to try and identify another motif.

I ran the iterative approach for finding motifs 4 times for each of the French and Taiwanese populations, and obtained one plausible PRDM9 motif for each of the French and Taiwanese populations, as well as 3 less plausible motifs. The best French

Figure 4.7. Best French motif from LD-hotspots From left to right, the panels are:a histogram across the LD-hotspots for the posterior probability the hotspot contains a motif; the prior probability on location within the LD-hotspots for the motif location; the forward PWM; the reverse complement PWM. On top of the posterior probability histogram, the number of hotspots sampled in the final iteration of the Gibbs sampler to construct the rightmost three panels is listed, as well as those with $> 50\%$ probability to contain a motif.

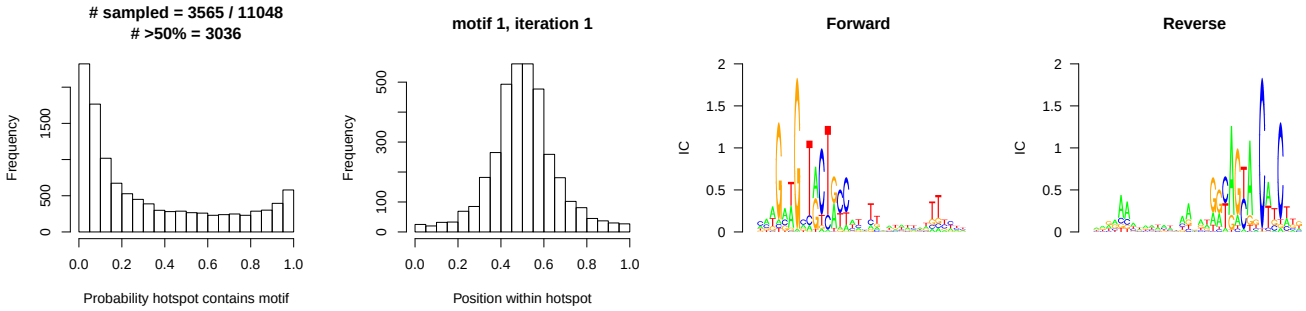


Figure 4.8. Example of bad French motif from LD-hotspots Panels as in Figure 4.7

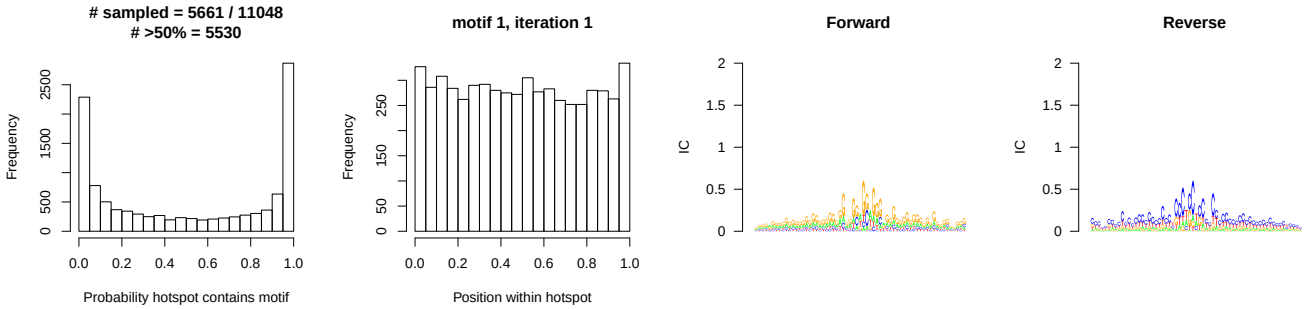
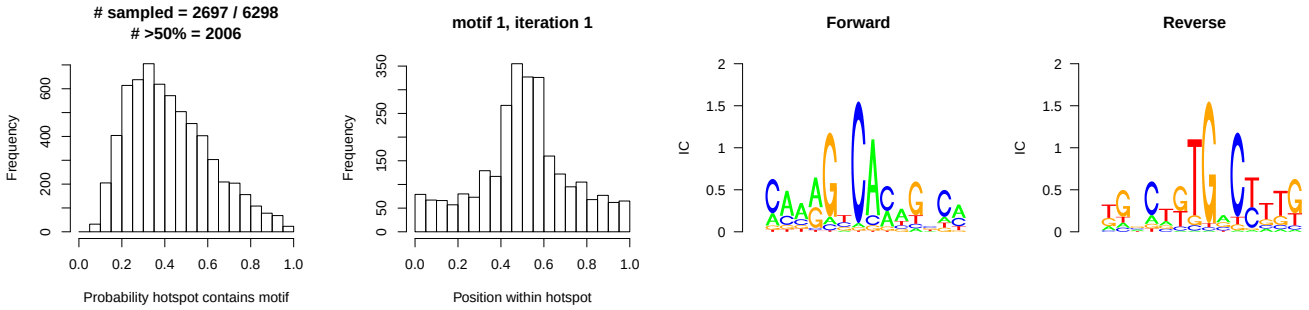


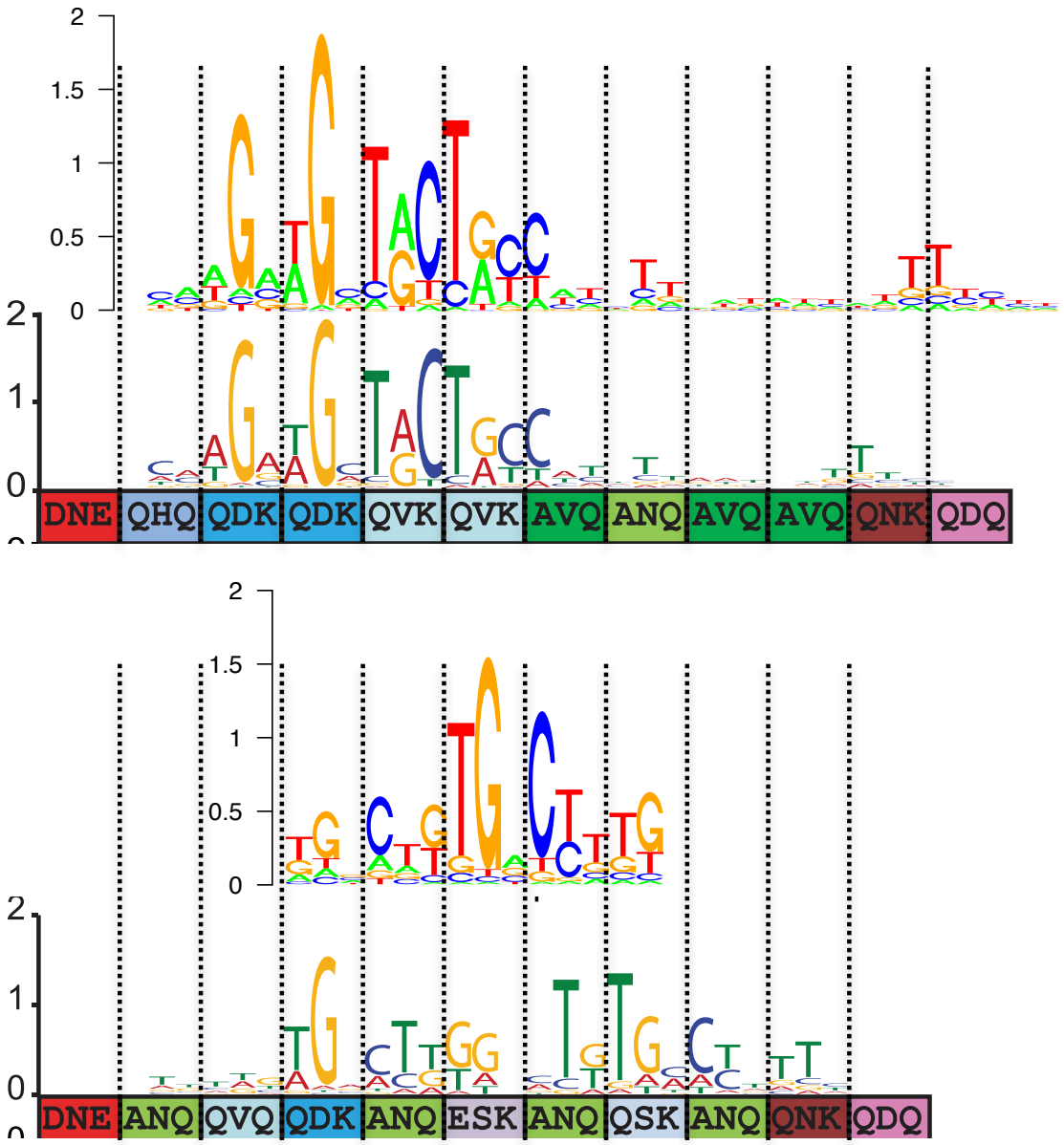
Figure 4.9. Best Taiwanese motif from LD-hotspots Panels as in Figure 4.7



motif is shown in Figure 4.7, where plausibility stems from the central enrichment of motif hotspot binding positions within the DNA sequences (due to crossovers being centrally distributed with peak probability at the motif location), shown in the second to left panel, as well the general complexity of the motif. Using this approach, the best French motif is found more likely than not ($> 50\%$ posterior probability) in 3,036 out of the 11,048 hotspots (27%). By contrast, some "motifs" found through the iterative procedure are low-complexity poly-A / poly-T or poly-C/poly-G, of which an example is shown in Figure 4.8. These low-complexity motifs also usually feature an absence of central enrichment, and are likely "real" insofar as they are more common then expected under the background model, due to deficiencies in its ability to handle higher order local DNA structure (a known deficiency of the motif finder). The best Taiwanese motif is shown in Figure 4.9, and much like the French motif, shows strong central enrichment, and high complexity motif. It is found in 2,006 out of 6,298 hotspots (32%). Incidentally, the poly-A k-mer was found to be enriched in both French and Taiwanese hotspots versus genome-wide, but without increased representation in hotspot centers versus edges.

I next compared the motifs found from the LD-hotspots to those previously found in 2015 in Baker *et al.* [119], as seen in Figure 4.10, and they show strong agreement, although the Taiwanese motif identified based on the LD-hotspots has a few

Figure 4.10. Best recombination motifs aligned to known motifs Known motifs reproduced from Baker *et al.* [119]. Top set is *M. m. domesticus* results, with French LD-hotspot motif above result from Baker *et al.*, while bottom set is *M. m. castaneus* results, with Taiwanese LD-hotspot motif above result from Baker *et al.*



discrepancies with the Baker *et al.* motif.

	No motif	Yes motif	Absent LD-hotspots
13R	3.2	0.7	96.1
9R	5.0	7.2	87.8
B6	4.7	6.8	88.5
F1	4.0	2.8	93.3
PRDM9_NULL	1.4	0.8	97.8
CAST	1.4	0.5	98.1
DOM	3.8	5.5	90.7

(a) French normalized against known-hotspots

	No motif	Yes motif
13R	5.7	3.4
9R	8.7	32.9
B6	10.1	38.2
F1	7.2	13.2
PRDM9_NULL	4.2	6.1
CAST	4.8	4.1
DOM	8.6	33.0

(b) French normalized to those counts of LD-hotspots with or without motif

	No motif	Yes motif	Absent LD-hotspots
13R	0.9	0.4	98.7
9R	0.9	0.4	98.7
B6	0.9	0.4	98.7
F1	0.8	0.4	98.8
PRDM9_NULL	0.6	0.6	98.8
CAST	4.0	2.6	93.4
DOM	0.6	0.2	99.2

(c) Taiwanese normalized against known-hotspots

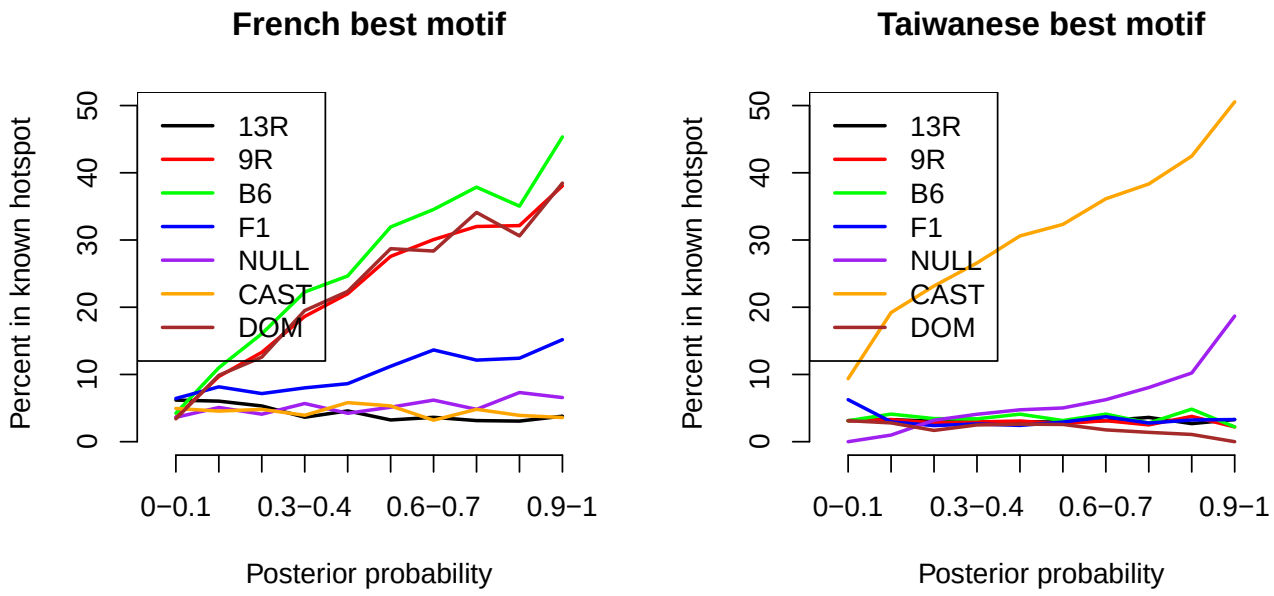
	No motif	Yes motif
13R	3.0	3.1
9R	3.0	2.8
B6	3.7	3.4
F1	2.6	3.1
PRDM9_NULL	3.5	7.0
CAST	25.4	36.1
DOM	2.3	1.9

(d) Taiwanese normalized to those counts of LD-hotspots with or without motif

Table 4.2. Intersecting known hotspots with LD-hotspots Normalization against known (a) stratifies the known-hotspots into three categories: those intersecting LD-hotspots with the motif, those intersecting LD-hotspots without the motif, and those that don't intersect LD-hotspots. Similarly, normalization against the counts of LD-hotspots with or without the motif (b) counts what percent of LD-hotspots also intersect a set of known-hotspots, given they either do, or do not, contain the motif. Known-hotspots sets are: upper 5, testes DMC1 ChIP-seq [29], 13R, 9R are strains, B6 = C57/BL6, F1=9R/13R, NULL=PRDM9 double knockout; lower; testes PRDM9 specific C57/BL6 and CAST/EiJ H3K4me3 marks [121]

I next took the LD-hotspots and motifs, and intersected them with the DMC1 and H3K4me3 known hotspots (Table 4.2). I see strong enrichment for the LD-hotspots containing the best motifs with the known-hotspots. For instance, for the known H3K4me3 results, about 10% of C57BL/6 known-hotspots intersect with French LD-hotspots, while only 2% of CAST/EiJ known-hotspots intersect French LD-hotspots. There is weak, but present, enrichment for the 13R motif defined hotspots, echoing the mild increase in recombination rate seen earlier in Figure 4.4. Results are similar

Figure 4.11. Intersecting LD-hotspots stratified by posterior motif probability against known hotspots The LD-hotspots are stratified by the posterior probability they contain a motif, and then for each bin of posterior probability, those hotspots are compared to those intersecting various classes of known-hotspots.

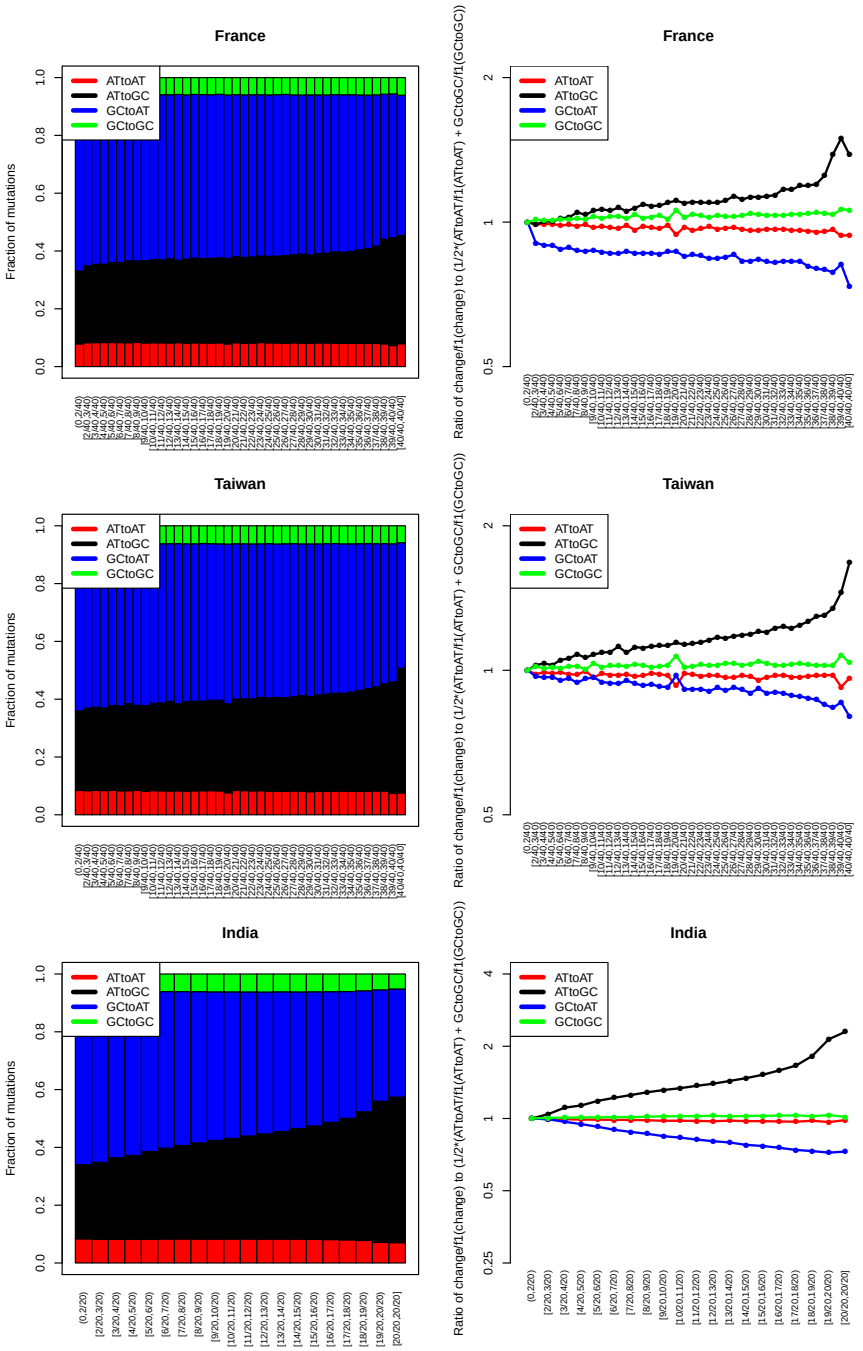


for the CAST/EiJ, albeit more moderately, for the Taiwanese hotspots. If we look specifically at what percent of the motif containing LD-hotspots are seen in known-hotspots, enrichment is more impressive, with about 30-40% of French LD-hotspots containing the best PRDM9 motif intersecting *M. m. domesticus* derived known-hotspots, with the same seen for the Taiwanese LD-hotspots against *M. m. castaneus* derived known-hotspots. This becomes clearer if we stratify this intersection by posterior probability of having the motif in each population, as shown in Figure 4.11. These results confirm that LD-hotspots that are more likely to contain a high quality match to the motif are more likely to intersect known-hotspots, with about 40-50% of the motifs most likely to contain the best motif intersecting known-hotspots. Intriguingly, a larger number of Taiwanese LD-hotspots that are a good match for the PRDM9 motif intersect hotspots defined by the PRDM9-null mouse from Brick *et al.* [29].

4.3.3 Effect of recombination on nucleotide substitution patterns

In Figure 4.6, we saw that GC-biased gene conversion operates in mice locally around PRDM9 defined hotspots. I next sought to more broadly characterize the role of GC-biased gene conversion in the mice. First, I stratified variants (SNPs) at different frequencies into one of four simple mutational classes, based on ancestral allele: weak to weak (WW) (A→T, T→A), weak to strong (WS) (A→C, A→G, T→C, T→G), strong to weak (SW) (C→A, C→T, G→A, G→T), or strong to strong (SS) (C→G, G→C). I then determined what proportion of variants at a given derived frequency were one of the four mutational classes (Figure 4.12). Results show a clear relationship between allele frequency and mutational class. For example, in the Indian population, for the lowest frequency variants, 60.0 % of variants are SW and 25.6% are WS, while at the highest frequency (but not fixed), 37.3% of variants are SW and 50.4 % of variants are WS.

Figure 4.12. Nucleotide skew by derived allele frequency Left plot shows fraction of SNPs at a given frequency in that population stratified by ancestral (weak or strong) to derived (weak or strong) substitution. Right panel is similar to left, except for each frequency, the SW and WS are normalized to the neutral (WW and SS) changes, which are then further normalized against those the leftmost, lowest frequency bin.



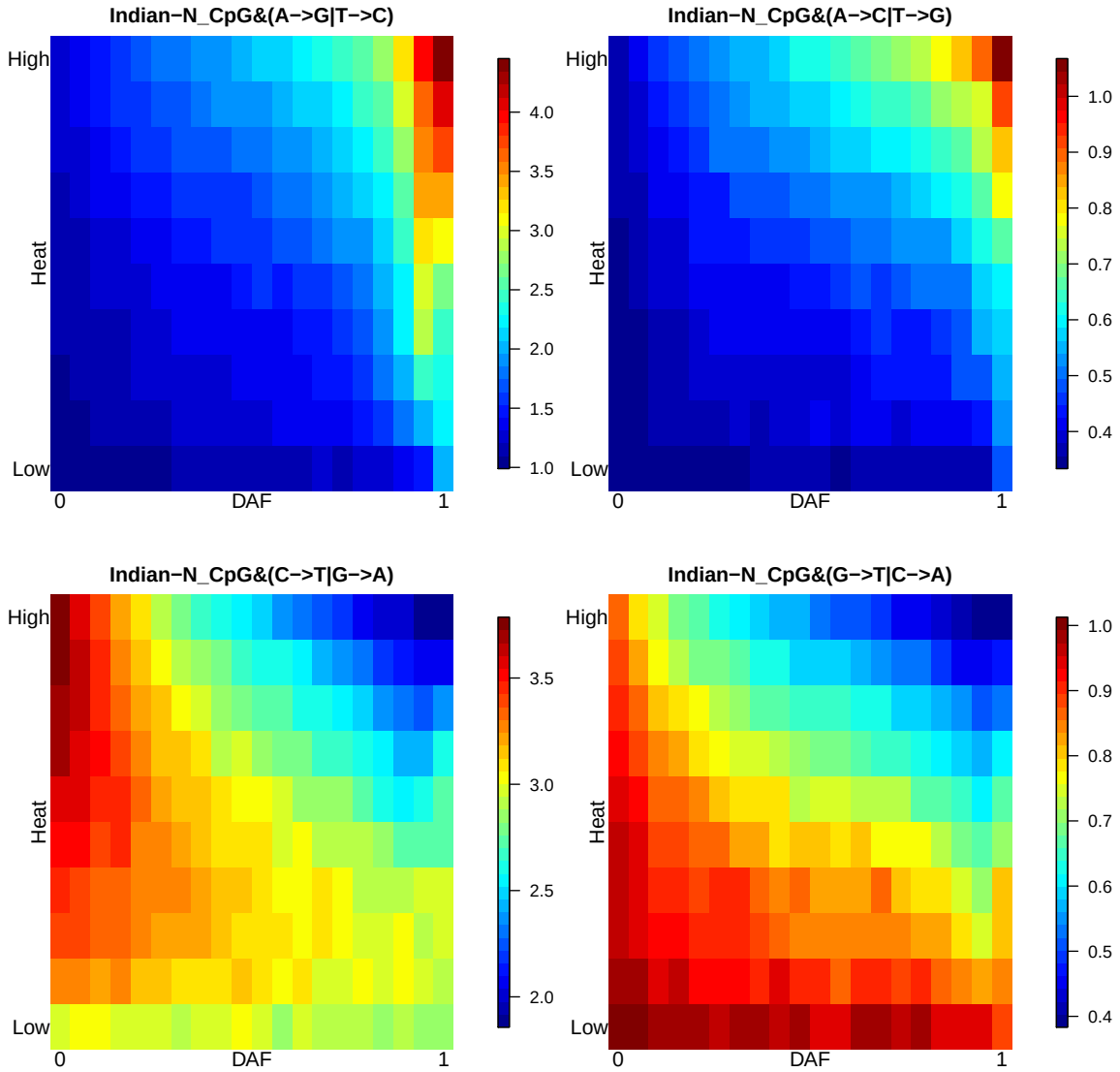
To get a better sense of the level of enrichment this represents, I then normalized results by dividing the SW and WS counts by the neutral counts (WW and SS), for each frequency bin. I then compared each of these normalized values against the normalized value for the lowest frequency bin. For example, if $n_{i,x}$ is the number of variants at frequency i with change x, and j is the lowest frequency, then I calculated the ratio

$$r_{i,x} = \frac{\frac{n_{i,x}}{n_{j,x}}}{\frac{1}{2}(\frac{n_{i,WW}}{n_{j,WW}} + \frac{n_{i,SS}}{n_{j,SS}})} \tag{4.1}$$

This allows for the determination for a given allele frequency of how enriched or depleted a given mutational class is compared to the lowest frequency bin, which will approximate the distribution of new mutations which are of each mutation type. Results are shown in the right panels of Figure 4.12. This more clearly highlights results than the unnormalized panel, as there is a smooth increase in the presence of WS mutations at high frequency among all populations. This is especially pronounced in the Indian population, where there are 213% as many WS variants at allele frequencies approaching fixation as opposed to when they originated, while for SW, there are 72% as many.

To gain better insight into the relationship between recombination and GC-bias, I made a series of plots comparing substitution type, transition versus transversion, recombination heat and allele frequency. To begin, I divided the genomes of the different populations into recombination heat deciles. Then, for each SNP in the genome that varied among *Mus musculus* mice, I recorded i) what recombination rate decile it was in ii) its derived allele frequency iii) its ancestral and iv) derived base. Then, for each pair of recombination heat decile and derived allele frequency, I could compare any desired substitution type against the neutral weak to weak and strong to strong changes. For example, for the Indian population, there were 84,443 weak to strong transversions in the coldest recombination decile in the genome at 5% frequency, while there were 123,032 weak to weak and 73,328 strong to strong in

Figure 4.13. Association between recombination heat and nucleotide skew for Indian population Panels represent a different substitution class, listed on top of the panel, normalized against neutral SS and WW changes. Left panels are transitions, right panels are transversions. Top panels are WS, bottom panels are SW. For a given panel representing a nucleotide substitution, for each rectangle representing a recombination decile of the genome and a derived allele frequency (DAF), the colour, given in the panel specific legend, represents the ratio of the number of SNPs of that substitution class against the neutral SS and WW changes, normalized against the number of available bases in the ancestral genome at that recombination rate.



the same bin, for a ratio of $\frac{84,443}{123,032+73,328} = 0.43$. These ratios are further adjusted by normalizing each of the different nucleotide substitution classes depending on the number of eligible ancestral bases in that recombination rate decile of the genome, to minimize the correlation between recombination rate and GC content in this inference. Calculating this ratio across all recombination rate deciles and frequencies, I could visualize results across multiple substitution types and recombination rates.

Results for the Indian population for four classes of substitutions are shown in Figure 4.13, while the French and Taiwanese plots are omitted for space reasons, although they are broadly similar to the Indian results, albeit with slightly more noise and reduced strength. These results are partly a more specific version of the general Figure 4.12, but also reveals some associations not visible previously. First, for WS, there is a smooth increase in the strength of GC-biased gene conversion promoting allele fixation at higher recombination rates, while the converse is true for SW, where allele fixation is suppressed. More interestingly, there is a marked difference in the relationship between recombination heat at low frequencies for SW transitions and transversions, with the transitions showing an increasing ratio to the neutral class with recombination heat at low frequencies, implicating recombination as being mutagenic for SW transitions, but not transversions.

4.4 Discussion

4.4.1 Recombination rate variation and hotspots

I this chapter, I began by presenting analysis of the recombination rate maps of 3 wild populations of mice. The recombination rate maps from the French and Taiwanese populations revealed largely flat landscapes punctuated with short, large increases in the recombination rate. These results are qualitatively similar to previously studied mammalian rate maps, such as those for humans [93] and chimpanzees [55]. By contrast, the Indian population had a very flat landscape, devoid of such obvious pop-

ulation hotspots, likely reflecting the actions of a large repertoire of PRDM9 alleles segregating in the population. At broad scales, recombination rate correlations between populations were similar between all three pairs of populations, indicating that factors influencing broad scale recombination rates are conserved between mice subspecies. Given their highly punctuate distributions, I was able to identify LD-hotspots in the French and Taiwanese, but not the Indian populations. These LD-hotspots were validated by showing they contain an increase in weak to strong nucleotide substitution rate down their respective subspecies lineages. By further comparing the hotspots to known hotspots, and looking in representative strains from all three *Mus musculus* subspecies, I could comment more specifically on hotspot use in the different subspecies. I was also able to identify a single dominant PRDM9 motif for both the Taiwanese and French mice that matched previously identified motifs, although this represented only a minority of hotspots in each population.

The lack of population level hotspots in the Indian population, despite the likely presence of hotspots in individual Indian mice, marks the first time a species with functional PRDM9 has been shown to have a recombination rate landscape devoid of population hotspots. This does suggest that practically speaking, consistent hotspot location between individuals in a species isn't necessary, and as such, speciation through PRMD9 isn't as simple as a new PRDM9 allele arising and quickly leading to reproductive isolation. On a separate note, the broad recombination landscape, and hence high PRDM9 diversity, in addition to the high levels of genetic diversity, does suggest that Indian mice would likely be a good resource for generating new strains of laboratory mice for genetic mapping studies. Also, on a more practical consequence, this empirical demonstration of a lack of population level hotspots in a population with a large effective population size suggests that the use of certain programs, such as ms or msHot, to model these populations should be done with caution, depending on the analysis goals, as several assumptions will certainly be

broken.

While the Indian population didn't have population level hotspots, I was able to find hotspots in the French and Taiwanese populations, and from these, one dominant motif for each population, which broadly matched known motifs from ChIP-seq. While the ability to find even one motif from these hotspots marks a methodological improvement from previous studies (discussed below), the inability to find additional motifs perhaps reflects population history of these populations. Both the French and Taiwanese populations went through recent, extreme bottlenecks. Combined with detailed inspection of sampling information on PRDM9 variants [12, 14], it is likely that over the last 10,000 years, recombination in these populations occurred principally through the dominant population variant, which I identified here. The remaining hotspots likely reflect a larger number of less frequently used PRDM9 alleles, results which are in concordance with the nucleotide substitution results at known and novel hotspots, assuming a low burden of false positive hotspots. It is also worth noting that the strongest k-mer observed in both the French and Taiwanese populations was poly-A, although this wasn't centrally enriched. This is similar to recent work in birds, which also found poly-A's enriched in hotspots, much like in yeast and *A. thaliana*, which might be related to nucleosome depletion or remodeling [124]. As for the comparison between known-hotspots and LD-hotspots, this was lower than might be expected. In particular, only about 10% of ChIP-seq hotspots were detected as LD-hotspots. This likely reflects the relatively small number of DNA sequences used to construct the rate maps, and that many PRDM9 binding motifs that could be bound in meiosis weren't used for crossovers in the history of the samples. For LD-hotspots, only 30-40% of them with a motif intersected known-hotspots, a proportion that increases to only around 50% with the best matches to the motif. This relatively low fraction might perhaps be due to lower H3K4me3 or DMC1 ChIP-seq sensitivity than thought, or only a partial match to the full motif being found, and

hence inefficient stratification into hotspots that contained a strong or weak match to the full motif obtained, or a combination of these and other factors.

We also saw in the results from the nucleotide substitution that the hotspots from the Taiwanese population were also active in *M. m. musculus*, based on an increased in weak to strong base fixation in PWK/PhJ. This furthers the *Prdm9* sampling work we saw in the Introduction [12, 14], where we now see direct evidence for sharing of hotspots, and hence motifs, between *M. m. musculus* and *M. m. castaneus*, to the exclusion of *M. m. domesticus*, with alleles likely present in the ancestral *Mus musculus*. By contrast, the alleles in *M. m. domesticus* likely arose after the split with *M. m. castaneus* and *M. m. musculus*. Further sampling, sequencing and recombination rate map construction would help clarify the relationship between active hotspots in *M. m. musculus* and *M. m. castaneus* mice.

This data may yet yield further insights with additional analyses. It would be beneficial to study known-hotspots from ChIP-seq either used or unused in the recombination rate maps, and whether any features can be discerned to explain why some motifs are able to be bound by PRDM9 and lead to crossovers during meiosis, and others aren't. Further, given the H3K4me3 and DMC1 maps are published, fine-scale localization around active hotspots with historical recombination patterns could comment on whether there exists any relationship between chromatin marks and variability in hotspot width, and hence crossover gene conversion tract length. Finally, in addition, it might prove interesting to examine the cold regions in multiple maps, especially the Indian map, in more detail. The subtlety of the overlap between cold regions in the Indian mice that are similarly cold in the French and Taiwanese implies that the majority of the signal is caused through stochasticity of the PRDM9 binding landscape in the Indian mice. Nonetheless, this slight enrichment may yield insight into finescale chromosomal features that lead to PRDM9 or other elements of the recombination machinery being unable to act at those locations.

It is also worth noting that the combination of seeding and refining motifs done here managed to find a strong motif for both the French and Taiwanese populations. This was done without any prior knowledge of the motifs. Further, attempts to find a motif using MEME were unsuccessful on this data. While ChIP-seq based datasets are often able to find motifs, they generate several times more data (for example 15,000 to 20,000 [29] hotspots versus the 5,000 to 10,000 seen here), with substantially narrower resolution, of about 300 bp versus about 2,000 bp. Researchers have struggled to find motifs before in human [93, 117] or chimpanzee data [55]. Therefore, the highly sensitive methods used here will be a powerful way to identify further *cis* influences on recombination activity, be they PRDM9 motifs or otherwise.

4.4.2 Nucleotide substitution bias

In the results, we saw that the landscape of recombination in Indian mice is widely distributed, and that it strongly promotes the fixation of strong variants at the expense of weak ones. Whether a substitution is weak to strong or strong to weak is the dominant influence on fixation probability, equivalent to a 3-4 fold difference, a level further increased in the most recombinogenic portions of the genome. This marks the first time that GC-biased gene conversion has been shown to operate at once over such a wide range in a mammalian genome, although similar results have been seen for insects, such as in honeybees [37]. Here, we see that not only is GC-biased gene conversion shown to have a stronger effect than seen previously in mammalian systems, due to the large Indian effective population size, but it shows that its effects will be distributed broadly genome-wide. This is important as it shows that a large fraction of the genome is susceptible to meiotic drive, and hence pressures mimicking selection, even if the substitutions have no direct effect on phenotype.

The results shown here also further implicate recombination in mutagenesis in mice, although the mutagenic effect is weaker than the effect of GC-biased gene

conversion. As of yet, no attempt has been made to survey the landscape of *de novo* variation in mice, or its causes, as it has in humans (*e.g.* [125]). In humans, results from studying *de novo* variation, have generally assessed a relatively small number of *de novo* variants, and have generally not implicated recombination in mutagenesis. More recently, Pratto *et al.*, by studying personalized recombination rate maps, generate results that strongly suggest recombination is mutagenic [39]. The results shown in this chapter similarly suggest that recombination is mutagenic in recombination in mice, specifically for strong to weak transitions.

This avenue of research also has important consequences on interpretation of the effects of selection on the genome. We see here that GC-biased gene conversion has a broad influence on allele fixation over the genome, however, I have made no effort to quantify its effect on, and relationship with, traditional selection on either coding or regulatory variation. Previous analysis, using the same Indian mouse sequencing data [3], has evaluated the strength and distribution of selection in mice, and shown that although the majority of selected sites are regulatory, the majority of phenotype evolution acts through changes to coding variation. However, this analysis did not account for the effect of GC-biased gene conversion, possibly due to a belief it would act on only a small fraction of the genome, like what is seen in the recombination landscape in other mammals, and possibly due to modeling simplicities. Given GC-biased gene conversion exerts an influence on allele fixation through a transmission distorting effect that is similar to selection, it will promote the fixation of weakly deleterious variants. Research by Galtier *et al.* in this area found that highly differentiated primate lineage specific functional substitutions often showed signatures of GC-biased gene conversion [126]. Therefore, given the strength, range, and known potential for GC-biased gene conversion to cause protein coding changes, it would be wise to model the effects of GC-biased gene conversion, either directly or indirectly through the recombination landscape, when estimating the broad scale effects of se-

lection in the genome. Further, both selection and GC-biased gene conversion are at their most efficient when population sizes are large, as they are proportional to effective population size through either $4N_e s$ or $4N_e b$, where s is the effect of selection and b the meiotic drive effect of GC-biased gene conversion. The diversity of recombination rate genome-wide should provide an opportunity to disentangle these effects, much like I was able to disentangle mutagenic effects on particular classes of nucleotide substitutions. Therefore, in the future, it would be informative to see to what extent estimates of the distribution of fitness effects in mice, and in other species, change once GC-biased gene conversion is accounted for in the model, and as a corollary, what percent of adaptive protein coding variation is weakly deleterious and driven by GC-biased gene conversion.

Chapter 5

The ancient landscape of recombination in mice

In Chapter 4, I used whole genome sequencing data for the 3 populations of wild mice to build linkage disequilibrium based recombination rate maps. I used these maps to show that the French and Taiwanese rate maps are organized largely into disjoint hotspots, and identify subspecies specific PRDM9 motifs. I also showed on a broad scale that GC-biased gene conversion is active in all populations, especially in the Indian population. Here, I use whole genome sequencing data on the 6 wild derived inbred strains representing different species and subspecies of mice. I use the principles of biased and GC-biased gene conversion to identify k-mers with unusual loss properties due to the two forces. I then cluster those k-mers into species lineage specific ancient PRDM9 motifs. I also examine the fine scale effect of GC-biased gene conversion and estimate parameters of interest.

The rest of this chapter will be organized as follows. First, I briefly re-introduce the topics of biased gene conversion and GC-biased gene conversion, and explain how they can be used to find k-mers forming part of ancient PRDM9 binding motifs. Second, I give detailed methodological information on how to do this, by looking for k-mers with either species specific accelerated loss or accelerated local GC fixation. I also explain how these k-mers can be clustered together into PRDM9 binding motifs. Third, I give results, including example PRDM9 motifs, and investigate nucleotide

substitution rates around k-mers forming part of the motifs. Finally, I discuss these results in light of other studies about the present, recent and ancient usage of PRDM9 in mice.

5.1 Introduction

As we saw in Chapter 1, Chapter 4, and has been shown elsewhere, recombination in mice, and other species such as humans, occurs in narrow regions of the genome referred to as hotspots [29, 93, 117, 118]. In mice and humans, this fine-scale localization is due to the zinc finger protein PRDM9 [25–27], which directs the recombination machinery to its binding motifs during meiosis. As we saw in the Introduction, during meiosis, repair of PRDM9 localized double strand breaks (DSBs) induced by SPO11 during meiosis is resolved as either a crossover, where one chromatid from each pair of homologous chromosomes are broken and rejoined, or a noncrossover, where the DSB is resolved without reciprocal exchange among the chromatids. Repair of DSBs is done through either the double strand break repair (DSBR) or synthesis-dependent strand annealing (SDSA) pathways. The former leads to either noncrossovers or crossovers, although there is evidence that only crossovers arise from DSBR in mice [38], while the latter leads exclusively to noncrossovers.

As we saw in the Introduction, it has been shown empirically that if an individual is heterozygous for a hotspot disrupting mutation, and hence has one hotter (higher affinity for PRDM9) and one colder (lower affinity) allele at that hotspot, then during gene conversion, the colder allele is likely to be over-transmitted, referred to as biased gene conversion [30,31]. Over time, this bias will lead to the rapid evolutionary erosion of active PRDM9 binding motifs from the genome [25]. Separately, in gene conversion, mismatches between SNPs on homologous chromatids lead to the overtransmission of strong (cytosine or guanine) bases versus weak (adenine or thymine) ones [32, 36]. This different form of meiotic drive, GC-biased gene conversion, which may arise from

favoured repair of strong versus weak bases in heteroduplex DNA during repair of the DSB, similarly promotes the fixation of strong bases near active PRDM9 binding motifs.

Together, these two processes, the erosion of PRDM9 binding motifs through biased gene conversion, and the accumulation of weak to strong substitutions near eroded PRDM9 binding motifs through GC-biased gene conversion, form the basis for the *de novo* search for ancient PRDM9 binding motifs I present here. First, I search for evidence of ancient PRDM9 binding motifs by characterizing each substitution in each mouse species lineage as causing the loss of k k-mers, or short sequences of DNA of length k. By comparing the rate of loss of each k-mer in one species lineage to other species lineages, and normalizing against an expected rate of loss based on its GC content and number of CpG's, I can determine which k-mers have been lost unusually quickly, which I presume is due to biased gene conversion. Second, I search for evidence of ancient PRDM9 motifs by examining whether the local weak to strong fixation rate is elevated around lost k-mers, again normalized to similar lost k-mers, which I presume is due to GC-biased gene conversion. This searching is done outside the region of the lost k-mer itself, and as such, presents an orthogonal source of information on each k-mer. The use of these two signals, based on biased gene conversion and GC-biased gene conversion, helps ensure results are driven through the actions of ancient PRDM9 binding motifs, and not mapping or other artefacts.

This idea is not totally new - the idea of using species lineage specific accelerated loss of k-mers to find eroded PRDM9 binding hotspots forms the basis of the DPhil of Afidalina Tumian [127], supervised by Simon Myers, and is a major idea in Myers *et al.* [25]. In her thesis, among other analyses, Afidalina used multiway alignments in primates to infer ancestral PRDM9 binding motifs using accelerated loss of k-mers. Apart from extending the use of the approach from primates to mice, I extend Afidalina's ideas by using VCF callsets, as opposed to multi-species alignments of

genome assemblies, to broaden utility of this method of investigation (since assemblies are harder to make than short read based SNP callsets), and enable it to be applied here. I also introduce the use of GC-biased gene conversion as a second test for each k-mer as to whether they've formed part of an ancient hotspot, which reduces artefacts considerably and hence increases robustness of the approach. Further, by recording nucleotide substitution around lost k-mers, I am able to get a highly sensitive picture of the fine scale nucleotide substitution landscape around previously active PRDM9 motifs, and from this, estimate parameters associated with gene conversion, which can be interpreted as gene conversion track lengths and crossover to noncrossover ratios.

The goals of this chapter were therefore as follows. First, using the strategy just described, I used the whole genome sequencing data on multiple mice to find ancient PRDM9 binding motifs in both living and ancient species of mice. Since motifs are known to sometimes contain strong matches to repeats [117], *e.g.* short tandemly repeating bases or longer interspersed transposable elements, and as they represent a highly non-random k-mer background, this analysis is performed by repeat class for each lineage. Second, I compared these results for the *Mus musculus* mice to those from Chapter 4, to see how ancient PRDM9 motifs compare to recent and presently active motifs. I then show that I can obtain a highly accurate picture of weak to strong substitutions surrounding lost k-mers, and present several models to interpret this results in terms of crossover and noncrossover tract lengths and proportions.

5.2 Methods

In this chapter, I use the following notation. In general, I refer to these and associated techniques as "hotspot death" methods, to distinguish them from results in Chapter 4 obtained with LD-based recombination rate maps. I use the term k-mer to refer to a string of k DNA bases. I will speak of sets of "similar" k-mers that have been

clustered together as either "clusters" of k-mers or "motifs" or "PWMs", the latter terms reflecting that I am identifying sets of k-mers underlying a PRDM9 binding motif, which can be represented as a position weight matrix (PWM). Finally, note that lineage is used both to specify a species lineage, for example *M. spretus*, as well as the name of a method (the "lineage" test) to calculate p-values based on the presence of accelerated loss of a k-mer down a particular species lineage versus all other species lineages.

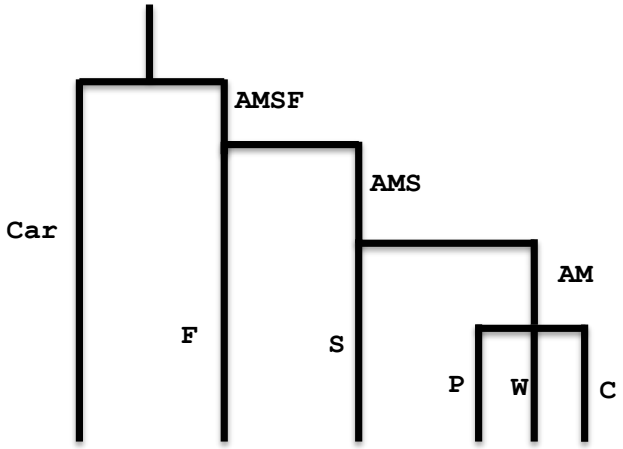
5.2.1 Lineage definitions and SNP filtering

I used sequence data from six wild derived inbred strains of mice from the joint SNP callset that includes the wild populations as well as the classical and wild derived strains of mice. I used: WSB/EiJ (abbreviated here W), representing *M. m. domesticus*; CAST/EiJ (C) representing *M. m. castaneus*; PWK/PhJ (P) representing *M. m. musculus*; SPRET/EiJ (S) representing *M. spretus*; the *M. famulus* sample (F); and CAROLI/EiJ (Car) representing *M. caroli*. I used the entire set of 140.8 M SNPs. I removed SNPs if they have any missingness among the 6 mice, are multi-allelic, or are heterozygous among any of the 6 mice.

I use lineages from a species tree as inferred in Chapter 3 of this thesis to assign SNPs to one of 10 lineages. A summary is shown in Figure 5.1, which also summarizes the assignment of SNPs to lineages. For the lineages, first, *M. caroli* (Car) branches off from the ancestor of *M. musculus*, *M. spretus*, and *M. famulus*, or AMSF. Next, *M. famulus* (F) branches off from the ancestor of *M. musculus* and *M. spretus*, or AMS. Next, *M. spretus* (S) branches off from the ancestor of *M. musculus*, or AM. Finally, I assume that we cannot resolve the split into the three subspecies of *Mus musculus*. Therefore, AM splits into six possibilities, for each of *M. m. musculus* (P), *M. m. domesticus* (W) and *M. m. castaneus* (C), and their binary combinations. Since we need Car and AMSF to orient all other SNPs, this leaves SNPs to be assigned uniquely

to one of 10 lineages: F, AMS, S, AM, P, W, C, P+W, P+C, P+D. Assignment of SNPs is as shown in Figure 5.1. Any SNP not conforming with the species tree in one of the shown possibilities is removed.

Figure 5.1. SNP assignment to lineages for hotspot death Top panel shows species tree for 6 mice used in this chapter. Bottom shows assignment of SNPs, labeled arbitrarily by 0's and 1's, to one of 10 lineages. Visible lineages are highlighted with boxes. Car = *M. carolis*, AMSF = ancestor of *M. musculus*, *M. spretus* and *M. famulus*, F = *M. famulus*, AMS = ancestor of *M. musculus* and *M. spretus*, AM = ancestor of *M. Musculus*, P = PWK/PhJ, W = WSB/EiJ, C = CAST/EiJ



Short name	Car	AMSF	F	AMS	S	AM	P	W	C	Alt name
F	0	0	1	0	0	0	0	0	0	FAM
AMS	0	0	0	1	1	1	1	1	1	AMS
S	0	0	0	0	1	0	0	0	0	Spretus
AM	0	0	0	0	0	1	1	1	1	AM
P	0	0	0	0	0	0	1	0	0	PWKPhJ
C	0	0	0	0	0	0	0	0	1	CASTEiJ
W	0	0	0	0	0	0	0	1	0	WSBEiJ
P+C	0	0	0	0	0	0	1	0	1	PWKPhJ.CASTEiJ
P+W	0	0	0	0	0	0	1	1	0	PWKPhK.WSBEiJ
W+C	0	0	0	0	0	0	0	1	1	WSBEiJ.CASTEiJ

SNPs are further filtered after assignment if they are too close together, as SNPs being too close together can be a sign of mapping artefacts. After experimentation with parameter values, I removed all SNPs affected if there were 12 SNPs across all lineages within 50 bases, and after removal of these SNPs, further removed SNPs

when there were 7 of more of them from a specific lineage within 50 bases. These thresholds were chosen as they were the smallest choices such that almost no SNPs would be expected to be removed by chance, as assessed by simulation (not shown).

To guard against artefacts driven by differences in callability, I next masked out regions of the genome with insufficient coverage in any of the 6 mice. I use the same procedure as used earlier in this thesis, guided by recommendations for usage of PSMC [63], where I mask out regions of the genome less than one third or more than twice the average coverage in any individual mouse. For repeats, I use the UCSC genome browser repeat masker track. I use the column "repName" to define repeat background, and non-repeat otherwise.

5.2.2 p-value calculations

Once SNPs are assigned, I make ancestral and derived sequences for each lineage, by putting SNPs into the reference sequence. I then count how many k-mers are lost (present in ancestral sequence, absent in derived sequence) or gained (the opposite) based on the SNPs present in that lineage. Any k-mer spanning a masked base was removed, as were any where the repeat background was not consistent throughout the length of the k-mer.

For each k-mer for each lineage, for either gained or lost, I used 2 different methods to make 2×2 tables to examine whether the k-mers exhibited unusual properties.

Table 5.1. 2×2 **table for lineage method** For a k-mer, summary of how 2×2 table was built for the "lineage" method. p-values and odds ratios follow from Fisher's Exact test

	This lineage	All other lineages
k-mer	n_1	n_2
Other k-mers with the same CpG, GC Content	n_3	n_4

The first method I refer to as "lineage", where for each k-mer, for each lineage, I test whether that k-mer is lost more frequently down that lineage versus all the other

Table 5.2. 2×2 **table for AT to GC test** For each k-mer, summary of how 2×2 table was built for the "AT to GC" test. p-values and odds ratios follow from Fisher's Exact test. Unchanged AT bases refer to those that are ancestral A or T and do not change to GC

	AT to GC bases	Unchanged AT bases
k-mer	n_1	n_2
Other k-mers with the same CpG, GC Content	n_3	n_4

lineages, compared to the other k-mers with the same GC content and number of CpG dinucleotides (Table 5.1). The second method I refer to as "AT to GC", where for each k-mer, for each lineage, I test whether there are more changes from weak ancestral to strong derived bases in a neighbourhood (100bp) surrounding each lost k-mer, by comparing the count of ancestral weak bases that become strong versus ancestral weak bases that do not become strong. I compare counts for a k-mer to the same counts for the other k-mers with the same GC content and number of CpGs (Table 5.2). For example, consider a k-mer lost with $k = 10$ from positions 1,000 to 1,009 along a chromosome. Then the following 100 bp are from 1,010 to 1,109, inclusive, and the preceding are 999 to 900, inclusive, and this represents the local neighbourhood of the lost k-mer.

Since results are generated for each repeat background, each k-mer and each lineage, the number of p-value calculations can be rather large. To reduce the number of calculations, for the lineage test, p-values for k-mers were only counted if across all lineages, that k-mer was lost at least 50 times. Within a lineage, a k-mer had to be lost 10 times for a p-value to be generated. Similarly, the same numbers were used for the AT to GC test based on the number of bases that had changed from weak ancestral to strong derived.

5.2.3 k-mer clustering

Since many similar k-mers are often highly statistically significant, I wrote a clustering method to group similar p-values into clusters, or "motifs". Clustering was performed

separately for each lineage, for each repeat background, and for each of the lineage test and the AT to GC test.

For a lineage, repeat background, and test, I first selected all k-mers that met a "study-wide" p-value significance threshold, where study-wide means that the same threshold was used for each lineage, repeat background and test. I then further required them to have an odds ratio (OR) > 1, not to have another lineage with an OR < 1 with a more significant p-value, and to have a p-value less than 0.05 and an OR > 1 on the other test. These are the k-mers that I would seek to use as seed k-mers for clustering, and after clustering, each one would uniquely be associated with a clustered motif. I set the study-wide p-value threshold not to be the Bonferonni significant number of tests across all species lineages, tests and repeats, but a slightly reduced threshold of the Bonferonni corrected maximum given the total possible number of tests that could be calculated on a given repeat background (for $k = 10$, $\frac{(4^{10}-5\times4^5)}{2} + 4^5 = 524,800$). Note that this more lenient threshold was thought to be reasonable given I require nominal significance from the p-value for the other test.

Given a set of eligible study-wide significant k-mers for a given lineage, test and repeat, I seeded a new cluster with the k-mer that had the most significant p-value. I then recursively added to this new cluster k-mers where, for each unexplored k-mer in the cluster, I found all k-mers within a certain distance (defined below), and kept those with a p-value that met a certain threshold (defined below) as well as having a nominal p < 0.05 and OR > 1 on the other test. Once I couldn't add new k-mers to the cluster I terminated the procedure. Note that I kept track of order and hence k-mer alignment to the cluster. Further, I kept track of whether a k-mer had been added to a cluster, and k-mers weren't allowed to contribute to more than one cluster.

During the clustering procedure, I considered only k-mers suitably close to the k-mer of interest (call it k^*-mer) as follows. I define as acceptably close for clustering all k-mers that align perfectly and are off by 1 base to $k*$-mer ($3k$ choices), and all

k-mers that are off by 1 base in the alignment left or right (2 choices for left vs right) with any new base in the gap (4 choices for the bases) with any one of the remaining $k-1$ bases allowed to change as well ($3(k-1)$ choices) or not (1 choice). In total, by this definition, for a given k^*-mer, I considered $3 \times k + 2 \times 4 \times (3 \times (k-1) + 1) = 254$ k-mers to be suitably close.

For the p-value thresholding for inclusion in the cluster, I started with a p-value threshold of the number of k-mers to be searched, $\frac{0.05}{254}$. Consider recursive iterations where on each iteration, the current list of new k-mers gives you a list of new k-mers for the subsequent round. For each subsequent recursive iteration when adding to the cluster, take a p-value threshold where the denominator is the number of tests already checked plus the number to be checked on this recursive iteration. For example, if on the first iteration, there were 3 k-mers close to the seed k-mer which were eligible and had a p-value less than $\frac{0.05}{254}$, then on the subsequent recursion iteration, when k-mers close to those 3 new k-mers are checked, the new p-value threshold for inclusion would be $\frac{0.05}{254+254\times3}$. Note that technically the denominator could be smaller as previously searched k-mers can't be added, but this should have only a relatively minor influence.

To generate a position weight matrix for a cluster of k-mers, I used the method of Schneider and Stephens (1990 [128]), which is also the same method used in the R package seqLogo. I built a sum of bases for each position in the cluster by adding together the influence of each k-mer in the cluster, using a weight of the odds ratio minus 1, and adding it to the bases in the motif. I added 0.1 to the sum of each entry in the counts to avoid dividing by 0. Finally, I divided by the column sums to get an entry between 0 and 1, and from this built a PWM. More specifically, if we label the four possible bases with index $a = 1, ..., 4$, and the bases $i = 1, ..., L$ for an alignment of length L, then given counts $n_{a,i}$ of the number of times base a is seen at position i in the alignment, then $f_{a,i} = \frac{n_{a,i}}{\sum_{a=1}^{4} n_{a,i}}$. Let the information content I_i for site i in a

2-bit system be

$$I_i = -\sum_{a=1}^{4} f_{a,i} \times \log_2(f_{a,i}) \tag{5.1}$$

Then, for plotting, the height of base a at position i is given by $H_{a,i}$

$$H_{a,i} = (2 - I_i) \times f_{a,i} \tag{5.2}$$

5.2.4 Estimating parameters associated with gene conversion

To characterize the substitution landscape surrounding lost PRDM9 motifs, I modelled the rate of change of weak ancestral to strong derived bases using a series of parameters (*i.e.* I fit the observed data with a function whose variables, or parameters, can be interpreted as having biologically relevant meaning). To estimate parameters associated with gene conversion, I summed weak to strong nucleotide changes surrounding lost k-mers for all loss events from significant clusters, and summed them into a single result. Since directionality is not discernible across different clusters, *i.e.* I don't know which of the forward or reverse PWMs is the real one, I summed nucleotide changes from before and after each lost k-mer, taking care so that lost k-mers couldn't overlap if they were from the same cluster. Over some range N, where i used here $N = 5000$, for some distance $x_i = 1, ..., N$ base pairs away from the motif edge, I counted the number of weak ancestral bases that remained unchanged ($n_{2,i}$) and the number that changed from ancestral weak to derived strong ($n_{1,i}$). Note that this distance is considerably larger than the local neighbourhood I used to discover k-mers as being part of a PRDM9 motif under the influence of GC-biased gene conversion.

Let $n_1 = \{n_{1,i} | i = 1, ..., N\}$ and $n_2 = \{n_{2,i} | i = 1, ..., N\}$. I model the likelihood of a set of parameters θ (to be explained) given observed data n_1 and n_2 as

$$L(\theta | n_1, n_2) = \prod_{i=1}^{N} P(n_{1,i}, n_{2,i} | \theta) \tag{5.3}$$

To calculate the likelihood, I considered a binomial model where for some distance x_i base pairs away, the probability of fixing a weak ancestral to strong derived base

using parameters θ is $p_i(\theta)$, giving

$$P(n_{1i}, n_{2i}|\theta) = \frac{(n_{1,i} + n_{2,i})!}{n_{1,i}!n_{2,i}!}(p_i(\theta))^{n_{1,i}}(1 - p_i(\theta))^{n_{2,i}} \tag{5.4}$$

For parameter estimation, I used maximum likelihood estimate on the log likelihood, represented as $l(\theta|n_1, n_2)$, given by

$$\begin{aligned} l(\theta|n_1, n_2) &= \log(L(\theta|n_1, n_2)) \\ &= \log\left(\prod_{i=1}^{N} \frac{(n_{1,i} + n_{2,i})!}{n_{1,i}!n_{2,i}!} p_i^{n_{1,i}}(1 - p_i)^{n_{2,i}}\right) \\ &= \sum_{i=1}^{N} \log\left(\frac{(n_{1,i} + n_{2,i})!}{n_{1,i}!n_{2,i}!}\right) + \sum_{i=1}^{N}\Big(n_{1,i}\log(p_i) + n_{2,i}\log(1 - p_i)\Big) \end{aligned} \tag{5.5}$$

The derivative of the log likelihood with respect to parameter θ_j is

$$\frac{\partial l(\theta|n_1, n_2)}{\partial \theta_j} = \sum_{i=1}^{N}\left[\frac{n_{1,i}}{p_i} - \frac{n_{2,i}}{1 - p_i}\right]\frac{\partial p_i}{\partial \theta_j} \tag{5.6}$$

The second derivative of the log likelihood with respect to parameters θ_j and θ_k is

$$\frac{\partial^2 l(\theta|n_1, n_2)}{\partial \theta_k \partial \theta_j} = \sum_{i=1}^{N}\left[-\frac{n_{1,i}}{{p_i}^2} - \frac{n_{2,i}}{(1 - p_i)^2}\right]\frac{\partial p_i}{\partial \theta_k}\frac{\partial p_i}{\partial \theta_j} + \sum_{i=1}^{N}\left[\frac{n_{1i}}{p_i} - \frac{n_{2,i}}{1 - p_i}\right]\frac{\partial^2 p_i}{\partial \theta_k \partial \theta_j} \tag{5.7}$$

For convenience and simplicity, I used the R package rSymPy to symbolically calculate the exact derivatives and Hessian of p_i with respect to the components of θ, for the different possibilities of p_i and θ (explained below).

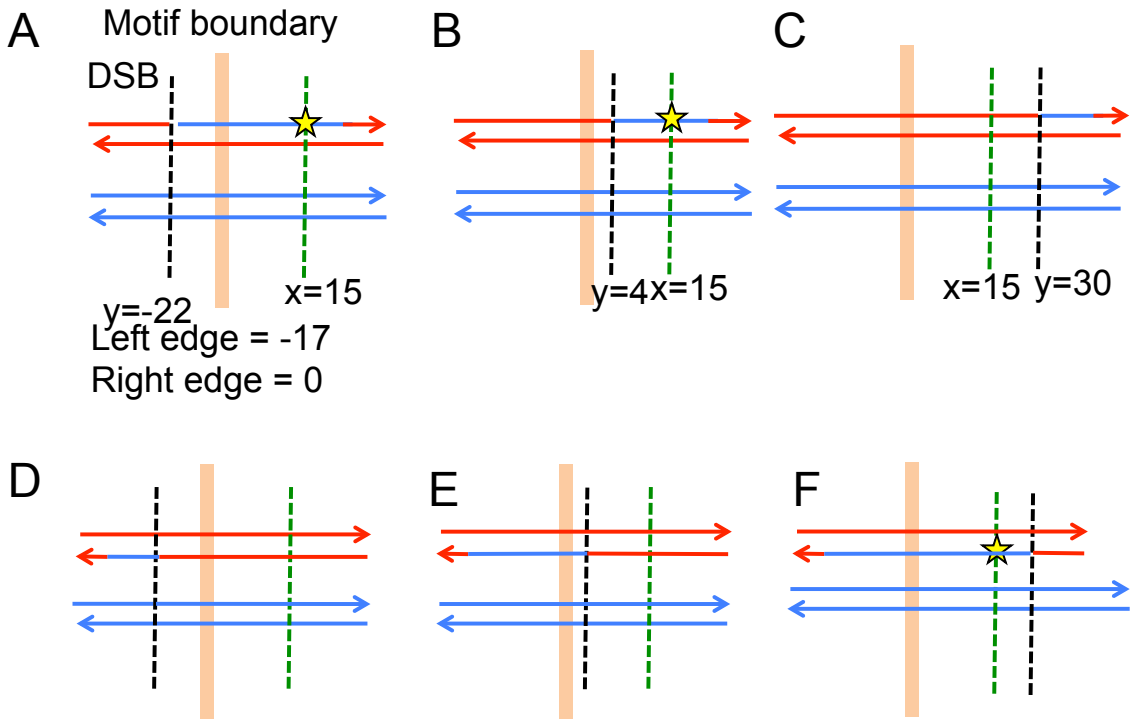
I considered five different biologically plausible models for p_i. The first model is a simple exponential decay, so that GC-biased gene conversion effect reflects exponentially distributed conversion tract lengths, for example

$$p_i = \tau_1 \lambda_1 e^{-\lambda_1 x_i} + \tau_3 \tag{5.8}$$

where τ_3 is the background rate, λ_1 is the tract length, and τ_1 is a scaler. The second model is a double exponential model, which might be appropriate if, for example, GC-biased gene conversion occurs for both crossovers and noncrossovers, and they have different tract lengths

$$p_i = \tau_1 \lambda_1 e^{-\lambda_1 x_i} + \tau_2 \lambda_2 e^{-\lambda_2 x_i} + \tau_3 \tag{5.9}$$

Figure 5.2. Example crossovers In the panels, examples are given to show in a two-exponential model, how a location can lie in a gene conversion tract. In this example, the PRDM9 motif is 18 bases long, from positions -17 to 0 inclusive, and we are considering base $x = 15$. Let y be the location of the DSB, which can occur in three locations: on the opposite side of the base compared to the motif ($y < 0$) (panel A and D); between the motif and the base $0 \leq y \leq 15$ (panel B and E); and further away than the base itself ($y > 15$) (panel C and F). We consider that extension occurs simultaneously in both directions away from the DSB (upper vs lower panels), which leads to 3 ways in which the base $x = 15$ can lie in a gene conversion tract (heteroduplex DNA in panels A, B and F).



where τ_2 and λ_2 parameterize the second distribution. The third model is more complex, and motivated by the idea that there might be uncertainty in where DSB's occur (*i.e.* not exactly at the PRDM9 motif), followed by an exponential tract length for the conversion tract length, as used in model 1. For the tract term, I need the probability of that base being in the tract given the DSB location, which is 1 minus the cumulative distribution function. An example of this model is given in Figure 5.2. Together, this yields

$$\begin{aligned} p_i =& \tau_1\lambda_3\left[\int_{y=-\infty}^{y=0} e^{(\lambda_3 y-\lambda_1 x_i+\lambda_1 y)}dy + \int_{y=0}^{y=x_i} e^{(-\lambda_3 y-\lambda_1 x_i+\lambda_1 y)}dy + \int_{y=x_i}^{y=\infty} e^{(-\lambda_3 y-\lambda_1 y+\lambda_1 x_i)}dy\right] \\ &+ \tau_3 \\ =& \frac{2\tau_1\lambda_3}{\lambda_1^2-\lambda_3^2}\left(\lambda_1 e^{(-\lambda_3 x_i)} - \lambda_3 e^{(-\lambda_1 x_i)}\right) + \tau_3 \end{aligned} \tag{5.10}$$

I also considered a fourth model that combines properties of the second and third models, which has an exponential decay parameter for DSB location, and two additional exponential decay terms that might model crossover and noncrossover tract lengths

$$\begin{aligned} p_i =& \tau_1\lambda_3\left[\int_{y=-\infty}^{y=0} e^{(\lambda_3 y-\lambda_1 x_i+\lambda_1 y)}dy + \int_{y=0}^{y=x_i} e^{(-\lambda_3 y-\lambda_1 x_i+\lambda_1 y)}dy + \int_{y=x_i}^{y=\infty} e^{(-\lambda_3 y-\lambda_1 y+\lambda_1 x_i)}dy\right] \\ &+\tau_2\lambda_3\left[\int_{y=-\infty}^{y=0} e^{(\lambda_3 y-\lambda_2 x_i+\lambda_2 y)}dy + \int_{y=0}^{y=x_i} e^{(-\lambda_3 y-\lambda_2 x_i+\lambda_2 y)}dy + \int_{y=x_i}^{y=\infty} e^{(-\lambda_3 y-\lambda_2 y+\lambda_2 x_i)}dy\right] \\ &+\tau_3 \\ =& \frac{2\tau_1\lambda_3}{\lambda_1^2-\lambda_3^2}\left(\lambda_1 e^{(-\lambda_3 x_i)} - \lambda_3 e^{(-\lambda_1 x_i)}\right) + \frac{2\tau_2\lambda_3}{\lambda_2^2-\lambda_3^2}\left(\lambda_2 e^{(-\lambda_3 x_i)} - \lambda_3 e^{(-\lambda_2 x_i)}\right) + \tau_3 \end{aligned} \tag{5.11}$$

Finally, I considered a fifth model, similar to model 4, except that the two exponential gene conversion tract lengths have their own separate exponential DSB terms

$$p_i = \frac{2\tau_1\lambda_3}{\lambda_1^2-\lambda_3^2}\left(\lambda_1 e^{(-\lambda_3 x_i)} - \lambda_3 e^{(-\lambda_1 x_i)}\right) + \frac{2\tau_2\lambda_4}{\lambda_2^2-\lambda_4^2}\left(\lambda_4 e^{(-\lambda_4 x_i)} - \lambda_4 e^{(-\lambda_2 x_i)}\right) + \tau_3 \tag{5.12}$$

Parameter estimates were obtained using the nlminb function in R, a box constrained multi-variable optimization method, using the likelihood, gradient and Hessian. Upper and lower bounds were supplied: the exponentials (λ) were bounded below by 0; the proportions (τ) bounded below by 0 and above by 1.

MLEs were obtained from the best of 11 different initializations of nlminb, with 11 different starting positions. Each of these used a different set of parameter estimates, coming either from all data, or from 10 different random subsets of 75% of the data *i.e.* 75% of the vectors n_1 and n_2. Let n_1^* and n_2^* be the counts over positions x_i^* that were used to initialize nlminb. Parameters were initialized from these as follows. τ_3 was estimated using the last 10% of the initialization data (furthest from the lost k-mers). λ_2 was estimated by fitting an exponential to the last 70% of the data (furthest from the lost k-mers) after subtracting out the effect of τ_3, while τ_2 is taken as the median of $\frac{n_{1i}^* - (n_{1i}^* + n_{2i}^*)\tau_3}{n_2^*}$ divided by $\lambda_2 e^{-\lambda_2 x_i^*}$. λ_1 and τ_1 were initialized in the same way on the first 10% of the data, after subtracting away the effects of τ_3, τ_2 and τ_1. λ_3 I initialized as one half of λ_1, and λ_4 I initialized as one tenth of λ_1.

To generate confidence intervals for θ, I considered each parameter separately. For each parameter, I started with a grid search over an equal number of points over a wide range of plausible values for that parameter. For each grid point within that range, I optimized the remaining parameters as before using the likelihood and its gradient. I took the parameter values for a 95% confidence interval as where twice the difference in the log likelihood between the maximum parameter estimate and the likelihood at the current parameter estimate exceeded the 95^{th} quantile from a chi-square distribution. I finally fine-mapped each parameter around the edges of the confidence interval by recursively considering more closely spaced grid points.

5.3 Results

5.3.1 *de novo* motif identication

From the entire autosomal genome (2.47 Gbp), I masked out regions that failed the coverage requirement in one or more mice (0.83 Gbp), leaving 1.64 Gbp of useable genome. Of this, I stratified the genome into repeat classes, leaving 1.11 Gbp of the genome as non-repeat sequence and the rest belonging to one of the 1,217 different

repeat classes. I then classified the 140.8M SNPs from the combined callsets into one of the 10 different lineages, or removed them, as outlined in the methods. The number of variants per species lineage is similar when compared to the ancestor of all samples, as shown in Table 5.3.

Table 5.3. Number of SNPs per lineage for hotspot death analyses Results are across all repeat backgrounds. The first column (N SNPs) gives the number of SNPs, while the second column gives the % of the alignable genome this represents. For the third and fourth columns, for the observed lineages, I count back all SNPs from that basal lineage to the common ancestor (AMSF), and repeat columns 1 and 2. Rows give lineage names (see Figure 5.1).

	N SNPs	% genome	N SNPs vs anc	% genome vs anc
FAM	17,913,173	1.089	17,913,173	1.089
AMS	8,758,108	0.532		
Spretus	11,716,370	0.712	20,474,478	1.245
AM	4,692,855	0.285		
PWKPhJ	3,710,202	0.226	19,771,430	1.202
CASTEiJ	3,560,700	0.216	19,620,275	1.193
CASTEiJ.PWKPhJ	1,381,827	0.084		
WSBEiJ	3,604,469	0.219	19,510,655	1.186
WSBEiJ.PWKPhJ	1,228,438	0.075		
WSBEiJ.CASTEiJ	1,226,785	0.075		

I calculated p-values for both the lineage test and AT to GC test, for each repeat background, for each species lineage. 4,795,903 p-values were generated for the lineage test and 2,476,625 for the AT to GC test; note that these are different due to filtering reasons, where p-values are not calculated due to insufficient events on a repeat background. Quantile-quantile plots for all 10 lineages for these two tests are shown in Figure 5.3, and are well behaved, with median p-values near 0.5, and the bulk of p-values conforming to an expectation under neutrality. Further, p-values generated from different tests were uncorrelated between lineages. As an example, Figure 5.4 shows as that for WSB/EiJ, rarely are significant k-mers on that lineage also significant on other lineages. Conversely, significantly associated k-mers in the lineage test in one lineage are often significantly associated in the AT to GC test

Figure 5.3. QQ plots for hotspot death p-values p-values for the lineage test are shown without filtering based on the AT to GC test p-values, and vice versa.

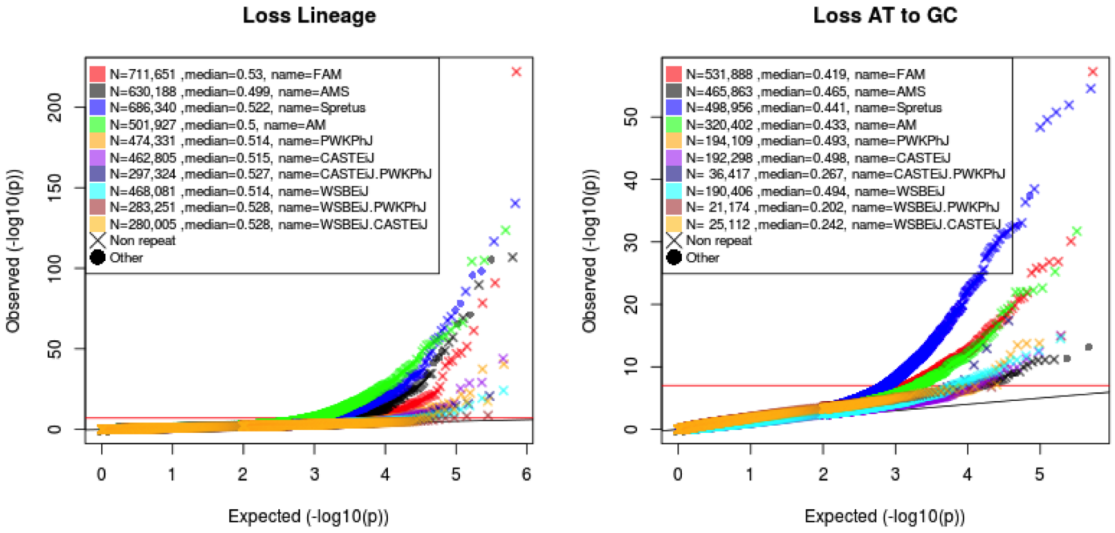


Figure 5.4. Comparing WSB/EiJ p-values from lineage test to other lineages Shown is each k-mer for which a p-value is calculated for WSB/EiJ by the lineage test compared to the p-value on each of the other 9 lineages, as indicated in the legend. A small minority are significant in both WSB/EiJ as well as another lineage

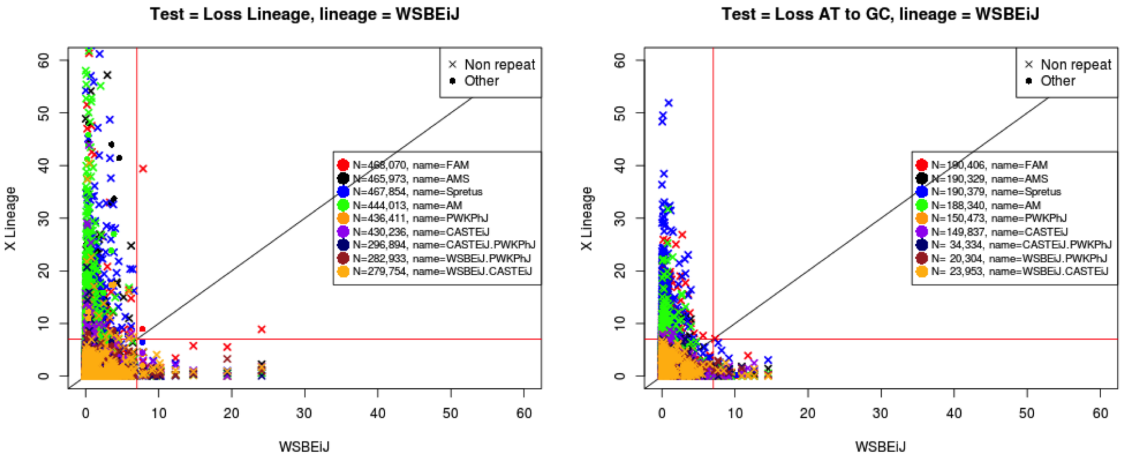
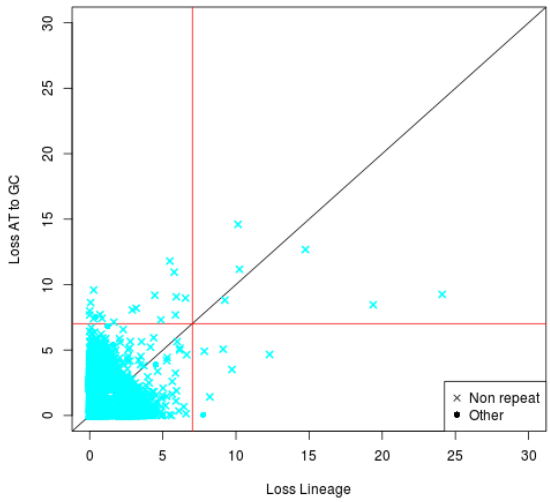


Figure 5.5. Comparing WSB/EiJ hotspot death lineage to AT to GC p-values Each point represents the p-value for a k-mer, with horizontal axis giving the p-value from the lineage test, and the vertical axis from the AT to GC test. Repeat versus non-repeat is stratified by pint type. Red bar denotes study-wide significance threshold.



in the same lineage, and vice versa. As an example, Figure 5.5 shows the excess of k-mers jointly significant for WSB/EiJ for both the lineage test and AT to GC test.

Table 5.4. Number of significant k-mers and PWMs per lineage Only included are 7 lineages with significant results. Number on the left indicates number of study-wide significant k-mers, while number on the right gives number of clustered motifs. Numbers are summed across both non-repeat and repeat backgrounds.

	lineage		AT to GC	
	k-mers	motifs	k-mers	motifs
FAM	63	15	181	30
AMS	87	16	26	9
Spretus	358	25	569	20
AM	437	15	140	9
PWKPhJ	0	0	5	2
CASTEiJ	5	1	3	2
WSBEiJ	11	2	16	3

Overall, the number of k-mers and clustered motifs for the two tests are shown in Table 5.4. Although not many motifs are identified for the different *Mus musculus*

subspecies, many are identified for the ancestral species as well as *M. famulus* and *M. spretus*. Disproportionately many motifs were found on the ancestral species lineages given the species branch length, which most likely reflects that these lineages only contain fixed derived substitutions, as compared to the species representing terminal branches, where many of the SNPs used here will still be segregating in that species or subspecies, and acted less upon by biased and GC-biased gene conversion. Further, there are generally a similar number of motifs identified in each species and subspecies between the lineage and AT to GC test.

Top non-repeat motif examples found based on the lineage test p-values are shown for Famulus, AMS, Spretus and AM in Figure 5.6. Generally, they are "well behaved" as opposed to the nonsense PWM we saw from Chapter 4 in Figure 4.8.

Specifically for the *Mus musculus* mice across all backgrounds, based on the lineage test, I observed two non-repeat results for WSB/EiJ and one non-repeat for CAST/EiJ (Figure 5.7). For the AT to GC results, the most noteworthy is a PWK/PhJ non-repeat example not present from the lineage test, which is also shown in Figure 5.7. I also confirmed that, for well behaved motifs, I was able to find similar motifs using only k-mers identified from the lineage test or the AT to GC test by relaxing the criteria that k-mers also had to be nominally significant from the other test when clustering. Results are shown for WSB/EiJ in Figure 5.8 for the top lineage test and AT to GC test motif identified in this fashion. The results are very similar, demonstrating that the two tests are largely picking up the same signal. The real benefit, which is not shown, is to help remove spurious motifs, likely driven by mapping and other artefacts.

I further compared the WSB/EiJ and CAST/EiJ top results to the previously published motifs and motifs from Chapter 4 shown in Figure 4.10 using manual alignment, with results shown in Figure 5.9. The alignment of the hotspot death WSB/EiJ with the French LD-hotspot motif and the Baker *et al. M. m. domesti-*

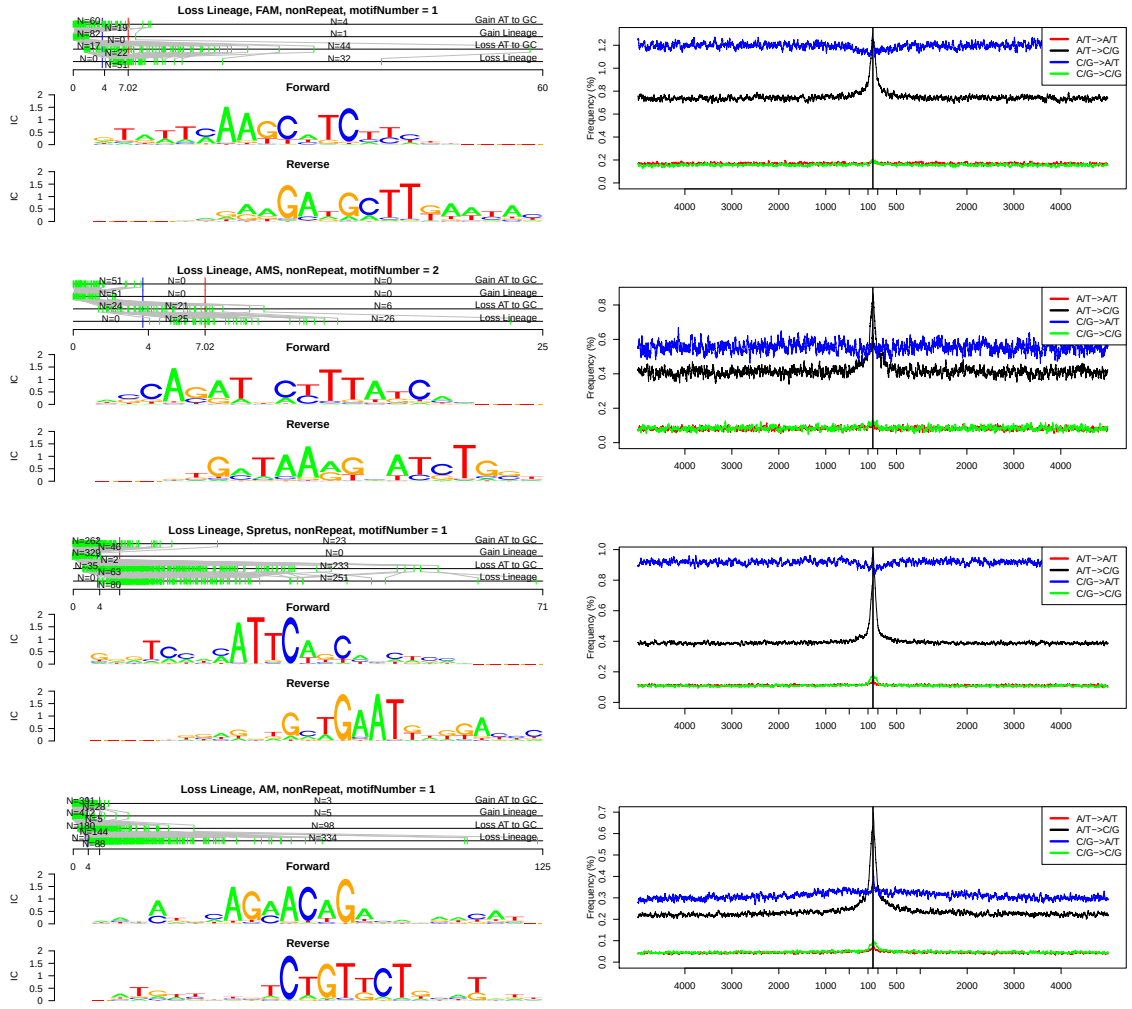


Figure 5.6. Some top motifs across different lineages from hotspot death Four sets of panels represent some of the top PWMs on the non-repeat background with, top to bottom, Famulus, AMS, Spretus and AM. Within each panel are three subpanels: upper left, $-\log10(p)$-values for each k-mer in the motif are represented as green vertical bars for either gain or loss, lineage or AT to GC test, with linking of the same k-mers between tests done by grey bar; bottom left panel is the forward and reverse PWM; right panel is nucleotide substitution bias around each lost k-mer, filtered such that k-mer loss positions are non-overlapping.

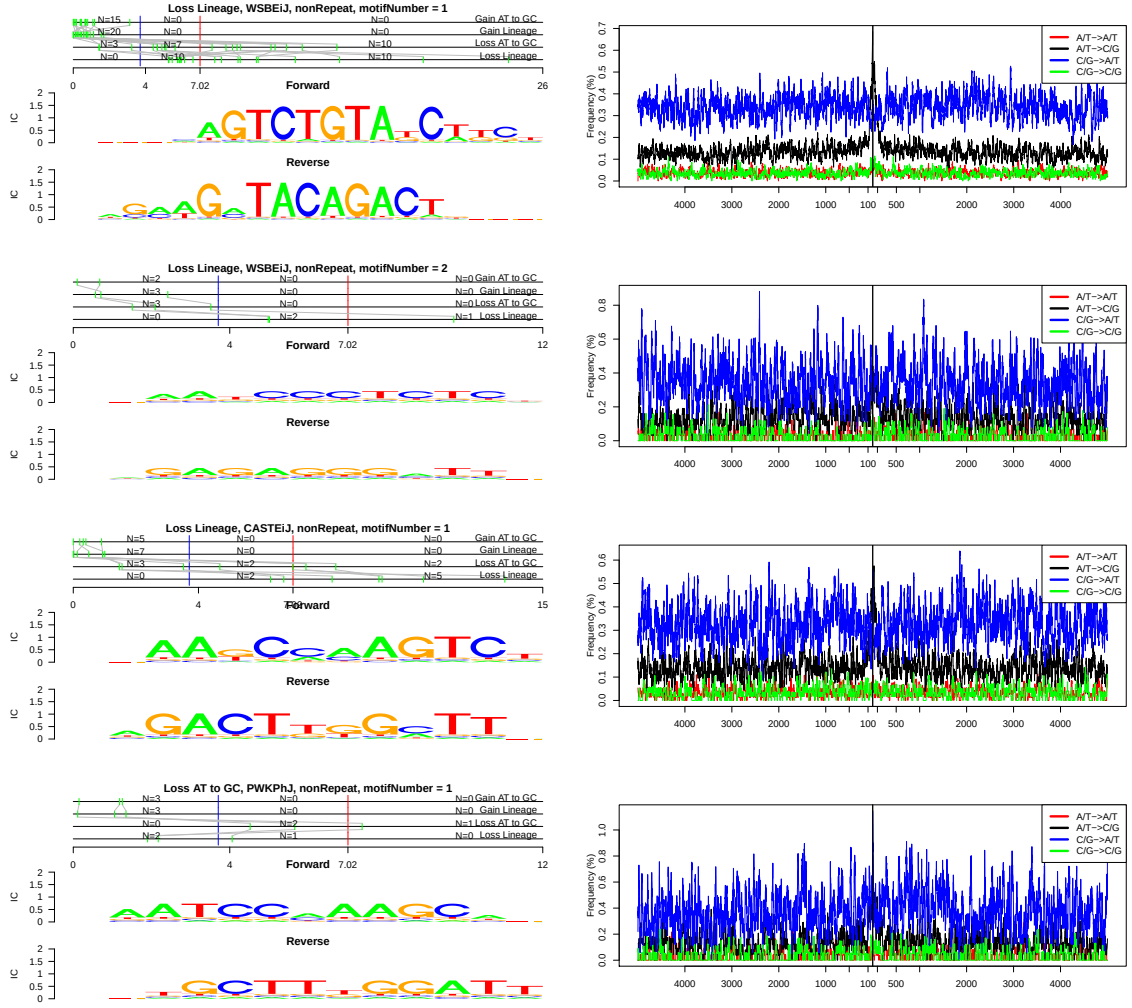


Figure 5.7. Top *Mus musculus* hotspot death motifs Sets of panels are as in Figure 5.6. Top two plots show WSB/EiJ results from lineage test, middle shows CAST/EiJ result from lineage test, bottom shows PWK/PhJ result from AT to GC test. All results are on non-repeat background.

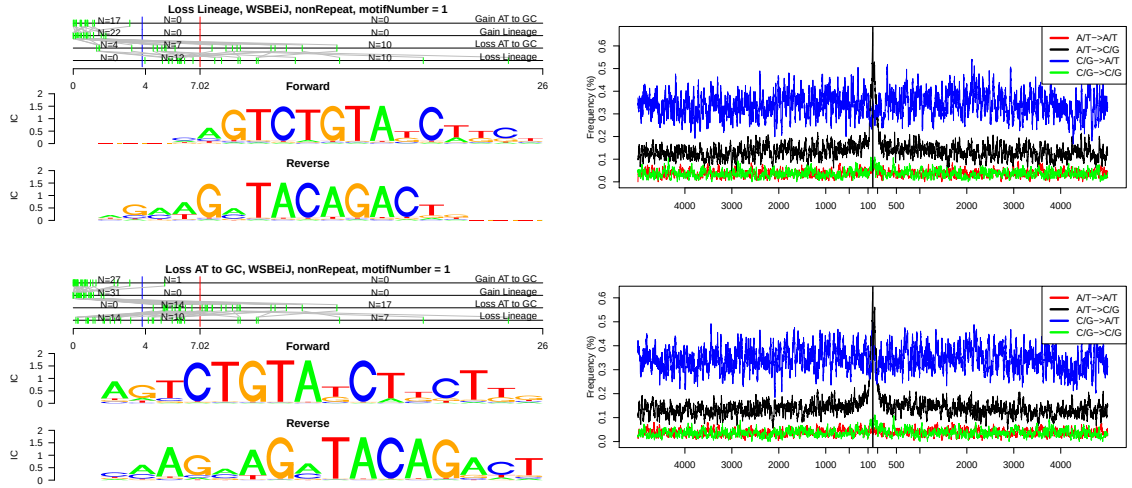


Figure 5.8. Motifs found without p-values from other test Sets of panels are as in Figure 5.6. Top two plots are top WSB/EiJ motifs found using lineage test or AT to GC k-mer p-values only, respectively. For comparison to the normal clustering procedure where nominal significance is required on the other test, compare the top panel here to the top panel of Figure 5.7

cus motif from [119] is quite strong, except at the second QVK zinc finger from the known PRDM9 allele, where the ancient hotspot death *M. m. domesticus* result differs from the recent LD-hotspot motif or the presently active Baker motif, suggesting perhaps a different zinc finger in the ancestral form of the allele. The alignement of the hotspot death CAST/EiJ with the Taiwanese LD-hotspot motif and the Baker *et al.* CAST/EiJ [119] motif is slightly more difficult, in that the Baker CAST/EiJ and Taiwanese LD-hotspot have a repeating CTKKG (where K is the IUPAC symbol for T or G), as in CTKKGNCTTKG, and so the hotspot death motif, with a similar CTKKG, can be aligned in two ways, as shown by the A and B alignments of Figure 5.9. The redundancy in the motif may explain why the identified motif from the hotspot death method is comparatively short for *M. m. castaneus*.

5.3.2 Estimating parameters associated with gene conversion

From the list of all significant motif clusters, for each k-mer in each cluster, for each time that k-mer was lost down its lineage, I counted all the substitutions over 5,000

Figure 5.9. Best recombination and hotspot death motifs aligned to known motifs Known motifs reproduced from Baker *et al.* [119]. Top set is *M. m. domesticus* results, with WSB/EiJ hotspot death motif at the top, French LD-hotspot motif in the middle, and the Baker *et al.* result at the bottom. Bottom set is *M. m. castaneus* results, with lineage test hotspot death result at the top (with A and B referring to two possible alignments), Taiwanese LD-hotspot motif in the middle, and the Baker *et al.* result at the bottom.

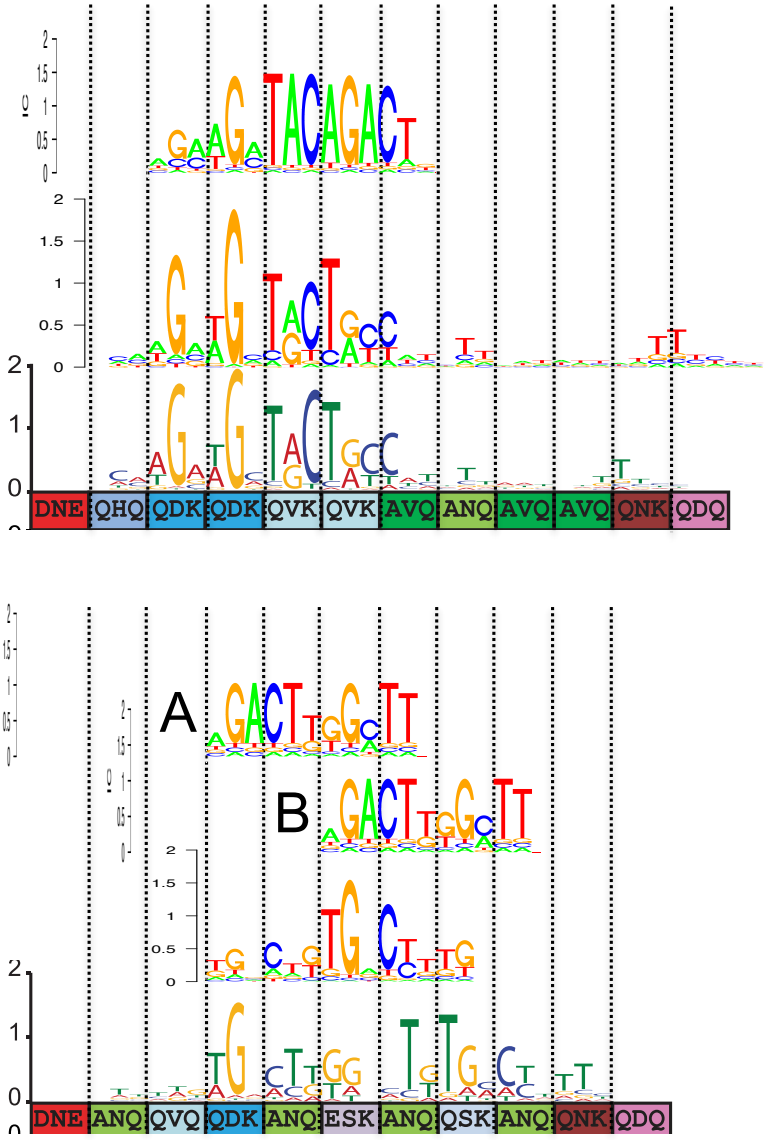


Figure 5.10. Nucleotide substitutions around loss events from clusters Horizontal axis shows distance from cluster boundary, with vertical bars near 0 indicating motif cluster edges. For example, if a motif cluster contained 2 k-mers, offset by one base with $k = 10$, then the motif would be 11 base pairs long, and distances on the plot for each k-mer would take into account the positions of the k-mer with respect to the motif. Vertical axis shows frequency of changes with respect to ancestral sequence. Each point shows smoothing of the raw rates of 10 base pairs forwards and backwards. Bottom plot is zoomed in on central region of top plot.

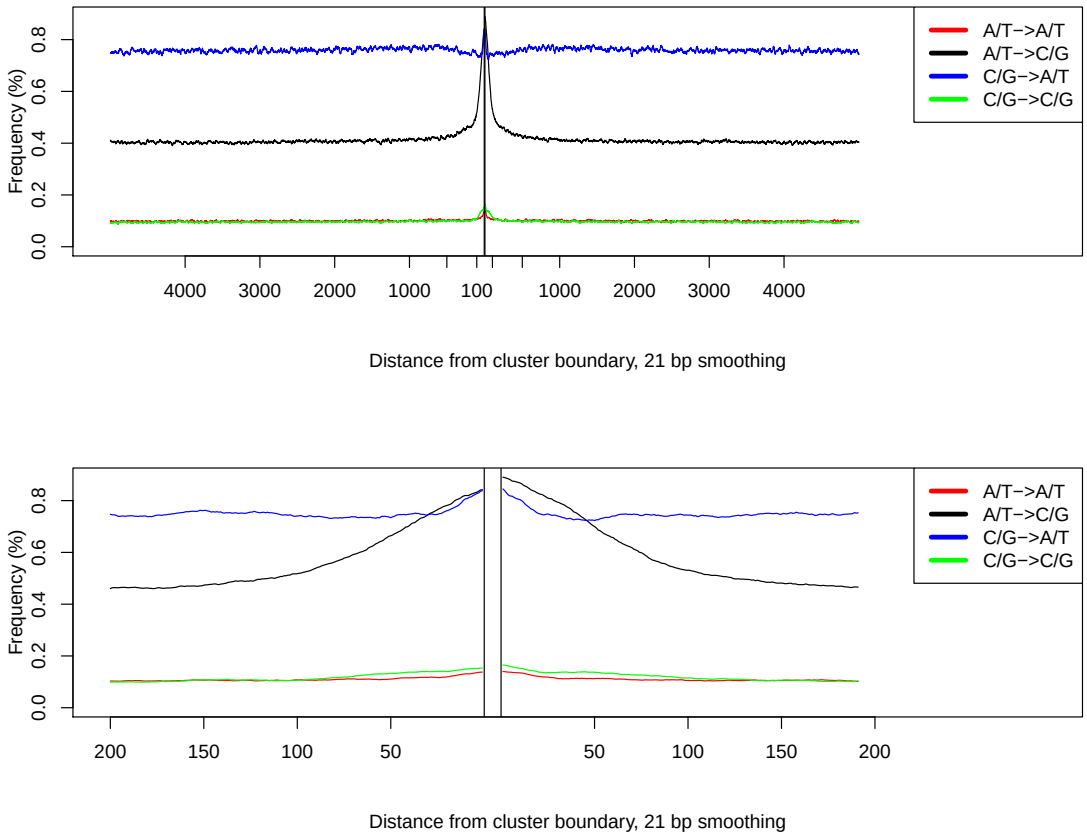
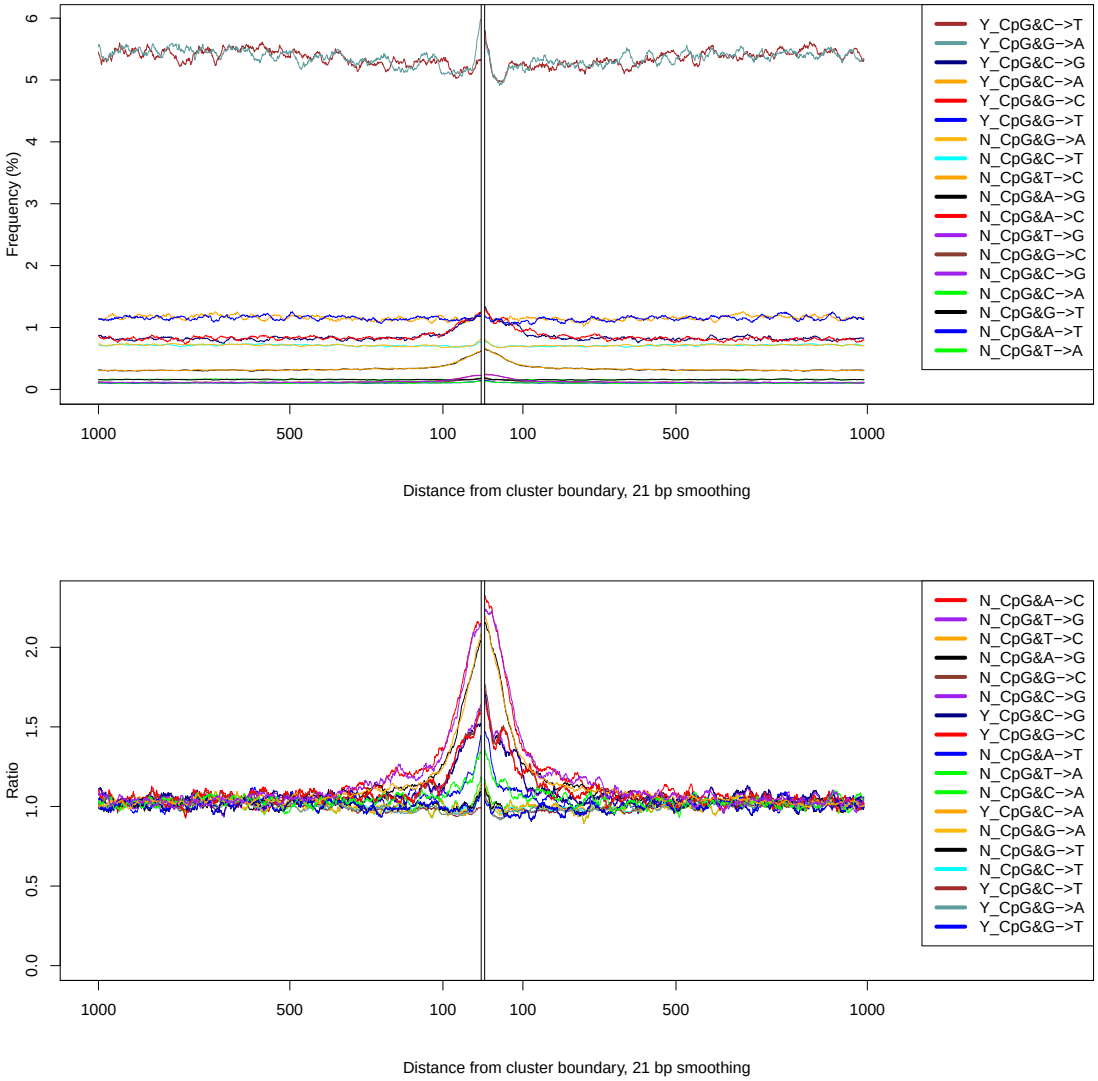


Figure 5.11. All nucleotide substitutions around loss events from clusters Figure is set up as in Figure 5.10, except with slightly different horizontal range, more substitutions, and the bottom plot being normalized by values from 1000 to 5000 before and after substitutions.



base pairs from the boundary of the motif cluster. Summed results as a fraction of the ancestral bases are shown in Figure 5.10. These results show that in general, the majority of changes occur within the first 100 base pairs of the k-mer boundaries, although there is a long tail extending out about 1,000 base pairs. For neutral substitutions, there is a small increase within about 20 base pairs of the motif cluster boundaries, as with the strong to weak substitutions. More broadly, there is a decrease in the strong to weak substitutions over about 1,000 base pairs. For weak to strong substitutions, the majority of the increase in fixation take place over about 100 base pairs, with a long tail out to about 1,000 base pairs. I also plotted out all possible substitutions in Figure 5.11, without symmetrizing results about the motif centers, for both the absolute changes and the ratios. Most notably, we see that weak to strong transitions and transversions show similar proportional rates, lending credence to the joint analysis that follows.

I next modelled the weak to strong substitutions from Figure 5.10, using the five models laid out in the methods. Model comparisons are shown for the models in Figure 5.12. The best models, as judged by the log likelihood, are the three and four exponential model, with $l(\theta|n_1, n_2) = -37126544$, as compared to the next best model, the two exponential model with $l(\theta|n_1, n_2) = -37126631$. The likelihood ratio between the two exponential model and the three exponential model is quite large, indicating the three exponential model has significantly better support. As the four exponential model showed no better fit despite an extra parameter, I pursued further analysis with the three parameter model . Residuals for the three exponential model show little evidence of trend, although there is evidence for minor deviations in the residuals over the first few hundred base pairs.

Finally, I estimated confidence intervals for the parameters in the three exponential model. Results for the exponential parameters and the ratio of the noncrossover term to the crossover term are shown in Table 5.5.

Figure 5.12. Comparison of models for estimating parameters associated with GC-biased gene conversion Left panels show the fraction of ancestral weak bases that become strong derived, as a function of distance away from motif cluster boundaries. Shown are raw data (grey x's), smoothed raw data (black line) and fitted curve (red line). Legend values give parameter estimates from fitting to the raw data. Right panels show residuals (observed minus real) data points (open black circles), smoothed raw data (red line, 20 bp smoothing) and a horizontal line at 0 (blue line). Top to bottom, panels show results for the four different models. Above the right panel for each model is the minus natural-log-likelihood

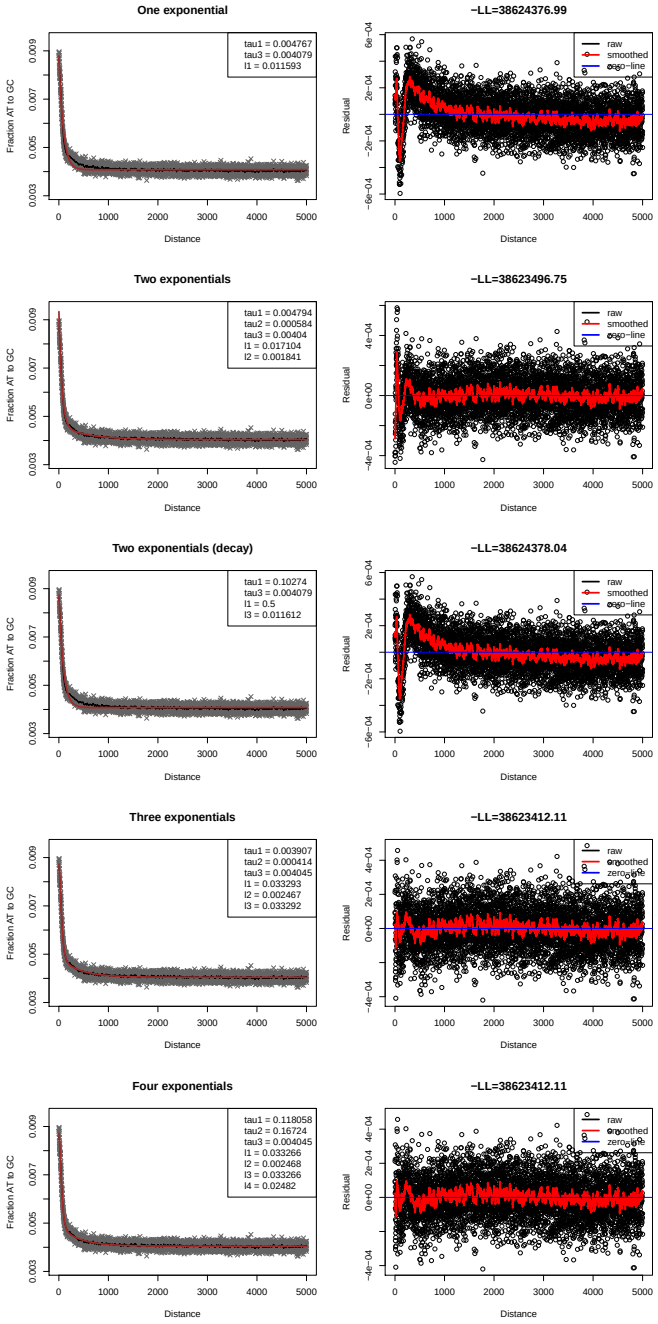


Table 5.5. Estimation of gene conversion parameters Fitted value and 95% confidence interval (CI) as determined by bootstrapping for the ratio of noncrossover to crossover gene conversions, as well as estimates of the initial exponential decay and noncrossover and crossover tract lenghts.

	Mean	CI Lower Bound	CI Upper Bound
Ratio	9.43	6.15	18.45
Exp 1	60.07	37.60	85.75
Exp 2	810.74	698.24	938.36
Exp 3	60.07	37.60	85.75

5.4 Discussion

In this chapter, I showed how k-mers could be identified based on two principles, biased and GC-biased gene conversion, as being associated with PRDM9 binding motifs. The p-values associated with these k-mers are well behaved, with negligible inflation, disjoint sets of k-mers identified in different species, but overlapping sets between the two tests within a species. Clustering the k-mers into PRDM9 motifs revealed striking similarity with PRDM9 motifs identified in *Mus musculus* through either LD rate map or ChIP-seq based methods, validating the method, and proving it to be a powerful method for studying recombination. Across all species and subspecies of mice studied here, dozens of clearly real ancient PRDM9 motifs are identified, usually on non-repeat backgrounds but sometimes on repeat backgrounds, representing hotspots active up to millions of years ago in the past. The abundant diversity of motifs in the different species are largely different, for instance between *M. spretus* and the ancestor of *Mus musculus* (AM) despite the capacity of the hotspot death method to find the same motif on different lineages, confirming the rapid evolutionary turnover of PRDM9. Longer species lineages typically have more motifs, although terminal branches, with more noise (*i.e.* fewer fixed SNPs), generally see fewer motifs. There is also an exceptionally consistent picture of increased weak to strong nucleotide substitutions around ancient PRDM9 motifs, which remains even if one doesn't require GC-bias in the discovery procedure. This suggests GC-biased

gene conversion is likely universal, and doesn't act differently in any of the species or subspecies studied here, or differently based on the PRDM9 allele.

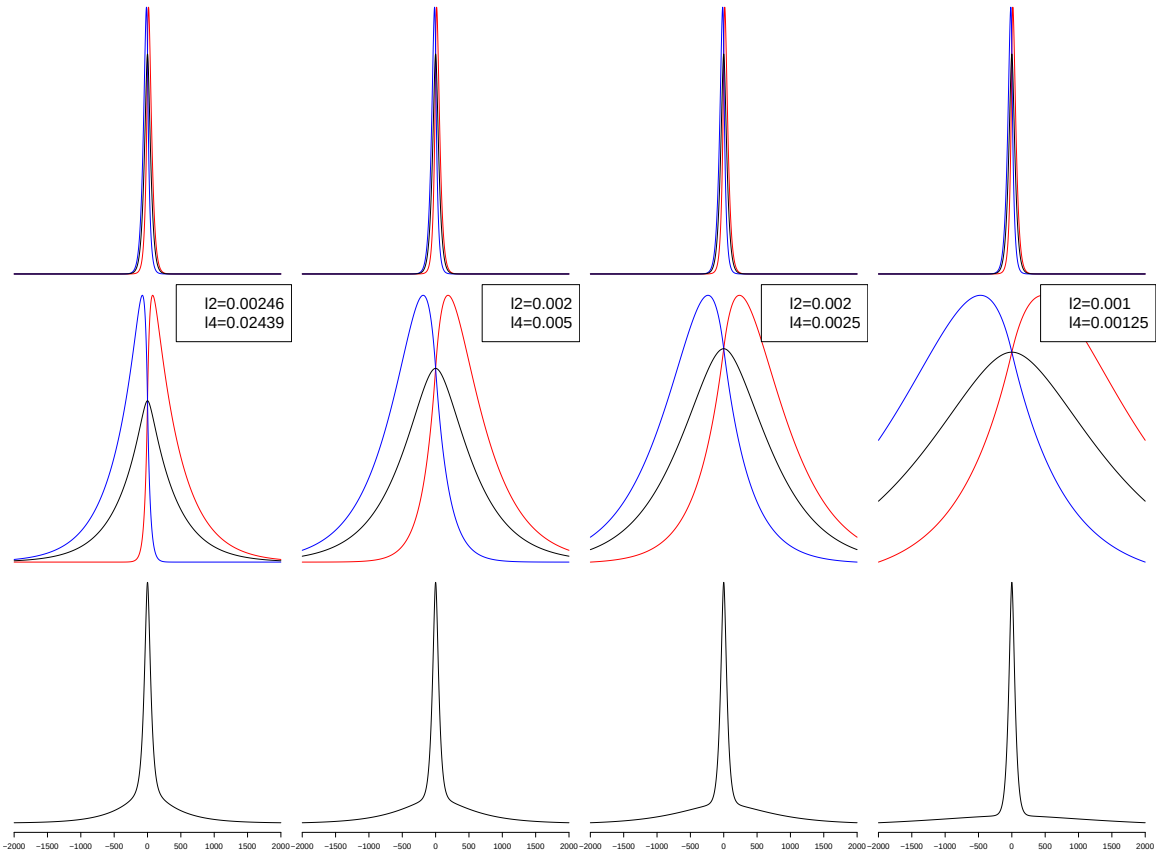
In terms of how the hotspot death motifs compared to the ChIP-seq and LD motifs in *M. m. domesticus* and *M. m. castaneus*, although they are exceptionally similar, there were still some subtle differences. The ChIP-seq motif, by definition, is generated from a single PRDM9 motif. Given derived populations of *Mus musculus* mice are often highly bottlenecked (like the French and Taiwanese), the PRDM9 motif picked up using the LD results may predominantly come from a single PRDM9 motif. By contrast, as we saw in Chapter 1, PRDM9 has considerable allelic diversity in *M. m. domesticus* and *M. m. castaneus*. The hotspot death approach used here might have been more susceptible to clustering together k-mers from similar but different PRDM9 alleles. The subtle difference between motifs identified from the three approaches might therefore represent a combination of either different ancestral zinc fingers, or a susceptibility of the hotspot death method for overclustering multiple related, but distinct, PRDM9 alleles, particularly for *M. m. castaneus*.

The identification of ancient PRDM9 motifs also led to the generation of summary level data on nucleotide subtitutions at ancient PRDM9 binding sites. Among the models fit, the three exponential model, which can be interpreted as DSBs occurring with an exponential distribution away from the PRDM9 motif, and noncrossover and crossover gene conversions occurring away from the side of the DSB and having separate tract lengths governed by exponential distributions, fit the observed data best. The parameters associated with this fit agree largely with the literature, for instance Cole *et al.* estimated the average length of crossover tracts to be about 500-600 bp and noncrossover tract length just below 100 bp [38], and others have observed about a 10:1 ratio of noncrossovers to crossovers [129] (although considerable variation in this estimate exists [130]).

However, while the summary level data does not require interpretation in itself, as

it is a distribution, it should be noted that the fitting of the model, and subsequent parameter estimation, are only one of several possible interpretations. Further model development, in combination with interpretations guided by the literature, may yet yield a better fit to the data and the real underlying biological processes.

Figure 5.13. Example of flexibility of four exponential hotspot death model Each column shows a different set of parameters for the four exponential model. The leftmost shows the original parameters, as shown in Figure 5.12, while the right 3 feature slightly different estimates of the crossover length and crossover DSB parameter, as shown in the legend of the middle plot. Within a column, the top plot shows noncrossover, the middle plot shows crossover, and the bottom plots shows (symmetrized) everything together. Colurs on the top two plots are red for tracts occurring to the right of DSBs, blue is for tracts corresponding to the left side of the DSBs, and black is their average.



For instance, I model crossovers as being initiated from a double strand break which occurs some exponential distance away from the ancient motifs, and non-crossovers and crossovers occuring with lengths governed by an exponential distri-

bution. In Figure 5.13, for the four exponential model fitted in the Results, I show a plot of the underlying tract distribution for the interpreted noncrossover (short exponential) and crossover (long exponential) rates, as well as the different shapes they can take with different parameter values. In Pratto *et al.* [39] in 2014, they suggest strong to weak SNP transitions are mutagenic by noting an strand-specific increase in their occurrence among rare SNPs at PRDM9 binding locations. If we believe that mutagenesis and GC-bias occur at the same uniform rate within gene conversion tract lengths, then we would expect that the longer (crossover) exponential distributions from Figure 5.13 would more closely match Figure 5D from Pratto *et al.* [39], *i.e.* show a more normal like distribution than what we are currently modelling. Yet despite the flexibility in the model to resemble a more normal distribution, as demonstrated in Figure 5.13, the fitted model does not have this shape. Therefore, it is not obviously a modelling deficiency that the longer exponential (crossover) term is non-normal. It is unclear what drives this difference but perhaps it relates to differences in the strength of GC-biased gene conversion depending on its location within the gene conversion tract.

As for noncrossovers versus crossover interpretation, while many individually analyzed hotspots show a tight noncrossover distribution, for example in humans in Odenthal-Hesse *et al.* [36], in Cole *et al.* (in mice), they show a hotspot with a substantially greater distribution of noncrossover tract locations [38] than would be expected based on Figure 5.13. It is also worth noting that in Odenthal-Hesse *et al.* [36], they only note GC-bias for noncrossovers very close to the motif, although they are less well powered to find them further away from the motif. Therefore, it seems possible that noncrossovers arise from two pathways, possibly the DSBR or SDSA pathways, have GC-biased gene conversion in both, but with different rates of occurrence in each, and crossovers are not subject to GC-biased gene conversion.

Finally, it is further important to note that for the ratio of noncrossover to

crossover as presented in the Results, I'm not measuring the biological rates themselves, but how much more likely a SNP is to undergo GC biased gene conversion given it is in either a crossover or noncrossover gene conversion. For instance, the estimates of track length could be accurate, but noncrossovers might convert variants twice as often given they are in a track as crossover, and in such a case I would overestimate by a factor of two the real underlying biological occurences of these events. This disctinction is less important for modelling and evolutionary perspectives, but is important for biological interpretation of these results. Regardless of whether the substitutions are caused by crossover or noncrossover, the summary distributions provided here should help guide our understanding as to the role of GC-biased gene conversion for both selection and genome evolution. For example, given the extreme base pair motif localization, and the wealth of data, I can confidently conclude in mice that the vast majority of GC-biased gene conversion occurs very close to PRDM9 binding hotspots, within 100 base pairs.

Beyond the analysis presented here, this work should also allow aid future research by easily allowing for similar analyses to be conducted in other species. The code to generate these results is robust and flexible. It requires a reference genome, masking of that genome, SNPs from a few species in VCF format, a model relating SNPs to species lineages, and estimate of callabilities genome-wide for the different species lineages. From the raw data, final results are generated in a few hours on a modest computational server, which further permits rapid optimization of the method. These "hotspot death" methods should prove fruitful for the analyses of other species with modestly well understood recombination landscapes, either with PRDM9, or in species such as birds [124] and dogs [131] that do not, using either the loss of k-mers, the accumulation of weak to strong changes near k-mers, or both, to understand *cis* determinants of recombination.

Finally, there are additional analyses which could be performed in addition to

those performed here. Perhaps of most interest is the construction of ancient recombination rate maps, generated by counting the rate of loss of k-mers at broad scales compared to their ancestral counts, possibly coupled with nucleotide substitution rates. These ancestral recombination rate maps, in species with non-uniform recombination rates, can help identify broad scale chromosomal rearrangements through differences in broad scale recombination rate when mapped to a single present day reference genome. Further, particularly in lineage rich species such as birds, they could help identify periods of time, if they exist, in which large changes in broad scale recombination rate occured. If coupled with knowledge of species specific genetic changes, this could suggest genes where different alleles are involved in the broad scale organization of recombination rate architecture, much like PRDM9 organizes the fine scale recombination rate landscape.

Chapter 6

Low coverage imputation in mice and humans

In the preceeding chapters, I presented an analysis of whole genome sequencing data on 69 mice. In this chapter, I present an algorithm I developed named "STITCH" unrelated to the wild mice data, for use in imputation of genotypes from low coverage sequencing data.

The rest of this chapter will be organized as follows. First, a brief introduction will provide context to the problem. Second, I will thoroughly detail the method, and describe two different datasets used to evaluate performance. Third, I will present results for the method and compare it to results from the imputation algorithm Beagle. Finally, I discuss the results and the potential future application of STITCH.

6.1 Introduction

Over the last decade, genome-wide association studies (GWAS) have detected thousands of loci associated with complex traits in the human genome [132]. Generally, these involve genotyping 0.5-1M SNPs on DNA genotyping microarrays, and then employing reference panels such as HapMap [133] or the 1000 Genomes Project [94] to infer genotypes at tens of millions of additional sites, employing algorithms for phasing [75] and imputation [134–137].

However, both high quality genotyping microarrays, and haplotype reference pan-

els, are unavailable for most other species. Where they do exist, such as for mice, imputation performance suffers when populations other than those used to design the chip are analyzed [54].

In place of arrays, it is possible to use low coverage next generation sequencing (NGS) for imputation when large haplotype reference panels are available [138]. Genotyping through low coverage NGS has several advantages: it can be less expensive than a similar array, is naturally customized to the population of interest, and even at modest depth samples most segregating variation. Furthermore, the read-based nature of NGS studies provides phasing information on nearby variants from a single (paired) read. However, in the absence of haplotype reference panels, existing methods for imputation may fail or perform poorly. While some approaches have started to use phase informative reads or similar procedures [77], there are benefits beyond phasing of multiple heterozygous SNPs. First, in the absence of reference haplotypes, phasing with reads may help to initialize the phasing procedure. Second, at high SNP density, it is inaccurate to treat genotypes from the same read as being independent.

Here, I describe a genotype imputation algorithm STITCH (Sequencing To Imputation Through Constructing Haplotypes) suitable for population samples sequenced at low coverage without requiring a haplotype reference panel. The only requirement is a high-quality reference assembly for read-mapping. STITCH models the population as a mosaic of K unknown founder or ancestral haplotypes which are inferred alongside the imputation process. I employ a hidden Markov model (HMM), updated using expectation maximization (EM), similar in spirit to the fastPHASE algorithm [44]. However, I also utilize key features of NGS data. At each iteration, in the expectation phase, ancestral haplotype probabilities are generated for each sample, while in the maximization step, ancestral haplotypes and other parameters are updated using sample haplotype membership; both of these steps are handled

in a read aware fashion based on the underlying sequencing reads. Like fastPHASE the time-complexity of the full algorithm scales quadratically with the number of founders K; however I also present an efficient yet accurate approximation I call "pseudo-haploid" which exploits the read-based nature of NGS and has only linear complexity in K. This makes it feasible to impute populations (including humans) where the haplotype complexity is large, as well as unknown. I demonstrate the utility of STITCH on two datasets from two species: first a set of 2,073 outbred Carworth Farms White (CFW) mice sequenced to 0.15X, and second a set of 11,670 Han Chinese samples sequenced to 1.7X [139].

6.2 Materials and Methods

6.2.1 Overview

In Section 6.2.2 I outline the model by describing how one would simulate (read) data from it, given knowledge of the underlying parameters. In Section 6.2.3 I show in detail how to calculate the probability of the complete data (observations and hidden states) given the parameters. In Section 6.2.4 I give details as to how parameters are initialized. In Section 6.2.5 I describe how parameter updates are calculated, dependent on the expecations of the number of hidden states used in various ways. In Section 6.2.6, I present bounds on the range of parameter values. In Section 6.2.7 I present how the diploid model can be approximated, which I term the "pseudo-haploid" model. In Section 6.2.8 I present useful variables as an extension to the introduction to hidden Markov models presented in Section 1.5.2. In Sections 6.2.9, 6.2.10 and 6.2.11 I lay out how expecations of sums are calculated for the haploid, pseudo-haploid and diploid models, respectively. In Section 6.2.12, I give calculations for efficient calculation of the forward and backward algorithms. In Sections 6.2.13 and 6.2.14, I give details for the CFW and CONVERGE studies. Finally, in Section 6.2.15, I present details of how Beagle was run on the same data.

Note that Table 6.1 summarizes notation commonly used in this chapter.

6.2.2 Simulation under the model

Here I outline the model by describing how one would simulate (read) data from it, given knowledge of the underlying parameters.

I consider a population of individuals that can be approximated as having been founded G generations ago with K unknown ancestral founding haplotypes. Start by considering a single chromosomal region with T SNPs from a present day individual drawn from the model. Begin by choosing starting state k according to π_k, the probability of starting in state k. Let d_t and p_t be the distance (in Mb) and average recombination rate (in M/Mb) between SNPs t and $t+1$, respectively, where 0.01M (1 cM) is defined as the physical distance where the expected number of crossovers in a single generation is 0.01. Therefore $\sigma_t = d_t p_t$ is the recombination distance between SNPs t and $t+1$ in one generation. For example, for a recombination rate of $p_t = 5 \times 10^{-3}$ M/Mb and $d_t = 10^{-3}$ Mb, then $\sigma_t = 5 \times 10^{-3} \times 10^{-3} = 5 \times 10^{-6}$. Because recombination events occur in each generation, I model recombinations at a rate of G per unit of genetic distance. As such, the probability of recombination between two SNPs is $1 - e^{-G\sigma_t}$.

Conditional on the location of recombinations, I sample ancestral haplotypes. I allow genetic drift to have played a sizeable role in the proportions of the ancestral haplotypes in each short genomic interval. As such, I model the probability of choosing ancestral haplotype k, given a recombination between SNPs t and $t+1$ as $\alpha_{t,k}$.

Finally, the reads are sampled conditional on the local haplotype background. Gene conversion, de novo mutation, read-mapping errors and other issues mean that not all chromosomes and reads descended from ancestral haplotype k will be an exact match to the ancestral sequence. I therefore model that for each read, for SNP t and

Table 6.1. Definitions for commonly used variables

Symbol	Definition
K	Number of ancestral or founder haplotypes
T	Number of SNPs in region
N	Number of sample individuals
G	Number of generations since population founding
R_r	Read with index r which spans J_r SNPs, with SNP indices u_r, sequenced bases s_r and base qualities b_r, or $R_r = \{u_r, s_r, b_r\}$
J_r	Number of SNPs spanned by read R_r
c_r	Central SNP for read R_r
O_t	Set of reads with central SNP t, $O_t = \{R_r \vert c_r = t\}$
O	Set of observations for each SNP t on the chromosome, $O = \{O_t \vert t = 1, ..., T\}$
$u_{r,j}$	For SNP j in read R_r, its index with respect to the chromosomal listing of SNPs (*e.g.* If the physical position of SNP t in the region is L_t for $t = 1, ..., T$, then SNP j in read R_r has physical position $L_{u_{r,j}}$)
$s_{r,j}$	Sequencing base for SNP j in read R_r, with $s_{r,j} = 1$ for the alternate base and 0 for the reference base
$b_{r,j}$	Base quality for SNP j in read R_r
$R_{r,j}$	Subset of read R_r for SNP j, or $R_{r,j} = \{u_{r,j}, s_{r,j}, b_{r,j}\}$
$\phi^i_{r,j}$	Probability of SNP j from read R_r coming from an underlying genotype i, or $P(s_{r,j} \vert g = i)$
I_t	Variable counting the number of recombinations that take place between SNPs t and $t+1$
H^j_r	Variable that takes value 1 if SNP j from read R_r is the alternate base and value 0 if it is the reference base
H_r	Variable that takes value 1 if read R_r comes from the maternal haplotype and 2 if it comes from the paternal haplotype
π_k	Probability of starting in state k at the first SNP
σ_t	Recombination distance between SNPs t and $t+1$
$\alpha_{t,k}$	Probability of switching into state k between SNPs t and $t+1$
$\theta_{t,k}$	Probability of haplotype k at SNP t
λ	Parameters of the model $\lambda = \{\pi, \sigma, \alpha, \theta\}$

ancestral haplotype k, that the alternate base will be drawn with probability $\theta_{t,k}$, and the reference base with probability $1-\theta_{t,k}$. Inherent in this is the assumption that the bases in different reads (from the same haplotype) are sampled independently from the same original fixed $\theta_{t,k}$; in reality, there would be a simple sampling of $\theta_{t,k}$ for each haplotype, and reads sampled conditional on these real underlying bases. This assumption is necessary for computational reasons, and has reduced impact for low coverage sequencing data.

Consider sampling the r^{th} read, R_r. To do this, I first choose read breakpoints and determine $u_{r,j}$, the indices of the SNPs in the read for $j = 1, ..., J_r$, where J_r is the number of SNPs in read R_r. I make the assumption that recombinations are infrequent enough that reads can be assumed to have constant state over their length; as such, each read has a central SNP, call it c_r, and state membership over the read is drawn from the central SNP. Therefore, the underlying "real" bases for the sequencing read are sampled according to $\theta_{u_{r,j},k_{c_r}}$. To sample "observed" bases, I then sample $b_{r,j}$, the base qualities of the SNPs in the read. For example, this could be done through their empirical distribution as could be obtained from a BAM file or from the GATK base quality score recalibrator. Once such a distribution is obtained, then observed bases can be sampled according to the real bases and the base qualities.

6.2.3 Complete data probability under the model

In the HMM, for the haploid model, let q_t be the hidden state at SNP t, *i.e.* $q_t \in \{1, ..., K\}$. For the diploid model, let $q_t = (k_{t,1}, k_{t,2})$ be the hidden states at SNP t.

Let $\lambda = \{\pi, \sigma, \alpha, \theta\}$ be the parameters of the model. Initial haploid state probabilities for the $k = 1, ..., K$ different states are defined as π_k. For the diploid model, probabilities for each starting haplotype are independent, so $\pi_{k_{1,1},k_{1,2}} = \pi_{k_{1,1}} \times \pi_{k_{1,2}}$.

For transitions, with probability $e^{-G\sigma_t}$, no recombination occurs between SNPs t and $t+1$, while with probability $1 - e^{-G\sigma_t}$, a recombination occurs and a new state

q_{t+1} is chosen at SNP t according to $\alpha_{t,k_{t+1}}$. This gives the haploid transition matrix

$$P(q_{t+1} = k_{t+1}|q_t = k_t, \lambda) = \begin{cases} e^{-G\sigma_t} + (1 - e^{-G\sigma_t})\alpha_{t,k_{t+1}} & \text{if } k_{t+1} = k_t \\ (1 - e^{-G\sigma_t})\alpha_{t,k_{t+1}} & \text{if } k_{t+1} \neq k_t \end{cases} \tag{6.1}$$

Assuming independence between the two chromosomes then the diploid transition matrix for state $q_{t+1} = (k_{t,1}, k_{t,2})$ at SNP t and $q_{t+1} = (k_{t+1,1}, k_{t+1,2})$ at SNP $t + 1$

$$\begin{aligned} P(q_{t+1} = (k_{t+1,1}, k_{t+1,2})|q_t = (k_{t,1}, k_{t,2}), \lambda) =& P(q_{t+1} = k_{t+1,1}|q_t = k_{t,1}, \lambda)\times \\ & P(q_{t+1} = k_{t+1,2}|q_t = k_{t,2}, \lambda) \end{aligned} \tag{6.2}$$

For the emission of reads, for read R_r, let c_r be the index of the most central SNP in that read, choosing at random when a read intersects exactly two SNPs. Reads that don't intersect any SNPs are removed as they are uninformative. Consider the probability of an observation of a set of reads whose central SNP is t, or in other words $O_t = \{R_r|c_r = t\}$. For SNP j in read R_r, $b_{r,j}$ is the Phred scaled base quality, and by definition is the probability that the base is called erroneously, so let $\epsilon_{r,j} = 10^{\frac{-b_{r,j}}{10}}$. Then

$$P(s_{r,j}|g) = \begin{cases} 1 - \epsilon_{r,j} & \text{if } s_{r,j} = g \\ \frac{1}{3}\epsilon_j & \text{if } s_{r,j} \neq g \end{cases} \tag{6.3}$$

For convenience, set $\phi^i_{r,j} = P(s_{r,j}|g = i)$. I disregard sequenced bases which are not the reference or alternate base.

Given there are J_r SNPs in read R_r, the probability of drawing read R_r from haplotype k is the product of the contribution of each SNP $j = 1, ..., J_r$ in that read. For the j^{th} SNP, the probability is the product the read sampled contained the alternate base $\phi^1_{r,j}$ times the probability $\theta_{u_{r,j},k}$ that ancestral haplotype k emitted the alternate base, added to the equivalent probability for the reference base. Taken together, this yields

$$P(R_r|q_t = k, \lambda) = \prod_{j=1}^{J_r} \left(\theta_{u_{r,j},k}\phi^1_{r,j} + (1 - \theta_{u_{r,j},k})\phi^0_{r,j})\right) \tag{6.4}$$

In the haploid model, I then get the probability of the observation at that locus as

$$P(O_t|q_t = k_t, \lambda) = \prod_{R_r \in O_t} P(R_r|q_t = k_t, \lambda) \tag{6.5}$$

In the diploid model, each read is equally likely to come from either the maternal or paternal chromosome, giving

$$P(R_r|q_t = (k_{t,1}, k_{t,2}), \lambda) = \frac{1}{2}P(R_r|q = k_{t,1}, \lambda) + \frac{1}{2}P(R_r|q = k_{t,2}, \lambda) \tag{6.6}$$

For every SNP, I get the probability of the observations at that locus as

$$P(O_t|q_t = (k_{t,1}, k_{t,2}), \lambda) = \prod_{R_r \in O_t} P(R_r|q_t = (k_{t,1}, k_{t,2}), \lambda) \tag{6.7}$$

Finally, note that for SNPs which are not covered by reads, I set $P(O_t|q_t = k_t, \lambda) = 1$ for all k_t.

6.2.4 Initialization

Haploid probabilities π_k are initialized with equal weights $\pi_k = \frac{1}{K}$, as are diploid priors $\pi_{k_1,k_2} = \frac{1}{K \times K}$. The state probabilities $\alpha_{t,k}$ are also initialized with equal weights $\alpha_{t,k} = \frac{1}{K}$. The recombination distance is initialized assuming a constant recombination rate multiplied by the physical distance between SNPs, for example assuming $d_t \times$ 0.5cM/Mb where d_t is the physical distance between SNPs t and $t+1$. Finally, given a lower bound δ on emission probabilities, for example $\delta = 0.0001$, $\theta_{t,k}$ are sampled from a $U(\delta, 1-\delta)$ distribution. Note that I leave G as a user set parameter, which can be approximated from Ne as $G = \frac{4\text{Ne}}{K}$.

6.2.5 Maximization and parameter updating

In the EM algorithm, a "complete dataset" D includes the observed data (O, the reads), as well as the hidden parameters (Q, the hidden states). Given a set of parameters λ, the log-likehood of the complete data is $L(\lambda) = \log(l(\lambda|D)) = \log(l(\lambda|D = (O, Q)))$. Given a current set of parameters λ^i, I generate a new set of parameters

λ^{i+1} to maximize the expectation of $l(\lambda^{i+1})$ with respect to the distribution of hidden parameters obtained by λ^i

$$\begin{aligned} U(\lambda^{i+1}, \lambda^i) =& \mathbb{E}[l(\lambda^{i+1})|O, \lambda^i] \\ =& \sum_Q P(Q|O, \lambda^i) \log(P(O, Q|\lambda^{i+1})) \end{aligned} \tag{6.8}$$

By choosing λ^{i+1} to maximize $U(\lambda^{i+1}, \lambda^i)$, I never decrease the likelihood of the observed data, $l(\lambda^{i+1}|O) > l(\lambda^i|O)$. This was shown in Section 1.5.1.

In applying the EM algorithm, I first initialize with a set of parameters λ^0. For each subsequent iteration $i = 1, 2, ...$, I then iteratively alternate between the "Expectation" phase, where I calculate $U(\lambda^{i+1}, \lambda^i)$, and the "Maximization" phase, where I calculate λ^{i+1} to maximize $U(\lambda^{i+1}, \lambda^i)$. In the Expectation phase, the crucial component is calculating the state probabilities $P(Q|O, \lambda^i)$ - these are calculated using the forward and backward algorithms. To calculate the updates in the Maximization stage, I must further augment the latent space to model how many recombinations occur between SNPs, whether emissions were due to occurences of an alternate base or a reference base, and whether observed reads were from the maternal or paternal haplotype. In this new augmented latent space, for some fixed set of hidden parameters for the N samples, consider some sums that can be calculated. Let n_k^1 be the number of sample haplotypes in state k at the first SNP, n_{stay}^t be the number of sample haplotypes which do not recombine between SNPs t and $t+1$, $n_{\text{switch},k}^t$ be the number of sample haplotypes which switch into ancestry k between SNPs t and $t+1$, and $n_{k,s}^t$ be the number of reads that have a reference $s = 0$ or alternate $s = 1$ base for SNP t that are in state k for their central SNP. Then the complete data log likelihood is

$$
\begin{aligned}
l(\lambda) =& \log(P(O, Q|\lambda)) \\
=& \sum_{k=1}^{K} n_k^1 \log(\pi_k) \\
&+ \sum_{t=1}^{T-1} n_{\text{stay}}^t \log(e^{-G\sigma_t}) + \sum_{t=1}^{T-1} \sum_{k=1}^{K} n_{\text{switch},k}^t \log((1 - e^{-G\sigma_t})\alpha_{t,k}) \\
&+ \sum_{t=1}^{T} \sum_{k=1}^{K} n_{k,1}^t \log(\theta_{t,k}) + \sum_{t=1}^{T} \sum_{k=1}^{K} n_{k,0}^t \log((1 - \theta_{t,k}))
\end{aligned} \tag{6.9}
$$

Calculating updates for a parameter is done by taking the derivative of $U(\lambda^{i+1}, \lambda^i)$ with respect to that parameter, setting it equal to 0 and solving. Employing the notation $\mathbb{E}[x|O, \lambda] = \mathbb{E}_\lambda[x]$, it is easy to calculate the following updates for $\lambda^{i+1} = (\pi^{i+1}, \theta^{i+1}, \alpha^{i+1}, \sigma^{i+1})$

$$
\pi_k^{i+1} = \frac{\mathbb{E}_{\lambda^i}[n_k^1]}{\sum_{j=1}^{K} \mathbb{E}_{\lambda^i}[n_j^1]} \tag{6.10}
$$

$$
\theta_{t,k}^{i+1} = \frac{\mathbb{E}_{\lambda^i}[n_{k,1}^t]}{\mathbb{E}_{\lambda^i}[n_{k,0}^t] + \mathbb{E}_{\lambda^i}[n_{k,1}^t]} \tag{6.11}
$$

$$
\alpha_{t,k}^{i+1} = \frac{\mathbb{E}_{\lambda^i}[n_{\text{switch},k}^t]}{\sum_{j=1}^{K} \mathbb{E}_{\lambda^i}[n_{\text{switch},j}^t]} \tag{6.12}
$$

$$
\sigma_t^{i+1} = \frac{1}{-G} \log\left(\frac{\sum_{k=1}^{K} \mathbb{E}_{\lambda^i}[n_{\text{switch},k}^t]}{\sum_{k=1}^{K} \mathbb{E}_{\lambda^i}[n_{\text{switch},k}^t] + \mathbb{E}_{\lambda^i}[n_{\text{stay}}^t]} \right) \tag{6.13}
$$

6.2.6 Parameter bounding

After parameter updating, newly calculated parameters are bounded with default but user tunable parameters. Prior probabilities π_k, new state parameters $\alpha_{t,k}$, and emission probabilities $\theta_{t,k}$ (and $1 - \theta_{t,k}$) whose values are less than a threshold are set equal to that threshold, and then probabilities re-normalized as appropriate to have sum 1. Under default conditions this bound is 1×10^{-4}. For the recombination distance, values of σ_t that exceed implied upper (default 100 cM/Mb) and lower (default 0.1 cM/Mb) bounds are reset to the bound value.

6.2.7 Pseudo-haploid approximation

The diploid model presented here and used in fastPHASE and other similar algorithms suffers from a quadratic computational complexity due to the need to sum over K^2 possible diploid states at each site. With sequencing reads, the observed data fundamentally comes from either the first (*e.g.* maternal) or second (*e.g.* paternal) haplotype. If I had labels for each read as to whether they came from the maternal or paternal haplotype, I would have separable likelihoods, and could use the maternal reads to infer the maternal states, and likewise for the paternal reads and paternal states, which would have computational cost proportional to $2 \times K$ as opposed to K^2.

In the diploid EM algorithm, I use the current set of parameters to generate the posterior probability of the pair of hidden states given the observations, and use these to generate a new set of parameters that maximize the likelihood. An alternative approach is to average over sampled hidden states realized through a hypothetical Gibbs sampler that i) samples labels conditional on states, observations, and parameters, and ii) samples states conditional on labels, observations and parameters. Implementing such a Gibbs sampler in reality would be computationally unwise, as it would likely take at least as long as the original diploid EM. However, with certain assumptions about the posterior distribution of the labels, I can approximate the posterior distribution of the hidden states quickly.

Let q_1 be the full hidden state for haplotype 1, the maternal haplotype. Let H_r be the label for read r with $H_r = 1$ corresponding to the maternal haplotype and $H_r = 2$ corresponding to the paternal haplotype. Let $O = \{R_r\}$ be the set of all reads, with $|O|$ reads in total, and let H correspond to an assignment of labels $H \in \mathcal{H} = \{1, 2\}^{|O|}$.

Let $\mathcal{R}_h = \{R_r | H_r = h\}$ be the set of reads with label h. Then I have

$$P(q_1|O,\lambda) = \sum_{H\in\mathcal{H}} P(q_1, H|O,\lambda) \tag{6.14}$$

$$= \sum_{H\in\mathcal{H}} P(q_1|H,O,\lambda)P(H|O,\lambda) \tag{6.15}$$

$$= \sum_{H\in\mathcal{H}} P(q_1|H,O,\lambda)\prod_{r=1}^{|O|} P(H_r|O,\lambda) \tag{6.16}$$

where the last equality requires the approximation that the probability of the labels are independent of each other. Now, the probability of a state given labels and reads can be further written as

$$P(q_1|H,O,\lambda) = \frac{P(O|H,q_1,\lambda)P(q_1|H,\lambda)}{P(O|H,\lambda)} \tag{6.17}$$

$$= \frac{\Big(\prod_{r:H_r=1} P(R_r|q_1,\lambda)\Big)P(\mathcal{R}_2|\text{hap2},\lambda)P(q_1|\lambda)}{P(\mathcal{R}_1|\text{hap1},\lambda)P(\mathcal{R}_2|\text{hap2},\lambda)} \tag{6.18}$$

where $P(q_1|H,\lambda) = P(q_1|\lambda)$, since labels don't affect state probabilities without observations. If I further approximate $P(\mathcal{R}_1|\text{hap1},\lambda) = \prod_{r:H_r=1} P(R_r|\text{hap1},\lambda)$, and approximate $P(R_r|\text{hap1},\lambda) = P(R_r|\lambda)$, I get that

$$P(q_1|H,O,\lambda) = P(q_1|\lambda)\prod_{r:H_r=1}\frac{P(R_r|q_1,\lambda)}{P(R_r|\lambda)} \tag{6.19}$$

Subbing Equation 6.19 into Equation 6.16, I get that

$$P(q_1|O,\lambda) = \Bigg(\sum_{H\in\mathcal{H}} P(q_1|\lambda)\Big(\prod_{r:H_r=1}\frac{P(R_r|q_1,\lambda)}{P(R_r|\lambda)}\Big)\Bigg)\Big(\prod_{r=1}^{|O|} P(H_r|O,\lambda)\Big) \tag{6.20}$$

$$= P(q_1|\lambda)\sum_{H\in\mathcal{H}}\prod_{r=1}^{|O|}\left(P(H_r|O,\lambda)\Big(\mathcal{I}\{H_r=1\}\frac{P(R_r|q_1,\lambda)}{P(R_r|\lambda)} + \mathcal{I}\{H_r=2\}1\Big)\right) \tag{6.21}$$

$$= P(q_1|\lambda)\prod_{r=1}^{|O|}\Big(P(H_r=1|O,\lambda)\frac{P(R_r|q_1,\lambda)}{P(R_r|\lambda)} + P(H_r=2|O,\lambda)\Big) \tag{6.22}$$

where the step from Equation 6.21 to Equation 6.22 follows as every possible labeling $H \in \mathcal{H}$ from Equation 6.21 corresponds in a one-to-one fashion to an expanded form of Equation 6.22. A generalized form of this step is proven in Theorem

6.1. Therefore, I get that read r contributes $P(H_r = 1|O, \lambda)P(R_r|q_1, \lambda) + P(H_r = 2|O, \lambda)P(R_r|\lambda)$ to the likelihood, after multiplying by the constant $P(R_r|\lambda)$, as opposed to $P(R_r|q_1, \lambda)$ as it would under a fully separable model. When testing on real data, I found that I achieved marginally but consistently better performance using $P(H_r = 1|O, \lambda)P(R_r|q_1, \lambda) + P(H_r = 2|O, \lambda)P(R_r|\text{hap2}, \lambda)$ instead, so this equation was used when calculating the state probabilities.

To use this, I need an estimate of the probability of a label given the data. To do this, consider a read R_r, with lead SNP c_r, and label H_r. Then I can calculate the following

$$\begin{aligned}
P(H_r = 1|O, \lambda) &= \sum_{q_1,q_2} P(H_r|q_1, q_2, O, \lambda)P(q_1, q_2|O, \lambda) && (6.23)\\
&= \sum_{q_1,q_2} P(H_r|q_1, q_2, R_r, \lambda)P(q_1, q_2|O, \lambda) && (6.24)\\
&= \sum_{q_1,q_2} \frac{P(R_r|q_1, \lambda)}{P(R_r|q_1, \lambda) + P(R_r|q_2, \lambda)} P(q_1, q_2|O, \lambda) && (6.25)\\
&= \mathbb{E}_{q_1,q_2}\Big[\frac{P(R_r|q_1, \lambda)}{P(R_r|q_1, \lambda) + P(R_r|q_2, \lambda)}|O, \lambda\Big] && (6.26)\\
&\approx \frac{\mathbb{E}_{q_1}[P(R_r|q_1, \lambda)|O, \lambda]}{\sum_{h=1}^{2} \mathbb{E}_{q_h}[P(R_r|q_h, \lambda)|O, \lambda]} && (6.27)
\end{aligned}$$

This uses a prior probability on labels of $P(H_r = 1) = P(H_r = 2) = \frac{1}{2}$. I also approximate the expectation of ratios as a ratio of expectations, to avoid a calculation with computational complexity of order K^2. Finally I calculate

$$P(R_r|\text{hap}h, \lambda) = \mathbb{E}_{q_h}[P(R_r|q_h, \lambda)|O, \lambda] \approx \sum_{k=1}^{K} P(R_r|q_h = k, \lambda')P(q_k|O, \lambda') \quad (6.28)$$

where λ' are the parameters from the previous iteration.

Therefore, in calculating the complete data probability for the pseudo-haploid model for haplotype $H = 1$, I use initial probabilities π_k, transition parameters from

Equation 6.1, and the probability of an observation given a state as

$$P_{H=h}(O_t|q_t = k_t, \lambda) = \prod_{r=1}^{|O_t|} P_{H=h}(R_r|q_t = k_t, \lambda)$$
$$= \prod_{r=1}^{|O_t|} P(H_r = h|O, \lambda)P(R_r|q_1, \lambda) + P(H_r \neq h|O, \lambda)P(R_r|\text{hap}h, \lambda) \tag{6.29}$$

where $P(R_r|q_1, \lambda)$ is from Equation 6.4, $P(H_r = h|O, \lambda)$ is from Equation 6.27, and $P(R_r|\text{hap}h, \lambda)$ is from Equation 6.28.

Theorem 6.1. *Let $N \in \mathbb{N}$, $n \in \{1, 2, ..., N\}$, $k \in \{1, 2\}$, and $A_{n,k} \in \mathbb{R}$. Then $\sum_{H=(H_1,H_2,...,H_N)\in\{1,2\}^N} \prod_{n=1}^{N} A_{n,H_n} = \prod_{n=1}^{N}(A_{n,1} + A_{n,2})$.*

Proof. Proof by induction. For $N = 1$, we have

$$\sum_{H\in\{1,2\}^N} \prod_{n=1}^{N} A_{n,H_n} = \sum_{H\in\{1,2\}} \prod_{n=1}^{1} A_{n,H} = A_{1,1} + A_{1,2} = \prod_{n=1}^{N}(A_{n,1} + A_{n,2}) \tag{6.30}$$

Now assume the equation holds for N, and consider $N + 1$. Then

$$\sum_{H\in\{1,2\}^{N+1}} \prod_{n=1}^{N+1} A_{n,H_n}$$
$$= \sum_{H_{N+1}=1}^{2} (\sum_{H\in\{1,2\}^N} \prod_{n=1}^{N+1} A_{n,H_n})$$
$$= \sum_{H_{N+1}=1}^{2} A_{N+1,H_{N+1}}(\sum_{H\in\{1,2\}^{N+1}} \prod_{n=1}^{N} A_{n,H_n})$$
$$= \sum_{H_{N+1}=1}^{2} A_{N+1,H_{N+1}}(\prod_{n=1}^{N}(A_{n,1} + A_{n,2}))$$
$$= A_{N+1,1}(\prod_{n=1}^{N}(A_{n,1} + A_{n,2})) + A_{N+1,2}(\prod_{n=1}^{N}(A_{n,1} + A_{n,2}))$$
$$= \prod_{n=1}^{N+1}(A_{n,1} + A_{n,2})$$

□

6.2.8 Useful variables

I use a standard forward backward HMM implementation with a set of parameters λ. Recall that q_t is the hidden state at SNP t. I use the following notations for states

k_t at SNP t and k_{t+1} at SNP $t+1$

$$\begin{aligned}
\alpha_t(k_t) &= P(O_1O_2...O_t, q_t = k_t|\lambda) \\
\beta_t(k_t) &= P(O_{t+1}O_{t+2}...O_T|q_t = k_t, \lambda) \\
\gamma_t(k_t) &= P(q_t = k_t|O, \lambda) = \frac{\alpha_t(k_t)\beta_t(k_t)}{P(O|\lambda)} \\
\xi_t(k_t, k_{t+1}) &= P(q_t = k_t, q_{t+1} = k_{t+1}|O, \lambda) \\
&= \frac{\alpha_t(k_t)P(q_{t+1} = k_{t+1}|q_t = k_t, \lambda)\beta_{t+1}(k_{t+1})P(O_{t+1}|q_{t+1} = k_{t+1}, \lambda)}{P(O|\lambda)}
\end{aligned}$$

The diploid version of these equations, where we go from state $(k_{t,1}, k_{t,2})$ at SNP t to state $(k_{t+1,1}, k_{t+1,2})$ at SNP $t+1$ is

$$\begin{aligned}
\alpha_t(k_{t,1}, k_{t,2}) =& P(O_1O_2...O_t, q_t = (k_{t,1}, k_{t,2})|\lambda) \\
\beta_t(k_{t,1}, k_{t,2}) =& P(O_{t+1}O_{t+2}...O_T|q_t = (k_{t,1}, k_{t,2}), \lambda) \\
\gamma_t(k_{t,1}, k_{t,2}) =& P(q_t = (k_{t,1}, k_{t,2})|O, \lambda) = \frac{\alpha_t(k_{t,1}, k_{t,2})\beta_t(k_{t,1}, k_{t,2})}{P(O|\lambda)} \\
\xi_t\Big((k_{t,1}, k_{t,2}), (k_{k+1,1}, k_{t+1,2})\Big) =& P(q_t = (k_{t,1}, k_{t,2}), q_{t+1} = (k_{t+1,1}, k_{t+1,2})|O, \lambda) \\
=& \frac{1}{P(O|\lambda)}\alpha_t(k_{t,1}, k_{t,2})P(q_{t+1} = (k_{t+1,1}, k_{t+1,2})|q_t = (k_{t,1}, k_{t,2}), \lambda)\times \\
& \beta_{t+1}(k_{t+1,1}, k_{t+1,2})P(O_{t+1}|q_t = (k_{t,1}, k_{t,2}), \lambda)
\end{aligned}$$

6.2.9 Parameter updating under haploid model

Initial probabilities

To update the prior parameters, I need the expectation of n_k^1, which I define as the number of sample haplotypes in state k at the first SNP. Denote the probability that the sample is in state k at SNP t by $\gamma_t(k)$. Let $\gamma_{n,t}(k)$ be $\gamma_t(k)$ for sample n. I can therefore calculate the required expectation for Equation 6.10 as

$$\mathbb{E}_\lambda[n_k^1] = \sum_{n=1}^{N} \gamma_{n,1}(k) \tag{6.31}$$

Transition matrix probabilities

To update the transition parameters, I use an augmented state space where I have knowledge of how many recombinations occured between two SNPs. Define a variable I_t as the count of the number of recombinations between SNPs t and $t+1$; in the haploid model, this takes value 0 or 1. This will allow for the calculation of the expectation of n^t_{stay}, the number of sample haplotypes that do not recombine between SNPs t and $t+1$, and $n^t_{\text{switch},k}$, the number that switch into state k between SNPs t and $t+1$.

We extend our transition probability to include I_t as follows

$$P(q_{t+1}=k_{t+1}, I_t | q_t = k_t, \lambda) = \begin{cases} e^{-G\sigma_t} & \text{if } k_t = k_{t+1} \text{ and } I_t = 0 \\ 0 & \text{if } k_t \neq k_{t+1} \text{ and } I_t = 0 \\ (1-e^{-G\sigma_t})\alpha_{t,k_{t+1}} & \text{if } I_t = 1 \end{cases}$$

Recall that $\xi_t(k_t, k_{t+1})$ is

$$\xi_t(k_t,k_{t+1}) = \frac{\alpha_t(k_t)P(q_{t+1}=k_{t+1}|q_t=k_t,\lambda)\beta_{t+1}(k_{t+1})P(O_{t+1}|q_{t+1}=k_{t+1},\lambda)}{P(O|\lambda)} \tag{6.32}$$

Denote the probability given the observed data O that across SNP t, the sample has states k_t, k_{t+1} and indicator I_t by $\xi_t(k_t, k_{t+1}, I_t)$. Then

$$\xi_t(k_t,k_{t+1},I_t) = \frac{\alpha_t(k_t)P(q_{t+1}=k_{t+1},I_t|q_t=k_t,\lambda)\beta_{t+1}(k_{t+1})P(O_{t+1}|q_{t+1}=k_{t+1},\lambda)}{P(O|\lambda)} \tag{6.33}$$

Let $\xi_t(k_t, k_{t+1}, I_t)$ be $\xi_{n,t}(k_t, k_{t+1}, I_t)$ for sample n. We can therefore calculate expectations as

$$\mathbb{E}_\lambda[n^t_{\text{stay}}] = \sum_{n=1}^{N}\sum_{k=1}^{K}\xi_{n,t}(k,k,I_t=0) \tag{6.34}$$

$$\mathbb{E}_\lambda[n^t_{\text{switch},k}] = \sum_{n=1}^{N}\sum_{i=1}^{K}\xi_{n,t}(i,k,I_t=1) \tag{6.35}$$

and since

$$\mathbb{E}_\lambda[n^t_{\text{stay}}] = N - \sum_{n=1}^{N}\sum_{i=1}^{K}\sum_{k=1}^{K}\xi_{n,t}(i,k,I_t=1) = N - \sum_{k=1}^{K}\mathbb{E}_\lambda[n^t_{\text{switch},k}] \tag{6.36}$$

it is therefore sufficient to calculate $\mathbb{E}_\lambda[n^t_{\text{switch},k}]$ to perform the EM updating in Equation 6.12 and Equation 6.13.

Emission matrix probabilities

To update the emission parameters, I use an augmented state space with knowledge of whether emissions were due to the alternate or reference base. Recall that $\phi^i_{r,j}$ is the probability SNP j in read R_r came from a read with underlying genotype i. Denote by H^j_r a variable which takes value 1 if the underlying base is the alternate base and 0 if it is the reference base. I will use this to calculate the expectation of $n^t_{k,s}$, the number of reads with a base at SNP t that contain the alternate ($s = 1$) or reference ($s = 0$) base where the sample was in state k at the central SNP of the read.

Recall that the original definition of the probability of read R_r given hidden state k at SNP t and parameters λ is

$$P(R_r|q_t = k, \lambda) = \prod_{j=1}^{J_r} P(R_{r,j}|q_t = k, \lambda) = \prod_{j=1}^{J_r} \left(\phi^1_{r,j}\theta_{u_{r,j},k} + \phi^0_{r,j}(1 - \theta_{u_{r,j},k})\right) \quad (6.37)$$

I extend the emission probability to include H^j_r as follows

$$P(R_r, H^j_r|q_t = k_t, \lambda) = \begin{cases} \left[\prod_{i\neq j} P(R_{r,i}|q_t = k_t, \lambda)\right]\phi^1_{r,j}\theta_{u_{r,j},k} & \text{if } H^j_r = 1 \\ \left[\prod_{i\neq j} P(R_{r,i}|q_t = k_t, \lambda)\right]\phi^0_{r,j}(1 - \theta_{u_{r,j},k}) & \text{if } H^j_r = 0 \end{cases} \quad (6.38)$$

For read R_r with central SNP c_r, the probability of the observation (set of reads) at SNP $t = c_r$ and H^j_r becomes

$$P(O_t, H^j_r|q_t = k_t, \lambda) = \begin{cases} P(O_t|q_t = k_t, \lambda)\frac{\phi^1_{r,j}\theta_{u_{r,j},k}}{\phi^1_{r,j}\theta_{u_{r,j},k} + \phi^0_{r,j}(1-\theta_{u_{r,j},k})} & \text{if } H^j_r = 1 \\ P(O_t|q_t = k_t, \lambda)\frac{\phi^0_{r,j}(1-\theta_{u_{r,j},k})}{\phi^1_{r,j}\theta_{u_{r,j},k} + \phi^0_{r,j}(1-\theta_{u_{r,j},k})} & \text{if } H^j_r = 0 \end{cases} \quad (6.39)$$

I expand $\gamma_t(k_t)$ as

$$\begin{aligned} \gamma_t(k_t) =& \frac{\alpha_t(k_t)\beta_t(k_t)}{P(O|\lambda)} \\ =& \frac{[\sum_{l=1}^K \alpha_{t-1}(l)P(q_t = k_t|q_{t-1} = l, \lambda)]P(O_t|q_t = k, \lambda)\beta_t(k_t)}{P(O|\lambda)} \end{aligned} \quad (6.40)$$

where we note that for $t = 1$, I substitute π_k for $[\sum_{l=1}^K \alpha_{t-1}(l)P(q_t = k_t|q_{t-1} = l, \lambda)]$. Denote the probability that for SNP j in read R_r with central SNP $t = c_r$, the sample

has a hidden state k_t and has indicator H_r^j given observed data O and parameters λ by $\gamma_t(k_t, H_r^j)$. Then

$$\begin{aligned}\gamma_t(k_t, H_r^j) &= \frac{[\sum_{l=1}^{K} \alpha_{t-1}(l) P(q_t = k_t | q_{t-1} = l, \lambda)]}{P(O|\lambda)} P(O_t, H_r^j | q_t = k, \lambda) \\ &= \begin{cases} \gamma_t(k_t) \frac{\phi^1_{r,j} \theta_{u_{r,j},k}}{\phi^1_{r,j} \theta_{u_{r,j},k} + \phi^0_{r,j}(1-\theta_{u_{r,j},k})} & \text{if } H_r^j = 1 \\ \gamma_t(k_t) \frac{\phi^0_{r,j}(1-\theta_{u_{r,j},k})}{\phi^1_{r,j} \theta_{u_{r,j},k} + \phi^0_{r,j}(1-\theta_{u_{r,j},k})} & \text{if } H_r^j = 0 \end{cases}\end{aligned} \tag{6.41}$$

Let $\gamma_{n,t}(k_t, H_r^j)$ be $\gamma_t(k_t, H_r^j)$ for sample n, and let A_n be the complete set of SNPs j from reads R_r for sample n such that $u_{r,j} = t$. We can therefore calculate the required expectations for Equation 6.11 as

$$\mathbb{E}_\lambda[n^t_{k,1}] = \sum_{n=1}^{N} \sum_{(r,j) \in A_n} \gamma_{n,c_r}(k, H_r^j = 1) \tag{6.42}$$

$$\mathbb{E}_\lambda[n^t_{k,0}] = \sum_{n=1}^{N} \sum_{(r,j) \in A_n} \gamma_{n,c_r}(k, H_r^j = 0) \tag{6.43}$$

6.2.10 Parameter updating under pseudo-haploid model

In the pseudo-haploid model, the only changes to the likelihood occur through the emissions (Equations 6.42 and 6.43). To update the emission parameters for the pseudo-haploid model, I use an augmented state space with knowledge of whether emissions were due to the alternate or reference base, and further have knowledge of whether the read came from the maternal or paternal haplotype. Recall that $\phi^i_{r,j}$ is the probability that observed base j in read R_r came from a read with underlying genotype i. Recall that H_r^j is an indicator variable which takes value 1 if the underlying base is the alternate base and 0 if it is the reference base. Let H_r take value 1 if the read came from the maternal haplotype and 2 if it came from the paternal haplotype. I will use these to calculate the expectation of $n^t_{k,s}$, the number of reads that emit the alternate base ($s = 1$) or reference base ($s = 0$) given they are in state k at the central SNP of the read.

Recall that for each individual, I make two forward backward passes of the algorithm, once for the maternal haplotype ($h = 1$), and a second time for the paternal haplotype ($h = 2$). I also attempt to probabilistically infer for each read which haplotype it came from. Let H refer to the haplotype we are currently modelling (maternal or paternal).

First, recall that the original definition of the probability while modelling haplotype h of read R_r given hidden state k at SNP t and parameters λ is

$$\begin{aligned} P_{H=h}(R_r|q_t = k_t, \lambda) = P(R_r|q_t = k_t, \lambda)P(H_r = h|O, \lambda) \\ + P(R_r|H_r \neq h, \lambda)P(H_r \neq h|O, \lambda) \end{aligned} \tag{6.44}$$

For notational convenience, set $G_{r,j,h} = P(H_r = h|O, \lambda)\left[\prod_{i \neq j} P(R_{r,i}|q_t = k, \lambda)\right]$. I therefore expand the emission probability to include H_r^j and H_r as follows

$$P_{H=h}(R_r, H_r^j, H_r|q_t = k, \lambda) = \begin{cases} G_{r,j,h}\theta_{u_{r,j},k}\phi_{r,j}^1 & \text{if } H_r^j = 1, H_r = h \\ G_{r,j,h}(1 - \theta_{u_{r,j},k})\phi_{r,j}^0 & \text{if } H_r^j = 0, H_r = h \\ P(H_r \neq h|O, \lambda)P(R_r|H_r \neq h, \lambda) & \text{if } H_r \neq h \end{cases}$$

Denote the probability that haplotype h of the sample is in state k_t at SNP t with H_r^j and H_r given observed data O and parameters λ by $\gamma_{t,h}(k_t, H_r^j, H_r)$. Then

$$\gamma_{t,h}(k_t, H_r^j, H_r) = \begin{cases} \gamma_{t,h}(k_t)\frac{G_{r,j,h}\theta_{u_{r,j},k}\phi_{r,j}^1}{P_{H=h}(R_r|q_t=k_t,\lambda)} & \text{if } H_r^j = 1, H_r = h \\ \gamma_{t,h}(k_t)\frac{G_{r,j,h}(1-\theta_{u_{r,j},k})\phi_{r,j}^0}{P_{H=h}(R_r|q_t=k_t,\lambda)} & \text{if } H_r^j = 0, H_r = h \end{cases}$$

Let $\gamma_{n,t,h}(k_t, H_r^j, H_r)$ be $\gamma_{t,h}(k_t, H_r^j, H_r)$ for sample n, and let A_n be the complete set of SNPs j from reads R_r for sample n such that $u_{r,j} = t$. I can therefore calculate the required expectations from Equation 6.11 in the main text as

$$\mathbb{E}_\lambda[n_{k,1}^t] = \sum_{n=1}^{N} \sum_{(r,j)\in A_n} \sum_{h=1}^{2} \gamma_{n,c_r,h}(k, H_r^j = 1, H_r = h) \tag{6.45}$$

$$\mathbb{E}_\lambda[n_{k,0}^t] = \sum_{n=1}^{N} \sum_{(r,j)\in A_n} \sum_{h=1}^{2} \gamma_{n,c_r,h}(k, H_r^j = 0, H_r = h) \tag{6.46}$$

6.2.11 Parameter updating under diploid model

Initial probabilities

To update the prior parameters, I need the expectation of n_k^1, which is defined as the number of sample haplotypes in state k at the first SNP. Denote the probability that sample n is in pairs of states $(k_{t,1}, k_{t,2})$ at SNP t given observed data O by $\gamma_{n,t}(k_{t,1}, k_{t,2})$. We can therefore calculate the required expectation for Equation 6.10 as

$$\mathbb{E}_\lambda[n_k^1] = \sum_{n=1}^{N} \sum_{j=1}^{K} \left(\gamma_{n,1}(k,j) + \gamma_{n,1}(j,k)\right) \tag{6.47}$$

Transition probabilities

To update the transition parameters for the diploid model, I use an augmented state space with knowledge of how many recombinations occured between two SNPs. Here I define a variable I_t which counts the number of recombinations that occur between SNPs t and $t+1$ for the two haplotypes of the diploid sample, and takes values 0, 1 or 2. This will allow for the calculation of the expectation of n_{stay}^t, the number of sample haplotypes that do not recombine between SNPs t and $t+1$, and $n_{\text{switch},k}^t$, the number of sample haplotypes that switch into state k between SNPs t and $t+1$.

I can therefore extend the diploid transition probability to include I_t by multiplying the haploid transition probabilities as follows

$$P(q_{t+1} = (k_{t+1,1}, k_{t+1,2}), I_t | q_t = (k_{t,1}, k_{t,2}), \lambda) =$$

$$\begin{cases} e^{-2G\sigma_t} & \text{if } I_t = 0 \text{ and } k_{t+1,1} = k_{t,1} \text{ and } k_{t+1,2} = k_{t,2} \\ e^{-G\sigma_t}(1 - e^{-G\sigma_t})\alpha_{t,k_{t+1,1}} & \text{if } I_t = 1 \text{ and } k_{t+1,1} \neq k_{t,1} \text{ and } k_{t+1,2} = k_{t,2} \\ e^{-G\sigma_t}(1 - e^{-G\sigma_t})\alpha_{t,k_{t+1,2}} & \text{if } I_t = 1 \text{ and } k_{t+1,1} = k_{t,1} \text{ and } k_{t+1,2} \neq k_{t,2} \\ e^{-G\sigma_t}(1 - e^{-G\sigma_t})(\alpha_{t,k_{t+1,1}} + \alpha_{t,k_{t+1,2}}) & \text{if } I_t = 1 \text{ and } k_{t+1,1} = k_{t,1} \text{ and } k_{t+1,2} = k_{t,2} \\ (1 - e^{-G\sigma_t})^2 \alpha_{t,k_{t+1,1}} \alpha_{t,k_{t+1,2}} & \text{if } I_t = 2 \\ 0 & \text{otherwise} \end{cases} \tag{6.48}$$

Denote the probability under the diploid model that the sample is in states $(k_{t,1}, k_{t,2})$ at SNP t and states $(k_{t+1,1}, k_{t+1,2})$ at SNP $t+1$ and has indicator vari-

able I_t given observed data O and parameters λ by $\xi_t((k_{t,1}, k_{t,2}), (k_{t+1,1}, k_{t+1,2}), I_t)$. Then

$$\xi_t((k_{t,1}, k_{t,2}), (k_{t+1,1}, k_{t+1,2}), I_t) = \frac{1}{P(O|\lambda)} \alpha_t(k_{t,1}, k_{t,2}) \beta_{t+1}(k_{t+1,1}, k_{t+1,2}) P(O_{t+1}|q_t = (k_{t,1}, k_{t,2}), \lambda) \times$$
$$P(q_{t+1} = (k_{t+1,1}, k_{t+1,2}), I_t | q_t = (k_{t,1}, k_{t,2}), \lambda) \quad (6.49)$$

Let $m^t_{\text{switch},k}$ be the number of haplotypes of the sample that switch into state k between SNPs t and $t+1$. I can calculate $\mathbb{E}_\lambda[n^t_{\text{switch},k}]$, and from this $\mathbb{E}_\lambda[n^t_{\text{stay}}]$, by summing across $\mathbb{E}_\lambda[m^t_{\text{switch},k}]$ for all N samples, and so we can calculate the required expectations from Equation 6.12 and Equation 6.13 by performing the calculations below. Note that I simplify the summation to give a formulation that enables quadratic versus quartic computational complexity in K. A similar approach is done for the haploid model to achieve linear versus quadratic computational complexity

(not shown).

$$\mathbb{E}_\lambda[m^t_{\text{switch},k}] = \sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\sum_{k_3=1}^{K}\sum_{j=0}^{2} j \times \Big(\xi_t\Big((k_1,k_2),(k,k_3),I_t=j\Big) + \xi_t\Big((k_1,k_2),(k_3,k),I_t=j\Big)\Big)$$
$$= \sum_{k_1=1}^{K}\sum_{k_2=1}^{K} 1 \times \Big(\xi_t\Big((k_1,k_2),(k,k_2),I_t=1\Big) + \xi_{n,t}\Big((k_1,k_2),(k_1,k),I_t=1\Big)\Big)$$
$$+ \sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\sum_{k_3=1}^{K} 2 \times \Big(\frac{1}{2}\xi_t\Big((k_1,k_2),(k,k_3),I_t=2\Big) + \frac{1}{2}\xi_t\Big((k_1,k_2),(k_3,k),I_t=2\Big)\Big) \tag{6.50}$$

$$= \sum_{k_1=1}^{K}\sum_{k_3=1}^{K} 2 \times \xi_t\Big((k_1,k_3),(k,k_3),I_t=1\Big)$$
$$+ \sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\sum_{k_3=1}^{K} 2 \times \xi_t\Big((k_1,k_2),(k,k_3),I_t=2\Big) \tag{6.51}$$
$$= 2\sum_{k_1=1}^{K}\sum_{k_3=1}^{K} \frac{\alpha_t(k_1,k_3)\beta_{t+1}(k,k_3)P(O_{t+1}|q_t=(k,k_3),\lambda)\alpha_{t,k}(1-e^{-G\sigma_t})e^{-G\sigma_t}}{P(O|\lambda)}$$
$$+ 2\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\sum_{k_3=1}^{K} \frac{\alpha_t(k_1,k_2)\beta_{t+1}(k,k_3)P(O_{t+1}|q_t=(k,k_3),\lambda)\alpha_{t,k}\alpha_{t,k_3}(1-e^{-T\sigma_t})^2}{P(O|\lambda)}$$
$$= \frac{2\alpha_{t,k}}{P(O|\lambda)}\sum_{k_3=1}^{K}\Big((1-e^{-G\sigma_t})e^{-G\sigma_t}\Big[\sum_{k_1=1}^{K}\alpha_t(k_1,k_3)\Big]$$
$$+ \alpha^t_{k_3}(1-e^{-G\sigma_t})^2\Big[\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\alpha_t(k_1,k_2)\Big]\Big)\beta_{t+1}(k,k_3)P(O_{t+1}|q_t=(k,k_3),\lambda) \tag{6.52}$$

Emission probabilities

To update the emission parameters for the diploid model, I use an augmented state space as in for the pseudo-haploid model where I have knowledge of whether emissions were due to the alternate or reference base, and further have knowledge of whether the read came from the maternal or paternal haplotype. Recall that: $\phi^i_{r,j}$ is the probability that observed base j in read R_r came from a read with underlying genotype i; H^j_r is an variable which takes value 1 if the underlying base is the alternate base and 0 if it is the reference base; and H_r is a variable that takes value 1 if the read came from the maternal haplotype and 2 if from the paternal haplotype. I will use these to calculate the expectation of $n^t_{k,s}$, the number of reads that emit the alternate base

$(s = 1)$ or reference base $(s = 0)$ given they are in state k at their central SNP.

Recall that the probability of an observation (set of reads) at SNP t in the diploid model is

$$P(O_t|q_t = (k_{t,1}, k_{t,2}), \lambda) = \frac{1}{2}P(R_r|q_t = k_{t,1}, \lambda) + \frac{1}{2}P(R_r|q_t = k_{t,2}, \lambda) \tag{6.53}$$

For notational convenience set

$$F_{r,j,H_r} = \frac{\frac{1}{2}P(R_r|q_t = k_{t,H_r}, \lambda)}{\frac{1}{2}P(R_r|q_t = k_{t,1}, \lambda) + \frac{1}{2}P(R_r|q_t = k_{t,2}, \lambda)} \left(\frac{1}{\theta_{t,k_{t,H_r}}\phi^1_{r,j} + (1 - \theta_{t,k_{t,H_r}})\phi^0_{r,j}}\right) \tag{6.54}$$

I can therefore calculate the probability that SNP j in read R_r with central SNP c_r has indicator variable H^j_r and H_r and observation for SNP $t = c_r$ of O_t given the pair of hidden states $(k_{t,1}, k_{t,2})$ and parameters λ as

$$P(O_t, H^j_r, H_r|q_t = (k_{t,1}, k_{t,2}), \lambda) = \begin{cases} P(O_t, |q_t = (k_{t,1}, k_{t,2}), \lambda)F_{r,j,H_r}\theta_{t,k_{t,H_r}}\phi^1_{r,j} & \text{if } H^j_r = 1 \\ P(O_t, |q_t = (k_{t,1}, k_{t,2}), \lambda)F_{r,j,H_r}(1 - \theta_{t,k_{t,H_r}})\phi^0_{r,j} & \text{if } H^j_r = 0 \end{cases}$$

Denote the probability for SNP j in read R_r that at the central SNP of the read $t = c_r$ is in the pair of states $(k_{t,1}, k_{t,2})$ given the observed data O and parameters λ by $\gamma_t(k_{t,1}, k_{t,2}, H^j_r, H_r)$. Then

$$\gamma_t(k_{t,1}, k_{t,2}, H^j_r, H_r) = \begin{cases} \gamma_t(k_{t,1}, k_{t,2})F_{r,j,H_r}\theta_{t,k_{H_r}}\phi^1_{r,j} & \text{if } H^j_r = 1 \\ \gamma_t(k_{t,1}, k_{t,2})F_{r,j,H_r}(1 - \theta_{t,k_{H_r}})\phi^1_{r,j} & \text{if } H^j_r = 0 \end{cases}$$

Let $\gamma_{n,t}(k_{t,1}, k_{t,2}, H^j_r, H_r)$ be $\gamma_t(k_{t,1}, k_{t,2}, H^j_r, H_r)$ for sample n, and let A_n be the complete set of SNPs j and reads R_r for sample n such that $u_{r,j} = t$. I can calculate the required expectations of Equation 6.11 in the main text as

$$\mathbb{E}_\lambda[n^t_{k,s}] = \sum_{n=1}^{N} \sum_{(r,j)\in A_n} \sum_{i=1}^{K} \Big(\gamma_{n,c_r}(k, i, H^j_r = s, H_r = 1) + \gamma_{n,c_r}(i, k, H^j_r = s, H_r = 2)\Big) \tag{6.55}$$

6.2.12 Efficient calculation of forward backward variables

I take the time here to write out the forward backwards calculations that are used for the diploid case, as symmetries in the transition matrix allow for the calculation to

be made in in quadratic, rather than quartic time with respect to K. Similar calculations (not shown) are used for the haploid model to ensure linear versus quadratic computational complexity in K. Note that these calculations are not original and are given in very similar form in the original fastPHASE paper, but are reproduced here as they represent important simplifications for computational reasons

$$\begin{aligned}
\alpha_{t+1}(k_3,k_4) &= \left[\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\alpha_t(k_1,k_2)P(q_{t+1}=(k_3,k_4)|q_t=(k_1,k_2),\lambda)\right]P(O_{t+1}|q_{t+1}=(k_3,k_4),\lambda) \\
&= \Bigg[\alpha_t(k_3,k_4)(e^{-G\sigma_t})^2 + \sum_{k=1}^{K} e^{-G\sigma_t}(1-e^{-G\sigma_t})\alpha_{t,k_3}\alpha_t(k,k_4) + \\
&\quad \sum_{k=1}^{K} e^{-G\sigma_t}(1-e^{-G\sigma_t})\alpha_{t,k_4}\alpha_t(k_3,k) + \\
&\quad \sum_{k_1=1}^{K}\sum_{k_2=1}^{K}(1-e^{-G\sigma_t})^2\alpha_{t,k_3}\alpha_{t,k_4}\alpha_t(k_1,k_2)\Bigg]P(O_{t+1}|q_{t+1}=(k_3,k_4)) \\
&= \Big[\alpha_t(k_3,k_4)(e^{-G\sigma_t})^2 + \alpha_{t,k_3}A_{t,1}(k_4) + \alpha_{t,k_4}A_{t,2}(k_3) + \alpha_{t,k_3}\alpha_{t,k_4}B_t\Big]\times \\
&\quad P(O_{t+1}|q_{t+1}=(k_3,k_4),\lambda)
\end{aligned}$$

where

$$A_{t,1}(k_4) = e^{-G\sigma_t}(1-e^{-G\sigma_t})\sum_{k=1}^{K}\alpha_t(k,k_4) \tag{6.56}$$

$$A_{t,2}(k_3) = e^{-G\sigma_t}(1-e^{-G\sigma_t})\sum_{k=1}^{K}\alpha_t(k_3,k) \tag{6.57}$$

$$B_t = (1-e^{-G\sigma_t})^2\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\alpha_t(k_1,k_2) \tag{6.58}$$

As such, the forward calcution can be done in quadratic time with respect to the number of ancestral haplotypes K.

Similarly, for the backward calculation

$$\begin{aligned}\beta_t(k_1,k_2) &= \sum_{k_3=1}^{K}\sum_{k_4=1}^{K} P(q_{t+1}=(k_3,k_4)|q_t=(k_1,k_2),\lambda)P(O_{t+1}|q_{t+1}=(k_3,k_4),\lambda)\beta_{t+1}(k_3,k_4)\\ &=(e^{-G\sigma_t})^2 P(O_{t+1}|q_{t+1}=(k_1,k_2),\lambda)\beta_{t+1}(k_1,k_2)+\\ &\quad (e^{-G\sigma_t})(1-e^{-G\sigma_t})\left(\sum_{k=1}^{K}\alpha_{t,k}P(O_{t+1}|q_{t+1}=(k,k_2),\lambda)\beta_{t+1}(k,k_2)+\right.\\ &\quad \left.\sum_{k=1}^{K}\alpha_{t,k}P(O_{t+1}|q_{t+1}=(k_1,k),\lambda)\beta_{t+1}(k_1,k)\right)+\\ &\quad (1-e^{-G\sigma_t})^2\sum_{k_3=1}^{K}\sum_{k_4=1}^{K}\alpha_{t,k_3}\alpha_{t,k_4}P(O_{t+1}|q_{t+1}=(k_3,k_4),\lambda)\beta_{t+1}(k_3,k_4)\\ &=(e^{-G\sigma_t})^2 P(O_{t+1}|q_{t+1}=(k_1,k_2),\lambda)\beta_{t+1}(k_1,k_2)+E_{t,1}(k_2)+E_{t,2}(k_1)+F_t\end{aligned}$$

where

$$E_{t,1}(k_2) = (e^{-G\sigma_t})(1-e^{-G\sigma_t})\sum_{k=1}^{K}\alpha_{t,k}P(O_{t+1}|q_{t+1}=(k,k_2),\lambda)\beta_{t+1}(k,k_2) \tag{6.59}$$

$$E_{t,2}(k_1) = (e^{-G\sigma_t})(1-e^{-G\sigma_t})\sum_{k=1}^{K}\alpha_{t,k}P(O_{t+1}|q_{t+1}=(k_1,k),\lambda)\beta_{t+1}(k_1,k) \tag{6.60}$$

$$F_t = (1-e^{-G\sigma_t})^2\sum_{k_3=1}^{K}\sum_{k_4=1}^{K}\alpha_{t,k_3}\alpha_{t,k_4}P(O_{t+1}|q_{t+1}=(k_3,k_4),\lambda)\beta_{t+1}(k_3,k_4) \tag{6.61}$$

6.2.13 CFW sequencing

CFW mice are a commercial outbred colony of mice, frequently used for testing. They are thought to have descended from two unknown founder animals about 100 generations ago. Sample pre-processing was done in accordance with best practice recommendations [48]. Sequencing reads were mapped to mm10 using bwa [50], remapped using Stampy [51], PCR duplicates were marked using Picard (`http://broadinstitute.github.io/picard/`), files were merged using Picard, indel realignment was performed using the GATK [53], and base quality score recalibration was performed using the GATK [53]. Variant calling was done using the GATK UnifiedGenotyper [53] and filtered by the GATK VQSR [53], using as training data a set of variants from the Mouse Genomes Project (17 Strains) [13] and a sensitivity threshold of 90%. Sites that failed VQSR but were in the training set and were polymorphic

were included. In total, 7.07 M SNPs were called on the autosomes and chromosome X. 4 mice were additionally sequenced at 10X. Genotypes for these mice were generated using the GATK UnifiedGenotyper using the GENOTYPE_GIVEN_ALLELES option at the 7.07 M SNPs. For comparisons with low coverage imputation, individual genotypes from the high coverage samples were set to missing if the read depth was less than 5 or more than 25, or if the genotype quality was less than 10.

CFW MegaMUGA Array Genotyping

DNA from 96 distinct CFW mice, of which 48 were included in the imputed dataset of 2,073 mice, were sent to Neogen on a single plate and genotyped using the Mega Mouse Universal Genotyping Array (MegaMUGA), an array built upon the Illumina Infinium platform with 77,808 SNPs (Neogen, Lincoln, Nebraska, USA). Genotype calling was performed by Neogen using GenCall and their proprietary workflows.

After genotyping, recorded genders were compared to X and Y chromosome marker information, revealing no gender mismatches on the arrays. Samples were further compared to imputation and array QC metrics (call rate and 10% GC score); this revealed 4 of 48 samples had poorly performing array metrics. These 4 samples were subsequently removed from further analysis.

For the 77,808 SNPs for which I had genotypes, 144 were not carried over from mm9 to mm10 using liftOver, and out of the remaining sites, 29,694 intersected between imputation and MegaMUGA. Out of those, I removed sequentially for the following reasons: 17 were removed for allele disagreements between sequencing and the array; 3,819 monomorphic array sites were removed; 3,160 SNPs with an imputed SNP within 25 bp of the array target SNP (as off-target variation can affect microarray genotyping [54]); 56 sites with an array Hardy-Weinberg Equilibrium p-value of less than 1×10^{-10}. Subsequent comparisons between CFW imputed dosages and array genotypes were made for the remaining 21,576 sites.

6.2.14 CONVERGE study

Full details for the processing of the CONVERGE data, including low coverage, high coverage, and Illumina array data, have been published elsewhere [139]. In brief, 11,670 low coverage (1.7X) Han Chinese samples were called using the GATK to yield a filtered set of 20.5M variants. 9 samples were separerately sequenced to 10X and variants were genotyped and filtered (5.9M). For the analysis, high coverage sample genotypes with a read depth of lower than 5, read depth greater than 25, or genotype quality of less than 10 were masked out. For the Illumina HumanOmniZhongHua-8 (v1.0B) BeadChip, of the 21057 sites present on chromosome 20 on the array and used for imputation, I removed 292 sites with $> 5\%$ missingness, 7642 sites with probes within 25bp of another site in the imputed dataset (as these are sites where off-target polymorphic variation in the sample may impair genotype microarray accuracy), and 0 sites with an array Hardy-Weinberg Equilibrium p-value of less than 1×10^{-10}.

6.2.15 Beagle

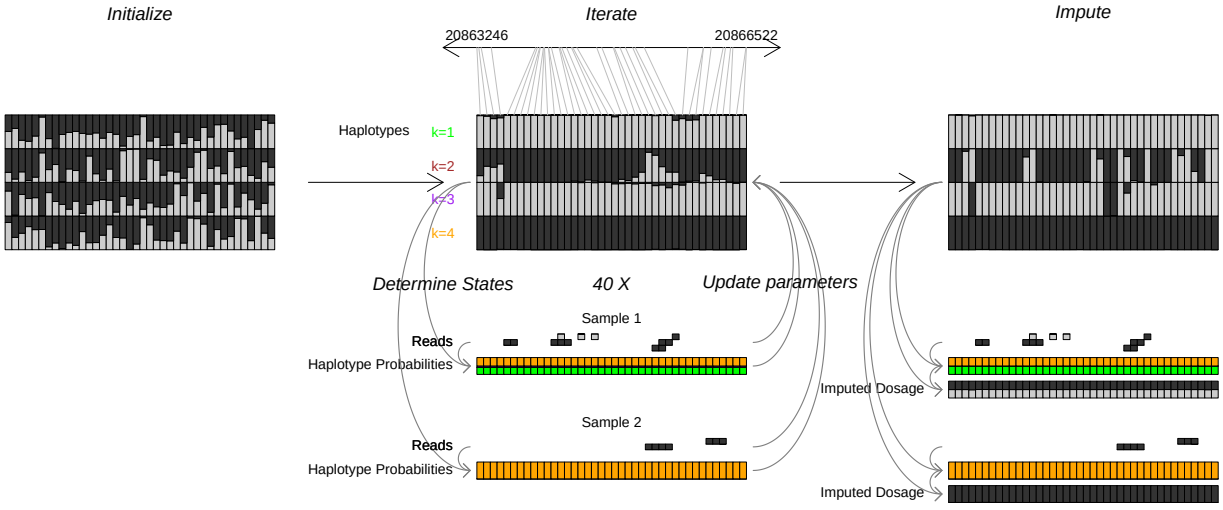
For both the CFW and CONVERGE samples, Beagle was run using default parameters [136]. Genotype likelihoods were used as inputs.

6.3 Results

6.3.1 Model description

An overview of the model used for imputation is presented in Figure 6.1. In this section, I start by providing results for the CFW mice in Section 6.3.2. In Section 6.3.3, I provide results for the CONVERGE study. In Section 6.3.4, I provide results under different downsampling scenarios, while in Section 6.3.5, I provide results under different variant filtration strategies.

Figure 6.1. Overview of STITCH STITCH begins by initializing various parameters (left), represented here by the ancestral haplotypes. The method then uses EM iterations (default 40) (middle) each of which involves i) determining states (going down, left side) using current parameters and sample reads, and ii) parameter updates (going up, right side) using sample reads and haplotype probabilities (latent states). Once the EM iterations are done, imputed genotypes are generated using the haplotype probabilities and ancestral haplotypes from the final iteration (right). This example uses real data from the CFW mice with $K = 4$ founder haplotypes for approximately 3000 base pairs on chromosome 19 containing 20 imputed SNPs, after removal of SNPs which imputed poorly. Each of the SNPs in the 4 reconstructed haplotypes are shown as a vertical bar split proportionally to the probability of emitting the reference (black) or alternate (grey). Sample reads are similarly coloured.



6.3.2 CFW outbred mice

I imputed a sample of 2,073 outbred CFW mice [140] that are thought to have descended from two outbred founders (i.e. $K = 4$) about $G \sim 100$ generations ago. Imputation results generated here were used to perform a GWAS in this population; further details, including characterization of the genetic architecture of this population, will be published elsewhere. The mice were sequenced to an average depth of 0.15X using multiplexed paired end 100bp Illumina reads.

Accuracy of imputation was assessed using 2 different validation datasets, 44 mice genotyped on the Illumina MegaMUGA array and 4 mice sequenced to 10X using an Illumina HiSeq. Correlations (r^2) between genotypes and imputed dosages were done either per-site for the array or aggregated across all SNPs in that frequency for the high coverage results, where aggregation was done to ensure major allele was consistent across all sites.

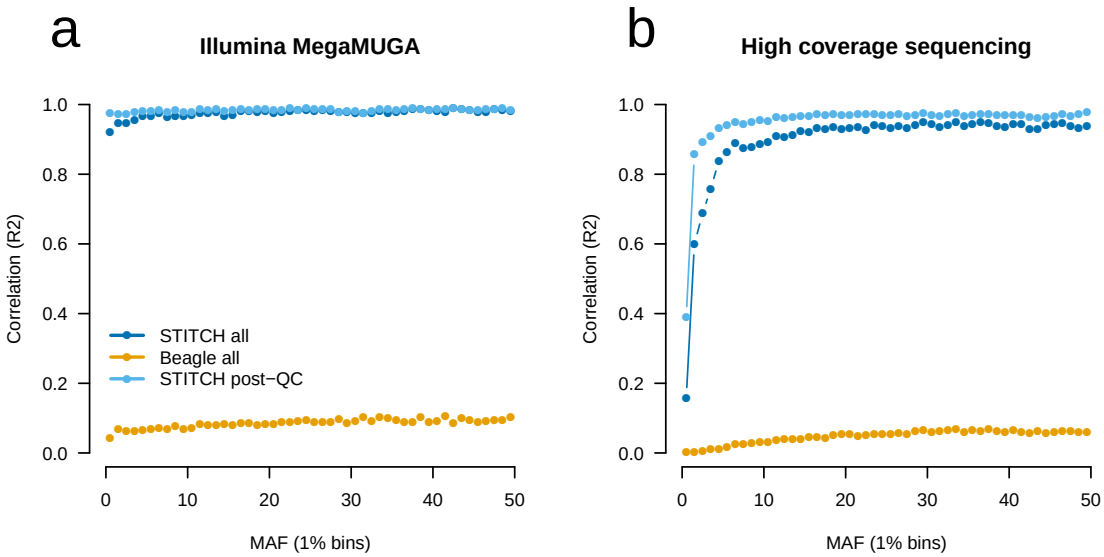
Table 6.2. Performance of CFW study under different programs and options Results are given for chromosomes 18 and 19. All STITCH results are for the diploid model. Results are on the same set of SNPs for each entry, without removal of any site for QC reasons. r^2 is against the Illumina array. RU refers to read unaware

Program	K	Av R2
Beagle		0.074
STITCH	2	0.622
STITCH	3	0.957
STITCH	4	0.972
STITCH	5	0.97
STITCH	6	0.966
STITCH	7	0.964
STITCH	8	0.967
STITCH RU	4	0.873

I first sought to characterize performance of STITCH and Beagle on a subset of the genome, chromosomes 18 and 19, with results shown in Table 6.2. In this setting, results from Beagle were poor - an average array r^2 of 0.074. For STITCH, I was able

to achieve good results for relatively low values of K, and the program could be run quickly in diploid mode. I found that $K = 4$ achieved optimal results at a minimal K among the range of K chosen. I also found that running the program in a read unaware mode led to a significant drop in imputation quality for $K = 4$, from and r^2 of 0.972 to 0.873.

Figure 6.2. Performance of method on CFW mice as assessed through external validation Validation dataset is the Illumina MegaMUGA array (a) and 10X Illumina sequencing (b). Results are shown for STITCH (K=4, diploid mode) and Beagle (default) genome-wide for n=2,073 mice featuring 7.07M SNPs before QC and 5.72M after QC. STITCH is run using K=4, diploid method, 40 iterations. Post-QC is SNPs with info> 0.4 and HWE p-value $> 1 \times 10^{-6}$.



I then ran STITCH genome-wide for $K = 4$ and Beagle. Note that a genome-wide imputation results for $K = 4$ took about 3 days when run per-chromosome on a server with 4 AMD Opteron(tm) Processor 6344 CPUs (48 cores total). Results as shown in Figure 6.2 when comparing against the arrays are generally consistent, with a mean r^2 of 0.972 and 0.948 for the Illumina MegaMUGA array and 10X sequencing at the 7.07M sites before QC, respectively. Results for Beagle on the same SNPs were 0.074 and 0.254. For STITCH, after applying post-QC imputation filters of an info score > 0.4 and Hardy-Weinberg p-value of $> 1 \times 10^{-6}$, 5.72M SNPs remained, and

accuracy on those variants improved to 0.981 and 0.968 against the array and 10X sequencing, respectively. Results are consistently good across all frequencies when compared against the MegaMUGA, except for variants between 0 and 1%, where there is a slight drop, and against sequencing, whrere there is a drop from about 0 to 5% frequency. Of note, for the former, these are situations where only a handful of sequencing reads are observed, while for the latter, the Ti/Tv of variants between 0 and 5% MAF in this dataset was 1.82, as compared to 1.99 for those greater than 5%, implicating false positive sequencing variants represent a component of the reduction in accuracy.

6.3.3 CONVERGE study

I performed imputation on a population of 11,670 Han Chinese women sequenced to 1.7X as a part of a genome wide association study [139]. Details of the read mapping, pre-processing and variant filtration are available elsewhere [139].

I first sought to characterize the performance of Beagle and STITCH on this dataset, using 10 Mbp of chromosome 20 broken into 500 Kbp chunks with 100 Kbp buffers on either side. Results are shown in Table 6.3. Beagle achieved an r^2 of 0.874. For STITCH, for $K = 20$, the diploid method outperformed the pseudo-haploid method. I then tried switching from pseudo-haploid to diploid towards the end of the EM procedure - this rescued performance with about 3 or 4 full diploid iterations, but with an added time cost. Given most of the rescue occured in the first few additional diploid iterations, and that I were particularly interested in exploring large K in this analysis and diploid iterations become prohibitively computationally expensive for large K, I pursued further analysis with 38 pseudo-haploid and 2 diploid iterations. Note that this means there are 38 pseudo-haploid iterations of first calculating hidden state probabilities then calculating updates, a 39th iteration of calculating state probabilities and updates in a diploid manner, and a final 40th iteration of calculating

Table 6.3. Performance of CONVERGE study under different programs and options with no reference panel Results are given for the first 10 Mbp region of chromosome 20, run in 20 0.5 Mbp regions with 0.1 Mbp buffers. Accuracy is average for Illumina array for high frequency (MAF 5% to 95%) variants. Time is average for each 0.5Mbp region, where all programs were run on 4 cores on 2.60 GHz Intel E5-2650 chips. Method is D for diploid and PH for pseudo-haploid. RU refers to read unaware. Note that times for STITCH do not include the generation of input data from BAMs, which was done at another site and took about 30 minutes per region, irrespective of other program options.

Program	Beagle Its	K	Iterations	Time	r^2
Beagle	5			12.5	0.874
STITCH		20	40D	24.5	0.922
STITCH		20	40PH	8.0	0.875
STITCH		20	34PH, 6D	10.6	0.920
STITCH		20	35PH, 5D	9.9	0.919
STITCH		20	36PH, 4D	9.6	0.918
STITCH		20	37PH, 3D	9.3	0.917
STITCH		20	38PH, 2D	8.8	0.911
STITCH		20	39PH, 1D	8.4	0.898
STITCH RU		20	38PH, 2D	9.4	0.910
STITCH		30	40D	52.2	0.927
STITCH		30	38PH, 2D	12.4	0.917
STITCH		40	38PH, 2D	16.5	0.920
STITCH		60	38PH, 2D	27.7	0.923
STITCH		80	38PH, 2D	42.2	0.925
STITCH		100	38PH, 2D	61.1	0.927

state probabilities and from this imputed dosages. To look at the effect of the read aware component of imputation, I performed the same imputation for $K = 20$ in a read unaware fashion, which led to a drop in r^2 from 0.911 to 0.910. I also looked at further values of K by analyzing $K = 30$ for both diploid and pseudo-haploid, as well as $K = 40$, 60, 80 and 100. As expected, large K in diploid mode became rapidly prohibitively computationally expensive, while in pseudo-haploid mode, much larger values of K could be run in the same amount of time.

I also tested the effect of running Beagle with a reference panel on a reduced set of variants which were also polymorphic in the 1000G Phase 1 ASN population. Results are shown in Table 6.4. When run with a reference panel, the number of iterations

Table 6.4. Performance of CONVERGE study under different programs and options with a reference panel Results are given for the first 10 Mbp of chromosome 20, run in 0.5 Mbp regions with 0.1 Mbp buffers. Accuracy is average for Illumina array for common (MAF 5% to 95%) variants. "Time" shows average in hours for each 0.5Mbp region, where all programs were run on 4 cores on 2.60 GHz Intel E5-2650 chips. Iterations reference how many pseudo-haploid (PH) and diploid (D) EM iterations were run. RU refers to read unaware. Note that times for STITCH do not include the generation of input data from BAMs, which took about 30 minutes per region, irrespective of other program options.

Program	Beagle Its	K	Iterations	Time	r^2
Beagle	4			114.4	0.946
Beagle	3			74.5	0.943
Beagle	2			39.7	0.939
Beagle	1			12.0	0.930
Beagle (no ref)	5			7.8	0.886
STITCH		20	38PH, 2D	5.4	0.911
STITCH		40	38PH, 2D	10.2	0.922
STITCH		60	38PH, 2D	16.6	0.925

used by Beagle was reduced due to computational burden, with a reduced number of burnin-its and phase-its due to their time costs (but with impute-its at the default of 5). With a reference panel and 4 burn-in and 4 phasing iterations, Beagle achieved an r^2 of 0.946, versus 0.922 for STITCH with $K = 40$ and 0.886 for Beagle without a reference panel. After removal of sites failing QC (info > 0.4, HWE p $> 10^{-6}$), these improved to 0.949, 0.939, and 0.911, respectively.

I also looked at the performance across the whole frequency spectra. I compared Beagle to STITCH without a reference panel, as well as with a reference panel on a subset of SNPs, as shown in Figure 6.3. Results were generally similar throughout between the different options.

In the two above examples, two sets of SNPs were examined - a larger set of SNPs (all SNPs polymorphic in CONVERGE), and a smaller set of SNPs both polymorphic in CONVERGE and in the 1000 Genomes ASN panel. In their original publication, the CONVERGE authors ran two different imputations - impute using Beagle at all SNPs without a reference panel, then impute at the subset of sites with a reference

Figure 6.3. Performance of STITCH on CONVERGE humans as assessed through external validation Validation dataset is the Illumina HumanOmniZhongHua-8 array (a, c) and 10X sequencing (b, d). Results are shown for STITCH (K=40, 38 pseudo-haploid iterations, 2 diploid iterations) and Beagle (default (a,b), 3 iterations with reference panel (c,d)) for the first 10 Mbp of chromosome 20 for N=11,670 Han Chinese samples, either for all SNPs (N=92,524 pre-QC, N=31,958 post-QC) (a,b), or for SNPs also present in the 1000 Genomes ASN reference panel (N = 52,875) (c,d). Post-QC is SNPs with info> 0.4 and HWE p-value $> 1 \times 10^{-6}$.

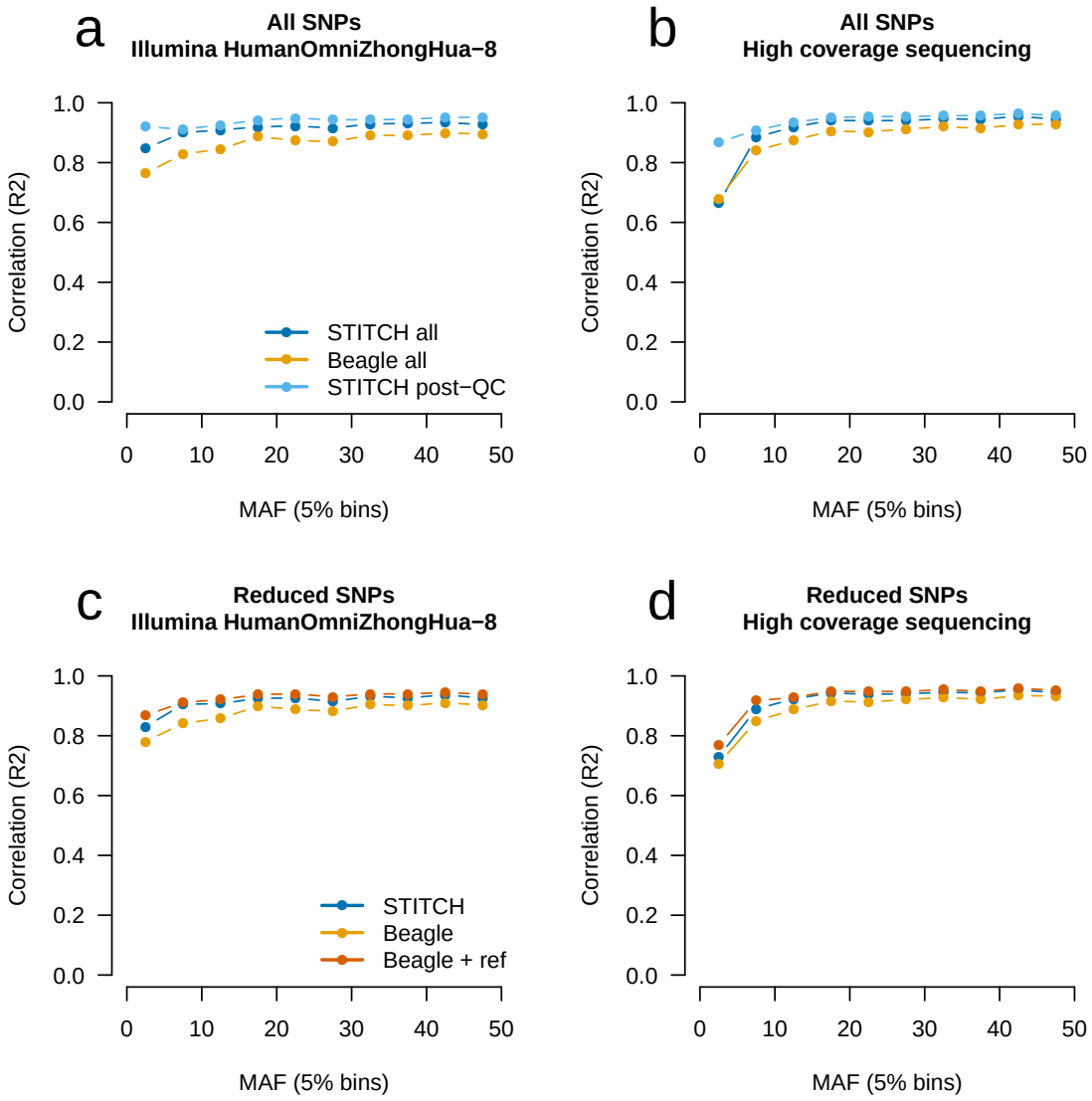


Table 6.5. Performance of STITCH on CONVERGE study original imputation Results are over the first 10 Mbp of chromosome 20. Beagle methodology was the same as done in the original CONVERGE paper and as explained in the text. STITCH results are for $K = 40$, 38 pseudo-haploid iterations, 2 diploid iterations. All sites with removal of SNPs failing QC also removed SNPs with Hardy-Weinberg p-value less than 10^{-6}. r^2 is the average r^2 for SNPs on the Illumina HumanOmniZhongHua-8 array for high frequency (MAF 5% to 95%) variants.

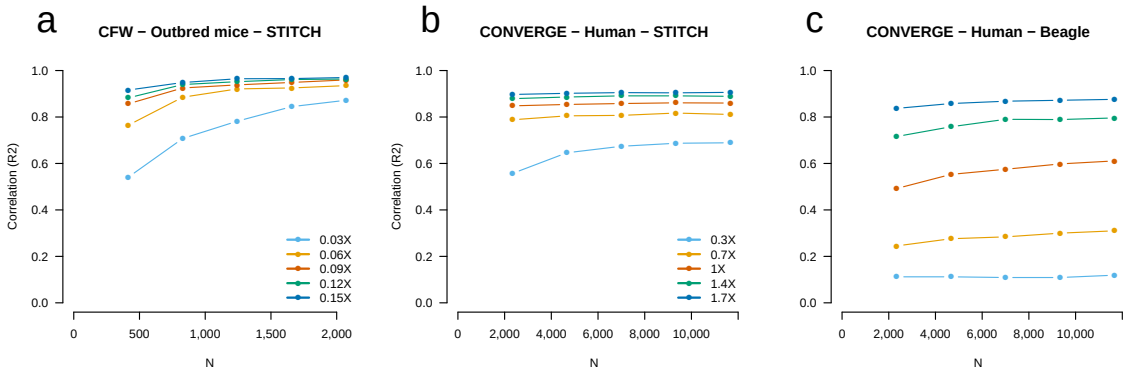
Method	SNP set	% SNPs	r^2
Beagle	All	100	0.933
STITCH	All	100	0.92
Beagle	info>0.4	90	0.939
STITCH	info>0.4	90	0.939
Beagle	info>0.9	78	0.968
STITCH	info>0.9	75	0.972

panel using the reference panel, finally making a combined dataset where sites in the former are replaced by sites in the later. I therefore also compared this dataset to what I get for the full dataset for STITCH $K = 40$ with 38 pseudo-haploid and 2 diploid iterations. Results are given for high frequency variants between 5% and 95% frequency (but imputing all variants) in Table 6.5. For Beagle, this led to r^2's of 0.933, 0.939 and 0.968, for all SNPs, with info> 0.4 and info> 0.9, respectively, while for STITCH the same values were 0.922, 0.939 and 0.972. The number of SNPs retained in each of the different QC settings was similar. Therefore, on the full dataset, results from STITCH were comparable to those from Beagle using the available reference panel. It is worth noting that the two Beagle runs would take approximately (12.4+86.4)/16.5 = 5.3 times longer than STITCH under this scenario, and require the existence of a reference panel.

6.3.4 Effect of subsampling reads and samples

To investigate the effect of sample size and sample coverage on imputation performance, I downsampled both the CFW mice and CONVERGE human results for both sample size and sequencing coverage. Results are shown in Figure 6.4. For the CFW

Figure 6.4. Effects of reduced sequence coverage Results are shown for CFW mice (a) and CONVERGE humans using STITCH (b) and Beagle run without a reference panel (c). Validation is using array data, with each value representing the average for common SNPs (allele frequency 5˜95%), without correction for post-imputation QC. Downsampling of samples and reads, as shown in the legends, was performed at random, except that samples necessary for accuracy assessment were always retained. STITCH settings are the same as for the full CFW, CONVERGE datasets. Colours representing downsampling sequence depth are the same for STITCH and Beagle.



mice, for the higher 0.15X, sample size has little relation to performance, while at lower X, increasing sample size leads to increased performance. Rather amazingly, at 0.06X for the full sample set, results are only marginally poorer than 0.15X. For the CONVERGE samples, sample size has less of an influence across the range of sequencing coverage considered here, although results did consistently get better with increasing sequencing depth.

6.3.5 Effect of variant filtration on imputation performance

Standard methods of genotyping from next generation sequencing employ a separate step of variant filtration so as to reject newly genotyped sites whose quality control metrics fall outside a distribution defined by known variant sites. One such method is the GATK Variant Quality Score Recalibrator. We examined whether we needed to perform variant filtration when using STITCH. We compared the original CFW mice results to a two-step strategy. In the first step, all discovered variants are imputed. In the second step, only those variants that pass quality control (QC) filters are re-

Table 6.6. Effect of filtering on imputation performance Results are given for chromosome 19. QC is defined per-run and reflects info> 0.4 and HWE p-value $> 1 \times 10^{-6}$. r^2 values are against the 4 10X mice.

Set	Description	SNPs	Number of SNPs	Ti/Tv	VQSR r2	No VQSR r2
1	VQSR	All	152,486	2.07	0.937	
2	VQSR	Post-QC	122,878	2.21	0.968	
3	No VQSR, Round 1	All	355,123	1.48		0.745
4	No VQSR, Round 1	Post-QC	136,164	2.08		0.945
5	No VQSR, Round 2	All	136,164	2.08		0.938
6	No VQSR, Round 2	Post-QC	128,054	2.14		0.952
7	Intersect Set 2 and Set 6		115,567	2.22	0.967	0.969
8	Present Set 2, absent Set 6		7,311	2.13	0.915	
9	Present Set 6, absent Set 2		12,487	1.55		0.930

imputed. Results indicate this to be a viable strategy (Results are shown in Table 6.6). For the one step strategy with variant filtration, 152K SNPs on chromosome 19 were imputed, with 122K SNPs passing QC at an r^2 of 0.968. For the two-stage approach, 355K variants were imputed, 136K passed QC and re-imputed, with 128K of those passing QC at an r^2 of 0.952. Overlap between the two approaches was 116K, with marginally better r^2 in the overlap from the two-step approach, with results specific to either set having lower r^2. These results indicate that a two-stage imputation strategy without variant filtration is at least as effective as a one-stage strategy with variant filtering.

6.4 Discussion

6.4.1 Algorithm performance

In this work, I described a method to accurately impute genotypes using low coverage sequencing without reference haplotype panels in a read aware fashion. I showed that the method is accurate on a real dataset of 2,073 CFW mice, more so than Beagle, and is sufficient for a GWAS or other downstream applications to be performed on the resulting variants. I also show that the method works using 11,670 Han Chinese

samples sequenced to 1.7X, showing it can work in humans and is likely to work in non-human outbred wild populations. In this context, performance is marginally better than Beagle without a reference panel, and comparable to Beagle with a reference panel on all available variants. A full discussion of why the CFW mouse population imputes so much better than the CONVERGE Han Chinese population, and how this relates to expected performance in future settings, is discussed in the next subsection. Before that, I briefly discuss issues relating to the imputation of low frequency variants, the effect of various program options like K and diploid vs pseudo-haploid on performance, the effect of read aware imputation, and the effect of the assumption that reads are emitted from point locations.

In terms of imputation performance, for both the mice and the humans, there is a slight reduction in imputation performance for variants of less than 5% MAF. For the CONVERGE wild human results, the reduction of accuracy at lower frequency variants likely stems more from an inability of the method to properly model a sufficiently large number of reference haplotypes such that low frequency variants correspond in a one to one fashion with a particular reference haplotype. Indeed, tests with larger values of K did help specifically to improve imputation results at low frequency variants, although at the expense of run time. For the CFW mice, I note that there are 2,073 of them sequenced to 0.18X, and as such, a 1% MAF SNP represents about 4 sequencing reads. Therefore, variant discovery and filtration becomes more susceptible to the effects of noise. Indeed, there is a notable reduction in the transition-transversion ratio at the lower frequency variants in the CFW population. As well, there is a notably smaller reduction in accuracy when comparing against the arrays, which have a more thorough variant discovery process, as compared to high coverage sequencing, indicating that part of the reduction in accuracy is due to false positive variants. Therefore, reduction at low frequency variants in the mice is more likely to be driven by a larger number of false positives, while the human results are

likely driven by modelling deficiencies.

In terms of computational issues, I also presented here a novel approximation of a diploid model, which I call a pseudo-haploid model. In a more traditional diploid mode, the method scales computationally linearly with the number of samples, linearly with the number of sites, and quadratically with the number of reference haplotypes, while in the pseudo-haploid mode, it is linear in the number of reference haplotypes. I found that I achieved results equivalent to the full diploid mode by running the algorithm in pseudo-haploid mode and switching to diploid on the last few iterations, which makes the algorithm nearly linear in the number of reference haplotypes. This should allow for very large values of K to be chosen and should broaden the use case of the algorithm considerably, for instance domesticated animals such as cows which might have gone through less modest bottlenecks. Furthermore, the use of a temporary directory, either on fast disks or a ram disk for storage of sample reads enables efficient multi-core computation with a low RAM overhead, so that imputation with hundreds of thousands of samples should be possible for larger K on normal computational clusters.

In terms of choice of K, for the CFW mice, this had relatively little effect beyond 4, while for the humans, it had a marginal effect, most notably at low frequency variants. For the human results, this is similar to previous work on human data [134], where K does have an influence on imputation accuracy. For future studies, I recommend that validation is obtained, either using high coverage sequencing on a few samples or array data, and that testing is done on a limited region of the genome to determine the minimal K at which imputation performance is maximized, subject to computational constraints. In the absence of validation data, a surrogate, such as the distribution of QC metrics such as the info score, should serve as a reasonable alternative.

While this has been used before by others, ours uses a fully read aware model. In real data, this had quite a large effect on the CFW mice, while its effect on the

CONVERGE study was more muted. I think there are several factors contributing to this. First, while there are fewer variants segregating in the mice, more of them are at high frequency, and as such, the number of reads spanning multiple high frequency variants is higher in the mice than the humans. Second, in a non-read aware fashion, multiple consecutive multi-SNP spanning reads drawn from the same haplotype can erreneously cause homozygous state inference when a sample is heterozygous. In the mice, due to the density of reads being 10 times smaller than in humans, such an abberation would occur over a distance 10 times larger. Finally, while there were more SNPs in the human data, marker density is more variable and concentrated when it occurs in the mice, meaning SNP containing reads will more often span multiple variants. Nevertheless, although this was less of an issue in the human data, I think it is an important consideration for future methods, especially as read length from sequencing increases.

One assumption I make in this method is that reads spanning multiple SNPs are emitted from a single SNP for the purpose of the HMM. This was done to preserve the HMM structure of the problem, namely that emissions at a particular SNP are dependent only on the current state and not on previous or future ones. In principle, I think this is not an overly restrictive assumption. For the CFW mice, at 0.18X in a 1 cM/Mb scenario for a 100 Mb chromosome, if I sequence a population 50 generations after founding, I would expect about 50 recombination breakpoints along the chromosome, of which only 10 might be covered by reads, and a smaller number still would be interesected by reads spanning multiple SNPs across the recombination breakpoint, depending on population heterogeneity. That being said, to help accomodate this problem, in the method, I do employ a step towards the end of the EM updating (iteration 25 out of 40), where for reads with at least 3 SNPs which are more likely to have been emitted by random haplotypes than any of the available founder haplotypes ($\theta = 0.5$ for all SNPs in the read), if a position exists along the

read that breaking it produces two new reads which are more likely to be derived from the available haplotypes than from random reads, I replace the original read with the two new reads split at the calculated breakpoint. In simulations with longer reads, I found this approach quite nicely split affected reads (not shown), although it is worth noting that imputation performance was negligibly improved by the application of this procedure on real data (not shown).

6.4.2 The effect of genetic history, sample size and sample coverage on performance

Although both the mice and human imputation results are good, and certainly good enough for downstream applications such as GWAS, to understand why they are different it is important to understand how genetic history, sample size and sample coverage relate, in terms of their effects on imputation performance. Here, I first discuss how sample coverage and sample size relate to algorithm performance for a given genetic history. Next, I discuss how these relate to genetic history, and why the CFW mouse population imputes far better than the Han Chinese human population for a given sample size and coverage.

As shown in the results, both sample size and sample coverage are important factors which may influence marginal imputation results. Imputation using this model can be thought of as having two components - the first is to build the founder haplotypes, and the second is to determine which of those haplotypes an individual carries at any genomic locus. Both sample coverage and sample size are necessary to build founder haplotypes. However, once the founder haplotypes are sufficiently well built, additional samples do not improve imputation performance, as they no longer contribute to making better reference haplotypes - this is why in the mouse and humans, beyond a certain sample size for a given level of coverage, imputation performance levels off. Similarly, sample coverage also helps improve the reconstruction of the founder haplotypes, which is why, particularly at low sample sizes, additional se-

quencing depth yields an improvement in imputation performance. However, once the founder haplotypes are effectively reconstructed, additional sample coverage still offers the ability to help resolve haplotype switches more readily, which is why for a given large sample size, increasing sequencing depth can still increase imputation performance.

Now, in terms of history, the mice come from a heavily bottlenecked population, with a limited number of founding haplotypes broken up by about 100 generations of recombination. As such, assuming 1 cM/Mb, for a chromosome of length 100 Mbp, this means each haplotype each mouse might contain 100 recombinations relative to the founding haplotypes, one every megabase or so, and effectively fewer when one takes into account that many of these recombinations will be invisible *i.e.* do not change the hidden state. By contrast, for a human population assuming an Ne of 20,000, a value used previously in other imputation settings, if one considers a small enough region of the genome over which recombination is unlikely to have acted, then under a coalescent model, one would need to go back approximately $4 \times 20,000/20 = 4,000$ generations into the past to perfectly represent the population using $K = 20$ haplotypes. Although in reality I could go less far into the past to obtain a reasonable approximation using $K = 20$, this still means that to model a human haplotype drawn today using $K = 20$ founder haplotypes with a recombination rate of 1 cM/Mb would require on the order of thousands of recombinations, meaning a recombination every 50-100 kbp. As such, the genetic history of the two populations influences two things, the number of founder haplotypes one has to model, as well as the average distance between recombinations when modelling an individual haplotype as a mosaic of ancestral haplotypes. Increasing sample size, K and sequencing depth helps improve the former, while sequencing depth improves the later.

The end result of the above is that for the CFW mice, increasing from 1,000 to 2,000 samples at a fixed coverage, like 0.15 X, did not increase performance, as the

haplotypes were already well resolved, while increasing sequencing depth from 0.1 to 0.15X at 2,000 also had only a very marginal effect on imputation performance, as haplotype switch resolution was already sufficiently good to resolve haplotype switches. For the CONVERGE humans, it was also true that increasing sample size at a fixed coverage had no marginal increase on performance. However, the effect of more sequencing depth continued to increase sequencing performance, highlighting the importance that sequencing depth was still playing on resolving haplotype switches in this population.

In terms of future algorithmic developments, for the CFW mouse population, methodological improvements may yet improve performance at the lowest frequency variants, the variants affected by actions since the founding of the colony, for instance mutation or GC biased gene conversion, or variants in individual mice near recombination breakpoints. I expect the same to hold for any similarly heavily bottlenecked population, as in common in domesticated animals, and as such one should be able to obtain very high levels of imputation performance in many domesticated settings without a reference panel. For the humans, increasing sequencing depth even at high sample sizes does continue to increase imputation performance. As such, it is likely that it would yet still increase if I were to obtain more sequencing depth, both for rare and high frequency variants. Further algorithmic developments are likely to be of more use here, as the much larger haplotypic diversity in outbred populations would be more effectively modelled by some method where you phase directly using information from other samples, as has shown in general to be the case using previous imputation algorithms.

6.4.3 Implication for study design

In previous work, Pasaniuc *et al.* showed both that low coverage sequencing could generate reasonably accurate imputation, and showed that it could offer a larger

effective study sample size for a fixed budget than genotyping microarrays [138]. However, their work is conditional on the availability of haplotype reference panels; in Figures 1 and 3 of their work, they show markedly reduced imputation performance without reference panels, and they note in their Discussion that haplotype reference panels are a critical ingredient for extremely low coverage sequencing studies [138]. Indeed, in the study, specifically for the mice, using available methods without a reference panel did not produce useable results. Here, I show the possibility of using reference panel free imputation, for both a bottlenecked mouse population as well as a diverse outbred human population.

A natural question to ask when designing a study using low coverage sequencing reads is what sequencing depth should one use for a given population and sample size. I note that this is a complex question, and will depend on many factors beyond the genetic history of the population of interest, including phenotyping and sequencing costs, fixed costs, and, if a GWAS, the heritability and desired power of the study. I do not attempt to answer this question here, but do note that the CONVERGE results, being from a non-bottlenecked, diverse population, represent a lower bound on imputation performance for a given population, although the CFW results are a more promising set of results which should be achievable for other domesticated animals. I note that it would likely be a wise idea for a population or species which has not yet been used for low coverage sequencing to obtain some form of validation data, like array data for a group or high coverage sequencing data for a few subjects, and then sequence the population, or at least 1,000 subjects, in a multi-plexed fashion one lane at a time. For the CFW mice, multiplexing was done with about 1,000 animals at a time, so one lane (96X for a 2.5Gbp genome) yielded about 0.1X per sample. Additional lanes could be added as necessary until a desired level of imputation performance was reached, ideally on the entire population at once, or perhaps if need be, on only a thousand samples at once. It is worth noting that for the CFW

mice, library preparation was available for 14 or 34 GBP per sample, with the 34 option chosen, and sequencing cost approximately 91 GBP per X. As such, although it seems possible that some studies could get away with lower than 0.2X, it doesn't seem necessary to go any lower than about 0.2X, as it wouldn't allow for a significant number of additional samples to be genotyped with any remaining cost savings.

In terms of other methods for genotyping populations, recently, Genotype-By-Sequencing (GBS) approaches have been suggested as an alternative to genotyping microarrays [62]. In brief, this approach involves using restriction enzymes to break the genome into fragments, then sequencing the ends of those fragments, which allows one to effectively sequence only a fraction of the genome at a desired high coverage. For example, perfectly done, one could use this to sequence 5% of the genome in small chunks at moderate coverage of 10X at approximately equivalent cost to sequencing the entire genome at 0.5X. Benefits over genotyping microarrays are similar to those mentionned here, with no need to design customized genotyping arrays, and at 10X, comparable genotyping accuracy between sequencing and genotyping microarrays. When comparing GBS to low coverage imputation performed here, I see two areas in which low coverage sequencing should be superior to GBS. First, the potential for a slight increase in mapping resolution, although I caution that I expect this to be only moderate, as non-human and non-wild populations often have poor mapping resolution anyway. Second, and more importantly, low coverage imputation in bottlenecked populations allows one to reconstruct near whole genome sequences of all study animals, as compared to GBS, which gives genotypes at a small subset of the genome. In GBS, once a target region is identified, which might be in excess of a megabase in length like in the CFW mice, considerable work remains to characterize variation in the region to identify causal variants and genes, particularly if mapping results overlap many target genes. As such, low coverage whole genome sequencing and imputation gives investigators a great head start for fine mapping studies, particularly

if they can be used in conjunction with ENCODE style projects, such as the Mouse ENCODE project, to identify variants affecting functional non-exonic sequence [141]. Lastly, it is worth mentionning that the bioinformatic cost of the approach mentioned here is only marginally higher than that of GBS alone; both require read mapping and variant calling, with this method requiring an additional step of imputation, a step which is, for the CFW mice, was of a similar computational burden as variant calling. I therefore believe low coverage sequencing is a superior approach to GBS in general, but recommend caution by pointing out that low coverage sequencing and imputation may fail if: the population being sequenced is too small, the sample coverage is too low, or if the complexity of the population is too great; and if the reference genome(s) of the species are too fragmented for imputation procedures based on linkage disequilibrium to work properly. In those instances, GBS would be a safer alternative.

This study also offers a reasonable alternative strategy to the deliberate population founding and sequencing approaches taken in previous papers. For example, the rat Heterogenous Stock (HS), founded in the 1980's, is descended from 8 inbred strains of rats, and was recently used in a large GWAS of 1,407 samples where each of the founder strains was sequenced to high coverage (22X) and sample rats from the HS were genotyped on an Affymetrix array with 803,485 SNPs (265,551 informative polymorphic variants), which allowed for simple and effective imputation of the full catalogue of 7.2M SNPs to be imputed into all 1,407 samples nearly perfectly based on the known founders [142]. While this strategy enables near perfect imputation, it has several downsides, such as requiring the ability to fully inbreed animals (not an option with all species), the high coverage sequencing of the founders, and most crucially, expensive housing of the colony for many generations to allow for sufficient breakdown of linkage disequilibrium due to recombination. Here, we show that it is possible to achieve genotype accuracy only marginally poorer using an existing

population without the need to sequence founding animals (in fact, sequencing the 8 founders at 22X alone represents almost $((8) \times (34 + 91 \times 22))/(2073 \times (34 + 91 \times 0.18)) = 16\%$ of the sequencing budget used to sequence the 2073 CFW mice at 0.18X, with 34 GBP library prep cost and 91 GBP per X. For mice, the mice used in the CFW study cost about 5 pounds each while other mice more similar to the rats mentioned above, the DO mice, cost about 100 . As the costs of genotyping both sets of animals will now be very similar, with both populations containing similar patterns of linkage disequilibrium, I expect this method will substantially increase the number of populations which can be studied cheaply. Therefore, moving forward, particularly for non-laboratory animals where colonies are not as well established, mapping studies might have greater success for a fixed budget using existing available populations rather than trying to use a specific breeding design based on known founders.

6.4.4 Effect of variant filtration

Variant discovery and subsequent filtration of those variants is an important part of any study using next generation sequencing data. Various strategies have been developed to filter candidate lists of variants, including the Variant Quality Score Recalibrator, part of the GATK software package, which uses a known list of true positive sites to filter candidate variation based on QC metrics related to the sequencing reads. However, the success of such a method is contingent on the existence of a large enough set of known variation to use for filtration, which doesn't exist for many populations.

Here I examined in the low coverage outbred CFW mice to what extent variant filtration could be avoided by using a two stage strategy where I first try to impute all variants regardless of quality, and then in a second round re-impute using only high quality variants from the first round. This result worked, with large agreement in final SNP sets from the two strategies. I also achieved a slight but noticeable increase in

imputation accuracy in the two step procedure as compared to the original procedure for variants in both sets, suggesting accuracy can be increased by running a multi-stage imputation procedure.

However, while certainly promising, I urge caution when using a similar approach, in that depending on the population demographics, variants at different frequencies will have fundamentally different degrees of imputability and hence be removed at different rates. Although this won't matter much for GWASs, it may matter more for population genetic or other similar studies, where one would seek to avoid differential filtration based on frequency. While this is at least partially true in general for the VQSR or other variant filtration strategies using non-parametric QC metrics, and is unlikely to negatively influence most studies, it should nonetheless be kept in mind while analyzing results from a study using an imputation based variant filtration strategy.

It's also worth mentioning there is scope that future methodological improvements could help to separate errors (bad SNPs) from simply hard-to-impute SNPs

6.4.5 Further considerations

Users of this method should be aware of the way errors will be related locally, and how this might affect downstream analysis. While this is true for other imputation methods already, I think it bears special attention in this case. When using arrays or GBS, except for factors present over longer range, for instance CNVs, genotype likelihoods should be independent once they are no longer spanned by the same reads. By contrast, with this method, one could expect errors at a large number of SNPs to be correlated. For example, a low frequency haplotype may be poorly modelled, and the subjects containing those variants on that haplotype have a large number of errors over the length of that haplotype, causing an error at one SNP to be related to an error at another SNP across the same set of samples. While we expect certain

applications, notably GWAS, to be relatively robust to this phenomenon, others, such as tree building procedures, may suffer. In those cases, either a more selective set of variants should be taken based on quality control metrics, so that the error rate is low at the expense of more (non-randomly) missing data, or the downstream methods should be fundementally re-written to work directly on the underlying low coverage read data, or an array or GBS used for genotyping, at the expense of the density of variants provided here.

I also offer caution in using this approach in insufficiently homogenous populations, or if a small number of more genetically unrelated samples are included. The examples presented here, Han Chinese and CFW mice, are all relatively uniform in their composition; with the CFW showing relatively little structure through principal component analysis, and the Han Chinese showing only a modest proportion of their variance explained by the first few PC's. It is easy to imagine a small number of genetically distinct individuals being modelled poorly when included in a large set of more related individuals.

Chapter 7

Future directions and concluding remarks

In this thesis, I used whole genome sequencing data on multiple populations of wild mice, as well as wild derived and classical strains, to analyze and interpret factors that influence patterns of genetic variation. I looked at population demography, inbreeding, and selection, as well as both recent and ancient recombination. I also presented STITCH, an algorithm for low coverage imputation. Alongside the presentation of the results in the various chapters, I discussed the implications of the results, both in terms of their relationships to each other, as well as with what was already known from the literature. As such, instead of further discussion of the results, or presenting a final summary of the thesis (one was given in Chapter 1), I instead end this thesis by highlighting, through recapping and expanding upon, three areas I feel are promising areas for future research. First, I talk about the need to better characterize patterns of migration and introgression among species and subspecies of mice. Second, I talk about the need for better joint modelling of the effects of GC-biased gene conversion and selection on nucleotide diversity and evolution. Finally, I talk about the need to simulate or attempt low coverage imputation in populations and species other than mice and human, and suggest thoughts for improving the method on outbred wild species.

Addressing introgression and gene flow

In Chapter 3, through a variety of measures, I showed that introgression and gene flow between diverged species and subspecies of mice is common, through measures such as derived allele frequency distributions, population histories, and selection. Notably, each of the strongest signals of selection in the wild populations was highly differentiated from the population it was detected in. Yet despite these efforts, it was difficult to fully characterize migration and introgression. Some simple future approaches, like the ABBA-BABA tests (seen in, for example [143]), might help quantify and resolve the level of support that exists for some of the more striking results, like introgression between *M. m. domesticus* and *M. caroli*.

A more thorough understanding would likely require further model development, or more promising, through better sampling. As discussed in Chapter 3, I spent the first few months of my DPhil in a largely unsuccessful effort to build upon the HapMix model to search for introgression in the wild mice, which was made difficult by small (reference panel) sample sizes, as well as large effective population size and hence incomplete lineage sorting. It may be possibly that further methodological developments may allow this approach to more successfully address the nature of migration and gene flow in wild mice. However, a more in depth understanding is likely to come through greater sampling and genotyping of populations of mice from different species and subspecies. Then, methods such as HapMix [42] or Globetrotter [144] could be used to characterize migration and introgression. This would allow for future attempts to understand population history and selection in mice to leverage better knowledge of the rates of gene flow between species and subspecies, as well as to enable better characterization of regions of the genome under selection as arising from mutations on a haplotype within that population, or from introgression from a different species or subspecies.

Effect of GC-biased gene conversion on protein coding substitutions

In Chapter 4, I showed that in the Indian mice, GC-biased gene conversion strongly influences the probability a weak to strong or strong to weak substitution will fix. This can lead to the fixation of variants that are mildly deleterious if the meiotic drive is stronger than the variant's effect on phenotype. Given that recombination was shown to act broadly on the Indian genome, this leads to the possibility that meiotic drive may have a pervasive and perverse influence on evolution. While some have sought to analyze the distribution of fitness effects across the genome [3], and others have sought to interrogate the strength of GC-biased gene conversion [145], joint modelling of these two effects would likely help calculate, to what extent, GC-biased gene conversion promotes the fixation of deleterious phenotypic variants. Perhaps this might shed light on whether GC-biased gene conversion is ultimately a "good" by-product of recombination, by counteracting the strong to weak mutational bias and hence increasing genome information density, or a "bad" by-product of recombination, by eroding the fitness of the species it acts in.

Low coverage reference panel free imputation

In Chapter 6, I presented STITCH, a method for reference panel free read-aware low coverage genotype imputation. This was shown to accurately impute genotypes in a bottlenecked population of outbred laboratory mice. At current prices, it was less expensive to genotype 5.7 million SNPs using low coverage sequencing and use STITCH than it would be to genotype only 25,000 SNPs using a commercial genotyping microarray. I also showed that STITCH worked on fully outbred wild populations, using 1.7X coverage of 11,670 Han Chinese samples.

While these two (somewhat extreme) scenarios showed that STITCH can work well, it would be useful to more thoroughly explore the accuracy of STITCH for various sequencing coverages, and hence cost, for a wider range of populations, par-

ticularly other model organisms or agricultural ones. This could be done using either simulation based on known histories, or through direct studies of a few populations of interest. Hopefully this will further clarify the extent to which low coverage sequencing and STITCH represent a viable, cost-effective alternative to alternative genotyping approaches.

Algorithmically, I believe that STITCH is very close to providing an optimal imputation model for populations that have experienced recent bottlenecks, like the CFW population. Within the current framework, there is also likely scope to accommodate some, if not all, of other genetic polymorphisms in the imputation, such as short insertions or deletions, or larger copy number variants. For SNP based imputation however, there is likely still room for further development in more outbred species. One such approach would be a further Gibbs sampler. For example, one could imagine a method where, after initiation using the current methodology in STITCH, one could alternate between assigning reads to haplotypes, phasing each haplotype based on other sample haplotypes, and updating per-haplotype (per-individual) emission probabilities. However, there would be several computational challenges, as this would likely require substantially more RAM, and key to such an approach would be whether one could efficiently, and accurately, determine a subset of sample haplotypes to copy from in the phasing procedure, as phasing one haplotype using all other haplotypes would be computationally prohibitive. To select a representative subset of haplotypes well would likely be very difficult, not just as any such search would naively have computational complexity of the square of the number of samples, but also, given the nature of recombination, is not as simple as choosing the nearest neighbours of each haplotype. Improving the speed of searching for similar samples with less than quadratic computational complexity would require clever algorithmics, while to best select a subset of neighbouring haplotypes, one might consider partitioning the region to be imputed into disjoint neighbouring sets of SNPs, and choosing

minimal sets of haplotypes that explain each of those sets to a certain degree. With all of these issues addressed, algorithms similar to STITCH may yet yield substantially improved imputation of more outbred samples using low coverage sequencing, without the need for a reference panel.

Conclusion

In this thesis, I analyzed factors that influence patterns of genetic variation in wild mice, improving our knowledge of what has shaped the genome of one of the most important model organisms. Through understanding patterns of variation in wild mice, we may gain greater insight into understanding important biological questions. Future studies, in mice as well as other species, will help fully disentangle the complicated relationship between recombination, demography, selection, and other factors, and help us better understand the relationship between genotype and phenotype.

Bibliography

[1] Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE, et al. Subspecific origin and haplotype diversity in the laboratory mouse. Nature Genetics. 2011 May;43(7):648–655.

[2] Guénet JL, Bonhomme F. Wild mice: an ever-increasing contribution to a popular mammalian model. Trends in Genetics. 2003 Jan;19(1):24–31.

[3] Halligan DL, Kousathanas A, Ness RW, Harr B, Eöry L, Keane TM, et al. Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents. PLoS Genet. 2013 Dec;9(12):e1003995.

[4] Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011 Oct;478(7370):476–482.

[5] Lilue J, Müller UB, Steinfeldt T, Howard JC. Reciprocal virulence and resistance polymorphism in the relationship between Toxoplasma gondii and the house mouse. eLife. 2013 Oct;2:e01298.

[6] Yonekawa H, Moriwaki K, Gotoh O, Miyashita N, Matsushima Y, Shi LM, et al. Hybrid origin of Japanese mice "Mus musculus molossinus": evidence from restriction analysis of mitochondrial DNA. Molecular Biology and Evolution. 1988 Jan;5(1):63–78.

[7] Prager EM, Orrego C, Sage RD. Genetic variation and phylogeography of central Asian and other house mice, including a major new mitochondrial lineage in Yemen. Genetics. 1998 Oct;150(2):835–861.

[8] Suzuki H, Nunome M, Kinoshita G, Aplin KP, Vogel P, Kryukov AP, et al. Evolutionary and dispersal history of Eurasian house mice Mus musculus clarified by more extensive geographic sampling of mitochondrial DNA. Heredity. 2013 Nov;111(5):375–390.

[9] Boursot P, Auffray JC, Britton-Davidian J, Bonhomme F. The Evolution of House Mice. Annual Review of Ecology and Systematics. 1993;24(1):119–152.

[10] Baines JF, Harr B. Reduced X-Linked Diversity in Derived Populations of House Mice. Genetics. 2007 Apr;175(4):1911–1921.

[11] Geraldes A, Basset P, Gibson B, Smith KL, Harr B, Yu HT, et al. Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Molecular Ecology. 2008 Dec;17(24):5349–5363.

[12] Buard J, Rivals E, Dunoyer de Segonzac D, Garres C, Caminade P, de Massy B, et al. Diversity of Prdm9 Zinc Finger Array in Wild Mice Unravels New Facets of the Evolutionary Turnover of this Coding Minisatellite. PLoS ONE. 2014 Jan;9(1):e85021.

[13] Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011 Sep;477(7364):289–294.

[14] Kono H, Tamura M, Osada N, Suzuki H, Abe K, Moriwaki K, et al. Prdm9 Polymorphism Unveils Mouse Evolutionary Tracks. DNA Research. 2014 Jan;p. dst059.

[15] Hardouin EA, Orth A, Teschke M, Darvish J, Tautz D, Bonhomme F. Eurasian house mouse (Mus musculus L.) differentiation at microsatellite loci identifies the Iranian plateau as a phylogeographic hotspot. BMC Evolutionary Biology. 2015 Feb;15(1):26.

[16] Forejt J, Iványi P. Genetic studies on male sterility of hybrids between laboratory and wild mice (Mus musculus L.). Genetical Research. 1974 Oct;24(2):189–206.

[17] Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J. A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase. Science. 2009 Jan;323(5912):373–375.

[18] Staubach F, Lorenc A, Messer PW, Tang K, Petrov DA, Tautz D. Genome Patterns of Selection and Introgression of Haplotypes in Natural Populations of the House Mouse (Mus musculus). PLoS Genet. 2012 Aug;8(8):e1002891.

[19] Song Y, Endepols S, Klemann N, Richter D, Matuschka FR, Shih CH, et al. Adaptive Introgression of Anticoagulant Rodent Poison Resistance by Hybridization between Old World Mice. Current Biology. 2011 Aug;21(15):1296–1301.

[20] Liu KJ, Steinberg E, Yozzo A, Song Y, Kohn MH, Nakhleh L. Interspecific introgressive origin of genomic diversity in the house mouse. Proceedings of the National Academy of Sciences. 2015 Jan;112(1):196–201.

[21] Baudat F, Imai Y, de Massy B. Meiotic recombination in mammals: localization and regulation. Nature Reviews Genetics. 2013 Nov;14(11):794–806.

[22] Handel MA, Schimenti JC. Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. Nature Reviews Genetics. 2010 Feb;11(2):124–136.

[23] Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, et al. Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa. PLoS Genet. 2009 Dec;5(12):e1000753.

[24] Allers T, Lichten M. Differential Timing and Control of Noncrossover and Crossover Recombination during Meiosis. Cell. 2001 Jul;106(1):47–57.

[25] Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, et al. Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. Science. 2010 Feb;327(5967):876–879.

[26] Parvanov ED, Petkov PM, Paigen K. Prdm9 Controls Activation of Mammalian Recombination Hotspots. Science. 2010 Feb;327(5967):835–835.

[27] Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, et al. PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. Science. 2010 Feb;327(5967):836–840.

[28] Neale MJ, Keeney S. Clarifying the mechanics of DNA strand exchange in meiotic recombination. Nature. 2006 Jul;442(7099):153–158.

[29] Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV. Genetic recombination is directed away from functional genomic elements in mice. Nature. 2012 May;485(7400):642–645.

[30] Jeffreys AJ, Neumann R. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nature Genetics. 2002 Jul;31(3):267–271.

[31] Jeffreys AJ, Neumann R. Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. Human Molecular Genetics. 2005 Aug;14(15):2277–2287.

[32] Webster MT, Hurst LD. Direct and indirect consequences of meiotic recombination: implications for genome evolution. Trends in Genetics. 2012 Mar;28(3):101–109.

[33] Coop G, Myers SR. Live Hot, Die Young: Transmission Distortion in Recombination Hotspots. PLoS Genet. 2007 Mar;3(3):e35.

[34] Boulton A, Myers RS, Redfield RJ. The hotspot conversion paradox and the evolution of meiotic recombination. Proceedings of the National Academy of Sciences. 1997 Jul;94(15):8058–8063.

[35] Jeffreys AJ, Cotton VE, Neumann R, Lam KWG. Recombination regulator PRDM9 influences the instability of its own coding sequence in humans. Proceedings of the National Academy of Sciences. 2013 Jan;110(2):600–605.

[36] Odenthal-Hesse L, Berg IL, Veselis A, Jeffreys AJ, May CA. Transmission Distortion Affecting Human Noncrossover but Not Crossover Recombination: A Hidden Source of Meiotic Drive. PLoS Genet. 2014 Feb;10(2):e1004106.

[37] Wallberg A, Glémin S, Webster MT. Extreme Recombination Frequencies Shape Genome Variation and Evolution in the Honeybee, Apis mellifera. PLoS Genet. 2015 Apr;11(4):e1005189.

[38] Cole F, Baudat F, Grey C, Keeney S, de Massy B, Jasin M. Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. Nature Genetics. 2014 Oct;46(10):1072–1080.

[39] Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. Recombination initiation maps of individual human genomes. Science. 2014 Nov;346(6211):1256442.

[40] Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proceedings of the National Academy of Sciences. 2015 Feb;112(7):2109–2114.

[41] Bonhomme F, Rivals E, Orth A, Grant GR, Jeffreys AJ, Bois PR. Species-wide distribution of highly polymorphic minisatellite markers suggests past and present genetic exchanges among house mouse subspecies. Genome Biology. 2007 May;8(5):R80.

[42] Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, et al. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLoS Genet. 2009 Jun;5(6):e1000519.

[43] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. In: PROCEEDINGS OF THE IEEE; 1989. p. 257–286.

[44] Scheet P, Stephens M. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. The American Journal of Human Genetics. 2006 Apr;78(4):629–644.

[45] Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, et al. Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals. Cell. 2013 Aug;154(3):530–540.

[46] Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. Evidence for Pervasive Adaptive Protein Evolution in Wild Mice. PLoS Genet. 2010 Jan;6(1):e1000825.

[47] Eisen EJ. The Mouse in Animal Genetics and Breeding Research. Imperial College Press; 2005.

[48] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;43(5):491–498.

[49] Yang H, Ding Y, Hutchins LN, Szatkiewicz J, Bell TA, Paigen BJ, et al. A customized and versatile high-density genotyping array for the mouse. Nature Methods. 2009 Aug;6(9):663–666.

[50] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 Jul;25(14):1754–1760.

[51] Lunter G, Goodson M. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Research. 2011 Jun;21(6):936–939.

[52] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug;25(16):2078–2079.

[53] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010 Sep;20(9):1297–1303.

[54] Didion JP, Yang H, Sheppard K, Fu CP, McMillan L, de Villena FPM, et al. Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. BMC Genomics. 2012 Jan;13:34.

[55] Auton A, Fledel-Alon A, Pfeifer S, Venn O, Ségurel L, Street T, et al. A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. Science. 2012 Mar;.

[56] Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, et al. Great ape genetic diversity and population history. Nature. 2013 Jul;advance online publication.

[57] Ai H, Fang X, Yang B, Huang Z, Chen H, Mao L, et al. Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. Nature Genetics. 2015 Mar;47(3):217–225.

[58] Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wgs500 Consortium, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics. 2014 Jul;advance online publication.

[59] Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nature Genetics. 2012 Jan;44(2):226–232.

[60] Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011 Feb;470(7332):59–65.

[61] Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015 Oct;526(7571):75–81.

[62] Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE. 2011 May;6(5):e19379.

[63] Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011 Jul;advance online publication.

[64] Suzuki H, Shimada T, Terashima M, Tsuchiya K, Aplin K. Temporal, spatial, and ecological modes of evolution of Eurasian Mus based on mitochondrial and nuclear gene sequences. Molecular Phylogenetics and Evolution. 2004 Dec;33(3):626–646.

[65] Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet. 2009 Oct;5(10):e1000695.

[66] Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, et al. Reconstructing Native American population history. Nature. 2012 Aug;488(7411):370–374.

[67] Do R, Balick D, Li H, Adzhubei I, Sunyaev S, Reich D. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nature Genetics. 2015 Feb;47(2):126–131.

[68] Frantz LAF, Schraiber JG, Madsen O, Megens HJ, Cagan A, Bosse M, et al. Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. Nature Genetics. 2015 Aug;advance online publication.

[69] Zhao S, Zheng P, Dong S, Zhan X, Wu Q, Guo X, et al. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nature Genetics. 2013;45(1):67–71.

[70] Wallberg A, Han F, Wellhagen G, Dahle B, Kawata M, Haddad N, et al. A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee Apis mellifera. Nature Genetics. 2014 Oct;46(10):1081–1088.

[71] Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of Population Structure using Dense Haplotype Data. PLoS Genet. 2012 Jan;8(1):e1002453.

[72] Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. Nature. 2015 Mar;519(7543):309–314.

[73] Sheehan S, Harris K, Song YS. Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. Genetics. 2013 Jul;194(3):647–662.

[74] Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007 Oct;449(7164):851–861.

[75] Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nature Methods. 2013 Jan;10(1):5–6.

[76] O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genet. 2014 Apr;10(4):e1004234.

[77] Delaneau O, Howie B, Cox A, Zagury JF, Marchini J. Haplotype Estimation Using Sequencing Reads. The American Journal of Human Genetics. 2013 Oct;93(4):687–696.

[78] McVean G. A Genealogical Interpretation of Principal Components Analysis. PLoS Genet. 2009 Oct;5(10):e1000686.

[79] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. 2009 Jul;.

[80] Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics. 1989;5:164–166.

[81] Pickrell JK, Pritchard JK. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genet. 2012 Nov;8(11):e1002967.

[82] Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008 Nov;456(7218):98–101.

[83] Revell LJ, Chamberlain SA. Rphylip: an R interface for PHYLIP. Methods in Ecology and Evolution. 2014 Sep;5(9):976–981.

[84] Charlesworth D, Willis JH. The genetics of inbreeding depression. Nature Reviews Genetics. 2009 Nov;10(11):783–796.

[85] Ceballos FC, Álvarez G. Royal dynasties as human inbreeding laboratories: the Habsburgs. Heredity. 2013 Aug;111(2):114–121.

[86] Joshi PK, Esko T, Mattsson H, Eklund N, Gandin I, Nutile T, et al. Directional dominance on stature and cognition in diverse human populations. Nature. 2015 Jul;523(7561):459–462.

[87] Bosse M, Megens HJ, Madsen O, Paudel Y, Frantz LAF, Schook LB, et al. Regions of Homozygosity in the Porcine Genome: Consequence of Demography and the Recombination Landscape. PLoS Genet. 2012 Nov;8(11):e1003100.

[88] Laurie CC, Nickerson DA, Anderson AD, Weir BS, Livingston RJ, Dean MD, et al. Linkage Disequilibrium in Wild Mice. PLoS Genet. 2007 Aug;3(8):e144.

[89] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007 Sep;81(3):559–575.

[90] Cox A, Ackert-Bicknell CL, Dumont BL, Ding Y, Bell JT, Brockmann GA, et al. A New Standard Genetic Map for the Laboratory Mouse. Genetics. 2009 Aug;182(4):1335–1344.

[91] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 2012 Apr;6(2):80–92.

[92] MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. Science. 2012 Feb;335(6070):823–828.

[93] McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. Science. 2004 Apr;304(5670):581–584.

[94] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov;491(7422):56–65.

[95] Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, et al. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. American Journal of Human Genetics. 2004 Jun;74(6):1111–1120.

[96] Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. Science. 2010 May;328(5979):710–722.

[97] Li M, Tian S, Jin L, Zhou G, Li Y, Zhang Y, et al. Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. Nature Genetics. 2013 Dec;45(12):1431–1438.

[98] Qanbari S, Pausch H, Jansen S, Somel M, Strom TM, Fries R, et al. Classic Selective Sweeps Revealed by Massive Sequencing in Cattle. PLoS Genet. 2014 Feb;10(2):e1004148.

[99] Carneiro M, Rubin CJ, Palma FD, Albert FW, Alföldi J, Barrio AM, et al. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. Science. 2014 Aug;345(6200):1074–1079.

[100] Axelsson E, Ratnakumar A, Arendt ML, Maqbool K, Webster MT, Perloski M, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature. 2013 Mar;495(7441):360–364.

[101] Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, et al. A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. Science. 2010 Feb;327(5967):883–886.

[102] Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002 Oct;419(6909):832–837.

[103] Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007 Oct;449(7164):913–918.

[104] Voight BF, Kudaravalli S, Wen X, Pritchard JK. A Map of Recent Positive Selection in the Human Genome. PLoS Biol. 2006 Mar;4(3):e72.

[105] Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, et al. Identifying Recent Adaptations in Large-Scale Genomic Data. Cell. 2013 Feb;152(4):703–713.

[106] Tajima F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics. 1989 Nov;123(3):585–595.

[107] Fay JC, Wu CI. Hitchhiking Under Positive Darwinian Selection. Genetics. 2000 Jul;155(3):1405–1413.

[108] Fu YX. Statistical Properties of Segregating Sites. Theoretical Population Biology. 1995 Oct;48(2):172–197.

[109] Fu YX. Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. Genetics. 1997 Oct;147(2):915–925.

[110] Kumar S, Tamura K, Nei M. MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. Computer applications in the biosciences: CABIOS. 1994 Apr;10(2):189–191.

[111] Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular Biology and Evolution. 2013 Oct;p. mst197.

[112] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007 Nov;23(21):2947–2948.

[113] Bole-Feysot C, Goffin V, Edery M, Binart N, Kelly PA. Prolactin (PRL) and Its Receptor: Actions, Signal Transduction Pathways and Phenotypes Observed in PRL Receptor Knockout Mice. Endocrine Reviews. 1998 Jun;19(3):225–268.

[114] Wiemers DO, Shao Lj, Ain R, Dai G, Soares MJ. The Mouse Prolactin Gene Family Locus. Endocrinology. 2003 Jan;144(1):313–325.

[115] Lord CJ, Ashworth A. RAD51, BRCA2 and DNA repair: a partial resolution. Nature Structural & Molecular Biology. 2007 Jun;14(6):461–462.

[116] Roy R, Chun J, Powell SN. BRCA1 and BRCA2: different roles in a common pathway of genome protection. Nature Reviews Cancer. 2012 Jan;12(1):68–78.

[117] Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. Science. 2005 Oct;310(5746):321–324.

[118] Smagulova F, Gregoretti IV, Brick K, Khil P, Camerini-Otero RD, Petukhova GV. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. Nature. 2011 Apr;472(7343):375–378.

[119] Baker CL, Kajita S, Walker M, Saxl RL, Raghupathy N, Choi K, et al. PRDM9 Drives Evolutionary Erosion of Hotspots in Mus musculus through Haplotype-Specific Initiation of Meiotic Recombination. PLoS Genet. 2015 Jan;11(1):e1004916.

[120] Duret L, Galtier N. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annual Review of Genomics and Human Genetics. 2009;10(1):285–311.

[121] Baker CL, Walker M, Kajita S, Petkov PM, Paigen K. PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration. Genome Research. 2014 May;24(5):724–732.

[122] McVean G, Awadalla P, Fearnhead P. A Coalescent-Based Method for Detecting and Estimating Recombination From Gene Sequences. Genetics. 2002 Mar;160(3):1231–1241.

[123] Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME Suite: tools for motif discovery and searching. Nucleic Acids Research. 2009 Jul;37(suppl 2):W202–W208.

[124] Singhal S, Leffler E, Sannareddy K, Turner I, Venn O, Hooper D, et al. Stable recombination hotspots in birds. bioRxiv. 2015 Jul;p. 023101.

[125] Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father/'s age to disease risk. Nature. 2012 Aug;488(7412):471–475.

[126] Galtier N, Duret L, Glémin S, Ranwez V. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends in Genetics. 2009 Jan;25(1):1–5.

[127] Tumian A. The evolution of recombination hotspots and their relationship with DNA sequences. University of Oxford; 2013.

[128] Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Research. 1990 Oct;18(20):6097–6100.

[129] Cole F, Keeney S, Jasin M. Comprehensive, fine-scale dissection of homologous recombination outcomes at a hot spot in mouse meiosis. Molecular Cell. 2010 Sep;39(5):700–710.

[130] Paigen K, Petkov P. Mammalian recombination hot spots: properties, control and evolution. Nature Reviews Genetics. 2010 Mar;11(3):221–233.

[131] Auton A, Rui Li Y, Kidd J, Oliveira K, Nadel J, Holloway JK, et al. Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. PLoS Genet. 2013 Dec;9(12):e1003984.

[132] Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research. 2014 Jan;42(D1):D1001–D1006.

[133] Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, et al. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature. 2007 Jul;448(7157):1050–1053.

[134] Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet. 2009 Jun;5(6):e1000529.

[135] Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. Genetic epidemiology. 2010 Dec;34(8):816–834.

[136] Browning SR, Browning BL. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. The American Journal of Human Genetics. 2007 Nov;81(5):1084–1097.

[137] Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through prephasing. Nature Genetics. 2012 Aug;44(8):955–959.

[138] Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nature Genetics. 2012;44(6):631–635.

[139] Converge Consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature. 2015 Jul;advance online publication.

[140] Yalcin B, Nicod J, Bhomra A, Davidson S, Cleak J, Farinelli L, et al. Commercially Available Outbred Mice for Genome-Wide Association Studies. PLoS Genet. 2010;6(9):e1001085.

[141] Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014 Nov;515(7527):355–364.

[142] Rat Genome Sequencing and Mapping Consortium. Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. Nature Genetics. 2013 Jul;45(7):767–775.

[143] The Heliconius Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature. 2012;.

[144] Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A Genetic Atlas of Human Admixture History. Science. 2014 Feb;343(6172):747–751.

[145] Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. Quantification of GC-biased gene conversion in the human genome. Genome Research. 2015 Aug;25(8):1215–1228.