COMPUTATIONAL TOOLS TO AID THE DESIGN AND DEVELOPMENT OF A GENETIC REFERENCE POPULATION

Catherine E. Welsh

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science.

Chapel Hill
2014

Approved by:

Leonard McMillan

Wei Wang

Jan Prins

Fernando Pardo-Manuel de Villena

William Valdar

ABSTRACT

CATHERINE E. WELSH. COMPUTATIONAL TOOLS TO AID THE DESIGN AND DEVELOPMENT OF A GENETIC REFERENCE POPULATION. (Under the direction of Leonard McMillan.)

Model organisms are important tools used in biological and medical research. A key component of a genetics model organism is a known and reproducible genome. In the early 1900s, geneticists developed methods for fixing genomes by inbreeding. First generation genetic models used inbreeding to create disease models from animals with spontaneous or stimulated mutations.

Recently, geneticists have begun to develop a second generation of models which better represent the human population in terms of diversity. One such model is the Collaborative Cross (CC), which is a mouse model derived from 8 founders. I have been involved in developing the CC since its early stages. In particular, I am interested in speeding up the inbreeding process, since it currently takes an average of thirty-six generations to achieve complete fixation.

To speed up the inbreeding process, I developed a simulator that replicates the breeding process and tested various breeding strategies before applying them to a CC. To apply the simulation techniques to live mice, a fast, low-cost way to monitor their genomes at each generation was needed. As a result, two genotyping arrays were designed, a first generation array with 7,851 markers called MUGA and a second generation array called MegaMUGA with 77,800 markers. Both arrays were designed specifically to be maximally informative for the CC population. Using these genotyping arrays, one can determine from which of the eight CC founders each part of a developing mouse lines genome is inherited. I refer to these as haplotype reconstructions, and they are used as the input into my simulations as well as various other monitoring tools. To determine the accuracy of these haplotype reconstructions, I used DNA sequencing data for three samples which were also genotyped, and compared the haplotype reconstructions from the DNA

sequencing data to solutions from the genotyping array data.

In loving memory of my mother,

whose support for me in all things made this possible.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Leonard McMillan, for his patience, his encouragement to pursue these ideas, for continually asking me to do a little more or a little better, and for his new perspectives that often seemed wrong to me at first but never were.

To my other committee members, thank you for your wonderful insights and ideas that have enriched this research and encouraged me to pursue angles and projects that I would not have otherwise. They are Wei Wang, Fernando Pardo-Manuel de Villena, Jan Prins and William Valdar.

I am also grateful for my truly supportive lab members throughout this experience. Thank you to Ping Fu, Jeremy Wang, James Holt, Katy Kao, and many others throughout the years with whom I interacted.

For my Bandidos crew over the years, Tabitha Peck, Laura Kassler, David Millman, Darrell Bethea, Alana Libonati, Robby Cochran, and the occasional guest appearances by Brittany Fasy, I am very thankful. These Friday lunchtime sessions kept me sane throughout graduate school and were often the highlights of my week!

I am thankful for my late mother's encouragement, without which I may not have pursued my PhD, and for my father for supporting me and for always knowing it would get done eventually. I'm eternally grateful to my sister for allowing me to vent my frustrations to her on numerous occasions. I'm also thankful to my brother-in-law for always being willing to discuss my research and offer new perspectives. I'm humbled by my family's willingness to put up with my poor graduate student status for so long and for visiting me as much as possible. I'm grateful that Linda made sure my father checked in with me throughout this process and also did so herself.

Finally, to the love of my life, my husband Drew, thank you for your constant and unwavering support for which I am especially grateful.

TABLE OF CONTENTS

LIST OF FIGURES . . . . **xi**

LIST OF TABLES . . . . **xvi**

1 INTRODUCTION **1**

1.1 Selective Breeding 2

1.2 Isogenics 2

1.3 Combining Inbreeding and Selection 3

1.4 Thesis Statement 4

1.5 Organization 5

2 BACKGROUND **6**

2.1 Genomic Data 6

2.1.1 SNPs 7

2.1.2 Microarrays 7

2.1.3 High-throughput sequencing 8

2.2 Recombination and Breeding 9

2.2.1 Genetic Reference Populations 10

2.2.2 Collaborative Cross 11

2.2.3 Diversity Outbred 12

2.3 Related Work 13

3 MARKER-ASSISTED INBREEDING THEORETICAL ANALYSIS 16

3.1 Approach 17

3.2 Experiments and Results 24

3.2.1 Nonmarker-assisted breeding schemes 25

3.2.2 Marker-assisted inbreeding 28

3.2.3 Selected advanced intercrosses 31

3.2.4 Low-resolution sampling 32

3.3 Conclusion 34

4 DESIGNING MICRO-ARRAYS FOR MAXIMUM INFORMATIVENESS . . . 36

4.1 MUGA 37

4.1.1 Database of Available Probes 38

4.1.2 Design 39

4.1.3 Samples Run on MUGA 41

4.1.4 Performance on MUGA 41

4.2 MegaMUGA 48

4.2.1 Design 48

4.2.2 Samples Run on MegaMUGA 54

4.2.3 Performance on MegaMUGA 54

4.3 Conclusion 54

5 MAI IN PRACTICE 57

5.1 Analysis Tools 57

5.1.1 Haplotype Reconstructions 57

5.1.2 Line Quality Control 59

5.1.3 Breeder Selection 65

5.1.4 Compute Union 68

5.2 Conclusion 70

6 HIGH-THROUGHPUT SEQUENCING DATA 75

6.1 Sequence Data 76

6.2 HMM Algorithm 77

6.2.1 Variants 77

6.2.2 Emission Probabilities 77

6.2.3 Transition Probabilities 80

6.2.4 Viterbi Solution 81

6.3 Founder-pair Resolution 83

6.4 Breakpoint Mapping of HTS data 84

6.4.1 Comparison with Refined Breakpoint Solution 84

6.4.2 Comparison to Genotype Solutions 88

6.4.3 Read Coverage Analysis 95

6.5 Conclusion 101

7 CONCLUSION AND FUTURE WORK 102

7.1 Theoretical Marker-Assisted Techniques 102

7.1.1 Future Work 103

7.2 Tools for Analysis of Live Mice 103

7.2.1 Future Work 104

7.3 High-Throughput Sequencing Data 104

7.3.1 Future Work 105

APPENDIX A SEQUENCE SIMILARITY MAPS 107

BIBLIOGRAPHY 123

LIST OF FIGURES

2.1 Depiction of a SNP. 7

2.2 Collaborative Cross breeding scheme. 12

2.3 Diversity Outbred breeding explanation. 14

3.1 The average number of founder segments in eight-way RILs at various generations of inbreeding. 17

3.2 The number of generations to complete fixation (A) and the number of resulting founder segments (B) in two-way and eight-way RILs. 18

3.3 This image shows all possible JH states between a potential mating-pair and illustrates my notion of a genomic segment. 19

3.4 A state diagram showing the transitions between all JH states in a single generation. 21

3.5 A histogram of segments colored according to their JH state as a function of generation. 22

3.6 This figure shows the pedigree diagrams for the alternating backcrosses: father-daughter backcross with the mother-son backcross (A) and the father-daughter with the random sibmating (B). 26

3.7 A comparison of five breeder selection alternatives for generating an eight-way RIL, showing the number of generations to reach complete fixation (A) and the total number of segments (B) found in the final inbred lines. 27

3.8 Compares the number of generations it takes to achieve complete fixation for 5 breeding schemes that make different assumptions about the available pool of breeders. 30

3.9 CHF as a function of number of generations. 31

3.10 Histogram of size of segments missed when conducting low resolution sampling. 33

4.1 Subset of all compatible triples and the chaining of compatible triples. 39

4.2 Distribution of the original 9,000 SNPs on MUGA in terms of location on chromosomes. 42

4.3 Distribution of the final 7,541 MUGA SNPs in terms of location on chromosomes (left plot) and spacing between SNPs (right plot). 43

4.4 Intensity plots of four markers, colored by genotype calls from Illumina. 47

4.5 Recombination events on a chromosome basis. 51

4.6 Distribution of the final 77,808 MegaMUGA SNPs in terms of location on chromosomes. 52

4.7 Break down of the final distribution of SNPs by type for which they were originally chosen. 53

5.1 Haplotype reconstruction of a CC mouse sample. 60

5.2 Quality Control Report screenshot. 61

5.3 Distribution of shared recombination events before and after identifying related samples. 63

5.4 Comparison table created by Choose Breeders tool. 66

5.5 Combined haplotype reconstruction images of proposed breeders as created by the Choose Breeders tool. 67

5.6 Simulate Matings tool output. 69

5.7 Partial view of the pedigree of the OR3252 CC line as well as the genome of obligate ancestors and the genome of the line based on MUGA genotypes. 71

5.8 Residual heterozygosity in distributable lines. 72

5.9 The UNC Systems Genetics Core web site. 73

6.1 HMM emission probabilities diagram for one bin. 79

6.2 HMM transition probabilities. 81

6.3 Four classes of transitions that can occur between HMM states. 82

6.4 Sequence similarity maps for three founder-pairs and histogram of spacing between informative SNPs for all founder-pairs. 85

6.5 Histogram of distance between HMM solution and refined recombination breakpoint. 87

6.6 Depiction of all informative SNPs between A/J and NOD/ShiLtJ on Chromosome 13 from 101.5Mb to 104.0Mb. 88

6.7 Histogram of distance between informative SNPS in the refined breakpoint solution. 89

6.8 Comparison of HTS full coverage solutions with MUGA and MegaMUGA solutions for OR867m532. 92

6.9 Comparison of HTS full coverage solutions with MUGA and MegaMUGA solutions for OR1237m224. 93

6.10 Comparison of HTS full coverage solutions with MUGA and MegaMUGA solutions for OR3067m352. 94

6.11 Comparison of HTS full coverage solution for OR867m352 with 4x and 0.25x coverage solutions. 98

6.12 Comparison of HTS full coverage solutions for OR1237m224 with 4x and 0.25x coverage solutions. 98

6.13 Comparison of HTS full coverage solutions for OR3067m352 with 4x and 0.25x coverage solutions. 99

6.14 Histograms of the delta between the maximum position and minimum position found at each recombination among the 10 runs at each coverage level. 100

A.1 Sequence similarity map for CC founders A (A/J) and B (C57BL/6J). 108

A.2 Sequence similarity map for CC founders A (A/J) and C(129S1/SvImJ). 108

A.3 Sequence similarity map for CC founders A (A/J) and D (NOD/ShiLtJ). 109

A.4 Sequence similarity map for CC founders A (A/J) and E (NZO/HlLtJ). 109

A.5 Sequence similarity map for CC founders A (A/J) and F (CAST/EiJ). 110

A.6 Sequence similarity map for CC founders A (A/J) and G (PWK/PhJ). 110

A.7 Sequence similarity map for CC founders A (A/J) and H (WSB/EiJ). 111

A.8 Sequence similarity map for CC founders B (C57BL/6J) and C (129S1/SvImJ). 111

A.9 Sequence similarity map for CC founders B (C57BL/6J) and D (NOD/ShiLtJ). 112

A.10 Sequence similarity map for CC founders B (C57BL/6J) and E (NZO/HlLtJ). . 112

A.11 Sequence similarity map for CC founders B (C57BL/6J) and F (CAST/EiJ). . . 113

A.12 Sequence similarity map for CC founders B (C57BL/6J) and G (PWK/PhJ). 114

A.13 Sequence similarity map for CC founders B (C57BL/6J) and H (WSB/EiJ). 114

A.14 Sequence similarity map for CC founders C (129S1/SvImJ) and D (NOD/ShiLtJ). 115

A.15 Sequence similarity map for CC founders C (129S1/SvImJ) and E (NZO/HlLtJ). 115

A.16 Sequence similarity map for CC founders C (129S1/SvImJ) and F (CAST/EiJ). 116

A.17 Sequence similarity map for CC founders C (129S1/SvImJ) and G (PWK/PhJ). 116

A.18 Sequence similarity map for CC founders C (129S1/SvImJ) and H (WSB/EiJ). 117

A.19 Sequence similarity map for CC founders D (NOD/ShiLtJ) and E (NZO/HlLtJ). 117

A.20 Sequence similarity map for CC founders D (NOD/ShiLtJ) and F (CAST/EiJ). 118

A.21 Sequence similarity map for CC founders D (NOD/ShiLtJ) and G (PWK/PhJ). 118

A.22 Sequence similarity map for CC founders D (NOD/ShiLtJ) and H (WSB/EiJ). 119

A.23 Sequence similarity map for CC founders E (NZO/HlLtJ) and F (CAST/EiJ). 119

A.24 Sequence similarity map for CC founders E (NZO/HlLtJ) and G (PWK/PhJ). 120

A.25 Sequence similarity map for CC founders E (NZO/HlLtJ) and H (WSB/EiJ). 120

A.26 Sequence similarity map for CC founders F (CAST/EiJ) and G (PWK/PhJ). . . . 121

A.27 Sequence similarity map for CC founders F (CAST/EiJ) and H (WSB/EiJ). . . . 121

A.28 Sequence similarity map for CC founders G (PWK/PhJ) and H (WSB/EiJ). . . . 122

LIST OF TABLES

2.1 Letter and color codes for CC Founders. 13

4.1 Comparison of biological replicate samples run on MUGA. 44

4.2 Entropy scores for micro-array comparison. 45

4.3 Comparison of biological replicate samples run on MegaMUGA. 55

5.1 Genome-wide CC founder contribution. 65

6.1 Comparison of HTS to genotype solutions. 90

6.2 Statistics for various coverage levels of sequencing and genotyping data for the three CC samples. 96

CHAPTER1: INTRODUCTION

The rediscovery of Gregor Mendel's Laws of Inheritance in 1900 launched a new wave of genetic studies. In particular, scientists wished to verify Mendel's laws in organisms other than plants since it was unknown if the laws would hold true for animals[40]. The mouse was chosen for these particular studies as it is ideally suited for genetic analysis since it has a relatively short generation time (about 10 weeks from birth to giving birth), it breeds easily in captivity and mice are easily housed in small cages[53]. An additional benefit is that due to mouse fanciers in the 1800s, numerous variants of mice had been derived which had visible phenotypes. A systematic analysis of inheritance and genetic variation in mice as well as other mammals ensued[53]. These early studies utilized selective breeding to verify recessive and dominant trait inheritance patterns. They also brought to light the existence of more than two alleles at a locus, recessive lethal alleles, and interactions among unlinked genes[53].

Shortly afterward, scientists realized the need to have inbred reproducible genetically homogeneous lines of mice, and these lines were soon developed. Inbred strains are easily reproducible and they are useful disease models[22]. In fact, the majority of inbred strains, from the most recent back to the first strain, were developed for use in cancer research, to prove or disprove the existence of genetic factors influencing the incidence of cancer and the independence of inheritance of different types of cancers[22]. By selection during inbreeding, various types of malignancies in predictable frequencies were established in several genotypes. As inbred strains became available and information about them began appearing in scientific literature, investigators recognized that these animals could contribute greatly to medical research. It became possible to use biological material in experiments with confidence as the only variables were those the investigator chose to include in the experimental design[22].

1.1 Selective Breeding

Quantitative genetics deals with the inheritance of complex traits that are controlled by many loci, each with relatively small effects, and by environmental influences such as diet and exercise[22]. In order to study quantitative genetics, selective breeding, which refers to the systematic breeding of animals in order to choose certain qualities in them, is often used. Breeders will select for quantitative phenotypes such as body weight, growth rate, feed efficiency, feed intake, body composition and litter size. By selecting for these traits, scientists can then see how far artificial selection can change a trait and how many generations are needed to reach a limit, if a limit can be reached[53]. During the selection process targeted traits are chosen intentionally, however, without specific controls, other traits may also be selected inadvertently, such as fertility and docility.

As technology advanced, it was possible to select for particular genotypes or genes. Through selection for quantitative traits, a number of mutations and variants were formed. In the analysis of these mutants, it is often not possible to distinguish between subtle effects due to the mutation itself and effects due to other genes within the background of the mutant strain. To make this distinction, it is essential to be able to compare animals in which differences in the genetic background have been eliminated as a variable in the experiment. This is accomplished through the placement of the mutation into a genome of another mouse strain. To do this, genotyping of the area surrounding the gene is done at each level of breeding and offspring with the gene of interest are selected for further breeding[53].

1.2 Isogenics

Inbreeding is the mating of related individuals, and leads to the creation of animals that are homozygous (same allele) at each locus, meaning both copies of each chromosome are identical to one another. The fastest way to create inbred strains is the continued mating of close relatives. In plants inbred strains can be achieved through a process of self-pollination, generally referred to as selfing. However, in laboratory animals, inbred strains are either achieved through a series of backcrossing, mating offspring back to their parent (or genetic equivalent), or through sibling matings. In mice, to create a new inbred strain from two outbred strains, repeated brother-sister matings are made for several generations to achieve fixation. A classic rule-of-thumb is that at least 20 generations are necessary to reach homozygosity for nearly all genetic loci[22]. These inbred strains are said to be isogenic because all individuals are genetically identical.

Since isogenic animals have fixed genomes, they are easily reproducible. Mating together two isogenic animals will always produce another isogenic animal, which enables reproducible studies and the integration of data over both time and space. The only way an inbred strain can change genetically is as a result of new mutations, which are relatively rare. Another advantage of inbred strains is that many can serve as disease models due to their lack of buffering alleles, which makes them susceptible to cancer, diabetes, obesity and other diseases. Therefore, inbred strains can be studied as models of these conditions.

1.3 Combining Inbreeding and Selection

Combinations of inbreeding (fixing the genome) and selection systems (selecting for a particular gene or trait) give geneticists a wide variety of methods for controlling the inherited characteristics of research animals. One particular breeding scheme variation is called a congenic strain, in which two inbred strains are bred together and then offspring of this mating are backcrossed to one of the original inbred strains (called the recipient strain). Typically offspring are selected based on the presence of a particular phenotype or genotype from the other inbred strain (donor strain), and backcrossed to the recipient strain for about 5-10 generations to achieve an inbred strain. A more specialized type of congenic strain is called a consomic, in which an entire chromosome is retained from the donor strain. Early on most inbred strains and congenics were selected based on the presence of particular phenotypes. However, as technology advanced, genotyping became useful in developing congenics and consomics to check for the existence of the donor strain DNA in the specified regions. This genotyping was done on a small scale, usually with a few loci chosen at points of interest.

Through advances in technology, it now costs about the same amount to get full-genome genotypes as the small scale genotyping used to cost. Full-genome genotyping allows us to have better control over the final inbred strains. In congenics, this means less donor strain in the genomic regions outside the chromosomal fragment of interest. By utilizing full-genome genotypes, marker-assisted or speed congenics were created, which only require about 5 generations of backcrossing. This is achieved by selecting offspring at each generation that not only retain the desired chromosomal fragment, but that also have a minimal amount of background genetic information from the donor strain in the other genomic regions. In the creation of new inbred strains, genotyping allows us to more closely track the amount of residual heterozygosity (loci that are still segregating).

1.4 Thesis Statement

Through monitoring of genome-wide genotypes over multiple generations, one can engineer user-specified genomic structures. This can be made more efficient, in terms of the number of generations, with accurate computational models. These computational models will lead to new breeding techniques, better breeder selection, and techniques for monitoring genomic structure.

To demonstrate the validity of my thesis statement I conducted a number of experiments and designed tools to enable the results of those experiments to be conducted on a live mouse population. The first experiment involved the design of a simulator to accurately model the breeding process of a mouse, the model organism of choice for these experiments. After validating the recombination model of the simulator, I ran a series of tests to determine the best breeding strategies to create user-specified genomic structures as described in Chapter 3 of this thesis. These accurate computational models created new breeding techniques as well as better techniques for breeder selection in the end-goal of creating a panel of inbred strains. To utilize these computational models on a live mouse population, techniques for monitoring the genomic structure of the population were developed, as described in Chapters 4 and 5 of this thesis. Chapter 6 discusses the process of validating the results from Chapters 4 and 5.

1.5 Organization

The rest of this thesis is organized as follows:

- **Chapter 2** presents biological background such as the basics of genetics and a description of the breeding population used throughout the studies in this thesis.
- **Chapter 3** presents a computational simulation model used to study various breeding schemes and examine techniques for speeding up the process of inbreeding through the use of marker-assisted techniques.
- **Chapter 4** presents the design and development of low-cost, low-density genotyping arrays for monitoring the genomes of a breeding population and includes analysis of the performance of these arrays.
- **Chapter 5** presents the application of our theoretical results from Chapters 3-4.
- **Chapter 6** presents a method for determining the underlying genomic structure of a breeding population using high-throughput sequencing data and compares those results to similar results obtained through the use of the genotyping platforms discussed in Chapter 4 and the results discusssed in Chapter 5.
- **Chapter 7** concludes with the major results of this thesis and discusses problem areas for future research.

CHAPTER2: BACKGROUND

2.1 Genomic Data

A gene is a molecular unit of heredity of a living organism. Genes hold the information to build and maintain an organism and pass its traits to its offspring. All organisms have genes corresponding to various biological traits, some of which are instantly visible, such as eye color and coat colors, and some of which are not, such as blood type, increased risk for specific diseases, or the thousands of basic biochemical processes that comprise life.

An organism's genome is comprised of the complete set of all its genes as well as the intergenic regions. Generally members of a species have a common set of genes and every cell within the organism carries the same genome. Each gene is a segment of deoxyribonucleic acid (DNA) and the genes are joined together to make up a set of very long DNA molecules called chromosomes. In diploid organisms like humans and mouse, there are two copies of each chromosome. One copy is inherited from each parent.

DNA is comprised of a sequence of nucleotides and the four primary DNA bases found in nucleotides are Adenine(A), Cytosine(C), Guanine(G), and Thymine(T). Each base binds with another specific base (T with A and C with G). A DNA molecule is comprised of a primary sequence and a "complementary" copy that allows it to self replicate, as each acts like a template for the other sequence.

Among humans, 99% of our DNA is identical. Individuals vary because although they all have the same set of genes, they have subtle sequence variants or alleles. An allele is a specific version of a gene, meaning the actual DNA sequence that forms a particular gene. Genetic variations are known as polymorphisms. The most common DNA variant is a substitution in a single base or a single nucleotide polymorphism (SNP).

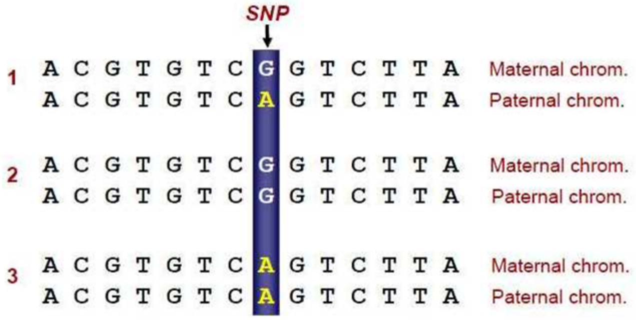

SNP
1
A C G T G T C G G T C T T A
A C G T G T C A G T C T T A
Maternal chrom.
Paternal chrom.
2
A C G T G T C G G T C T T A
A C G T G T C G G T C T T A
Maternal chrom.
Paternal chrom.
3
A C G T G T C A G T C T T A
A C G T G T C A G T C T T A
Maternal chrom.
Paternal chrom.

Figure 2.1: Depiction of a SNP. This image shows a subset of DNA from three different organisms. While the majority of base pairs for these three organisms are identical, at the denoted SNP there are three possible combinations of G/A alleles.

2.1.1 SNPs

A SNP is a DNA sequence variation occurring when a single nucleotide differs between two sequences, resulting from a substitution of one nucleotide for another. For example, two sequenced DNA fragments from different individuals, AAGCCTA and AAGCTTA, contain a difference in a single nucleotide. In this case we say that there are two alleles. Almost all common SNPs have only two alleles or are biallelic. Figure 2.1 is depicting a SNP, and shows a subset of the DNA from three different organisms. While the majority of base pairs for these three organisms are identical, at the denoted SNP there are three possible combinations of G/A alleles. To determine the allele at a SNP location for a particular sample, genotyping needs to be done. In genome-wide studies, often a DNA microarray is used for this purpose.

2.1.2 Microarrays

A microarray is a collection of DNA sequence probes attached to a solid surface and is typically used to capture a complementary DNA sequence. When exposed to extracted DNA from lysed cells it is referred to as a genotyping array or a SNP array and is used to probe multiple regions of a genome at the same time. Some of the research described in this thesis relies heavily on the use of SNP arrays, and I describe in Chapter 4 the design techniques used to build two genome-wide SNP arrays. In a popular model organism like the mouse, designing an array consists of selecting a subset of known and reliable SNPs at specified genomic locations that segregate between strains of interest.

A number of companies including Affymetrix and Illumina offer the ability to design custom arrays. While both companies use slightly different technologies, the basic idea is the same. The customer selects known SNPs and reports the surrounding DNA sequence for that SNP. Since a complementary strand of DNA will be created for each of the chosen SNP locations, it is important that the area surrounding the SNP has a unique DNA sequence. A complementary strand of DNA is then designed for each of the SNPs and placed on the microarray. A sample's DNA is then washed over the plate and the complementary DNA strands will hybridize or "stick" to the sample DNA in the correct locations. Fluorescently labeled target sequences that bind to a probe sequence generate a signal and the total strength of the signal depends upon the amount of target sample binding to the probes present on that spot. Microarrays use relative quantitation in which the intensity of a feature is compared to the intensity of the same feature under a different condition, and the identity of the feature is known by its position. This intensity ratio is then normalized and used to determine the genotype (A, T, C, or G) for a particular SNP location.

2.1.3 High-throughput sequencing

An alternative way to procur genome-wide genotypes is to use high-throughput sequencing (HTS). While microarray genotyping samples every strain at the same loci, DNA sequencing constructs a genomic sequence based on random fragments of DNA that get spliced together. High-throughput sequencing speeds up this process by parallelizing the sequencing process and producing a large number of sequences at once. Categorizing genetic differences in HTS data requires a database of known sequence variants, while microarray-based genotyping is based on a set of reliable variants that were selected previously as part of the array design.

Before a sample is sequenced, its DNA is replicated and cleaved into small pieces (about 50-1000 base pairs). Each of these DNA pieces is then run through the sequencer and the nucleotides are determined. Once the nucleotide sequences are determined, it is necessary to assemble the short reads in their correct order to determine the whole genome DNA sequence. This can be done using de novo alignment (meaning the DNA sequence is not previously known) or the reads can be aligned to a known reference genome. The coverage level of the HTS data is determined by the number of times the DNA was replicated before being sequenced. The coverage refers to the average number of reads that should pileup at each genomic position. Areas of the genome that are highly repetitive often have very high pileups. Since HTS samples the genome randomly, occasionally no reads will align to particular areas of the genome making it difficult to resolve the genotypes for certain regions. However, once HTS data has been properly aligned, it can be used to ascertain genotypes, determine recombination event locations, and accurately infer ancestry as I show in Chapter 6.

2.2 Recombination and Breeding

Genetic recombination is an important process that occurs during reproduction and can produce new combinations of alleles. Most recombination occurs naturally either during meiosis (sexual reproduction) or mitosis (asexual reproduction). During meiosis, genetic recombination involves the pairing of homologous chromosomes (the set of one maternal chromosome and one paternal chromosome that match up with each other). These homologous chromosomes have the same set of genes in the same locations or loci. These loci provide points along the chromosome which enable a pair of chromosomes to align correctly with each other before separating during meiosis. Meiosis creates genetic diversity through two processes, the independent segregation of chromosomes or recombination. Recombination occurs when either the pairs of homologous chromosomes randomly segregate into two different daughter cells or by cross-over events where homologous chromosomes exchange lengths of their genetic material. While recombinations can also occur during mitosis, this process does not create new allele combinations except by mutation. Moreover, recombinations during mitosis do not impact future generations (i.e. the germ line).

The main impact of meiosis is the creation of genetic diversity by allowing offspring to inherit different allele combinations from their parents. Organized breeding schemes are utilized in biological experiments to create desired allele combinations in offspring. Often the desired result are isogenics or inbred strains, which can be replicated indefinitely as both parents will have identical copies of all chromosomes (except for X and Y). When a panel of such animals are derived from the same founder strains, this is referred to as a genetic reference population.

2.2.1 Genetic Reference Populations

Genetic reference populations (GRPs) are defined as sets of individuals with fixed and known genomes that can be replicated indefinitely. Typically they consist of dozens to hundreds of inbred lines derived from a set of common ancestors (founders). GRPs have been developed for many organisms, including yeast, plants, flies, and mammals [4, 18, 11, 3, 31, 19]. GRPs are popular for the study of complex traits and biological systems in both medical and life science applications because genotyping is required only once (described as the "genotype once, phenotype many times" paradigm); replicate individuals can be produced with the same genotype allowing for optimal case/control and gene-by-environment designs, and custom analysis tools[59, 13, 29]. GRPs are also attractive because the phenotypic, genetic, and genomic data associated with each line can be integrated across labs, experiments, and time.

Most mouse GRPs are collections of inbred lines derived from pairs of inbred strains. In mice, these include panels of chromosome substitutions strains (i.e., consomics), recombinant inbred lines (RIL), and subcongenics [4, 56, 28, 21, 39]. Alternative GRPs include panels of extant inbred lines with complex population structures and nonuniform genetic relationships among the lines, such as the Laboratory Strain Diversity Panel derived from the Mouse Phenome

Project [41] and combinations of diversity panels and pairwise panels [6]. Key parameters that determine the usefulness of GRPs for the analysis of complex traits are the number of lines; the density, distribution, and functional significance of the genetic variation present in the GRP; the number and distribution of unique recombination sites; the presence of population structure; and the level of inbreeding and genetic drift.

2.2.2 Collaborative Cross

An example of a mouse GRP is the Collaborative Cross (CC)[17]. The CC is a multi-parental recombinant inbred panel derived from a set of eight genetically diverse inbred laboratory mouse strains. The set of founders consists of five classical inbred strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HlLtJ) and three wild-derived inbred strains which were selected to represent three Mus musculus subspecies(CAST/EiJ, PWK/PhJ, WSB/EiJ). They were chosen to capture a high level of genetic diversity, representing on average 90% of known genetic variation in laboratory stocks across all 1-Mb intervals[49].

The CC lines were generated via a funnel breeding scheme that combined the eight founder genomes in three intercross generations prior to repeated generations of inbreeding through sibling mating (Figure 2.2).The eight founder strains capture a much greater level of genetic diversity than existing RIL panels or other extant mouse GRPs, and the genetic variants are more uniformly distributed across the genome than in other GRPs [49, 30, 63, 65]. As seen in Figure 2.2, the eight CC founders have been assigned letters A-H as well as a color. These letter and color codes are used consistently throughout all publications about the CC, including this thesis, and are shown in Table 2.1.

In 2008, there were 650 CC lines in production, with over 200 of the lines past seven generations of inbreeding[14]. However, the extinction rate in early generations of the CC population was greater than 50% [14], possibly because of the presence of incompatible combinations of alleles originating in different subspecies, and as a result, as of 2012, there existed 350 unique

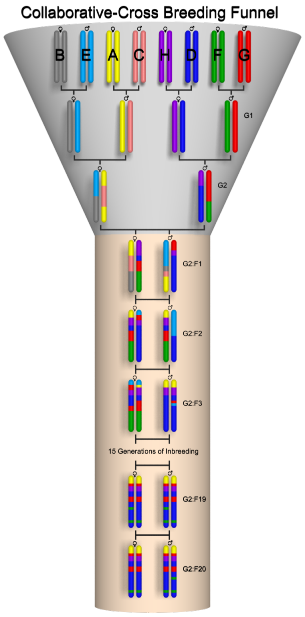

Collaborative-Cross Breeding Funnel
B
E
A
C
H
D
F
G
G1
G2
G2:F1
G2:F2
G2:F3
15 Generations of Inbreeding
G2:F19
G2:F20

Figure 2.2: Collaborative Cross breeding scheme. Each independent CC strain begins with a funnel breeding stage that mixes eight founders, which are crossed for two generations, G1 and G2. The lines are then inbred for at least 20 generations to obtain recombinant inbred lines. CC lines are regularly genotyped after their 6th generation of inbreeding to monitor their residual heterozygosity, detect breeding errors, and to accelerate the inbreeding of selected lines.

CC lines[17]. Of those 350, about 70 of these lines are currently considered completed and are available for distribution. As reported[17], there is little long-range linkage disequilibrium in the CC population and the recombinations are independent.

2.2.3 Diversity Outbred

Another recently developed mouse population is the Diversity Outbred (DO)[55]. Unlike the CC, the DO is not an inbred population, but instead crosses are randomized between non-related individuals to create an outbred population. Animals from CC lines at early stages of inbreeding were used to establish the DO population, which is maintained by a randomized outbreeding strategy (see Figure 2.3). The founders of the DO, 144 partially inbred CC lines taken from the CC breeding colony at Oak Ridge National Laboratory[14], were at generations ranging from F4 to F12 of inbreeding. This allowed for the capture of recombination events that

CC Founder	Letter	Color
A/J	A	Yellow
C57BL/6J	B	Black
129S1/SvImJ	C	Pink
NOD/ShiLtJ	D	Dark Blue
NZO/HlLtJ	E	Light Blue
CAST/EiJ	F	Green
PWK/PhJ	G	Red
WSB/EiJ	H	Purple

Table 2.1: Letter and color codes for CC Founders.

occurred in the early generations of CC breeding to effectively jump-start recombination density in the DO population.

Since the DO and CC populations are derived from a common set of eight founder strains, they both capture, in theory, the same set of alleles. However, while each CC inbred strain represents a fixed and reproducible genotype, each DO animal is a unique individual with one of an effectively limitless combination of the segregating alleles. This makes the DO an ideal resource for high-resolution genetic mapping.

2.3 Related Work

Recombinant inbred lines (RILs), first developed in 1971 [56, 4], have long been an important resource for genetics. Typically, RILs are derived by crossing two inbred strains followed by repeated generations of selfing or sibling mating to produce an inbred line whose genome is a mosaic of its parental lines. More recently, panels of multiway RILs have been developed that combine the genomes of multiple founder lines via an initial mixing stage followed by successive generations of inbreeding. Examples include mouse [57, 16, 14, 17], maize [11], Drosophila melanogaster [27], and Arabidopsis thaliana [44, 31, 26]. For all species, inbreeding via either selfing or sibling mating is the primary process used for fixing the genetic background. RILs derived by sibling matings from two parental backgrounds require multiple generations to fix their genome as homozygous, and the number of generations depends on the diploid number. In

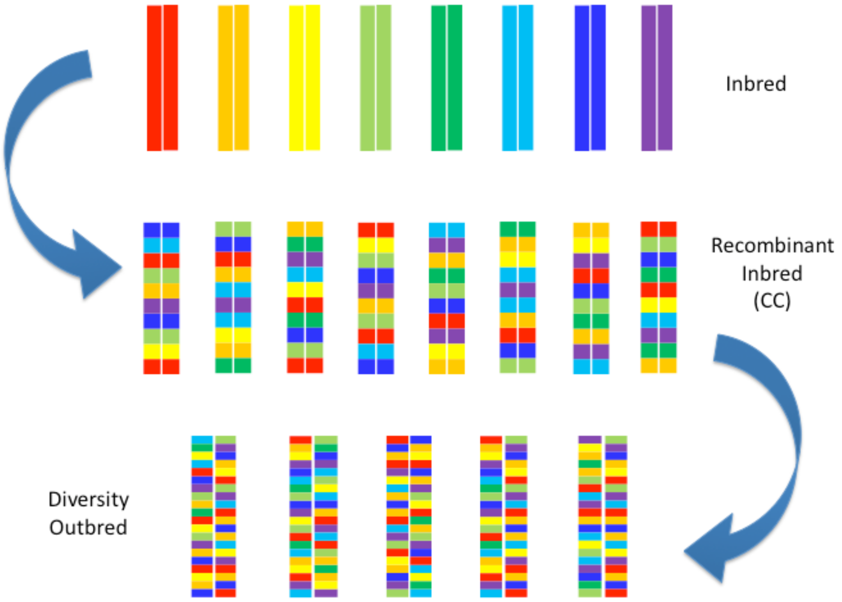


Figure 2.3: Diversity Outbred (DO) breeding explanation[55].

mice, this requires, on average, a minimum of 20 generations [22] and assuming an average of four generations per year, it takes a minimum of 5 years to create a new RIL. Moreover, a large fraction of the started RILs fail, presumably as the result of genetic incompatibilities affecting survival and reproduction [53].

Many recent efforts to generate RILs have focused on multiway crosses where more than two parental lines are initially mixed before inbreeding. In 2005, Broman [8] ran simulations to determine the average number of generations required for two-way and eight-way RILs to reach 99% fixation and complete fixation. He also tracked the number of segments generated through recombination in inbred lines and used it as a comparison between the genetic diversity of two-way and eight-way sib-mating RILs.

Marker-assisted breeding techniques have been used to fix a selected haplotype interval against a fixed background in congenic strains [37]. In mouse, marker-assisted speed congenics have demonstrated a reduction in the number of generations of backcrossing from 10 generations to five. This reduction was achieved by selecting the progeny with the lowest residual heterozygous fraction to cross back to the background strain. These selection criteria have evolved overtime, as technology has allowed for more rapid and specific genotyping [23].

By monitoring genome-wide genotypes over multiple generations, one can engineer user-specified genomic structures. I will show in Chapter 3 how this can be made more efficient, in terms of the number of generations, with accurate computational models, by using a simulator to test different breeding strategies as well as different breeder selection metrics.

CHAPTER3: MARKER-ASSISTED INBREEDING THEORETICAL ANALYSIS

Broman [8] showed through simulation that eight-way RILs take on average 26.7 generations of sib-mating to reach 99% fixation, and 38.9 generations, on average, to reach complete fixation. Although a major source of genetic variation in a RIL is derived from the choice of founder strains, I focus on the additional genetic variations introduced by mixing of allele combinations via recombinations between founder genomes. This is the primary source of genetic variation between RILs. Therefore, the number of distinct founder segments, defined as the regions between recombination breakpoints on the RIL chromosomes, can be used as a measure of genetic diversity. From now on, I refer to these distinct founder segments simply as segments.

Recombinations in early generations increase diversity, but eventually diversity peaks and the process of inbreeding leads to a loss of segments. To verify this, I simulated 100,000 eight-way RILs and tracked the number of segments in each line at every generation until the simulated lines reached complete fixation. In an eight-way cross, the peak in diversity is reached at the seventh generation of inbreeding on average and before 10 generations of inbreeding for 75% of line starts (Figure 3.1). Therefore, I will consider 10 generations of inbreeding as past the point of peak diversity. If inbreeding acceleration is started before this peak is reached, the resulting inbred lines are likely to see a reduction in the number of segments. Therefore, unless otherwise specified, I use traditional methods for constructing RILs in the first 10 generations, after which I apply various methods for accelerating the inbreeding process.

Just as marker-assisted techniques have been used to improve mapping resolution in self-pollinated species [7] and have been adapted for consomics [1], I adapt them for multiparental RILs. Rather than attempt to fix one specific genomic region or one complete chromosome, my goal is to achieve complete fixation of the genome in fewer generations than random sib-matings,

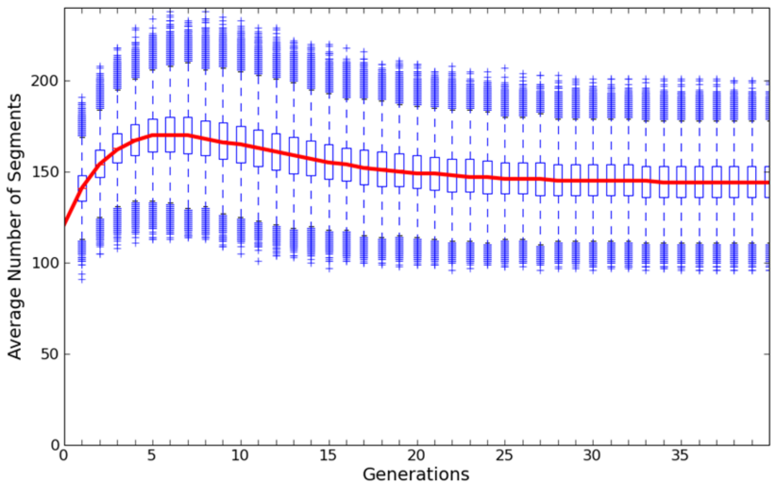


Figure 3.1: The average number of founder segments in eight-way RILs at various generations of inbreeding. This figure is based on 100,000 simulations, and the number of segments was tracked until they reached complete fixation. The average peak in the number of segments occurs at generation 7 and before generation 10 for 75% of all lines. Therefore, I consider generation 10 to be past the point of peak diversity.

without substantially impacting the overall genetic architecture of the inbred lines.

In this chapter, I address accelerating the inbreeding process of RIL creation by using a combination of alternative breeding strategies and marker-assisted inbreeding (MAI) techniques.

3.1 Approach

I developed a simulator that represents a genome as a collection of intervals whose boundaries can be resolved at the resolution of a base pair rather than a string of alleles as is common in many breeding simulators [8, 58]. The interval representation has the advantage of implicitly representing every base pair in the genome while explicitly tracking every recombination. This approach provides a conservative estimate of homozygosity because it treats every founder sequence as a separate genotype without taking into account regions of sequence identity among founders. Moreover, my interval model can be trivially converted to a string of alleles representation if given the founder sequences or markers from any platform.

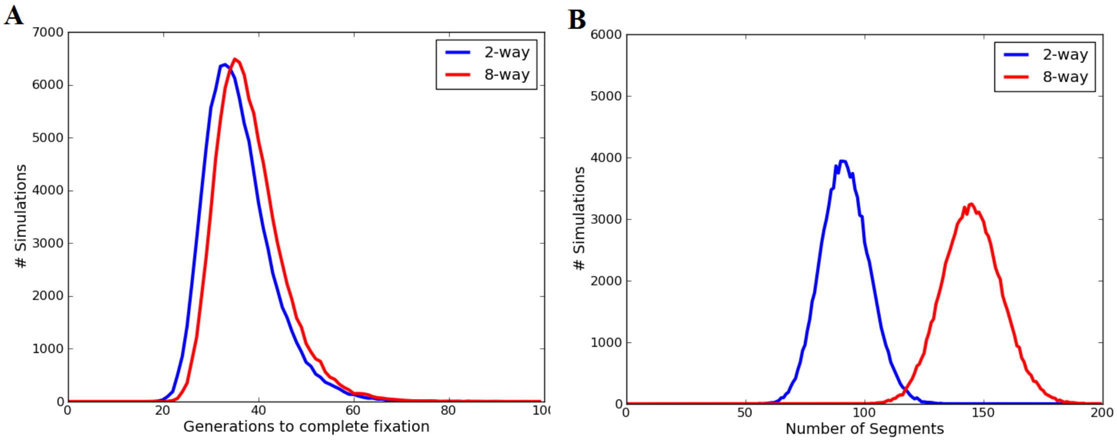


Figure 3.2: The number of generations to complete fixation (A) and the number of resulting founder segments (B) in two-way and eight-way RILs. On average, two-way RILs take 35.92 generations to reach complete fixation and have 91.95 segments. Eight-way RILs take 38.21 generations and have 145.12 segments on average. These figures are based on 100,000 simulations and are consistent with previous simulations [8].

Despite the differences in the underlying representation, my simulator produces results nearly indistinguishable from those presented by Broman [8]. Figure 3.2 shows the distribution of the number of generations to complete fixation and number of segments for both the two-way and eight-way sib-mating RILs based on the simulation of 100,000 RILs. For a randomized eight-way RIL my simulations show that it takes an average of 38.21 ± 7.1 (SD) generations of sib-matings to reach complete fixation. The genomes of the resultant inbred lines have an average of 145.1 ± 12.48 segments in their mosaic structure. Furthermore, 25.72 ± 3.16 generations of sib-mating on average are needed to reach 99% fixation. These baseline metrics are used for comparison against my accelerated inbreeding simulations. My analysis is based on an initial funnel-breeding scheme like that used in the eight-way Collaborative Cross (CC) [16, 17], where the mixing of eight inbred lines occurs in three initial crossing stages, followed by successive generations of sib-matings until the line becomes fully inbred.

I introduce a notion of joint heterozygosity (JH) to express four possible states between the homologous alleles of a potential breeding pair. Figure 3.3 shows two homologous chromosomes

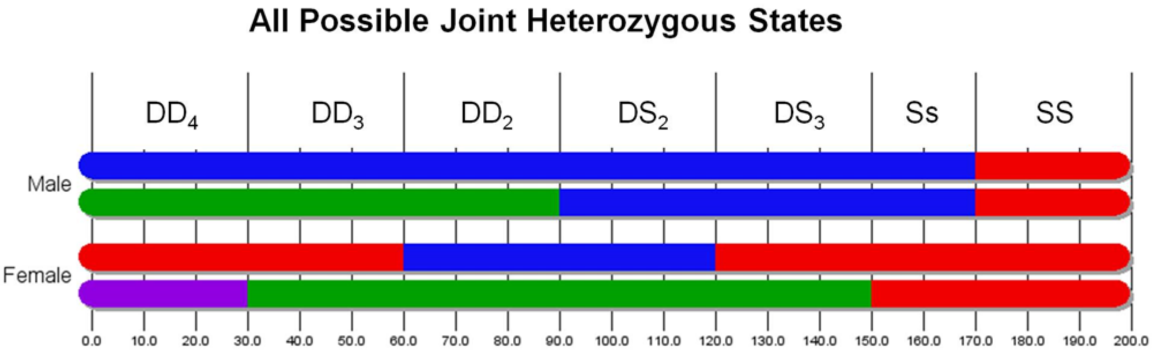


Figure 3.3: This image shows all possible JH states between a potential mating-pair and illustrates my notion of a genomic segment. DD stands for different-different and occurs in three variations. DD_4 occurs when both breeders are heterozygous and do not share any founder alleles among them. DD_3 occurs when both breeders are heterozygous and share one founder allele, whereas DD_2 refers to both breeders being heterozygous for the same two founder alleles. DS stands for different-same and occurs in two variations. DS_3 occurs when the heterozygous gene shares no founder alleles with the homozygous allele of its mate. DS_2 refers to when the heterozygous gene shares one founder allele with its mate. Ss is opposite same, where the male is homozygous for one founder allele and the female is homozygous for another allele. The final state, SS (same-same), is achieved when both male and female are homozygous for the same founder allele. All JH segments are depicted with a chromosome fraction of 0.15, except for Ss, with 0.10.

from each parent of a potential breeding pair and depicts all possible JH states. The inbred state is achieved when both male and female samples are homozygous for the same founder state. I call this state same-same (SS). Another possible state involves a breeding pair that is heterozygous with alleles from two founders while the mate is homozygous. I call this different-same (DS). This state occurs in two forms, DS_2 when the heterozygous gene shares a founder allele with the homozygous allele of its mate, and DS_3, when the heterozygous gene shares no founder alleles with its mate. The third state is opposite-same (Ss), where the male is homozygous for one founder and the female is homozygous for another. The final state is different-different (DD), where both male and female are heterozygous. This state comes in three variations, involving, two, three, and four founders, respectively. The two-founder state, called DD_2, occurs when both male and female are heterozygous between the same founder alleles. DD_3 refers to when both male and female are heterozygous but share one common founder allele. DD_4 occurs when the male and female are heterozygous and do not share any founder alleles. Figure 3.4 shows a state diagram with these four states and their forms depicting all possible transitions between them in a single generation. The directed edge weights represent the probability of transitioning between JH states. A similar transition matrix, which uses thirteen states instead of my seven, has also been derived by Broman [9]. It is a simple matter to extend our JH model to two generations by finding every path of length two within the graph and inserting an edge with weight equal to the product of the two edges along its path. The weights of edges from a common source to a common destination, but passing through different intermediate states, can be added and combined into a single edge. This approach can be extended to n generations, and as n increases all of the heaviest edges eventually lead to the inbred (SS) state. For analytical expressions for extending our JH model for n generations, see [9, 17].

In early generations the CC lines include genomic intervals in JH states involving three or more founders (DD_3, DD_4, DS_3), but in later generations these intervals eventually transition to states with two or less founders (DD_2, DS_2, SS, and Ss; Figure 3.5). To determine this trend,

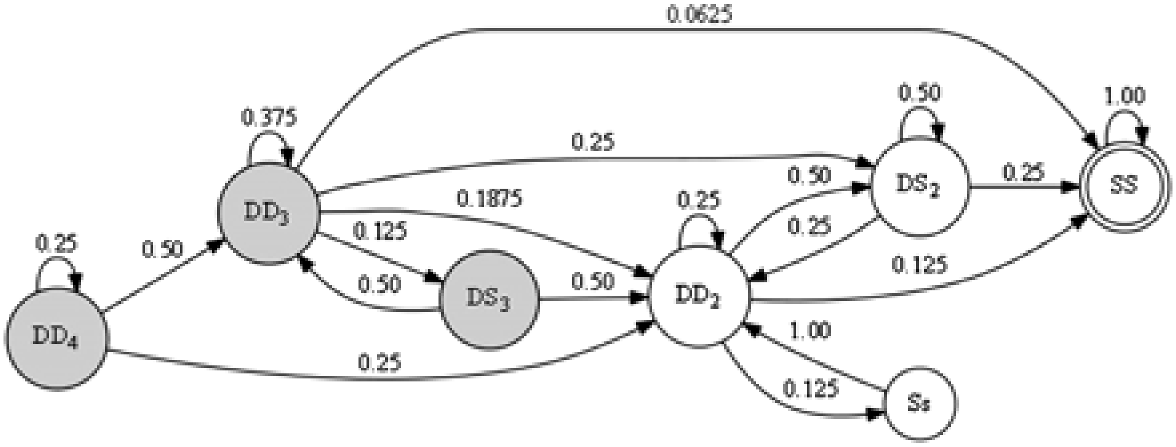


Figure 3.4: A state diagram showing the transitions between all JH states in a single generation. The directed edges are labeled with the transition probability. The grayed-out nodes represent transient states; once a segment moves away from these three states, there are no returning edges. Transient states tend to go away after a few generations and are rarely seen past the point of peak diversity (as shown in Figure 3.5). CC lines begin inbreeding in one of the states, DD_4, DD_3,, and DD_2. The desired inbred state for all intervals is SS. DS_2 is the most likely to become SS. DD_2 is the next most likely state to become fixed. It takes at least two generations to transit from Ss to SS, as there is no direct path between these two states.

I simulated 100,000 eight-way crosses and tracked the JH states between breeder pairs at each generation, as shown in Figure 3.5. By generation 10 (after the point of peak diversity), all segments have contributions from two or fewer founders. DD_4, DD_3, and DS_3 are transient states (see Figure 3.4), meaning that once this group of three states is left, there are no returning edges. In two-way RILs, the three transient states do not occur because there are at most two founders present. When selfing, the model further reduces to only two JH states, DD_2 and SS. The transition probabilities to reach the inbred state are incorporated into my metric for selecting the best mating pair at each generation, which is discussed later in this section.

Using the notion of JH state, I split the genome into intervals according to state and track the genomic fraction of each type. I combine these fractions to arrive at several useful measures. Adding the genomic fraction of all regions in the same-same state (SS) gives the fixed genomic fraction (FGF). I call the complement of this, or 1-FGF, the mating pair's combined heterozygous fraction (CHF). FGF and CHF can be used to assess how inbred a line is, such that FGF = 1 refers

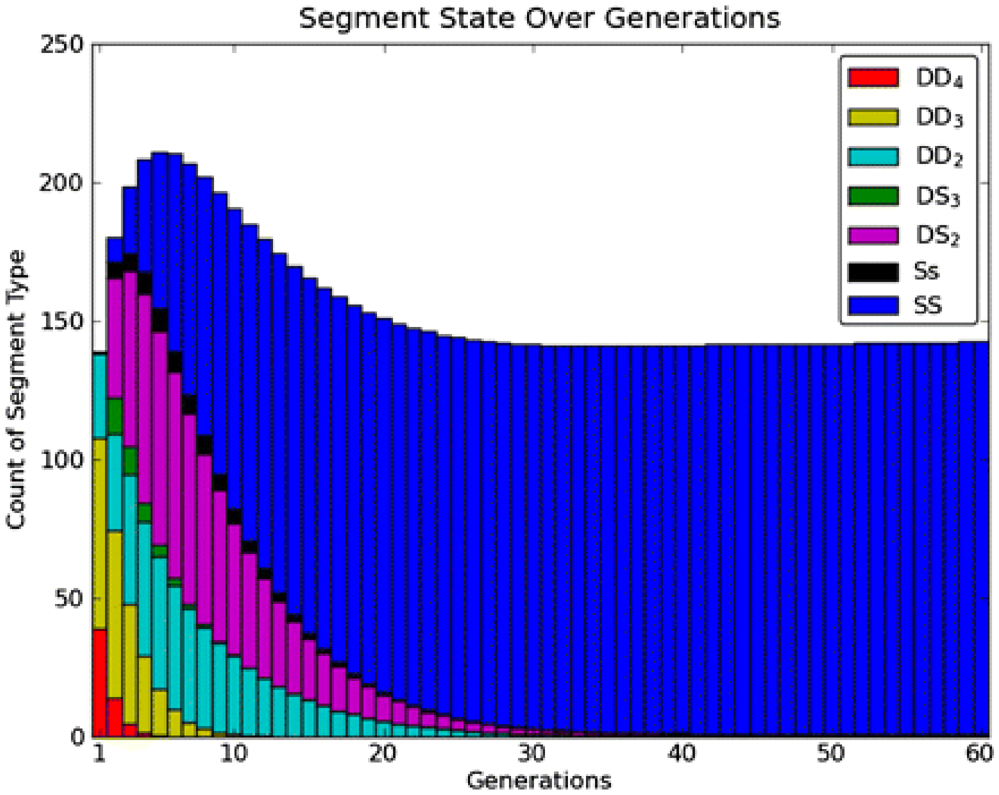


Figure 3.5: A histogram of segments colored according to their JH state as a function of generation. In early generations, most segments have contributions from three or more founders, but by generation 10 (after the point of peak diversity), segments have contributions from two or fewer founders. This plot was created by tracking the JH states between breeder pairs and finding the average contribution of each state over 100,000 simulations.

to fully inbred.

I choose the "best" breeding pair, by considering a weighted genomic mix of the JH types of all candidate mating pairs. The best pair is selected as the maximum of a weighted combination of transition probabilities for all JH segments of a given mating pair considering all chromosomes. For each distinct JH segment of a chromosome the probability that it will become inbred in the next generation (i.e., the weight of the edge from the current JH state to the SS state) is multiplied by the chromosome fraction of the segment, and the sum is accumulated over all segments on the chromosome. This calculation results in a chromosome score ranging from 0, when the entire chromosome is Ss, DD_3, DD_4, or DS_3, to 1 when the entire chromosome is SS. This approximation ignores the relative ordering of segments, and, therefore, does not consider linkage. The individual chromosome scores are then multiplied together, modeling their independent segregation, to arrive at the total pair score. Therefore, I assign a score for a given mating pair as:

$$Score(n,m) = \prod_{i=1}^{N} \sum_{JHSeq_{n,m} \in Chr_i} p(JHSeq_{n,m} \rightarrow SS) \frac{\|JHSeq_{n,m}\|}{\|Chr_i\|} \tag{3.1}$$

This score is an approximation of the actual likelihood that the entire genome will become inbred in the next generation. I refer to this score as the weighted state metric (WSM). $JHSeq_{n,m}$ represents a JH segment on the specified chromosome i induced by the pairing n,m, and the best pair is the maximum of this score over all possible pairs n,m. In self-pollinated species, my score simplifies to a scaled version of the FGF because the only relevant states are DD_2 and SS, which has been described previously [7].

3.2 Experiments and Results

I explored two marker-assisted breeding schemes. The first of these is MAI, which modifies the breeding scheme only after the point of peak diversity is reached. Once the peak is reached, the WSM discussed previously is applied to choose the best breeding pairs. The second is a marker-assisted advanced intercross, which modifies the breeding scheme to choose sib-pairs to increase segments until either a specified generation or a desired number of segments is reached; it then reverts to choosing sib-pairs to accelerate inbreeding. Through simulations, I track the average number of generations to fully inbred and to 99% inbred as well as the average number of segments present in the inbred lines to compare the different breeding schemes.

The simulator is written in Python and runs on a Dell Studio XPS with 8GB RAM, with dual-threaded quad-core processors. It takes approximately 5.5 hours to complete 100,000 simulations of eight-way RILs.

For the purposes of this analysis, the eight-way CC funnel breeding scheme was used, but my simulator also supports the input of any breeding scheme using pedigree files. It has also been used to simulate two-way RILs, F2 crosses, and outbred populations.

To test my MAI methods, I used the developing CC [17] and a low-density genotyping platform I codesigned, referred to as the Mouse Universal Genotyping Array (MUGA)(see Chapter 4). The SNPs on MUGA are uniformly distributed with an average spacing of 325 Kb and a standard deviation of 191 Kb. In an eight-way cross, the genotypes at multiple markers (at a minimum three) are needed to distinguish among the founders. The founder assignments and recombination breakpoints are inferred from the genotypes using a hidden Markov model similar to the ones described by Mott et al. [38], Zhang et al. [47], and Liu et al. [36]. Because multiple markers are needed to distinguish each founder, the effective founder-ascertainment resolution of MUGA is approximately 1 Mb.

3.2.1 Nonmarker-assisted breeding schemes

In simulation, I tested a number of modified breeding schemes in an attempt to accelerate the inbreeding process. These nonmarker-assisted breeding schemes minimally impact the traditional RIL generation process and require no genotyping. I considered several variations of backcrosses. The use of backcrosses was motivated by two main ideas. The first was the analysis of Broman [8], which identified a substantial advantage for selfing when compared to sib-mating. Selfing in two-way plant RILs takes on average 10.5 generations to reach complete fixation, which is a substantial reduction from the 35 generations needed for two-way sib-mating. Since offspring are exactly 50% related to their parents, but only on average 50% related to their siblings, when selecting breeders at random, a backcross guarantees a level of similarity that sib-matings cannnot. A second motivation for using backcrosses is that loss of fertility in the creation of RILs is a major issue. Valuable time can be lost when unproductive sib-matings are set up. Therefore, backcrossing allows the use of known-fertile samples and has been a useful fallback for preserving lines.

The first breeding scheme examined was alternating backcrosses in successive generations, father-daughter in one generation followed by mother-son in the next (Figure 3.6A). This scheme has many practical advantages in that it leverages known-fertile samples. Furthermore, this strategy also serves as a useful fallback for preserving lines. I simulated this approach starting after the point of peak diversity, with a backcross between a father and daughter followed by a backcross between a mother and son in the next generation (each breeder is used in two successive generations, alternating dam and sire). This process was repeated for each subsequent generation until complete fixation was achieved. Alternating backcrosses achieves a reduction in the number of generations to complete fixation with an average number of generations of 33.45 ± 5.88 (Figure 3.7). This represents a reduction of nearly five generations over randomized mating and a substantial reduction in variance. It decreases the number of segments in the resulting inbred lines to 141.21, a loss of about four segments on average. The alternating backcross also reduces the number of generations to 99% fixation to 23.45 ± 3.11, a reduction of two generations.

There are several practical limitations to the alternating backcrossing approach. For instance, female fertility often spans a limited window that might not allow for mother-son backcrossing. Therefore, I also explored, through simulation, a modified breeding scheme involving only father-daughter backcrosses. Starting after the point of peak diversity, a father-daughter

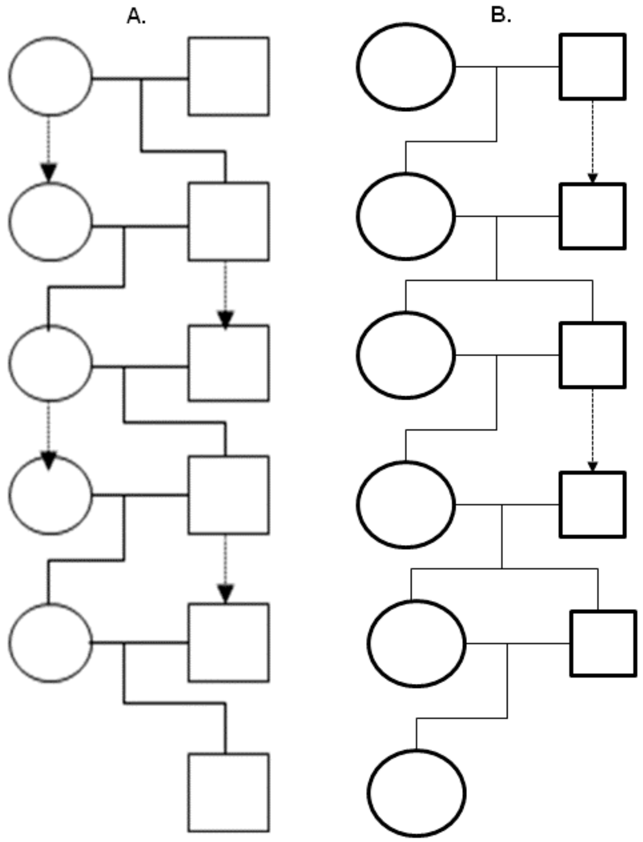


Figure 3.6: This figure shows the pedigree diagrams for the alternating backcrosses: father-daughter backcross with the mother-son backcross (A) and the father-daughter with the random sib-mating (B).

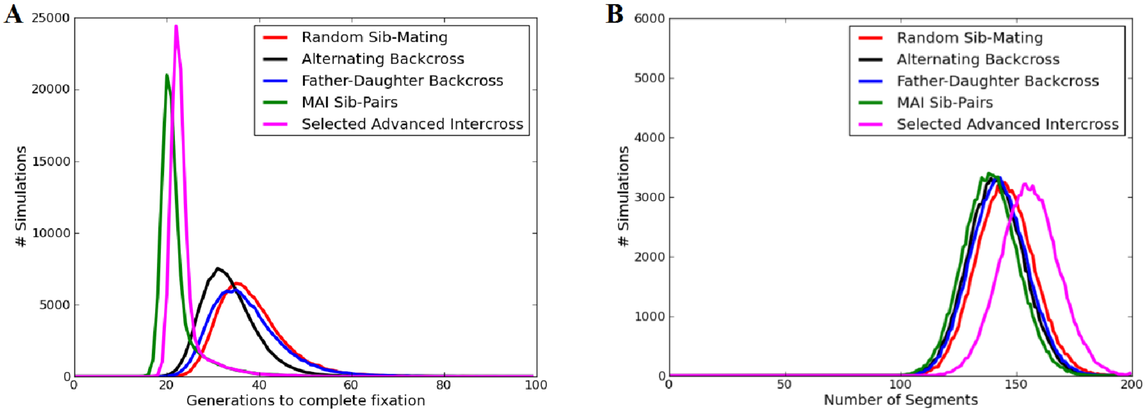


Figure 3.7: A comparison of five breeder selection alternatives for generating an eight-way RIL, showing the number of generations to reach complete fixation (A) and the total number of segments (B) found in the final inbred lines. Random sib-pair mating is used as my baseline. The alternating backcross swaps between father-daughter and mother-son matings in successive generations. The father-daughter scheme alternates between father-daughter and random sibling matings in successive generations. MAI uses my weighted state metric to choose between 16 breeding pairs after the point of peak diversity. The selected advanced intercross modifies early stages of the breeding scheme to choose sib-pairs that maximize diversity, and then at a pre-established generation (10), it reverted to choosing sib-pairs to accelerate the inbreeding process.

backcross is followed in the next generation by a random sib-mating. This breeding scheme (Figure 3.6B) is repeated for each subsequent generation until complete fixation is achieved. The father-daughter backcross takes 37.06 ± 7.55 generations to reach complete fixation, and the inbred lines contain on average 142.39 ± 12.24 segments (Figure 3.7). This breeding scheme also takes 24.70 ± 3.54 generations to 99% fixation. Although the benefits of father-daughter mating are modest relative to random sib-mating, in practice they are compensated for by a reduction in generation time resulting from a mature and known fertile sire.

3.2.2 Marker-assisted inbreeding

The steadily decreasing cost of full-genome genotyping combined with the advantages of considering each sample's individual full genetic makeup motivated the decision to also explore MAI techniques. The ability to compare potential breeding pairs based on full-genome genotypes allows one to choose breeding pairs with the greatest likelihood of producing inbred offspring. The Ss (opposite same) is the least-preferred state in a breeding pair because it has no chance of becoming inbred in the next generation, as shown in Figure 3.4. In contrast, of the noninbred states, DS_2 has the greatest probability of becoming fixed in the next generation, and DD_2 is the next most likely. For all MAI techniques, random sib-matings were simulated until the point of peak diversity was passed. This was followed by subsequent generations of selecting the best breeding pair, until the line reached complete fixation.

Using the WSM, I selected the best pair from sib-pairs, parent-child backcrosses, or a combination of both. To see what other pair relationships were worth considering, I simulated 100,000 lines such that random sib-matings were used for 15 generations, at which time three mating pairs were generated, producing two male and two female offspring each. The best breeding pair was then chosen by comparing every female to every male (both parents and offspring). The pair with the lowest combined heterozygous fraction (CHF) was selected. Sib-pairs were selected 63% of the time, whereas backcrosses were chosen 23% of the time. Cousin-pairs (offspring from different mating pairs of the same generation) were the next most likely, being selected 6.9%. The remaining 7.1% included mating combinations such as aunt-nephew, uncle-niece, or grandparent-grandchild. I concluded that non-sib, non-backcross matings should be used sparingly, except in the case of preserving a line.

Because sib-pairs were most often the best option, I limited subsequent simulations to selecting the best sib-pair and report those statistics in Figure 3.7 and Table 1. For the MAI sib-pairs breeding scheme, random sib-matings were simulated until the point of peak diversity was reached. After this point, four female and four male offspring were simulated (4-4), all pairs were considered, and the best pair was chosen as the breeders. This process was continued until the line reached complete fixation. My model is based on generation number and may require multiple litters to achieve the four females and four males assumed in simulation.

A potential shortcoming of my model is that I report the time to inbred as a function of generations, not the number of litters or calendar time required to produce enough viable offspring. However, I did perform additional simulations assuming smaller litter sizes (two females, two males), and unbalanced sex-ratio (eight total offspring with one to seven females), and compared all three sets of assumptions (4-4, 2-2, 8) to the greedy approach of setting up breeders as soon as any sibling mating pairs are available(Figure 3.8). Each of these forms of MAI was able to considerably reduce the number of generations to achieve inbred status regardless of sex balance or litter size. Moreover, waiting for a sufficiently large breeder-candidate set always outperformed the greedy approach of setting up matings as soon as any pair was available. These tests were done by simulating each of the above breeding schemes 100,000 times and plotting the results in terms of number of generations to complete fixation as well as the number of intervals in the final inbred lines.

Using this MAI breeding scheme, it was found that 99% fixation can be reached in an average of 16.44 ± 1.00 generations, whereas complete fixation can be reached in 22.10 ± 4.41 generations on average. These inbred lines have an average of 138.83 ± 11.83 segments. Figure

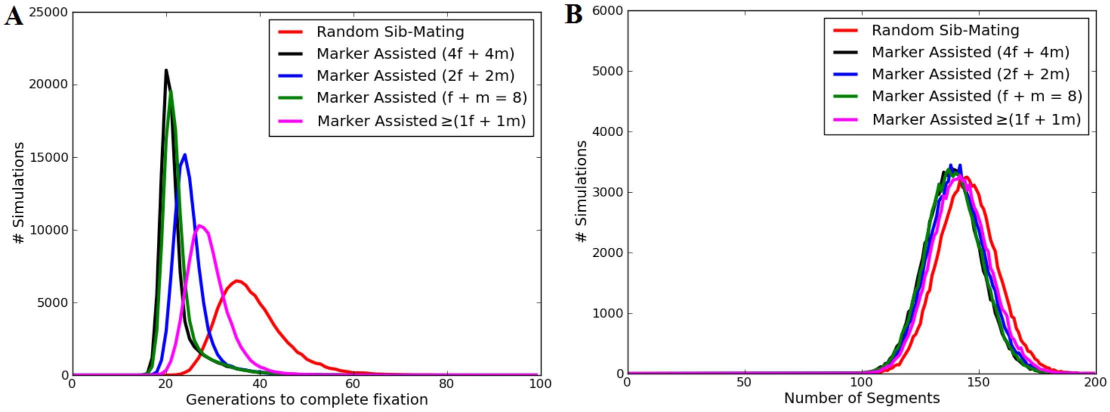


Figure 3.8: Compares the number of generations in takes to achieve complete fixation for 5 breeding schemes that make different assumptions about the available pool of breeders. The standard random sib-mating is provided for comparison. MAI Sib-Pairs assumes a pool of 8 breeders with 4 of each sex and takes an average of 22.10 ± 4.41 generations to reach complete fixation. The MAI Sib-Pairs assumes a pool of 4 (2 each sex) takes an average of 25.05 ± 3.89 generations to reach complete fixation. The MAI Sib-Pairs Unbalanced Sex-Ratio assumes 8 offspring with varying sex-ratios at each generation. These sex-ratios range from 1 female and 7 males to 7 females and 1 male. This breeding scheme requires an average of 22.63 ± 4.25 generations to reach complete fixation. Finally, the Greedy Sib-Pairs (Marker Assisted $\geq$ (1f+1m)) breeding scheme creates small litters of 1-3 offspring and sets up the best breeder pair as soon as at least 1 female and 1 male offspring exist. The Greedy Sib-Pairs breeding selection depicts the natural inclination to set up breeders as soon as possible; however, our simulations indicate that it does not reduce the number of generations required to reach complete fixation as much as waiting until 8 offspring are available for comparison. In fact, it requires an average of 28.97 ± 4.46 generations to reach complete fixation. The overall impact of each breeding scheme on the genetic diversity is negligible.

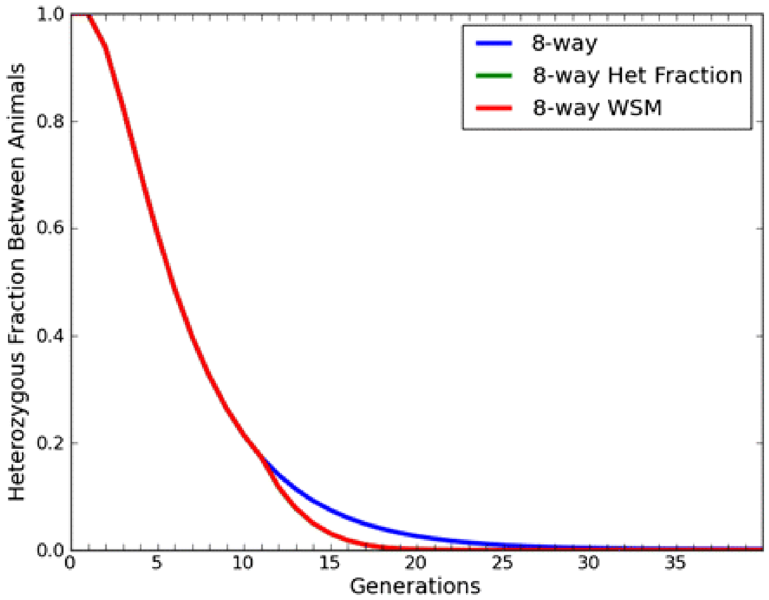


Figure 3.9: CHF as a function of number of generations. This plot shows that MAI reduces the CHF among breeding pairs much faster than random sib-matings. We can see the effect as soon as the breeding scheme is modified (at the point of peak diversity).

3.9 shows that MAI reduces the CHF among mating pairs much faster than random sib-matings. As soon as the breeding scheme is altered at the point of peak diversity, the effect is apparent.

3.2.3 Selected advanced intercrosses

Although MAI achieves a substantial reduction in the number of generations required to fix a RIL, it does so with an average loss of approximately seven segments per line. This result is unfortunate because the number of segments determines the resolution of a RIL panel for quantitative trait mapping [2]. Therefore, I attempted to overcome this loss by using marker-assisted techniques in the first 10 generations of inbreeding to select mating pairs most apt to increase the number of recombination segments. I refer to these lines as selected advanced intercrosses [20] in that they attempt to increase the number of segments on every chromosome by maximizing diversity until a designated generation is reached. This is similar to work done in self-pollinating populations to maximize mapping resolution [7]. After the designated generation, the same MAI techniques as discussed previously are used to select the breeding-pairs until the line is fixed. I found that it took on average 23.5 ± 3.82 generations to become inbred. At the point of peak diversity, the lines had an average of 196.1 ± 15.44 segments, compared with 167 segments in randomized sib-pair matings. The average number of segments in the final inbred animals was 155.6 ± 12.53. On the basis of my analysis, if genotyping is done at every generation, the lines will become inbred in approximately the same number of generations as the MAI breeding strategy but will have approximately 17 more segments per animal. This could lead to increased mapping resolution in the final population.

3.2.4 Low-resolution sampling

In my MAI analysis, I assumed that one is able to accurately assign genomic regions to founders at single base-pair resolution. In reality, genotyping platforms have a limited resolution with which they can ascertain a founder's genomic sequence. This limited resolution creates two main obstacles to the use of MAI methods: the possibility that small recombination intervals might escape detection, and the imprecision with which the cross-over points of recombination can be detected. The impact of both of these limitations can, however, be modeled in a simulation.

I modeled this reduced resolution by sampling the JH state at 1-Mb intervals. I ran 100,000 simulations of breeding using the MAI breeding strategy discussed earlier, but modified the WSM to consider the JH state only at sample points. Furthermore, I declared lines inbred on the basis of the 1-Mb sampling (when all sample points were SS). I then inspected each declared "inbred" mouse to see if, at a base-pair resolution, all intervals were truly fixed, and found them to be actually inbred only 38.3% of the time. On average I missed three nonfixed segments per line, and these segments were on average 327 ± 234 Kb. Figure 3.10 shows a histogram of the sizes of the missed segments, where all segments are less than 1Mb and most of the missed segments are <500Kb. I also found that the lines were considered inbred approximately 2.5

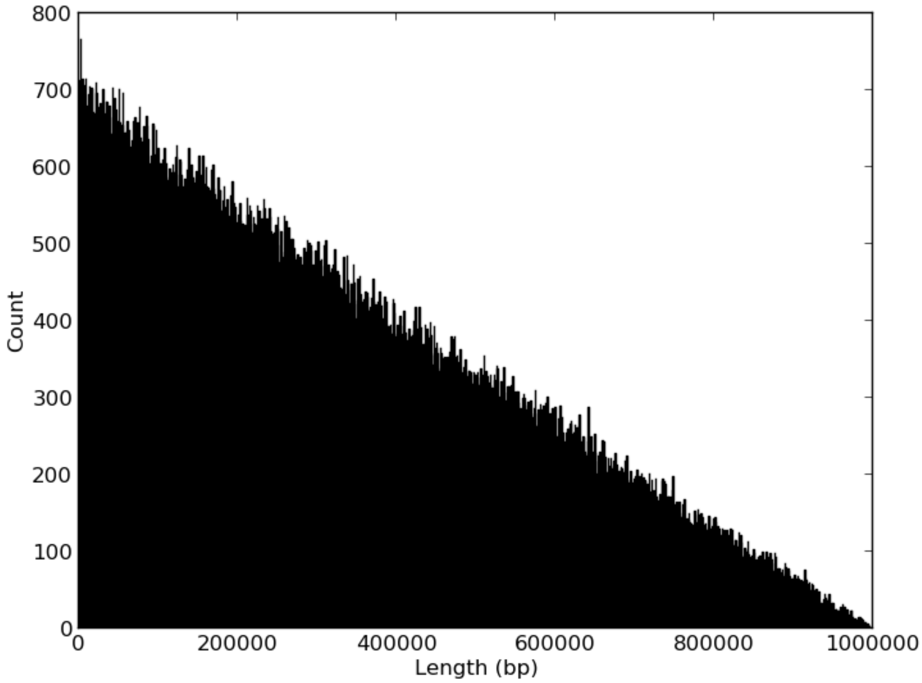


Figure 3.10: Histogram of size of segments missed when conducting low resolution sampling.

generations earlier than MAI with complete observability. This finding implies that the inability to detect small recombinants might require additional inbreeding generations to attain the desired level of fixation.

3.3 Conclusion

Through simulations, I have developed several alternatives to random sib-matings to dramatically accelerate the creation of RILs by as much as 16 generations. These include the judicious use of parental backcrossing and the selection of mating pairs based on genotypes from genome-wide SNPs. Both of these techniques, when applied after the point of peak diversity is reached, result in a negligible reduction in the number of segments. I also propose an advanced intercross variant in which MAI is applied during the early generations to increase the number of haplotype segments for better mapping resolution.

In simulation I also have the luxury of assuming uniform litter sizes and equal sex ratios, but in reality the fecundity of a RIL and the sex-balance of litters are complicating issues. As lines become more inbred, fertility generally decreases [53]. One way to address this is to use backcrosses as discussed previously. However fertility issues might override the choice of “best breeding pair”. To address this problem I calculate backups that, when used, may extend the number of generations required to achieve fixation.

Taking fertility into account and prioritizing for the preservation of the lines, how do I select the final breeders? WSM optimizes for becoming inbred in one generation, but it might be more advantageous in the early MAI generations to select for animals whose probability to become inbred in two or more generations is maximized. However, in simulations, the two-generation metric generally chooses the same breeding pairs as the single-generation model, leading to the same number of generations to achieve fixation. Once lines reach small levels of residual heterozygosity, it might also be advantageous to maintain multiple breeding pairs selected to produce compatible offspring, which are more like sib-pairs than cousin-pairs. This provides more pair options, as well as a chance to compensate for uneven sex ratios or small litter sizes. Although it seems best to choose the optimal breeding pairs early on, finding good pairs near the end-game (in order to fix the last 1%-2% of the genome) is a harder problem. The last few heterozygous regions can take several generations to fix if compatible breeding pairs do not exist. Trying to fix the last 1%-2% of the genome is difficult since it may take 1-2 generations for each residual heterozygous region to become fixed. It is unlikely that two compatible breeders will exist that are able to produce offspring in which each of the remaining regions is fixed.

The simulation software used in this analysis is available for download from http://sourceforge.net/p/breedingsim/. It has been adapted for many uses other than marker assisted inbreeding such as estimating the significance of measured statistics in the developing CC [17].

In the next chapter I will further discuss the genotyping microarrays used in the experimental side of this work, MUGA and MegaMUGA. I will discuss their design principles as well as their performance metrics.

CHAPTER4: DESIGNING MICRO-ARRAYS FOR MAXIMUM INFORMATIVENESS

Genotyping arrays have long been used to characterize the underlying DNA within particular regions of interest in model organisms. More recently, it has become cost-effective to use full-genome genotyping arrays, rather than targeted arrays. These microarrays have the benefit of only needing to be designed once, but can be used for many experiments. When designed properly, these genotyping arrays can be used to distinguish between most population diversity within each area of the genome. However, to make these arrays useful in most experiments they need to be cost-effective and widely-available as well as informative. To be cost-effective, these arrays can only contain a set amount of SNPs based on the cost of the technology at the time of design.

The laboratory mouse is a popular model organism in biomedical research that complements the strengths of many human studies. As a result, a number of these arrays have been designed for use with mouse[64, 14, 52, 34]. However these arrays have either too few markers to be informative genome-wide [14, 52], too many markers to be cost-effective for large experiments[64] or are not widely available [34]. In each case, one of the crucial components for an ideal genotyping platform was missing.

Existing mouse strains exhibit evidence of a population structure which makes them less than ideal.[65] The Collaborative Cross (CC)[17], described in Chapter 2, is an ongoing effort to create a more genetically diverse panel of inbred mouse strains to provide a more useful model for mapping complex genetic traits. In the later generations of inbreeding of the Collaborative Cross, the CC progeny are genotyped to select breeders with the least residual heterozygosity. This genotyping requires an efficient and low-cost platform for determining residual heterozygosity genome wide, as well as the CC founder origin of fixed regions of the genome.

In response to this need for a genotyping platform to use with the CC, two cost-effective, maximally informative, widely available full-genome genotyping arrays were designed. At the time of the original design in 2010, it was determined that cost effective meant a price point of $100/sample, which allowed for the selection of 9,000 SNPs. Two years later when the second generation genotyping array was designed, it was determined that for the same cost, 80,000 SNPs could now be chosen. The first generation genotyping array is called the Mouse Universal Genotyping Array (MUGA) and the second generation array is called MegaMUGA, as it has 10x more SNPs on it than MUGA. Both custom arrays were developed using the Illumina iSelect platform for the Infinium system. In this chapter, I describe the design criteria for each of these two genotyping arrays, as well as the number of samples genotyped on each and the performance of the arrays on these samples.

4.1 MUGA

MUGA was designed to optimize the identification of founder contribution and detection of residual heterozygosity among CC strains at any stage of inbreeding. In particular, probe sequences were chosen to efficiently discriminate between the eight CC founders and their resulting heterozygous combinations. A typical technique for designing informative genotyping arrays is to choose a series of singleton SNPs where only one founder has the minor allele and all other founders have the majority allele. This allows a researcher to quickly determine if an area of the genome was inherited from the founder with the minor allele. You can imagine a series of eight of these singleton SNPs allowing a researcher to determine exactly which CC founder an area of the genome was inherited from. However, if SNPs are selected to instead maximize the entropy on a per SNP basis, the same level of informativeness can be achieved with only three consecutive SNPs (still assuming eight founders). To maximize the per SNP entropy, a biallelic model is assumed for each SNP and SNPs are chosen to maximize the minor allele frequency, which in an eight-way cross amounts to four strains with minority allele and four strains with the majority. Each SNP was then encoded in binary such that the first strain was always 0 and all other strains were either a 0 or a 1 depending if they were the same allele as the first strain (0) or not (1). Since there are eight founder strains, there is always a binary string or SNP diversity pattern (SDP) of exactly eight characters per SNP. This creates $\binom{7}{4}$ or thirty-five unique SDPs (since the first strain is always a 0). In order to divide groups of SNPs into eight unique patterns, at least three SNPs are required. The eight unique haplotypes are then 000, 001, 010, 011, 100, 101, 110, and 111. Sets of three SDPs that create eight unique haplotypes are referred to as compatible triples, and there exist 5040 compatible triples. For any two SDPs that create exactly four equal groups of haplotypes (00,01,10,11), there exist eight unique SDPs that will divide the pair into exactly eight unique haplotypes. This can be shown by noting that the first two SDPs must break into four unique haplotypes, with exactly two of each type of haplotype. Since all binary codes begin with a 0 in this case, the other three binary codes (01, 10, 11) can be divided up in 2 different ways each, creating 2^3 unique SDPs that will create compatible triples with the given beginning two SDPs. By linking together compatible triples, such that the last two SDPs of the previous triple match the first two SDPs of the next triple, one can achieve maximum informativeness in any window of three SNPs genome wide. Figure 4.1 depicts a subset of the compatible triples graph. You can see from this image that the indegree and outdegree of any node in the compatible triples graph is eight.

4.1.1 Database of Available Probes

In order to select maximally informative SNPs in the pattern described above, a large database of SNPs is required. For mouse, there exists a database of 8.27 million SNPs for eighteen common inbred mouse strains that is available from the Wellcome Trust/Sanger Institute[24]. This set of SNPs was further enhanced by using the whole genome sequences of fourteen inbred mouse strains, including the eight CC founder inbred strains. This was done as part of the Mouse Genome Project from the Wellcome Trust/Sanger Institute [30, 63]. For MUGA, the array design was based on an early release of this sequencing data from the Wellcome Trust/Sanger Institute

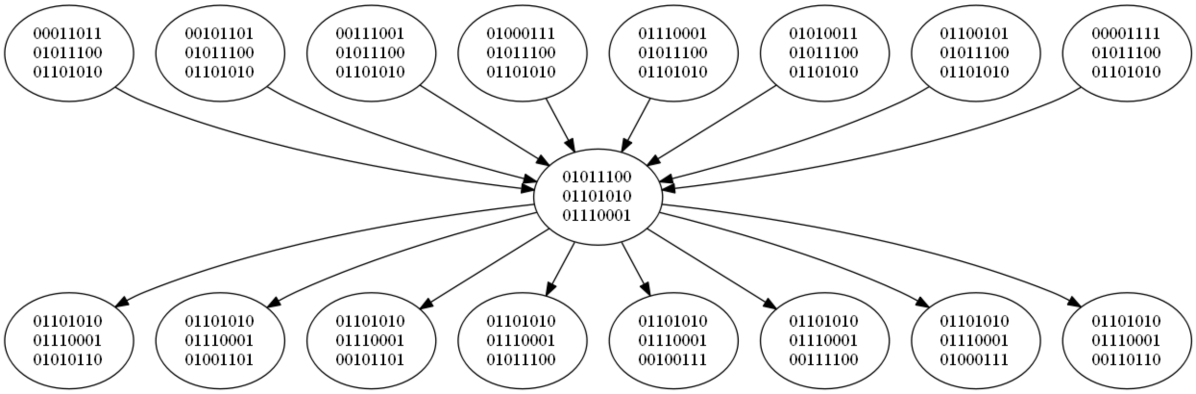


Figure 4.1: Subset of all compatible triples and the chaining of compatible triples. One can get to a compatible triple in one of eight ways, and one can choose the next SNP in the sequence from one of eight SDPs.

in September 2009, and the array verification and subsequent analysis were based on the final sequencing data published in September 2011.

4.1.2 Design

In practice, despite having 8.27 million SNPs to choose from, it was not possible to evenly cover the entire genome using only compatible triples. There are a number of areas of the genome with few or no segregating SNPs among the eight founders, and also a number of areas of sequence identity among founders, which creates areas of ambiguity between particular founder strains. Therefore, when designing MUGA, at times SNPs were chosen with minor allele frequencies of 3/5 as well as the 4/4 SNPs. However, the goal of maximizing the information content remained intact and the greedy algorithm utilized attempted to differentiate among the eight founders in as few SNPs as possible. The segregation pattern of each chosen SNP was also required to be different than its immediate neighbors, so that in the ideal case, three continuous SNPs can differentiate between each of the eight CC founders.

In addition to SNP selection based on SDP, we also applied additional filters. Since the Illumina Infinium technology uses 50-bp probes, any SNPs with known variants within 50 bps on either side of a variant were filtered out to remove potential off-target effects. The genetic sequence of each CC founder was obtained from the Wellcome Trust/Sanger Institute, and at the time of array creation, none of the selected SNPs had other known variants within 50 bps. A final SNP filter was conducted to remove all markers that segregate between C/G and A/T alleles since the Illumina Infinium genotyping array technology requires the construction of two beads to correctly differentiate between these allele combinations, rather than one bead as needed for all other allele combinations. This means that a single SNP would cost two beads, so that less locations in the genome could be genotyped overall.

As the total mouse genome is about 2.7 billion base pairs long and 9,000 total SNPs were allotted for the design, the final filtered set of SNPs was then binned into 300Kb bins. A SNP was selected from each non-empty bin so that it maximized the number of founders that could be differentiated when combined with previously selected SNPs. Using the selection technique described, 9,000 SNPs were selected and run through the Illumina scoring software. All SNPs that received low quality scores from Illumina were replaced with another SNP from the same bin with the same SDP. The final score file was submitted to Illumina by GeneSeek.

Of the original 9,000 SNPs selected, 7,851 were converted and placed on the Illumina Infinium platform (see Figure 4.2). The SNP markers have an average spacing of 325Kb (SD 191Kb); the distribution of the gaps between consecutive SNPs is shown in Figure 4.3. In genotyping array design, it is fairly typical that some of the SNPs will not be included on the final array for various reasons. Illumina guarantees a success rate of converting SNPs in the design file to beads in the final array of at least 80%. Therefore, the conversion rate of 87% on MUGA was acceptable.

When MUGA was designed, the 8.27 million SNP database had not yet been annotated by Sanger with quality scores for the SNPs, and it is believed that some of the SNPs chosen originally were low quality. Also, as previously mentioned, SNPs with no off-target variants within 50 bps were chosen, however, when intensity-based clusters of each SNP were examined, it was observed that many SNPs had unexpected intensity clusters outside the traditional AA, BB, or AB clusters. Many of these additional clusters are believed to be caused by previously unannotated SNPs within the probe sequence and with the most recently published SNPs from the Sanger Institute[30, 63], this hypothesis was confirmed. The final distribution of included SNPs can be seen in blue in Figure 4.2.

4.1.3 Samples Run on MUGA

Due to the CC founders' genetic diversity and MUGA's maximally informative design, MUGA's utility extends beyond CC animals. To date, 8,265 samples have been genotyped on MUGA, including a large number of mice from the Collaborative Cross (CC), the Diversity Outbred population (DO)[55], mice from the Mutant Mouse Regional Resource Centers (MMRRC), and wild mice. MUGA's design allows for accurate ancestry inference of not only mice from developing CC lines but also mice with non-CC ancestors, such as those from the MMRRC repository. MUGA and its accompanying multiallelic genotyping algorithm provide a versatile and low-cost genotyping platform for laboratory mice from the CC population and beyond.

Included in the 8,265 samples genotyped on MUGA were eight copies of each CC founder, at least two copies of all viable F1 crosses between the CC founders, and 1,833 CC samples genotyped at various stages of inbreeding. There were also about 1,100 DO mice as well as some F2 crosses of inbred strains, congenics, consomics and wild mice. Performance tests were done on MUGA using a panel of controls to ensure that it was producing quality genotype calls.

4.1.4 Performance on MUGA

The first performance test conducted on MUGA was to compare the results from biological replicates to ensure that MUGA was producing consistent genotype calls on a per SNP basis. Table 4.1 shows the results of these comparisons for thirteen pairs of biological replicates. The first seven pairs in this table are inbred animals for which one would expect to have a low number

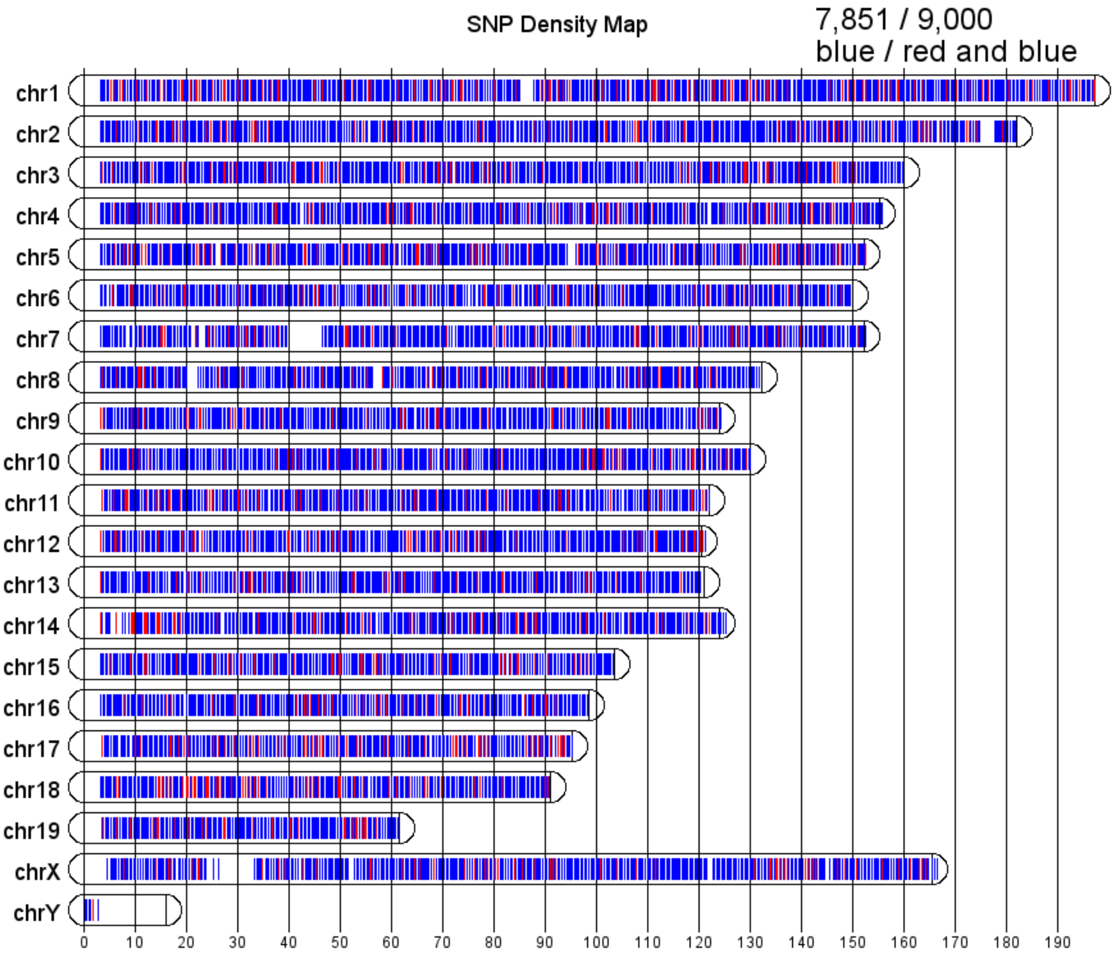


Figure 4.2: Distribution of the original 9,000 SNPs on MUGA in terms of location on chromosomes. Blue depicts the SNPs that ended up on the final MUGA array, while red shows SNPs that did not make it onto the final array. From this plot it can be seen that SNPs were lost fairly uniformly throughout the genome and there are no clusters of missing SNPs, showing that overall, the principles for selection of the SNPs should hold true with about 4-5 SNPs being necessary to completely differentiate among the 8 founder strains.

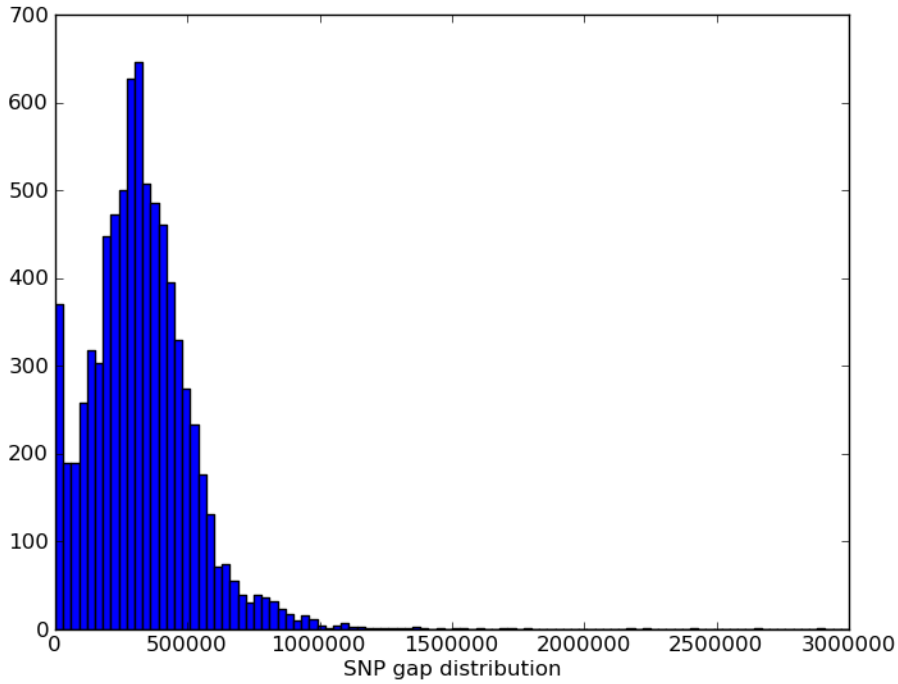


Figure 4.3: Distribution of the final 7,541 MUGA SNPs in terms of location on chromosomes (left plot) and spacing between SNPs (right plot).

Sample 1	Sample 2	N=N	H=H	A=A	# Discordant Calls
A/Jm111	A/J	127	91	7595	41
C57BL/6J	C57BL/6J	30	87	7612	125
129S1/SvImJm212	129S1/SvImJ	145	75	7581	53
NZO/HlLtJ	NZO/HlLtJm51	162	80	7572	40
CAST/EiJm42	CAST/EiJ	282	93	7398	81
PWK/PhJ-F11	PWK/PhJm175-C08	291	95	7368	100
WSB/EiJ-H06	WSB/EiJ-F09	168	85	7555	46
WSBxPWKf003	WSBxPWKm001	108	2737	4841	168
NODxPWKm004	NODxPWKm003	112	3438	4286	18
CASTxWSBf015	CASTxWSBm001	107	2678	4888	181
CASTxAJm005	CASTxAJm005	90	3426	4306	32
AJxPWKm006	AJxPWKf001	90	3430	4081	253
129xPWK037m	129xPWK 040F	85	3333	4124	312

Table 4.1: Comparison of biological replicate samples run on MUGA, where N=N depicts the number of times both samples had an N call at the same SNP, H=H depicts number of times both samples had an H call at the same SNP, A=A depicts the number of times the pair of samples had the same allele call (A, T, C, or G) at a particular SNP, and # of Discordant Calls depicts the total number of SNPs for which the two samples received different genotype calls.

of heterozygous calls and very high concordance between the two samples. The remaining samples in the table are F1s (first generation cross between two inbreds). One would expect these samples to have a high number of heterozygous calls but still have high concordance between the samples. Over the entire table, there is 98.6% concordance among biological replicates, and if one only considers inbred concordance where the genotype call is an A, T, C, or G and that both samples tested had the same call, there is 95.8% concordance.

The next performance test performed on MUGA was testing how consistent F1 genotype calls were with the founder calls. Overall, it was found that the F1 genotype was 91.9% concordant with their founder parentals. Statistics were also collected on individual SNPs and any SNP that was found to be underperforming (getting N for all samples) was flagged in the database.

Once it was established that MUGA performed well on various types of samples (inbreds, F1s, non-CC mice), the next step was to measure the information content of MUGA on a per SNP basis and to compare the overall information content of MUGA with other previously designed

Array	# SNPs	Avg. Entropy	Avg. bits per SNP
MUGA	7851	0.75555	0.31077
Sanger	13457	0.80189	0.18622
MDA	550000	0.56828	0.081011

Table 4.2: Entropy scores for micro-array comparison. The Avg. Entropy column represents the average entropy per SNP, while the inverse of the Avg. bits per SNP shows the average number of SNPs it took to differentiate among all eight founders.

genotyping arrays. Based on the founder genotypes of the eight CC founders, it was determined that when strictly using genotype calls, 1,426 of the 7,851 SNPs on MUGA have an entropy score of 0 (no information content in terms of founder assignment). However, by using the intensities of the probes rather than the genotype calls, this entropy score can be much improved[25].

By clustering biological replicates of the eight CC founders and their F1 crosses, one can ascertain which founders and F1s fall into the same clusters and which founders and F1s create their own clusters. One expects to see three distinct clusters for each SNP; one for the majority allele, A, one for minority allele, B, and one cluster representing a heterozygous call of AB. It is also expected that all eight CC founders will fall into one of the two inbred clusters. In reality, when this experiment was done, it was found that the eight CC founders segregated into a single cluster for 1,104 markers (no information content). The eight CC founders segregated into two homozygous clusters with a reference allele cluster, an alternate allele cluster and a single heterozygous cluster with F1 samples for 5,500 markers, as expected. However, the remaining 1,200 markers exhibit three or more clusters among the eight inbred founder. It is this last group of markers that allows for more entropy on a per SNP basis. Figure 4.4 depicts intensity plots of four markers, colored by genotype calls obtained from Illuminas GenomeStudio. Figure 4.4a shows a typical biallelic marker with two homozygous clusters and one heterozygous cluster, while Figure 4.4b shows a non-hybridizing marker with arbitrary H calls. Figures 4.4c shows a multiallelic SNP with several heterozygous clusters, one of which is uniformly called N, and Figure 4.4d shows a multiallelic SNP with one heterozygous cluster alternately called both N and H due to batch effects in the calling algorithm. Both (c) and (d) represent SNPs with higher entropy scores than the typical biallelic SNP shown in (a), since if genotypes alone were used in (c), then there are only three possible calls (A, G, or H, since N calls are ignored)), but if intensity clusters are utilized, then five calls are possible[25].

In comparing entropy scores of genotyping arrays, only the genotype calls of the founder mice were used so that each array had the same possible information. The genotype information for the eight CC founders was obtained for the Sanger array [52] with 13,457 SNPs on it, as well as the Mouse Diversity Array(MDA)[64] with about 550,000 SNPs. Table 4.2 shows the calculated entropy scores for each of these arrays as well as MUGA. The Avg. Entropy column represents the average entropy per SNP, while Avg. bits per SNP shows the average number of SNPs it took to differentiate among all eight founders. To calculate this value, I divided three (the minimum number of SNPs with which it is possible to differentiate among all eight foundes) by the actual number of SNPs it took to create eight unique haplotypes. This number was then averaged over all sliding windows of the genome and shown in Table 4.2.

In terms of the average entropy per SNP, the Sanger array does slightly better than MUGA and MDA with an average entropy of 0.80. MUGA isn't far behind at 0.76, and MDA does fairly well, with an average entropy per SNP of 0.57. Overall, MUGA has the best average effective entropy of the three arrays at 0.31, meaning that it takes the fewest consecutive SNPs to differentiate among the eight founders. By taking the inverse of the average effective entropy, you can see that it takes an average of 9.67 SNPs to differentiate among the eight founders on MUGA, but about 16 SNPs to do so on Sanger, and 37 consecutive SNPs to differentiate among all eight founders on MDA. This analysis was conducted genome-wide and results can be skewed by the beginning and ends of chromosomes, as well as areas of sequence identity between founders.

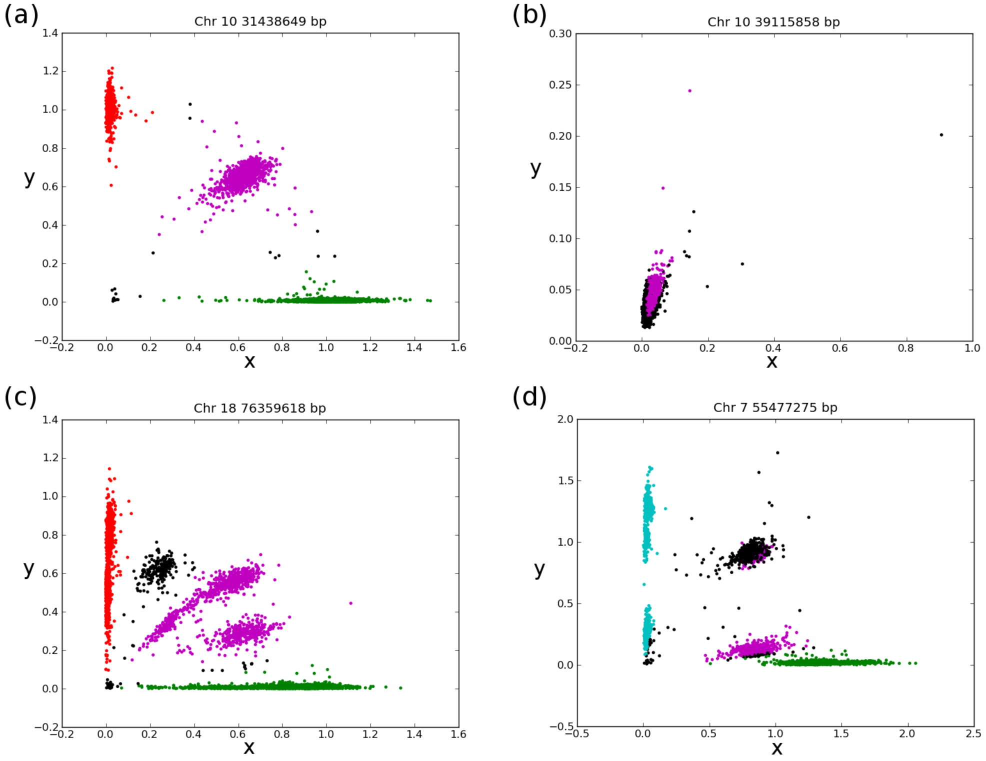


Figure 4.4: Intensity plots of four markers, colored by genotype calls obtained from Illuminas GenomeStudio. Each point represents a single MUGA sample with its reference probe intensity on the x-axis and its alternate probe intensity on the y-axis. H calls are colored magenta, N calls are colored black, and the four nucleotides A, C, G, and T are colored green, cyan, red, and blue, respectively. (a) A typical biallelic marker with two homozygous clusters and one heterozygous cluster. (b) A non-hybridizing marker with arbitrary H calls. (c) A multiallelic SNP with several heterozygous clusters, one of which is uniformly called N. (d) A multiallelic SNP with one heterozygous cluster alternately called both N and H due to batch effects in the calling algorithm.[25].

4.2 MegaMUGA

A second generation genotyping microarray, MegaMUGA, is also built on the Illumina Infinium platform and was designed to expand the number of markers and versatility of the successful Mouse Universal Genotyping Array (MUGA). It extends MUGA from 7.8K to 77.8K markers, and includes all MUGA markers as a subset. There are three types of probes on MegaMUGA. In addition to traditional SNP probes, a second probe type for tracking known structural variants (insertions, deletions and duplications) has been introduced. A third probe type was designed to detect the presence of sequences present only in genetically engineered mice (GEM) (Cre, Luciferase, etc). MegaMUGA was designed to not only optimize the identification of founder contribution and detection of residual heterozygosity among CC strains at any stage of inbreeding, it was also designed to correctly identify the founder pairs present in areas of residual heterozygosity.

The vast majority of MegaMUGA probes ascertain traditional biallelic SNPs. SNPs were selected to be distributed across the entire genome including the mitochondria and the Y chromosome with an average spacing of 33 Kb. For the autosomes, these probes were distributed as evenly as possible based on a new linkage map[35] for the mouse with a slight excess of probes in the telomeric regions to facilitate detection of recombination events in these regions. SNPs were selected to be informative in most mouse populations (including wild mice and multiple Mus species) with a special emphasis for markers that are informative in the CC and DO populations.

4.2.1 Design

In the selection of the 80,000 markers to be placed on MegaMUGA, a number of different criteria were considered. The majority of the SNPs, about 65,000, were chosen since they were maximally informative for the CC/DO mice, while 14,000 were chosen to work well with wild mouse strains (domesticus, musculus, castaneous), 750 were chosen to identify Mus spretus species, 150 were chosen to differentiate between C57BL/6J and C57BL/6N, 102 were selected for GEM, 58 were selected for a chromosome evolution project, and 14 were selected for an X-inactivation mapping experiment [12]. To work well with the Illumina Infinium technology, these SNPs were all selected to not have any off-target variations within 50 base pair on at least one side. Also, the 49-mer either immediately preceding or following the SNPs must be unique and not occur elsewhere in the genome. All SNPs were also selected to include as few markers as possible that segregate beteween C/G or A/T alleles. This was done since as described in the MUGA section above, the technology used in Illumina genotyping arrays requires two beads to be developed to correctly differentiate between the aforementioned allele combinations. This means that a single SNP would cost two beads, so that less locations in the genome could be genotyped overall. In some regions of the genome, there were no alternative SNPs. Therefore, the final design of the array includes about 200 of these two-bead allele combination SNPs.

Of the 65,000 CC/DO SNPs, about 60,000 SNPs were selected to be uniformly distributed by recombination events over all autosomes and Chromosome X. About 20 invariants in the PAR region were included, as well as 45 SNPs on Chromosome Y, and 31 SNPs on the mitochrondria. The remaining SNPs were distributed in and beyond the last interval of each chromosome to obtain better resolution in a known high recombination area. Unlike in MUGA, the bin sizes for selecting SNPs are not uniform by genomic distance. Instead, the bin sizes were determined by using a mouse recombination map[35] so that the genome is binned into intervals with like-sized number of recombinations. To obtain this information, a series of plots were made, as shown in Figure 4.5, and it was determined that in order to choose 60,000 SNPs this way, 2.75 SNPs should be chosen per recombination (using only recombinations that were less than 25% of the length of the chromosome). The midpoints of the recombinations were used to determine the final bins to be used for SNP selection.

In choosing the CC/DO SNPs, SNPs that were maximally informative between not only the eight CC founders, but also the twenty-eight F1 combinations were desired. This enables the user to better determine founder pairs in regions of residual heterozygosity. To achieve this goal, within each bin, SNPs with the same strain diversity pattern (SDP) were binned together. We looked at each chromosome in 5-bin sliding windows and used a dynamic programming algorithm to find one SDP in each interval that maximized the number of founder/F1 combinations that could be distinguised. The sliding window size was set to five since five SNPs is the fewest number that could potentially differentiate among all thirty-six founder and F1 combinations. Since all possible paths in the solution can increase exponentially, the possible solutions were pruned at each step. Occasionally a bin did not contain one of the eight candidate SDPs, which created a penalty for this path. Once the number of penalities along a particular path exceeded some threshold above the current best path, it was pruned. While this greedy pruning metric may not have produced the optimal path, it was able to produce a reasonable path in a short time period. As previously mentioned, the telomeres of chromosomes are very recombination rich, so therefore an additional 500 SNPs were selected at the ends of each chromosome in and beyond the last interval found for each chromosome. These SNPs were selected in the same sliding window fashion as described above.

The final array has 77,808 unique SNPs on it. Of the originally selected 79,797 SNPs, 98% were converted and present in the final design. This is an extremely high rate of conversion. The final distribution of SNPs on MegaMUGA can be seen in Figure 4.6. It is obvious from this figure that SNPs are evenly distributed over the entire genome in all areas where variants exist. Any white regions in this figure illustrate regions of the genome where there are known gaps. To drill down on the SNP distribution further, a colleague created Figure 4.7 that breaks down the final distribution of SNPs by type for which they were originally chosen. The purple represents the CC/DO SNPs, the blue represents SNPs chosen to differentiate the wild-derived strains, and the red dots depict SNPs chosen to differeniate between strains of C57BL/6 mice.

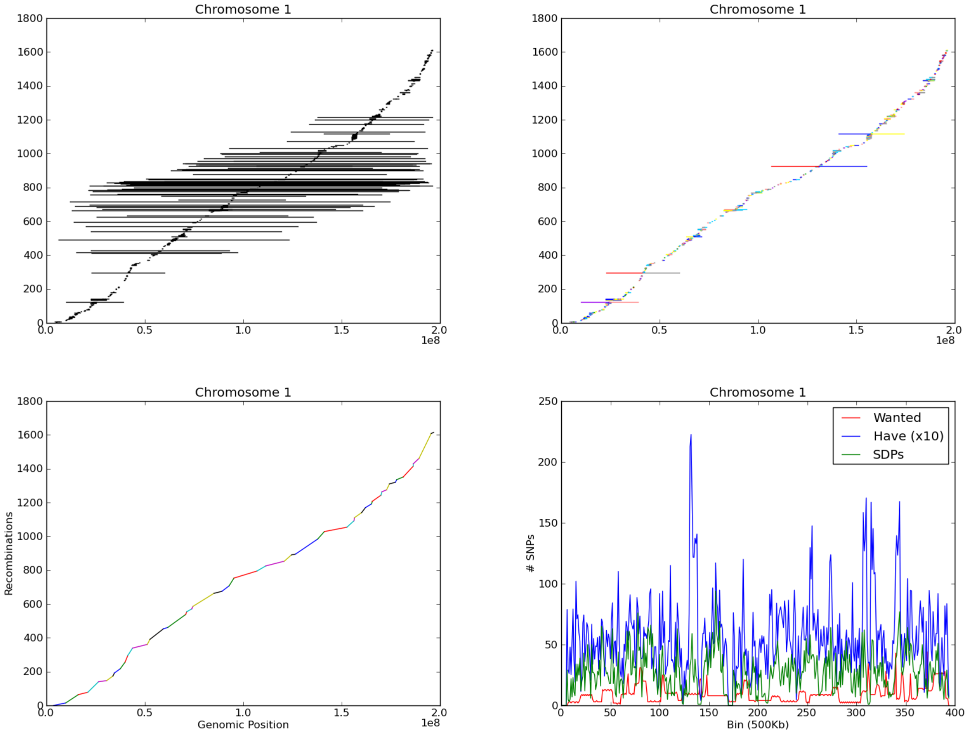


Figure 4.5: Recombination events on a chromosome basis. Plot A shows the recombinations events ordered by the midpoint of each recombination, the start and end points of each event is plotted. The x-coordinate equals the genomic position, and the y-coordinate shows the number of the event. In Plot B, since there are a number of large intervals in the data set, I removed them and replotted all intervals that are less than 25% of the length of the chromosome. I also colored the intervals such that the colors correspond to the founders for the proximal and distal ends of the intervals. For plot C, I show a smoothed out linear fit curve using 50 line segments to describe the recombination distribution, and in plot D, using a fixed bin size of 500Kb, I show the distribution of number of SNPs that we want (red) to see on MegaMuga as well as the distribution of the SNPs we have available (blue), and the SDPs (green) of those available SNPs. In order to plot these on a similar scale, I had to first divide the total SNPs we have per bin by 10.

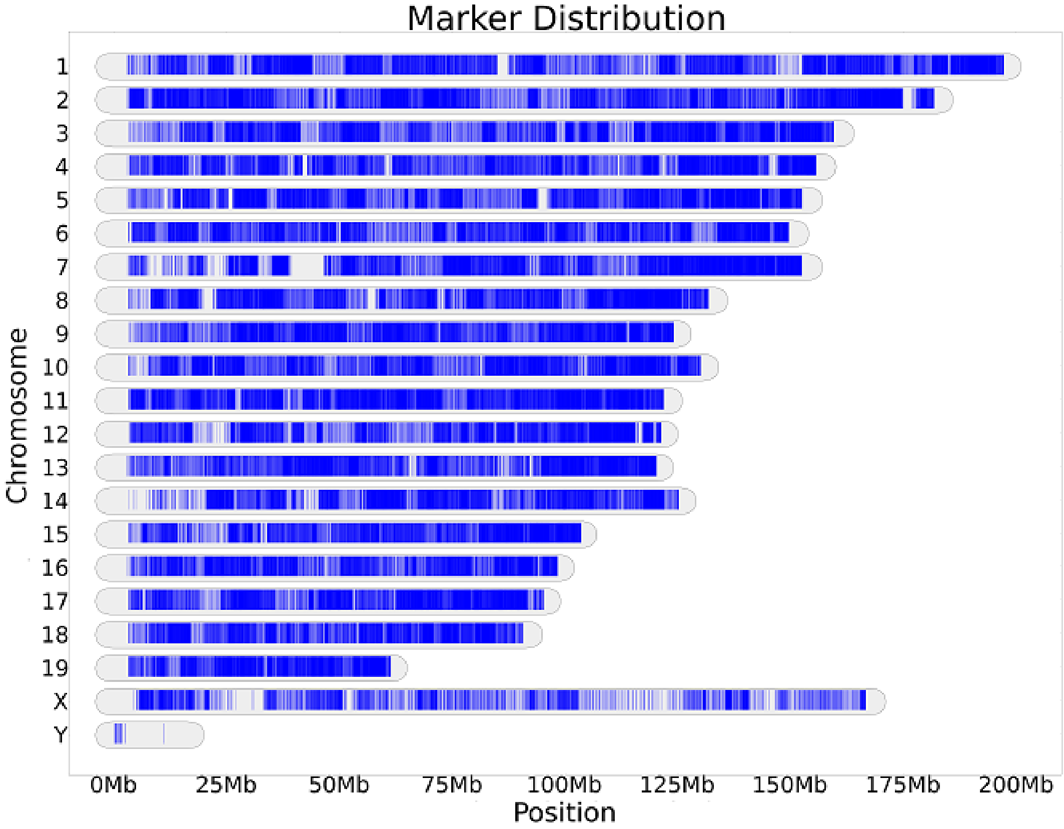


Figure 4.6: Distribution of the final 77,808 MegaMUGA SNPs in terms of location on chromosomes.

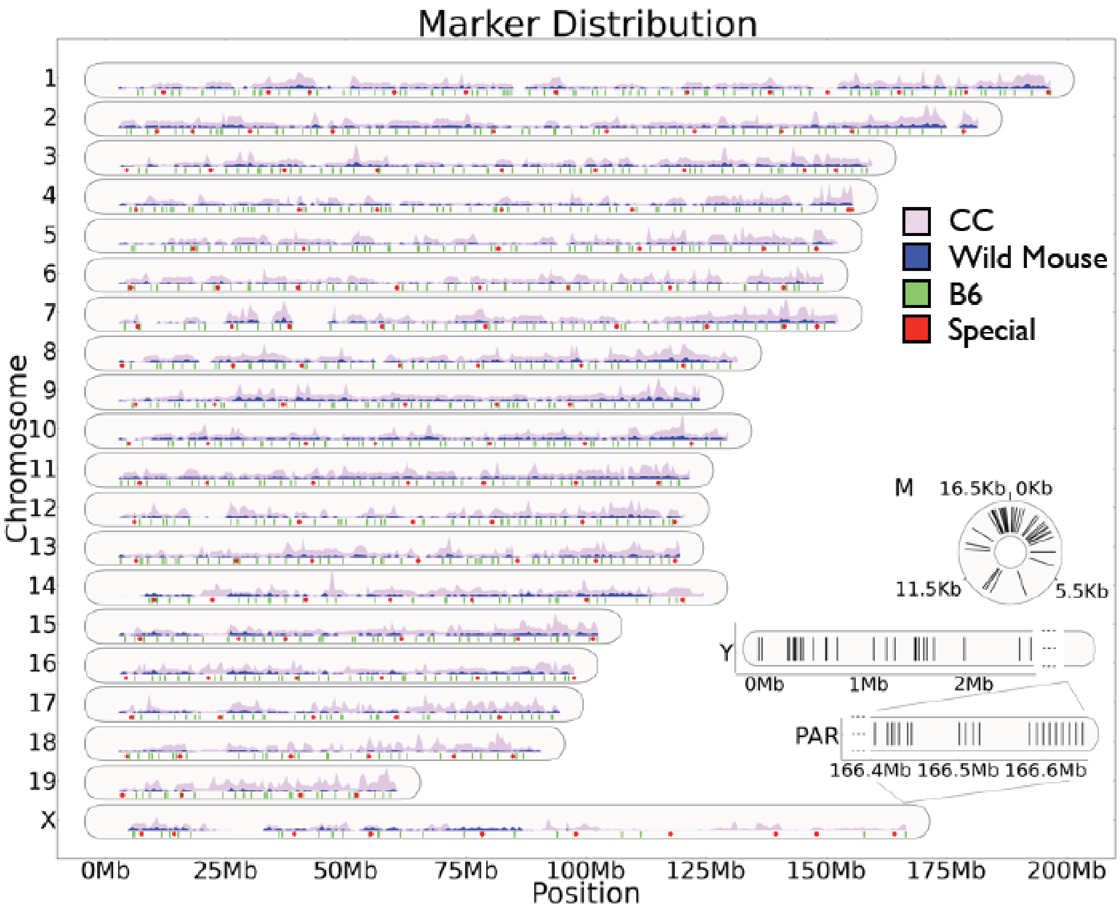


Figure 4.7: Break down of the final distribution of SNPs by type for which they were originally chosen. The purple represents the CC/DO SNPs, the blue represents SNPs chosen to differentiate the wild-derived mouse strains, and the red dots depict SNPs chosen to differeniate between strains of C57BL/6 mice. (Credit to Chen-Ping Fu for the image)

4.2.2 Samples Run on MegaMUGA

In total, more than 6,600 samples have been genotyped on MegaMUGA. Of these samples, eight copies of each CC founder were genotyped, as well as all viable F1 crosses betweeen the CC founders. There were also 462 CC samples genotyped at various stages of inbreeding, as well as a number of DO mice, F2 crosses of inbred strains, congenics, consomics, and wild mice.

4.2.3 Performance on MegaMUGA

As with MUGA, similar performance tests were run to ensure the quality of the SNP genotyping calls. The first performance test was to compare the results from biological replicates to ensure that MegaMUGA was producing similar genotype calls on a per SNP basis. Table 4.3 shows the results of these comparisons for fifteen pairs of biological replicates. The first eight pairs in this table are inbred animals for which one would expect a low number of heterozygous calls and very high concordance between the two replicate samples. The remaining samples in the table are F1s (first generation cross between two inbreds). One would expect these samples to have a high number of heterozygous calls but still have high concordance between the replicate samples. Over the entire table, there is 98.9% concordance among biological replicates, and if one only considers inbred concordance where the genotype call is an A, T, C, or G (where A=A) and both samples have the same call, there is 96.2% concordance. The consistency of the F1 genotype calls with the founder calls was also tested. Overall, it was found that the F1 genotype was 95.2% concordant with their founder parentals. Statistics were also collected on individual SNPs and any SNP that was found to be underperforming (getting N for all samples) was flagged in the database.

4.3 Conclusion

As more genetic reference populations (GRPs) are created, there is a need for low-cost methods for ascertaining the underlying genomic structure of these populations. By using a

Sample 1	Sample 2	N=N	H=H	A=A	# Discordant
129S1/SvImJm1314	129S1/SvImJm35370	2194	237	75149	228
A/Jm0111	A/Jm0417	2196	242	75197	173
C57BL/6Jm1957	C57BL/6Jm36826	2109	246	75273	180
CAST/EiJm0042	CAST/EiJm0538	2807	208	74395	398
NOD/ShiLtJm0150	NOD/ShiLtJm1214	2212	214	75176	206
NZO/HILtJm0051	NZO/HILtJm0591	2217	225	75029	337
PWK/PhJm0175	PWK/PhJm1090	2749	191	74456	412
WSB/EiJm0993	WSB/EiJm1345	2303	208	73968	1329
(129S1xB6)F1f15916	(129S1xB6)F1m15914	2045	23889	51071	803
(129S1xPWK)F1f040	(129S1xPWK)F1m037	2105	37458	36567	1678
(A/JxNOD)F1f15432	(A/JxNOD)F1m15427	2076	19784	55200	748
(A/JxPWK/PhJ)F1f001	(A/JxPWK/PhJ)F1m006	2124	38214	35732	1738
(CAST/EiJxPWK/PhJ)F10123	(CAST/EiJxPWK/PhJ)F1f0163	2360	23094	51247	1107
(NOD/ShiLtJxWSB/EiJ)F1f0141	(NOD/ShiLtJxWSB/EiJ)F1m0143	2097	32130	42255	1326
(WSB/EiJxPWK/PhJ)F1f0284	(WSB/EiJxPWK/PhJ)F1m0276	2132	32807	41144	1725

Table 4.3: Comparison of biological replicate samples run on MegaMUGA, where N=N depicts the number of times both samples had an N call at the same SNP, H=H depicts number of times both samples had an H call at the same SNP, A=A depicts the number of times the pair of samples had the same allele call (A, T, C, or G) at a particular SNP, and # of Discordant depicts the total number of SNPs for which the two samples received different genotype calls.

large database of available probes and making careful design decisions, low-density full-genome genotyping arrays can be designed that gain insight into the underlying structure of these large populations. Maximizing the entropy on a per SNP basis allows the array to be maximally informative genome-wide and give the most population information in the fewest number of SNPs. Since mouse populations are a popular model organism, a number of arrays have been designed for use in mouse. However, none of these arrays had the necessary characteristics needed for use with the Collaborative Cross. To work well for the CC experiment, an array needed to be cost-effective, highly informative and widely available as this population is being widely distributed. Therefore, to meet this need, two cost-effective, maximally informative, widely available full-genome genotyping arrays were designed.

This chapter described the design of two highly informative genotyping arrays for use with the CC lines as well as other mouse strains. Both arrays attempt to maximize the information content over a pre-determined number of SNPs. However, MUGA was designed to determine the amount of residual heterozygosity genome-wide as well as correctly distinguish between the eight CC founders in all homozygous regions of the genome, while MegaMUGA was designed to not only correctly infer the CC founder ancestry within homozygous regions, but also within heterozygous regions of the genome. In MUGA, the bin sizes for selecting SNPs was uniform by genomic distance, but in MegaMUGA, the bin sizes were determined by using a mouse recombination map[35] so that the genome is binned into intervals with like-sized number of recombinations. Both arrays had numerous samples genotyped on them and performance tests conducted to ensure the accuracy of their genotyping calls. In the next chapter, I will discuss the type of analysis that can be done by combining the theoretical aspects of marker-assisted inbreeding as described in Chapter 3 with the genotyping arrays described in this chapter.

CHAPTER5: MAI IN PRACTICE

This chapter is about using the theory explained in Chapter 3 and the genotyping platforms introduced in Chapter 4 to apply the MAI theory to real mice. All tools were written primarily to analyze the Collaborative Cross (CC) mice as described in Chapter 2. However, all mice that are genotyped on one of the two platforms discussed in Chapter 4 (MUGA or MegaMUGA) can easily be analyzed with these tools as well. To apply MAI techniques to real mice, I needed to first obtain haplotype reconstructions of the mice based on the genotyping platforms. These haplotype reconstructions are then used to perform line quality control analysis, choose breeders, and simulate matings of our live mice. For lines that met certain thresholds as described in this chapter, I then merged together haplotype reconstructions of multiple samples within the same line to predict the final genotypes of these lines. This information is available publically since the CC lines are now available for distribution.

5.1 Analysis Tools

All analysis tools are available on the UNC Systems Genetics website, although most tools do require a login to use since not all samples genotyped on MUGA and MegaMUGA are publically available. All tools are written using a Python web framework tool and will work on any sample run using either of the genotyping platforms described in Chapter 4.

5.1.1 Haplotype Reconstructions

The genome of an individual from an admixed population is a mosaic of segments inherited from its ancestors. Mapping populations, in particular, are often derived from a set of inbred founders where the genome of each individual is a mixture of founder haplotype segments. Ancestry inference on such an admixed individual refers to the problem of partitioning the individual's genome into haplotype blocks labeled with the contributing ancestor, with or without a given pedigree. A haplotype reconstruction refers to the most likely representation of this ancestry mosaic. More specifically, when speaking about the CC or DO mice, a haplotype reconstruction refers to assigning the most likely of the eight founders to each area of the genome. As you can see in Figure 5.1, eight colors are used to represent the eight founder strains and the images are colored according to which of the eight founders each part of a developing mouse lines genome is most likely inherited. Each founder is assigned a color and these colors are used consistently throughout all analysis done on the CC or DO mice. Figure 5.1 is depicting the mouse genome and shows that mice have 19 autosomes, mitochondria, and since this sample is a male, there is one copy of Chromosome X and one copy of Chromosome Y. If this sample were female, one would see two copies of X and no Y.

There are numerous methods for inferring ancestor mosaics when given the genotypes of an individual and a set of ancestral haplotypes. Such methods generally rely on bialleic SNP genotype calls obtained from genotype calling algorithms that classify each marker as belonging to one of four states (reference allele, alternate allele, heterozygous, or no call) based on probe hybridization intensity signals. In humans, mapping ancestry is an essential step in admixture mapping, and methods such as HAPMIX [46], HAPAA [54], and LAMP [51] use HMM-based methods to infer the most likely ancestral blocks for each individual. However, most of these methods accept genotypes from calling algorithms as ground truth and seldom discuss the artifact of calling errors, although LAMP does attempt to improve accuracy by analyzing sliding windows and taking a majority vote. Algorithms for inferring ancestry in model organisms with known ancestors have also been proposed, such as HAPPY [38], a package for QTL mapping designed for outbred crosses. Methods for ancestry inference in recombinant inbred strains include two designed for the Collaborative Cross [36, 48]. GAIN [36], which was designed with the CC in mind, is an HMM-based algorithm that uses knowledge of the pedigree to efficiently infer ancestry probabilities. One assumption of GAIN and other existing methods is the use of high density genotypes. SNPs from high density arrays are often heavily filtered based on non-performing markers or questionable genotype calls. However, studies using low density arrays do not have the luxury of filtering out a significant percentage of SNPs and keeping only reliable genotype calls. Therefore, Fu et al. [25] developed a method for inferring ancestry without first converting the probe intensity data into genotype calls. This method works by minimizing the intensity difference between a target individual and one or more of its ancestors. By using the probe intensities directly, rather than first converting the intensities into genotype calls, there is less information loss and therefore less errors produced from incorrect genotype calls[25].

Since the haplotype reconstructions used in this work were derived mainly from low-density genotyping arrays (MUGA and MegaMUGA), the majority of the haplotype reconstructions were computed using the last method by Fu et al. However, any of the above mentioned methods could be used to create the haplotype reconstructions that were used as input for the tools in this chapter.

5.1.2 Line Quality Control

For MUGA samples, I ran standard quality controls on all lines before the genotypes were used for any other analysis. For all samples run on MegaMUGA, it is highly recommended that individual scientists run the following quality control tools for their samples before using the other analysis tools. There are a number of different types of errors that may have occurred including incorrect labeling of the samples when genotyped, insufficient DNA present at time of genotyping, or breeding errors resulting in either an under/over-abundance of one or more founders DNA within a sample or a high number of shared recombinations between samples in different CC lines. During the earlier stages of the CC breeding, 458 CC samples were genotyped using MUGA and the following quality control tools were run on those samples. Therefore,

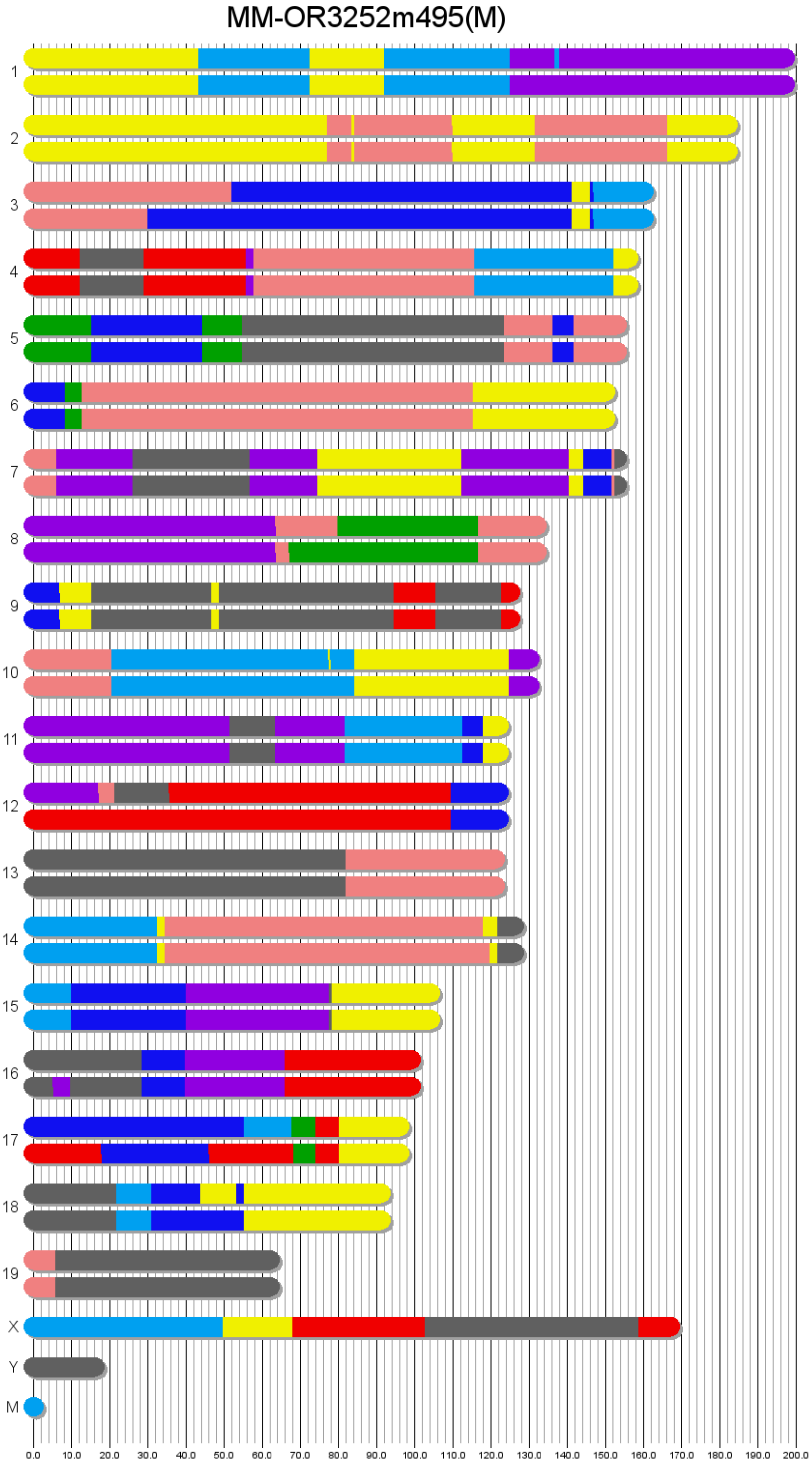
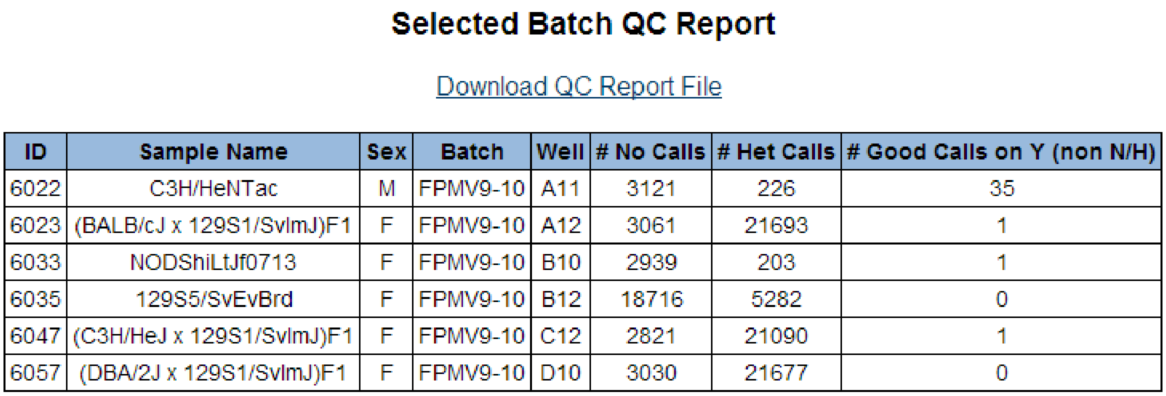


Figure 5.1: Haplotype reconstruction of a CC mouse sample.



Selected Batch QC Report

Download QC Report File

ID	Sample Name	Sex	Batch	Well	# No Calls	# Het Calls	# Good Calls on Y (non N/H)
6022	C3H/HeNTac	M	FPMV9-10	A11	3121	226	35
6023	(BALB/cJ x 129S1/SvImJ)F1	F	FPMV9-10	A12	3061	21693	1
6033	NODShiLtJf0713	F	FPMV9-10	B10	2939	203	1
6035	129S5/SvEvBrd	F	FPMV9-10	B12	18716	5282	0
6047	(C3H/HeJ x 129S1/SvImJ)F1	F	FPMV9-10	C12	2821	21090	1
6057	(DBA/2J x 129S1/SvImJ)F1	F	FPMV9-10	D10	3030	21677	0

Figure 5.2: Quality Control Report screenshot.

breeding errors have been ruled out and lines have been culled if any of these errors were found. As a result, the statistics table and shared recombination tools are less likely to find breeding errors within the extant CC lines and more likely to be used to determine characteristics about any mouse genotyped on one of our genotyping platforms.

Quality Control (QC) Report

The first quality control tool that should be run on all samples simply counts the number of no calls, heterozygous calls, and Y chromosome good calls. While all other tools in this chapter assume that haplotype reconstructions have already been computed, this tool uses the genotype calls that come directly from the Illumina calling software. This test should be run on all samples to determine if any errors occurred during the genotyping process. A good quality sample should have fewer than 2000 "no call" genotypes, but samples with fewer than 4000 might still be usable depending on the genetic makeup of the sample. For example, mice derived from classical laboratory mouse strains should have fewer "no calls" than a wild caught specimen. Figure 5.2 shows a screenshot of some inbred and F1 samples run through the QC Report.

An inbred sample should have fewer than 1000 heterozygous genotypes. Remember, that these calls probably do not really represent heterozygosity in an inbred sample, but are most likely due to one of our multiallelic markers, which are not called correctly by Illumina's software. F1 crosses should have from 10,000 to 40,000 Het calls depending on the genetic diversity between the parentals. A male sample is indicated by 30 or more good calls on Y, whereas female samples generally have fewer than 10. If a sample does not meet the above criteria, the sample labels should be rechecked, as well as the sample quality metrics from the Illumina calling software.

Once a sample has passed the metrics from the QC Report, a haplotype reconstruction is generated as described above. This haplotype reconstruction is then used as the basis for all further analysis about the line.

Shared Recombinations

A recombination refers to an area of the genome at which a recombination event must have occurred. When looking at a haplotype reconstruction, a recombination occurs wherever a color change is shown and can be described by its genomic location as well as the two founder colors that flank it. Shared recombination events are defined as those involving the same two strains in the same proximal-to-distal orientation at the same chromosome position. I determined the number of shared recombination events in the autosomes between all pairwise combinations of 458 genotyped CC samples. Events that are fixed in a strain were counted only once. As expected, most pairs of lines do not share any recombination events (mean 0.0653 ± 0.7552) but a subset of pairs had a significantly higher rate of shared events as shown in Figure 5.3. All known related lines have at least three shared events, while not a single pair of independent lines with three shared recombination events exists, and only 5% of 47,278 pairwise combinations between independent lines have one or two shared events (Figure 5.3). Using the shared recombination tool, I identified 99 related CC samples defining 46 sets of related lines. This left 405 indepedent CC lines.

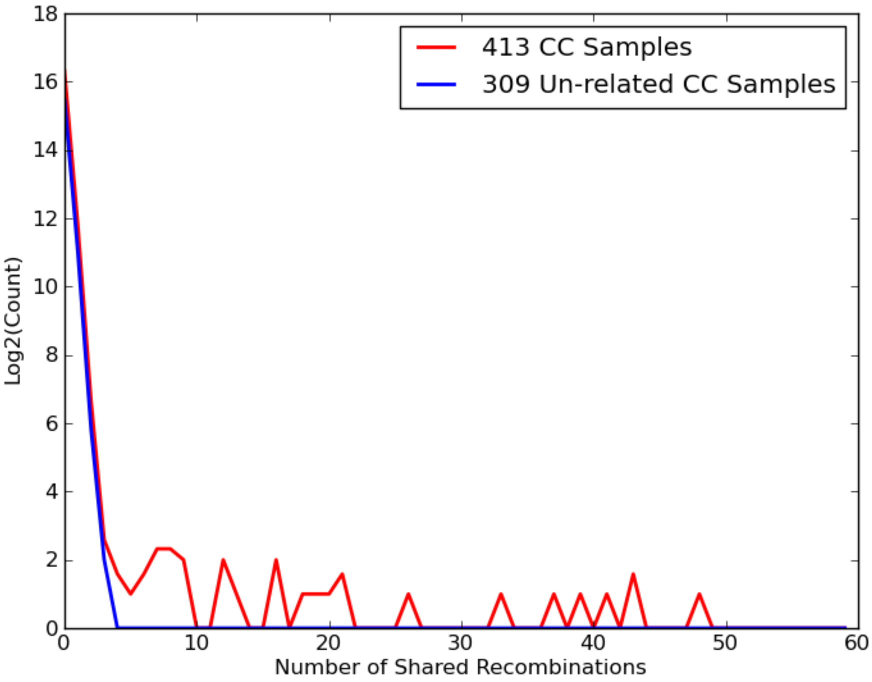


Figure 5.3: Distribution of shared recombination events before and after identifying related samples.

The shared recombination tool allows the user to select a set of samples run on either genotyping array and it outputs a table with every pairwise combination of the samples. Because it computes all pairs, this tool slows down expontentially as more samples are selected. Therefore, it is recommended that the user choose a small number of samples. The data in the table represents the total number of recombinations and line segments for each sample, as well as the total number of shared recombinations for each pair. Each line of the table is a pair of samples.

Statistics Table

For CC and DO lines, the expectation is that all eight founders should be represented in each line and that the percent contribution from each founder should be approximately equal. Particularly in early stages of the CC breeding, lines that did not have evidence of all eight founders had been subject to some breeding error and were therefore no longer standard CC lines. Among the 405 independent lines, only 330 have alleles from each of the eight founder strains present in the autosomes. Based on the simulation of 7 million CC lines, I estimate that 0.05% will have $<$1% of any given founder. The rate of CC lines missing one or more founders was significantly higher than the results of the simulation, and I eliminated any line with more than one founder missing.

The statistics table tool is also useful in determining if a CC or DO sample has an over-representation of one or more founders. An over-representation of a particular founder may have biological significance, and could lead to some interesting results when broken down by chromosomal regions. The statistics table tool allows the user to select a set of samples that have been previously genotyped and it outputs a table of data that includes the sample name, the percent contribution of each of the eight founders (columns A-H), the percentage of residual heterozygosity found within that sample, the number of founders present, and any founders that are missing or over-represented. A founder is considered to be missing if there is less than 0.32% contribution from that founder and a founder is considered to be over-represented if there is more than 24.68% contribution. These numbers represent two standard deviations away from the mean value expected based on simulations.

Using the results of the statistics table, overall statistics were collected for the CC population, using 350 CC lines that were determined to be unique and independent[17]. It was found that the eight founder strains alleles were similarly represented when averaged across the genome of the CC lines, and their contribution varied between 11.02% for CAST/EiJ and 13.52% for 129S1/SvImJ (Table 5.1). Genome-wide statistics of founder contribution were also collected and it was found that there is little population structure among the extant CC lines, with few exceptions. In particular, there is a region of Chromosome 2 with a high contribution from WSB/EiJ, as well as overrepresentation of NZO/HlLtJ on chr 5 and overrepresentation of

A/J	C57BL/6J	129S1/SvImJ	NOD/ShiLtJ	NZO/HlLtJ	CAST/EiJ	PWK/PhJ	WSB/EiJ
12.38%	13.26%	13.52%	12.68%	12.98%	11.02%	11.13%	13.04%

Table 5.1: Genome-wide CC founder contribution.

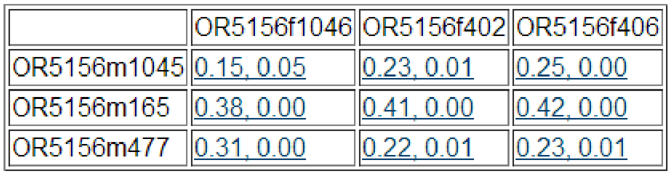
WSB/EiJ and 129S1/SvImJ on chr 7. There is also underrepresentation of CAST/EiJ on Chromosome X[17].

5.1.3 Breeder Selection

In the original design of the CC, sibling mating pairs were chosen at random at each generation by a software package so as not to add any population structure or accidently select for particular phenotypes. In order to incorporate the MAI techniques discussed in Chapter 3, I needed to create similar tools that allowed the mouse technicians to choose more compatible breeding pairs without affecting the overall structure of the inbred lines. Using the tools described below allows the user to make informed decisions both about which breeding pair/pairs to choose as well as the probability that a proposed breeding pair will generate an inbred pup. These tools use the same metrics and simulator as the experiments discussed in Chapter 3.

Choose Breeders

The choose breeders tool allows the user to select a set of samples and it does a pairwise comparison of the samples. If the user chooses to "Compare All Pairs", then all pairs, regardless of sample sex, are compared. If the user does not check the box, then all females are paired with all males, but are not compared with other females and males are not compared to other males. The initial comparison yields a table containing two metrics for each pair (see Figure 5.4). The first metric tells the user the total residual heterozygosity within the two samples. This residual heterozygosity metric uses the notion of joint heterozygosity I introduced in Chapter 3, meaning that it calculates the total genomic distance in which either one sample is not fixed, both samples are not fixed, or the samples differ from one another within particular regions of the genome



	OR5156f1046	OR5156f402	OR5156f406
OR5156m1045	0.15, 0.05	0.23, 0.01	0.25, 0.00
OR5156m165	0.38, 0.00	0.41, 0.00	0.42, 0.00
OR5156m477	0.31, 0.00	0.22, 0.01	0.23, 0.01

Figure 5.4: Comparison table created by Choose Breeders tool. The first metric is the total residual heterozygosity between the sample pair, and the second metric shows the multiplied probability of becoming inbred in 1 generation.

and divides that distance by the total genomic distance of a mouse. The second metric shows the multiplied probability of becoming inbred in 1 generation, otherwise known as the weighted state metric (WSM) that I also introduced in Chapter 3. Ideally a pair will minimize the first metric and maximize the second metric. These metrics are both shown as a hyperlink for each pair. When the user clicks on a hyperlink, they are shown an image of the combined haplotype reconstruction images from both samples, as shown in Figure 5.5. Sample 1 is shown as the top 50%of each chromosome and Sample 2 is the bottom half. This image gives the user a visual of which regions of the genome might be fixed in the next generation and which regions cannot be fixed in one generation of inbreeding. Between the image and the metrics, a user can make an informed decision about which breeding pair would be the best based on all possible breeding pairs shown.

Simulate Matings

Much like the choose breeders tool, the simulate matings tool allows the user to select a set of samples for pairwise analysis. However, instead of showing you the combined haplotype reconstructions of the two parents, it instead runs the simulation software utilized in the Chapter 3 experiments to simulate 10,000 offspring (5,000 female and 5,000 male) from that breeding pair. It then creates a plot of the distribution of the residual heterozygosity of simulated offspring, as shown in Figure 5.6. It also shows the combined residual heterozygosity of the breeding pair as a red line on the plot as a point of comparison. If the residual heterozygosity of the breeders is

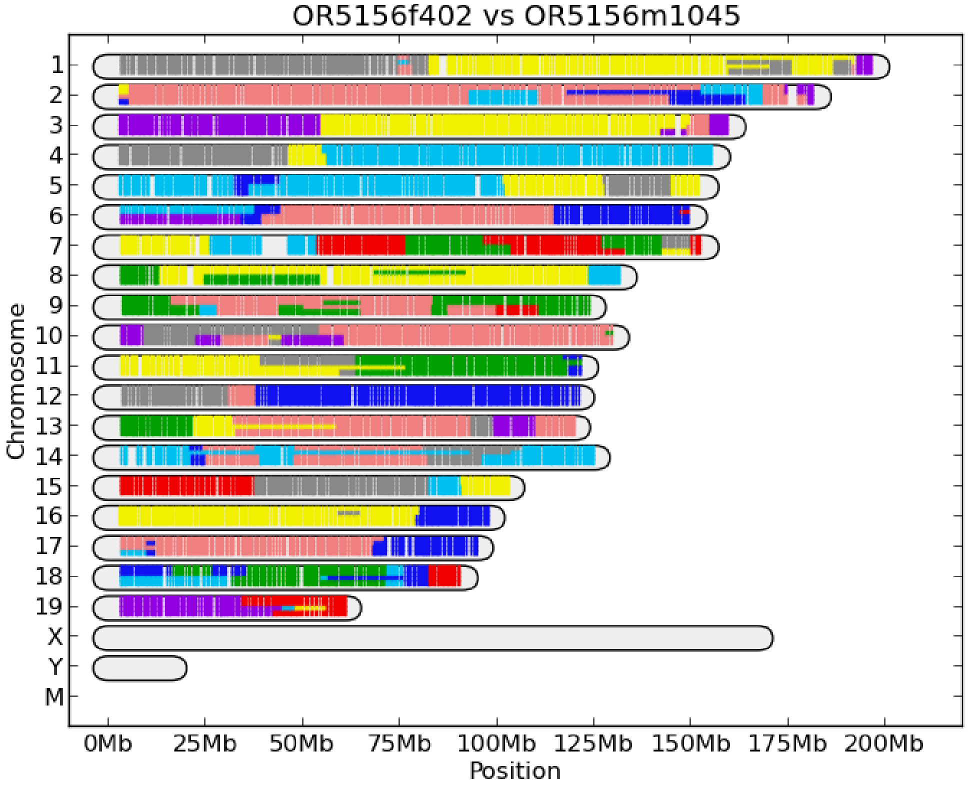


Figure 5.5: Combined haplotype reconstruction images of proposed breeders as created by the Choose Breeders tool.

greater than 50% though, it is not shown since the image would need to be skewed to fit both lines on the plot. If more than 2 samples are selected, the pairwise combination of all females with all males are simulated and plots for each potential breeding pair are shown. It is recommended that the user select small numbers of samples when running this tool.

Both the simulate matings and the choose breeders tools can be used to make informed decisions about which breeding pairs to select as well as to understand the level of residual heterozygosity that will most likely be present in the next generation of a line.

5.1.4 Compute Union

Once a CC line reaches a certain level of residual heterozygosity, it is considered a distributable line and made publically available for other scientists to order online. A key step in determining whether a line has reached distributable or completed status is the identification of obligate ancestors in the line of all extant mice or subsets of extant mice with limited heterozygosity (Figure 5.7). Genotypes from MUGA or MegaMUGA are used for haplotype reconstruction as described previously[17]. The haplotype reconstructions of the obligate ancestors are jointly considered to determine the maximum heterozygosity of a distributable line (Figure 5.7b). Calculating the joint heterozygosity involves two steps: establishing recombination breakpoints and determining segregating regions within and between the obligate ancestors (Figure 5.8). Recombination breakpoints are estimated by the midpoint of any ambiguous region found by the haplotype reconstruction (these tend to be no longer than 2-3 SNPs). Ambiguous heterozygous regions within an ancestor begin and end at the closest heterozygous genotype call. Ambiguous heterozygous regions between ancestors begin and end at the nearest informative genotype. When genotype calls are consistently inconsistent with the intensity-based founder assignment[17], I assume that this is a feature of the line's haplotype and treat the region as fixed. I then compute the genomic length of all segregating regions divided by the full genomic length to determine the maximum residual heterozygosity within a line. If lines have reached the

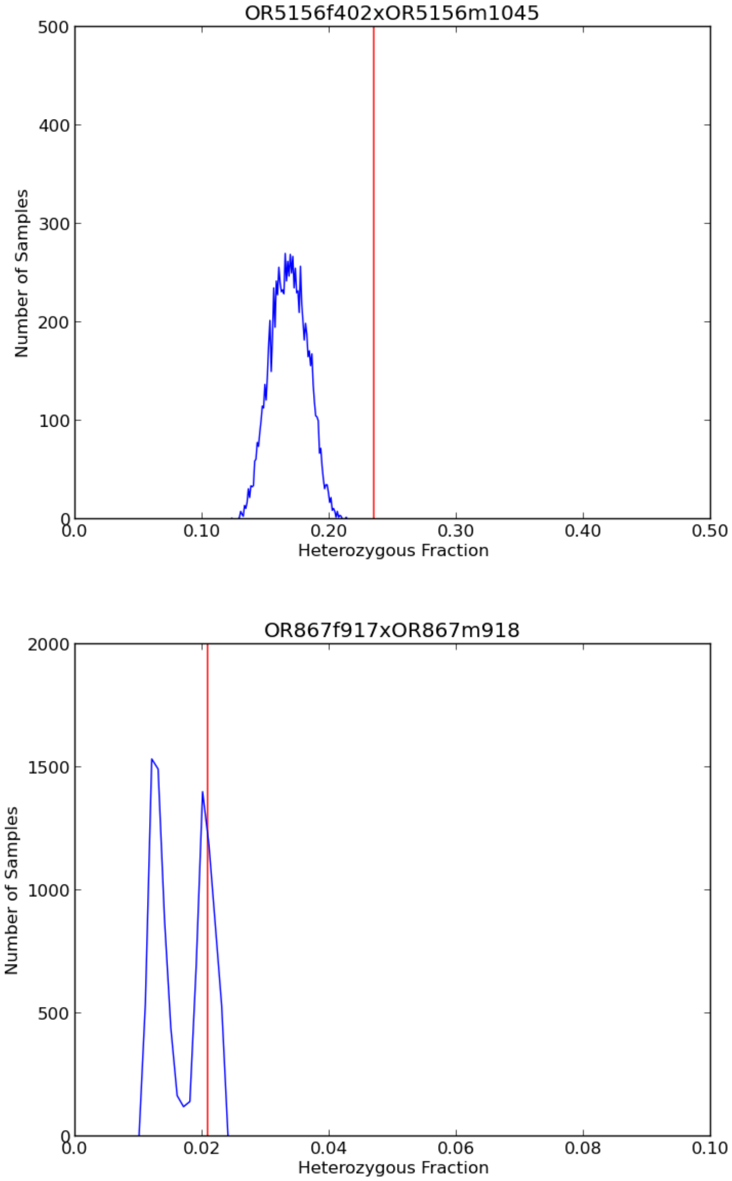


Figure 5.6: Simulate Matings tool output. Residual heterozygosity fraction of 10,000 simulated offspring of a given set of breeders (blue line) compared with the combined residual heterozygosity of the breeding pair (red line).

required thresholds, I generate a special haplotype file for the entire distributable line indicating regions fixed for a specific CC founder and regions that are still segregating and between which CC founders they are segregating (Fig 5.7c). All computations are based on Mb distances of the NCBI m37 version of the mouse assembly. The haplotype assignments for each line can be visualized (as shown in Fig 5.7c) or downloaded as text files (as shown in Figures 5.8 and 5.9) from the UNC Systems Genetics Core web site (Figure 5.9).

The genomes of the Collaborative Cross lines are available at http://csbio.unc.edu/CCstatus/. The menu bar offers links to the CC resource ("CC Mice") and specific pages for information on "Available Lines", "Ordering", and the "CC Viewer".

5.2 Conclusion

To implement the MAI techniques described in Chapter 3 on live mice a number of tools were needed. The first of these tools is a way to receive relatively cheap ($100/sample) genotype information about the mice. These genotypes needed to be genome-wide and be able to capture the vast majority of the residual heterozygosity within a sample. This led to the development of first MUGA and then MegaMUGA, as described in Chapter 4. Once there was a fast, cost efficient way to learn about the underlying genomes of the mice in the CC, it was then possible to build a series of tools on top of these genotyping platforms to facilitate the implementation of the MAI techniques. Since quality control is very important with any experiment, the first few tools focus on making sure that samples are labeled correctly, that the array data is good for each sample, and that the underlying genome of each sample is reasonable for the type of mice that are being bred. In order to efficiently analyze the mice, it is necessary to first compute a haplotype reconstruction of each sample. This creates a uniform data structure that can be used in all further analysis; the haplotype reconstruction data is more efficient to analyze as it describes the data in terms of founder segments rather than needing to go back to the probe level for each analysis. This means that instead of using 7,851 different data points for MUGA, and 77,808 data points

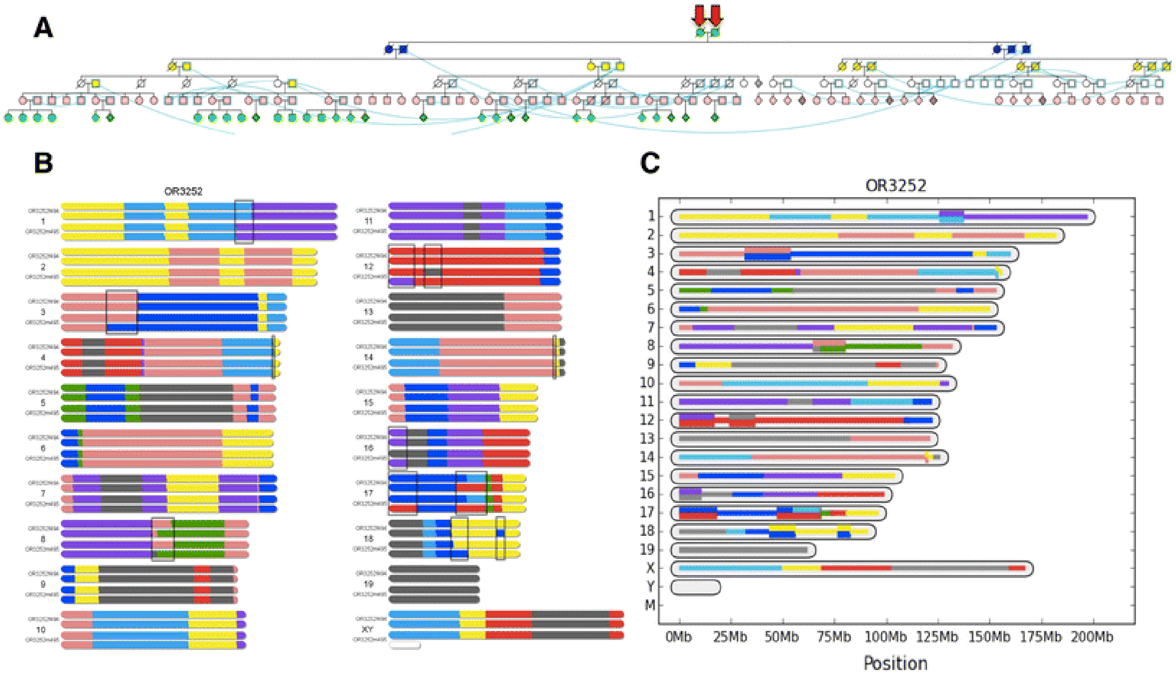


Figure 5.7: **a** Partial view of the pedigree of the OR3252 CC line. Mice are represented using standard symbols for human pedigrees. Mice that are present multiple times (because they participate in multiple matings) are linked by blue curved lines. Colors represent different generations of inbreeding. Mice shown at the top of the pedigree with arrowheads are the obligate ancestors of this line used to determine whether it passes the threshold for distribution (most recent obligate ancestors). **b** Genome of obligate ancestors based on MUGA genotypes. We use standard colors and a single-letter code to represent the contribution of the eight CC parental strains [17] to the genome of the two most recent obligate ancestors. Briefly, A/J, A, yellow; C57BL/6J, B, gray; 129S1/SvImJ, C, pink; NOD/ShiLtJ, D, dark blue; NZO/H1LtJ, E, light blue; CAST/EiJ, F, green; PWK/PhJ, G, red; and WSB/EiJ, H, purple. The two autosomes and the corresponding complement of X chromosomes for each mouse are drawn to illustrate the regions that are fixed (all four autosomes or three X chromosomes have the same haplotype) or segregating (shown in boxes). **c** The genome of the OR3252 line. The figure represents fixed regions as single lines and segregating regions as double lines of the appropriate colors.

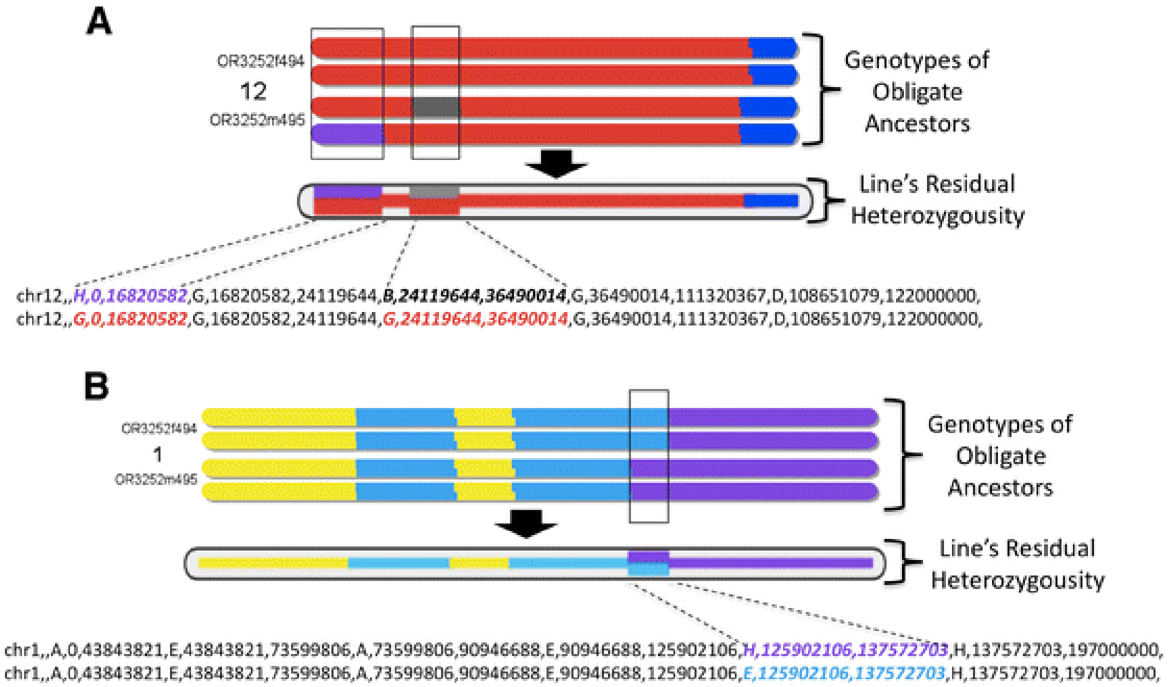
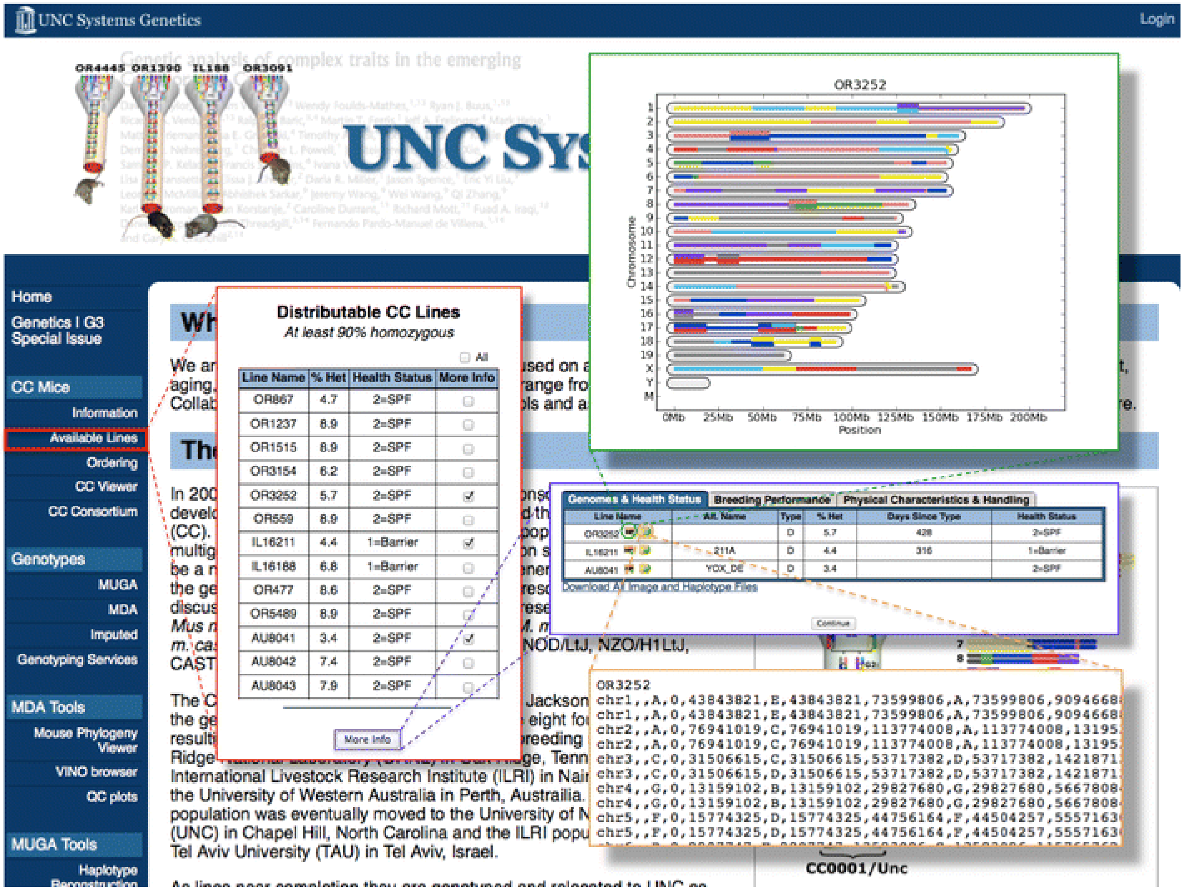

A
OR3252f494
12
OR3252m495
Genotypes of Obligate Ancestors
Line's Residual Heterozygousity
chr12,,H,0,16820582,G,16820582,24119644,B,24119644,36490014,G,36490014,111320367,D,108651079,122000000,
chr12,,G,0,16820582,G,16820582,24119644,G,24119644,36490014,G,36490014,111320367,D,108651079,122000000,
B
OR3252f494
1
OR3252m495
Genotypes of Obligate Ancestors
Line's Residual Heterozygousity
chr1,,A,0,43843821,E,43843821,73599806,A,73599806,90946688,E,90946688,125902106,H,125902106,137572703,H,137572703,197000000,
chr1,,A,0,43843821,E,43843821,73599806,A,73599806,90946688,E,90946688,125902106,E,125902106,137572703,H,137572703,197000000,

Figure 5.8: Residual heterozygosity in distributable lines. The figure shows two chromosomes from line OR3252 (shown in Figure 5.7) to illustrate the identification of segregating regions in distributable lines. The figure follows the conventions detailed in Figure 5.7, with the top part of each subheading representing the contribution of the eight CC parental strains to the genome of the two most recent obligate ancestors and the midsection and lower sections representing the haplotypes of the line as provided in the CC website as figure or as text, respectively. **a** Chromosome 12 illustrates two segregating regions in which one of the most recent ancestors is homozygous while the other is segregating. The figure also illustrates that in some cases the most recent obligate ancestors may appear to have slightly different boundaries between parental contributions. We suggest that investigators rely on the haplotype reconstruction provided in the text file rather than on visual inspection of most recent ancestors. We expect these discrepancies to be resolved in the near future with use of MegaMUGA. **b** Chromosome 1 illustrates a segregating region in which each of the most recent ancestors' parents was homozygous for a different parental allele.



Distributable CC Lines

At least 90% homozygous

□ All

Line Name	% Het	Health Status	More Info
OR867	4.7	2=SPF	□
OR1237	8.9	2=SPF	□
OR1515	8.9	2=SPF	□
OR3154	6.2	2=SPF	□
OR3252	5.7	2=SPF	☑
OR559	8.9	2=SPF	□
IL16211	4.4	1=Barrier	☑
IL16188	6.8	1=Barrier	□
OR477	8.6	2=SPF	□
OR5489	8.9	2=SPF	□
AU8041	3.4	2=SPF	☑
AU8042	7.4	2=SPF	□
AU8043	7.9	2=SPF	□

More Info

Genomes & Health Status | Breeding Performance | Physical Characteristics & Handling

Line Name	Alt. Name	Type	% Het	Days Since Type	Health Status
OR3252		D	5.7	428	2=SPF
IL16211	211A	D	4.4	316	1=Barrier
AU8041	YOX_DE	D	3.4		2=SPF

Download All Image and Haplotype Files

Continue

```
OR3252
chr1,,A,0,43843821,E,43843821,73599806,A,73599806,9094668
chr1,,A,0,43843821,E,43843821,73599806,A,73599806,9094668
chr2,,A,0,76941019,C,76941019,113774008,A,113774008,13195
chr2,,A,0,76941019,C,76941019,113774008,A,113774008,13195
chr3,,C,0,31506615,C,31506615,53717382,D,53717382,1421871
chr3,,C,0,31506615,D,31506615,53717382,D,53717382,1421871
chr4,,G,0,13159102,B,13159102,29827680,G,29827680,5667808
chr4,,G,0,13159102,B,13159102,29827680,G,29827680,5667808
chr5,,F,0,15774325,D,15774325,44756164,F,44504257,5557163
chr5,,F,0,15774325,D,15774325,44756164,F,44504257,5557163
```

International Livestock Research Institute (ILRI) in Nair
the University of Western Australia in Perth, Austrailia.
population was eventually moved to the University of N
(UNC) in Chapel Hill, North Carolina and the ILRI popu
Tel Aviv University (TAU) in Tel Aviv, Israel.

Figure 5.9: The UNC Systems Genetics Core web site. Screenshots of the main pages associated with the distribution of CC lines are shown. The “Available Lines” tab is highlighted on the left side of the web site as well as inserts of the pages associated with information on the number, genome, and characteristics of the available CC lines. Links from the menu take you to the ordering page and the CC viewer.

for MegaMUGA, only about 80-120 genomic segments needed to be analyzed and compared.

Once haplotype reconstructions are computed for the samples a user can choose breeders from all genotyped samples within a line, simulate matings to determine an approximate level of residual heterozygosity after another generation of inbreeding, compute the union of a set of samples to determine maximum residual heterozygosity in a line, as well as learn more about the underlying genomic structure of each sample. These tools allow mouse technicians to utilize the MAI techniques within the lab with the flexibility necessary to ensure that the selected breeders are still viable and will create the results that the experiment hoped to achieve.

This chapter described the tools necessary to implement MAI techniques on real mice. As one might notice, despite the haplotype reconstructions being an integral piece of this framework, this input does not need to come from any particular algorithm or even from genotyping data. Haplotype reconstructions simply need to assign the most likely founder call to each genomic region. As high-throughput sequencing technology becomes more cost efficient, it is plausable that it may become more cost efficient to sequence the CC mice rather than genotype them. Therefore, in Chapter 6 I describe a technique for using high-throughput sequencing data to derive haplotype reconstructions.

CHAPTER6: HIGH-THROUGHPUT SEQUENCING DATA

High-Throughput Sequencing (HTS) of short reads is rapidly becoming cost competitive with full-genome genotyping using microarrays. A key difference between HTS and microarray genotyping is that microarrays sample specific genomic locations, whereas HTS samples the genome randomly. Categorizing genetic differences in HTS data requires a database of known sequence variants, while microarray-based genotyping is based on a set of reliable variants that were selected previously as part of the array's design. A common application of full-genome genotyping is to determine the ancestral origin of genomic segments arising from recombination. In this chapter, I contrast the resolution and accuracy of determining recombination boundaries using genotyping microarrays with HTS. In addition, I consider the impacts of sequence coverage and genetic diversity on localizing recombination boundaries.

As discussed in previous chapters, the genomes of the Collaborative Cross (CC)[17] have been monitored throughout its development. This is being done to ascertain the level of heterozygosity in various developing RILs as well as to decrease the number of generations of inbreeding required to achieve fully inbred animals[60]. Two different genotyping arrays have been used to monitor the CC genomes, which were designed specifically to be informative for the CC[64, 17]. As discussed in Chapter 5, for each of these genotyping platforms, algorithms have been designed to assign founders and estimate recombination breakpoints[36, 25]. Versions of these founder assignment algorithms have been demonstrated to work on a number of different mouse resources, including the Diversity Outcross (DO) [55] and other outbred populations.

Recently others have considered using HTS technologies to determine ancestral origins[45] and have also used sparse sequence data for this same analysis[50]. Sequencing data from four pooled samples were used to establish that the genetic variants and haplotypes of commercial outbred mouse stocks are largely shared with common laboratory strains[62]. I perform a similar analysis with eight-founder CC RILs, which leverages high-throughput sequencing data for three samples from nearly inbred CC lines (OR867m532, OR1237m224, and OR3067m352) that have also been previously genotyped on two genotyping platforms (MUGA and MegaMUGA as described in Chapter 4).

Beissinger et al.[5] addresses determining the necessary read coverage needed to genotype-by-sequencing in order to perform Quantitative Trait Loci (QTL) mapping. In this chapter, I perform a similar determination of the read coverage necessary to map recombination breakpoints and compare this resolution to that obtained using the genotyping arrays described in Chapter 4. I do this analysis using the same three CC lines mentioned previously and sampling the reads to simulate various coverage levels. This chapter validates the accuracy of the haplotype reconstructions described in Chapter 5 and describes a technique for creating haplotype reconstructions using high-throughput sequencing data.

6.1 Sequence Data

Whole-genome sequencing for three extant CC lines was completed by the Washington University School of Medicine Genome Sequencing and Analysis Center using Illumina sequencing technology with 30x haploid coverage. DNA was extracted from the spleen of a single male sample from each of the three extant CC strains. The resulting 100 base pair paired-end sequence fragments were aligned to a CC-specific consensus reference genome using Bowtie (v 2.0.5)[33, 32]. The consensus genome was created by inserting the majority allele of the eight CC founders at all known variant positions into the NCBI37/mm9 mouse genome[15]. The genetic variants were provided by the Wellcome Trust/Sanger Institute's Mouse Genome's Project[30]. I applied my techniques to these three extant lines since MUGA and MegaMUGA genotypes and sequencing data existed for all three samples.

6.2 HMM Algorithm

A Hidden Markov Model (HMM) algorithm is used to determine the founder mosaic for the three sequenced animals. Since CC animals have eight founders and each loci can be heterozygous or homozygous, my HMM has thirty-six possible states (eight inbred and twenty-eight founder-pair combinations). To help alleviate some of the noise inherent in sequencing data, I binned the genome into uniform sized genomic windows, so that each bin would contain sufficient evidence to discriminate between thirty-six possibilities using primarily biallelic variants. I then used a standard Viterbi algorithm to solve for the most likely founder mosaic represented in the HMM as described below.

6.2.1 Variants

A database of 65 million variants in 17 laboratory mouse strains has recently been produced by the Wellcome Trust/Sanger Institute[30]. They included the eight Collaborative Cross founder strains. Of these 65M SNPs, 31M high-confidence SNPs are informative among the eight CC founders. The majority allele at each of these 31M SNPs was used to construct the consensus genome used for alignment. I further filtered these down to a subset 29M SNPs such that there are no unknown genotypes among all eight founders, eliminating any need for imputation.

6.2.2 Emission Probabilities

The aligned reads were then examined at each of these 29M SNP positions and binned using uniform-sized non-overlapping bins. The bin size is a user specified parameter which should be set based on the amount of genetic diversity between the founders. Unless otherwise specified, it was set to 1000bp in this chapter. Within each bin, emission probabilities are computed for each of the twenty-eight heterozygous founder-pair combinations and the eight inbred founders by counting the number of variants consistent with each of the thirty-six possible states, as shown in Figure 6.1. Counts for each of the thirty-six states were converted to a likelihood score based on the number of reads supporting each genotype call, and subsequently adjusted to compensate for the likelihood of the same counts occurring by chance as modeled by a binomial distribution. A noise model of 1 sequencing error per 100 sequenced bases was assumed, so that the binomial distribution of a homozygous call is 0.99, while the assumed split for a heterozygous call is 0.495. Three possible probabilities (homozygous for each allele and heterozygous) are calculated at each SNP based on the number of reads that supports each allele, and applied appropriately to each bin. The probabilities of all SNPs in a bin were combined, and then the values are normalized so that the sum of all probabilities in the thirty-six states sum to 1. When there are no SNPs or no reads present in a bin, the emission probabilities are assumed to be equal for all thirty-six states.

I also reweighted the informativeness of each bin based on the average number of reads and the total number of SNPs within each bin modeled as a Poisson distributed random variable, as shown in the formulas below, where R_{avg} is the average number of reads in all bins, R_{std} is the standard deviation of reads in all bins, and h_R is the number of reads in the current bin. Similarly, N_{avg} is the average number of SNPs per bin, N_{std} is the standard deviation, and h_N is the current bin count of SNPs.

$$\alpha_1 = e^{-\frac{(h_R - R_{avg})^2}{R_{std}}} \tag{6.1}$$

$$\alpha_2 = e^{-\frac{(h_N - N_{avg})^2}{N_{std}}} \tag{6.2}$$

$$\alpha = min(\alpha_1, \alpha_2) \tag{6.3}$$

$$P_{s'} = P_s * \alpha + \frac{1}{36} * (1 - \alpha) \tag{6.4}$$

This was done so that bins with a large number of reads (typical of highly repetitive regions of the genome) and bins with a small number of SNPs would not overly influence the solution. Parameters R_{avg}, R_{dev}, N_{ave} and N_{dev} are based on the reads and SNPs per bin for each given

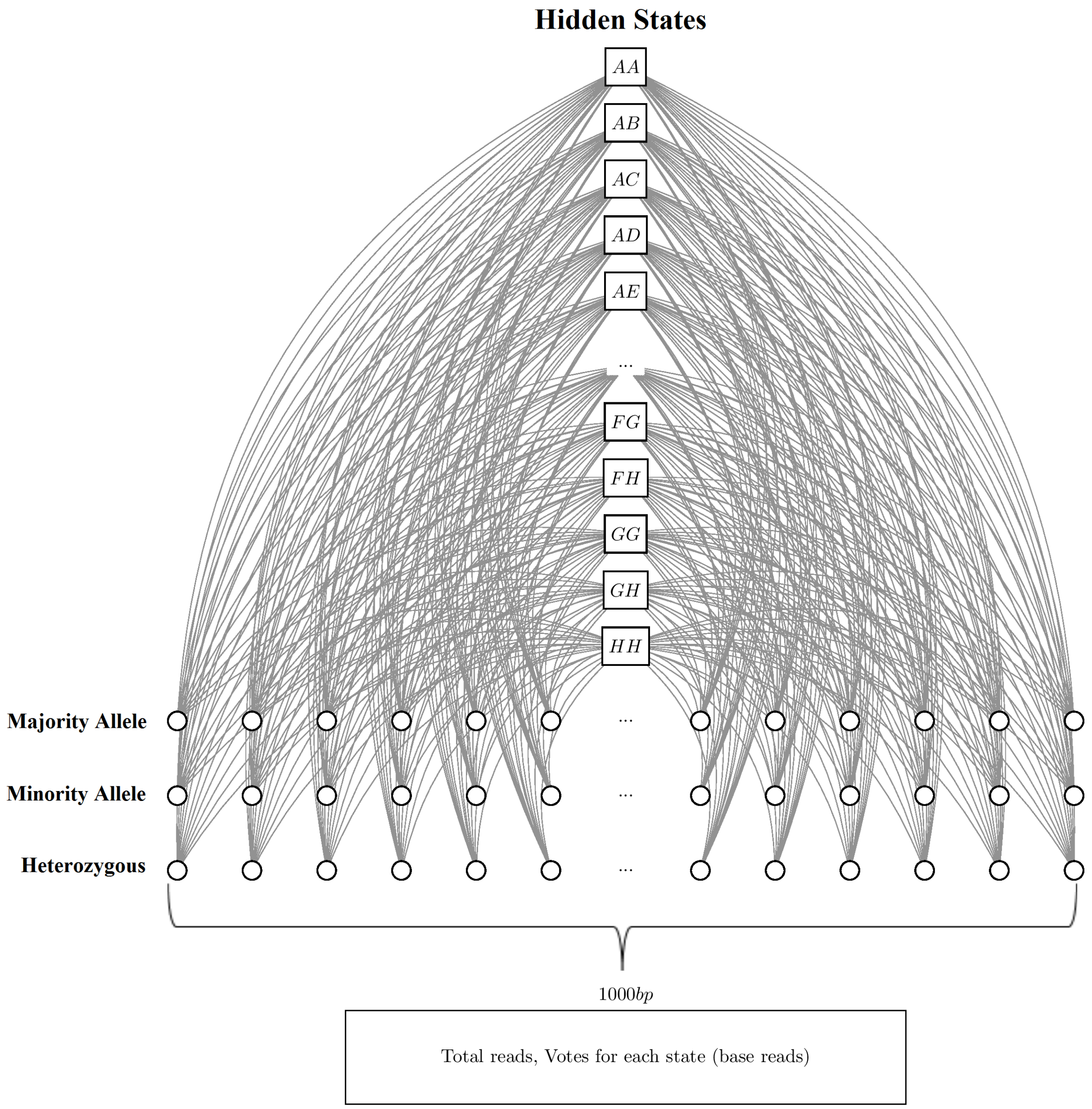


Figure 6.1: Calculating the emission probabilites for the HMM is done by first binning the genome into uniform-sized non-overlapping bins, and was set to 1000bp in this experiment. Within each bin, emission probabilities are computed for each of the twenty-eight heterozygous founder-pair combinations and the eight inbred founders by counting the number of variants consistent with each of the thirty-six possible states. Counts for each of the thirty-six states were converted to a likelihood score based on the number of reads supporting each genotype call, and subsequently adjusted to compensate for the likelihood of the same counts occurring by chance as modeled by a binomial distribution.

data set.

6.2.3 Transition Probabilities

The transition probabilities for the HMM are estimated based on observed recombinations seen in previous MUGA haplotype reconstructions for 350 unique, emerging CC lines[61]. There are four classes of transitions that can occur between states, as shown in Figure 6.3. The most likely class of transition is that the state remains the same between two adjacent bins. This is because over a genome of about 2,470 Mb, I observed an average of 100 recombinations among the 350 genotyped CC samples when founders were assigned using the intensity-based algorithm described by Fu et al[25]. A similar number of recombinations were found using the Liu et al[36] algorithm as reported by Fu et al[25]. Another class of transitions occurs when a recombination on one chromosome generates a heterozygous state, or when a recombination on a single chromosome causes a transition from one heterozygous state to another. The homozygous to heterozygous transitions appear in two versions: either the homozygous founder is included in the heterozygous state (more likely) or the transition from a homozygous to a heterozygous state involves simultaneous transitions on both chromosomes. The heterozygous to heterozygous states have two variants as well, such that either 1 or 0 of the founders remain the same between the two states.

To determine the transition probabilities, I used the observed recombinations from 350 independent CC lines with varying levels of residual heterozygosity ranging from 0.21% to 66.96%, with an average residual heterozygosity of 25.38% [17]. Based on these observed recombinations, I calculated the expected transition probabilities at a specified bin size. I assumed that 100 bins on average should contain a transition, and the rest should maintain the same state between consecutive bins. Therefore, the probability of remaining the same is (total bins - 100) / total bins. Of the 100 transitions, I observed that 41.85% of them are between a homozygous

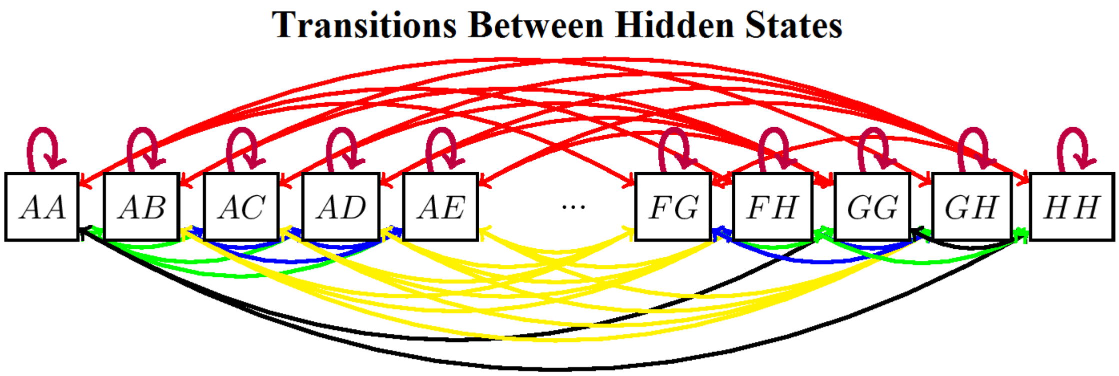


Figure 6.2: Transition probabilities for the HMM. Based on observed recombinations from 350 independent CC lines, I calculated the expected transition probabilities at a specified bin size. I assumed that 100 bins on average should contain a transition, and the rest should maintain the same state between consecutive bins. Therefore, the probability of remaining the same is (total bins - 100) / total bins (purple self loops). Of the 100 transitions, I observed that 41.85% of them are between a homozygous state and a heterozygous state that contained the homozygous states founder (green), 37.14% were between two different homozygous states (black), 17.92% were between heterozygous states that share a founder(blue), and the remaining 2.89% was between a homozygous state and a heterozygous state (red) with no shared founder.

state and a heterozygous state that contained the homozygous state's founder, 37.14% were between two different homozygous states, 17.92% were between heterozygous states that share a founder, and the remaining 2.89% was between a homozygous state and a heterozygous state with no shared founder (Figure 6.2).

6.2.4 Viterbi Solution

I initially assume that all states were equally likely and set the priors to reflect that. The Viterbi algorithm then proceeds to find the path maximizing the sum of log-likelihoods, thus computing the most probable sequence of founder assignments. This process is repeated for each chromosome independently.

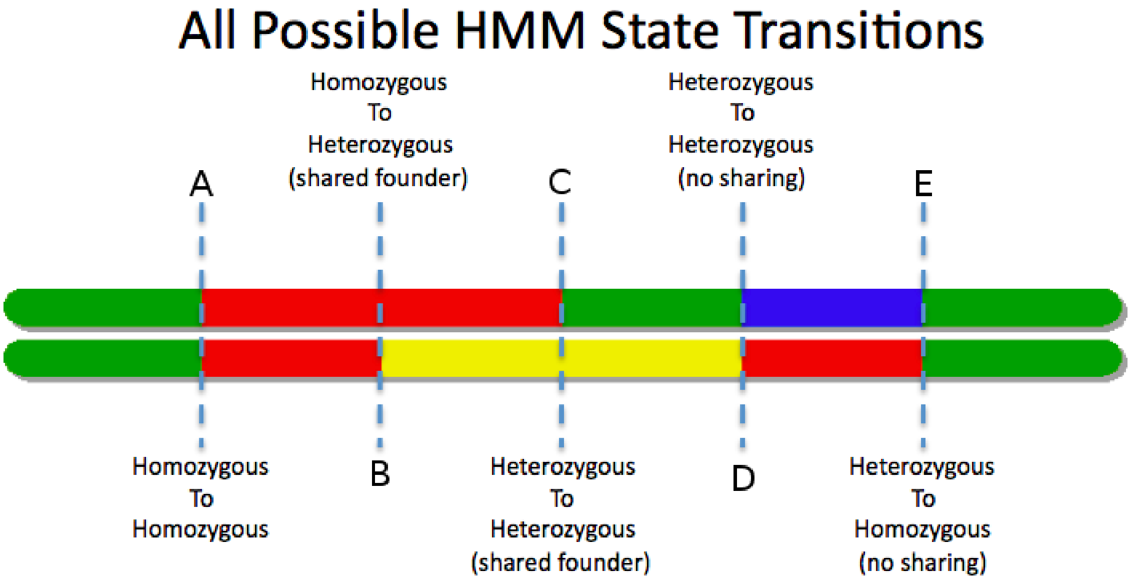

All Possible HMM State Transitions
Homozygous To Heterozygous (shared founder)
Heterozygous To Heterozygous (no sharing)
A
C
E
Homozygous To Homozygous
B
Heterozygous To Heterozygous (shared founder)
D
Heterozygous To Homozygous (no sharing)

Figure 6.3: There are four classes of transitions that can occur between HMM states. The most likely transition is to remain the same founder state between two adjacent bins. In inbred animals, a shared recombination on both chromosomes generates the typical homozygous to homozygous transition (A). Another class of transitions occurs when a recombination on one chromosome transitions a homozygous state to heterozygous state (B), or causes a transition from one heterozygous state to another (C). Rare heterozygous to heterozygous transitions occur as a result of recombinations on both chromosomes (D) and are usually due to a recombination hotspot. Likewise, rare transitions from heterozygous to homozygous states can result from two aligned but separate recombinations (E).

6.3 Founder-pair Resolution

To separate the degree to which resolving recombination boundaries depends on sequencing depth versus sequence similarity between the two sequences on either side of the recombination event, I developed a pairwise sequence similarity map. Sequence similarity varies throughout the genome and serves as a fundamental limit in my ability to resolve recombination boundaries. No amount of additional read coverage can improve the localization of a recombination breakpoint beyond the resolution determined by a sequence similarity map. To measure the accuracy of a given recombination boundary estimate, it is necessary to factor in the extent to which genomic variations exist near the region in question. A sequence similarity map provides such a gauge. It can also be used to normalize accuracy measures of recombination breakpoint positions.

Sequence similarity maps were constructed between all twenty-eight founder pairs and are available in Appendix A. Figure 6.4 shows visualizations of three of these sequence similarity maps. These images depict the number of 1000bp bins that have at least one informative SNP within each 100Kb bin. The sequence similarity map indicates where in the genome there are few or no informative SNPs distinguishing a particular founder-pair. The frequency of informative SNPs in a genomic region places a fundamental limit on the resolution for which a recombination breakpoint can be mapped using sequence information, regardless of read coverage. The density of informative SNPs varies significantly between founder pairs from the CC. The three wild-derived CC founder strains (CAST/EiJ, PWK/PhJ, and WSB/EiJ), include many variants and result in very few regions of ambiguity (areas with few or no informative SNPs) when they participate in a founder-pair. However, there is considerable sequence similarity among the five classical CC founder strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, and NZO/H1LtJ) and therefore there are many regions of ambiguity in founder-pairs involving two classical CC strains. Figure 6.4 depicts three sequence-similarity maps, one between two classical strains, a second between a wild-derived strain and a classical strain and a third between two wild-derived strains. Also shown is the distribution of distances between informative SNPs genome-wide for all twenty-eight founder-pairs, which peaks at 512 base pairs, thus justifying my choice of bin size (i.e. most 1000 base pair bins are likely to include a informative SNP variant between most founder pairs). The sequence similarity maps also depict regions of the genome where there are few annotated variants due to lack of sequence complexity, such as the large gaps on chromosomes 7, 12, and X. In these regions I would also expect to be limited in capability to resolve recombination breakpoints. These sequence similarity maps are used to assess the possible localization accuracy of a specific recombination event as determined by experiments with variable read coverage.

6.4 Breakpoint Mapping of HTS data

I considered the recombination breakpoint mapping accuracy attainable from the full 30x coverage sequence data. Accuracy depends both on sampling density and the genetic diversity between the founders surrounding each breakpoint. My HMM solution pools evidence within regions of a user specified window size (1000 bases for 30X coverage) to infer the most likely source of the genome within a window. HMM transitions, which are suggestive of a recombination breakpoint, occur between window boundaries.

6.4.1 Comparison with Refined Breakpoint Solution

The HMM solution at best determines a recombination boundary to the resolution of a bin (typically 1Kb). In a post process, I utilize all informative SNPs between the most likely two founders identified on each side of the recombination by the HMM solution to refine the recombination breakpoints down to the distance between two consecutive informative SNPs. This becomes more complicated in regions of high sequence similarity, leading to regions where the resolution of the recombination boundary depends on the pair of founders on each side of the event. However, in most cases, I was able to determine the two informative SNPs between which the recombination occurs and these SNP positions are then used to bound the recombination

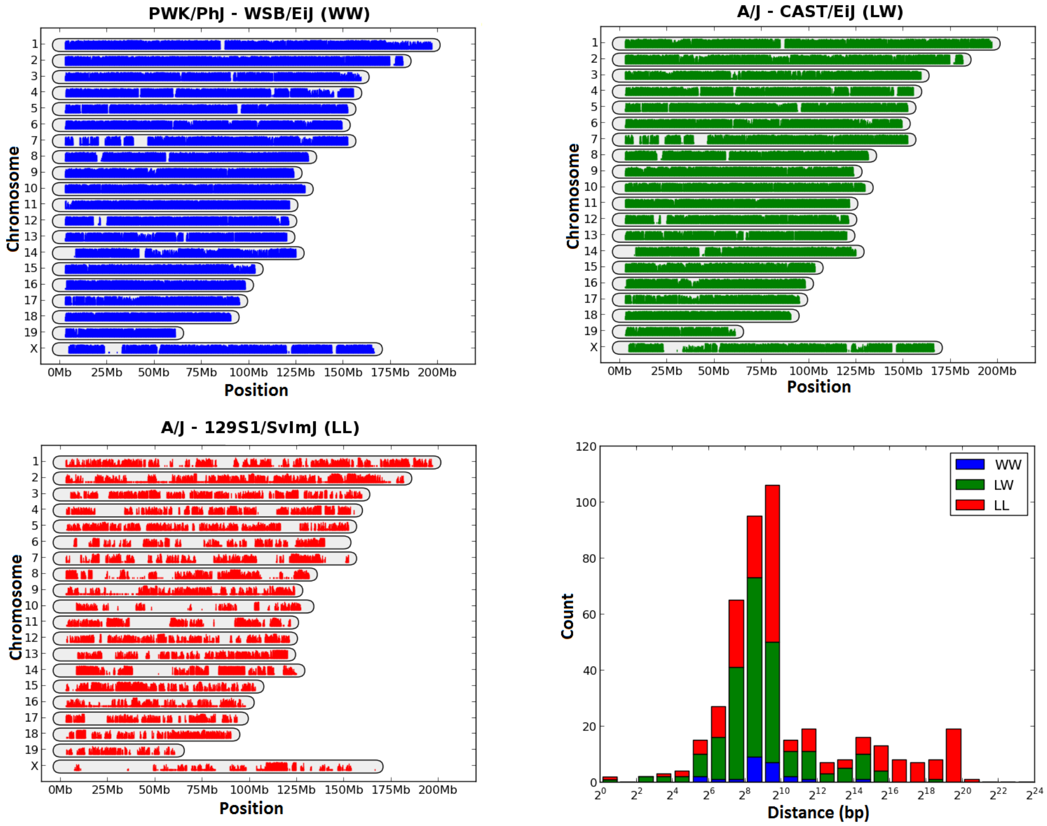


Figure 6.4: Sequence similarity maps for three founder-pairs and histogram of spacing between informative SNPs for all founder-pairs. The red, green, and blue subplots illustrate the percentage of 1000 base pair bins within a 100 kilobase window for which there is at least 1 informative SNP distinguishing the founder pair. Because CC founders fall into two categories, classical lab strains and wild-derived lab strains, there are three possible categories of founder-pair combinations. WW, shown in blue, occurs when both founders are wild-derived. These founder-pairs typically have low sequence similarity and many informative variants as seen by the relative density of the blue plot. LW, shown in green, occurs when one founder is wild-derived and the other is a classical lab strain. These founder-pairs also typically have many informative SNPs, but less than WW pairs. LL, shown in red, is an example where both founders are classical lab strains. These combinations typically have significantly more sequence similarity. As shown in the red plot there are many 1000 bp bins with no informative SNPs. Regions shown in white on these three sequence similarity maps indicate areas of the genome that will be difficult to detect recombinations between these founder-pairs. The stacked histogram plot shows the distance between all informative SNPs genome wide. It is divided into the 3 founder-pair categories to illustrate the larger distances between informative SNPS in red founder-pairs, as compared to the distances between informative SNPs for the blue and green founder-pairs.

event. Where these informative SNPs are far apart, areas of uncertainty are drawn and I assume that the founders are Identical-By-State (IBS) within the determined interval.

I refined my HMM estimates by expanding the region surrounding each transition, and then consider only the informative markers between the two founders identified on each side of the breakpoint. Generally, there is a clear transition where every marker distal to a boundary marker is consistent with one founder and every marker proximal to a second boundary marker is consistent with the other founder. For most recombinations, I was able to find two consecutive informative SNPs that were obviously on different sides of the recombination breakpoint. The actual breakpoint is most likely to have occurred between these two SNPs. I then found the distance between each of these flanking informative SNPs and the HMM solution at each recombination for each of the three samples. Figure 6.5 depicts a histogram of these distances for each recombination.

Next I analyzed the mapping accuracy of the HMM solution relative to the refined informative-marker solution. For the three samples given, the HMM transition occurred at a median distance of 527 base pairs from the midpoint of the surrounding informative markers, with the first quartile falling 284 base pairs from the median, and the third quartile falling 899 base pairs from the median. A summary histogram of the distance of my HMM solution from the refined solution is shown in Figure 6.5. This histogram shows that the majority of the breakpoints were actually in the bin that the HMM transitioned, but there were some instances where there were no informative SNPs and the breakpoint estimation could not be narrowed down to within 2Mb of the HMM solution. In 61.8% of the recombinations, the HMM solution fell between the informative SNPs, while 18.6% transitioned before the informative SNP pair and 19.5% transitioned after. Transitions that occured before the informative SNP pair tended to occur within a median distance of 546 base pairs, while transitions that occurred after, were a median distance of 233 base pairs.

I estimated the precision of the recombination-breakpoint localization using the gap spacing

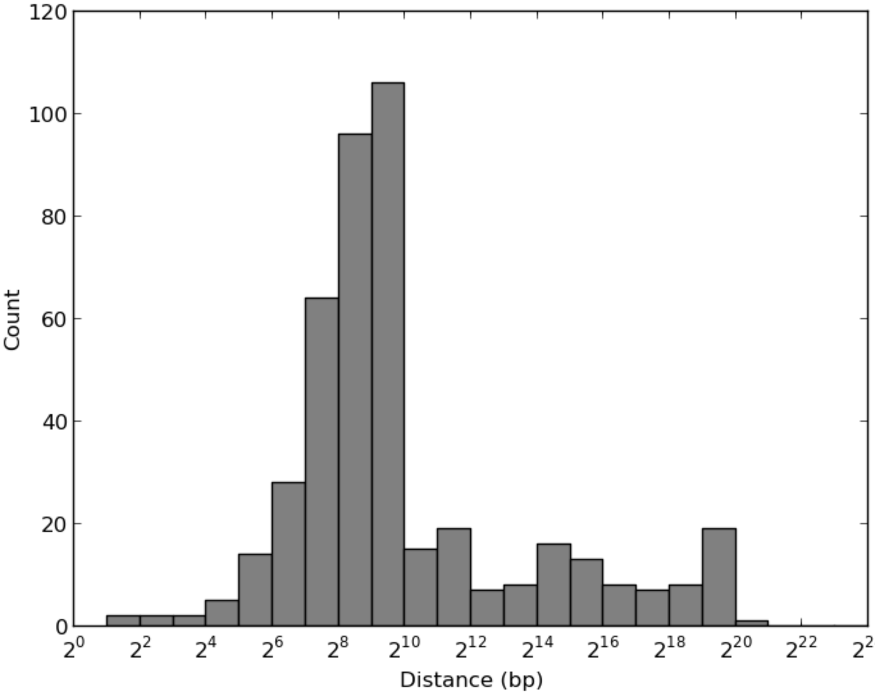


Figure 6.5: Histogram of distance between HMM solution and refined recombination breakpoint. The HMM solution can at best determine a recombination breakpoint to the resolution of a bin, which in this case is 1Kb. In a post process, I further refine these breakpoints by searching for informative SNPs within the region of the transitions and determining between which two consecutive SNPs the breakpoint actually occurs. I calculate the distance between each of these SNPs and the HMM solution and plot a histogram of the frequency at which each distance occurs. The high peak at 1Kb and the large number of distances less than 1Kb indicate that the HMM solution is typically within the range of the informative SNPS.

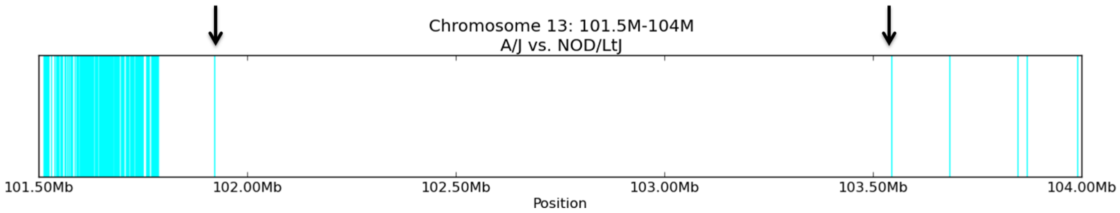


Figure 6.6: Depiction of all informative SNPs between A/J and NOD/ShiLtJ on Chromosome 13 from 101.5Mb to 104.0Mb. Informative SNPs are shown in cyan and arrows are used to depict the beginning and end of the ambiguous region between these two strains, since there are no informative markers between 101.9Mb and 103.5Mb.

between the two informative markers of the refined solution. Over the 220 detected recombinations, I was able to localize each to a median region of 1,022 bases with the first quartile falling within 749 bases of the median, and the third quartile falling within 26,412bp of the median. The closest that a recombination breakpoint was determined was 5bp between strains NZO/H1LtJ and PWK/PhJ on Chromosome 1 around 29Mb in sample OR867m532. The poorest precision that I could assign an observed breakpoint was to 1,623,010 bases between A/J and NOD/ShiLtJ on Chromosome 13 between 101.9Mb and 103.5Mb. This poor mapping is consistent with the sequence similarity map for A/J and NOD/ShiLtJ [65], as is shown in Figure 6.6, where all informative SNPs between A/J and NOD/ShiLtJ are shown in cyan and arrows depict the start and end of this ambiguous region. A summary histogram of distance between recombination breakpoint informative SNPs is shown in Figure 6.7.

Throughout the rest of the analysis I use the full-coverage HTS solution with refined breakpoints as the standard with which to evaluate alternative genotyping approaches and lower-coverage solutions.

6.4.2 Comparison to Genotype Solutions

Next I compared the recombination breakpoints determined from the whole-genome sequence data to the breakpoints estimated from the 7K MUGA and 77K MegaMUGA genotyping platforms. Given the relatively low sampling density of microarray based genotyping when

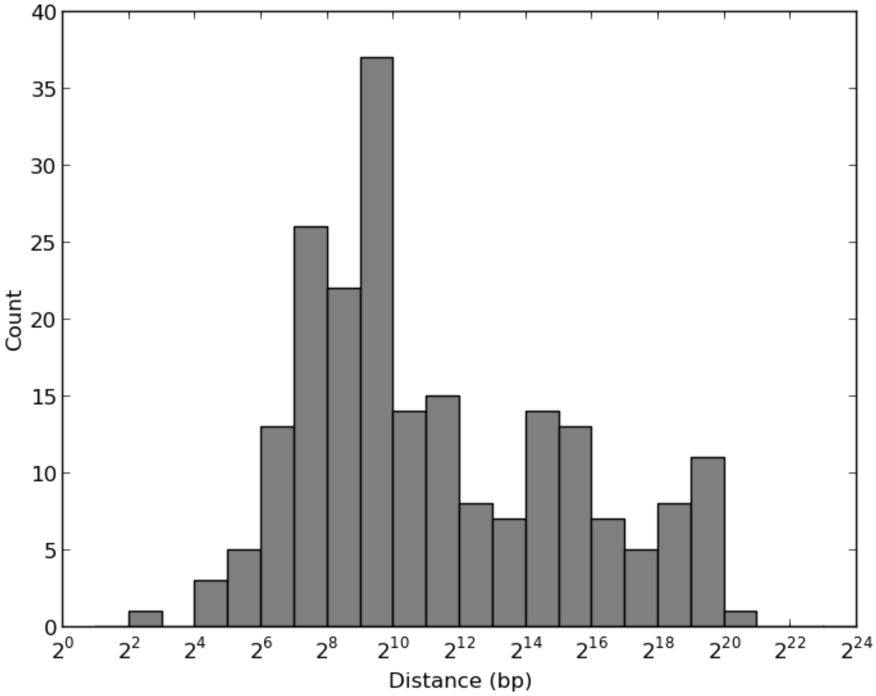


Figure 6.7: Histogram of distance between informative SNPS in the refined breakpoint solution. Starting from each HMM transition, I found the two consecutive SNPs informative for the different founders on each side of the breakpoint as determined by maximum likelihood founder-pair. The separation between these informative SNPs was used to compute the histogram shown. I consider the true recombination breakpoint to have occurred somewhere between these two SNPs. This indicates that the precision of recombination breakpoint mapping varies from a few bases to over a megabase in dense read coverage.

	Number of Intervals			Concordance with HTS	
Sample	HTS	MUGA	MegaMUGA	MUGA	MegaMUGA
OR867m532	117	108	117	95.56	98.12
OR1237m224	116	102	115	95.97	98.47
OR3067m352	112	102	112	96.76	98.93

Table 6.1: Comparison of HTS to Genotype Solutions, showing both the number of intervals found using each algorithm as well as the concordance between the HTS solution and the genotyping solutions. The concordance is measured such that at every base pair in the genome, I find the total number of base pairs where the genotyping solution is the same as the HTS solution divided by the total number of base pairs genome wide.

compared to whole-genome sequencing, it is possible that some small genomic intervals (regions between two recombination breakpoints attributable to a single founder) can be missed entirely. The size of the minimum detectable genomic intervals was a design consideration for both MUGA and MegaMUGA. MUGA was designed to detect haploid founder intervals larger than 1Mb on average, whereas MegaMUGA was designed to detect both homozygous haploid or heterozygous diploid intervals larger than 160Kb on average. For the three samples, OR867m532, OR1237m224, and OR3067m352, MegaMUGA missed 1, 2, and 0 small genomic regions respectively. On OR867m532, MegaMUGA missed a 106K heterozygous region on Chromosome 8 from 19.68Kb - 19.79Kb, while on OR1237m224, it missed a 102Kb heterozygous region on Chromosome 8 from 19.68Kb-19.79Kb, and a 394Kb heterozygous region on Chromosome 11 from 97.50Kb-97.89Kb. On OR3067m352, there were no missing regions on the autosomes. The two missing heterozygous regions on Chromosome 8 of OR867m532 and OR1237m224 are in the same range, and examination of the sequence similarity maps shows that this region is adjacent to an area of very few informative SNPs for all founder-pairs (see Figure 6.4). MUGA solutions for the three samples tended to miss 7-11 intervals ranging in size from 102Kb - 3.1Mb.

A second aspect of recombination breakpoint accuracy is whether the two sequences on either side of the recombination breakpoint are consistent with the HTS predictions. MegaMUGA chose a different founder-pair in 2-3 intervals per sample and also had about 0-2 false positives (extra recombinations) per solution. The extra recombinations all occurred at the ends of chromosomes however, which is most likely explained by the high number of extra SNPs placed at the end of each chromosome on MegaMUGA. Since these extra intervals are very small, it is possible that one or two SNPs created the false recombination. MUGA results had only 1 false positive total among the three samples and it occurred at the beginning of a chromosome. MUGA also only chose a different founder-pair in two instances total for the three samples. On both MegaMUGA and MUGA, the regions with different founder-pair calls were in areas of high sequence identity between the HTS solution and the genotyping array solution.

The final aspect of comparison is the breakpoint accuracy, which applies only to genomic intervals that are both detected and whose genomic intervals have founders consistent with the whole-genome sequence solution on both sides. On average, MegaMUGA localized the recombination breakpoint to within 161Kb-320Kb while MUGA's breakpoints were within 820Kb-870Kb. Based on the resolution of the genotyping arrays, one would expect MUGA to be able to refine breakpoints to within 1Mb of the actual location and MegaMUGA to be within 160Kb on average. MUGA performed slightly better than anticipated on average, while MegaMUGA is not quite as good as expected at this point, but it is still 3 to 5 times more accurate than the MUGA platform it replaced. A comparison of the founder solutions for each of the three CC samples is shown in Figures 6.8, 6.9, 6.10 and Table 6.1. In Table 6.1, the concordance between the HTS solution and the genotype solution is measured such that for every base pair in the genome, I find the total number of base pairs where the genotyping solution is the same as the HTS solution divided by the total number of base pairs genome wide. Where one solution is found to be inbred and the other is heterozygous but includes the inbred solution, I consider this to be half right, and count it accordingly.

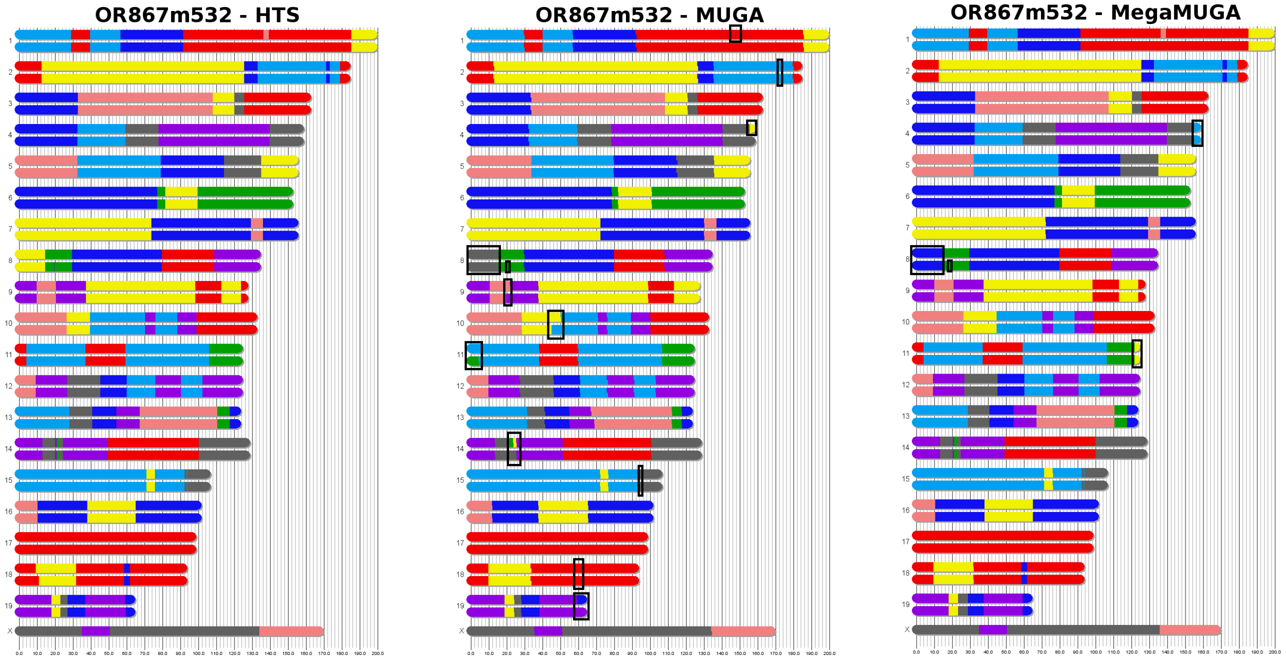
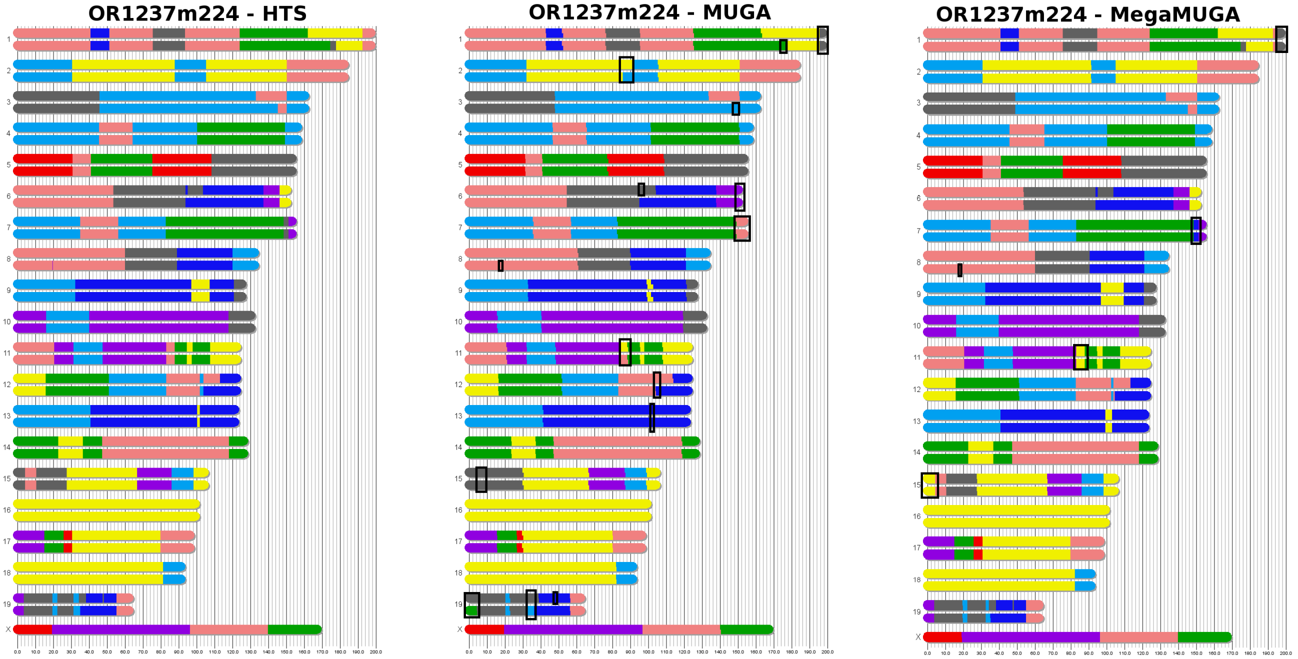


Figure 6.8: Comparison of HTS full coverage solutions with MUGA and MegaMUGA solutions for OR867m532. The HTS solution is shown first, followed by the MUGA solutions and then MegaMUGA solution and black boxes are drawn on the MUGA and MegaMUGA solutions to depict differences between each and the HTS solution. For this sample, MUGA missed 8 recombinations, had 1 false positive, and chose a different founder 3 times. MegaMUGA missed 1 intervals (causing 2 missed recombinations), had 2 false positives (at the ends of chromosomes), and chose a different founder once.



OR1237m224 - HTS

OR1237m224 - MUGA

OR1237m224 - MegaMUGA

Figure 6.9: Comparison of HTS full coverage solutions with MUGA and MegaMUGA solutions for OR1237m224. The HTS solution is shown first, followed by the MUGA solutions and then MegaMUGA solution and black boxes are drawn on the MUGA and MegaMUGA solutions to depict differences between each and the HTS solution. For this sample, MUGA missed 11 recombinations, had 0 false positives, and chose a different founder twice. MegaMUGA missed 1 heterozygous intervals (causing 2 missed recombinations), had 2 false positives, and chose a different founder twice.

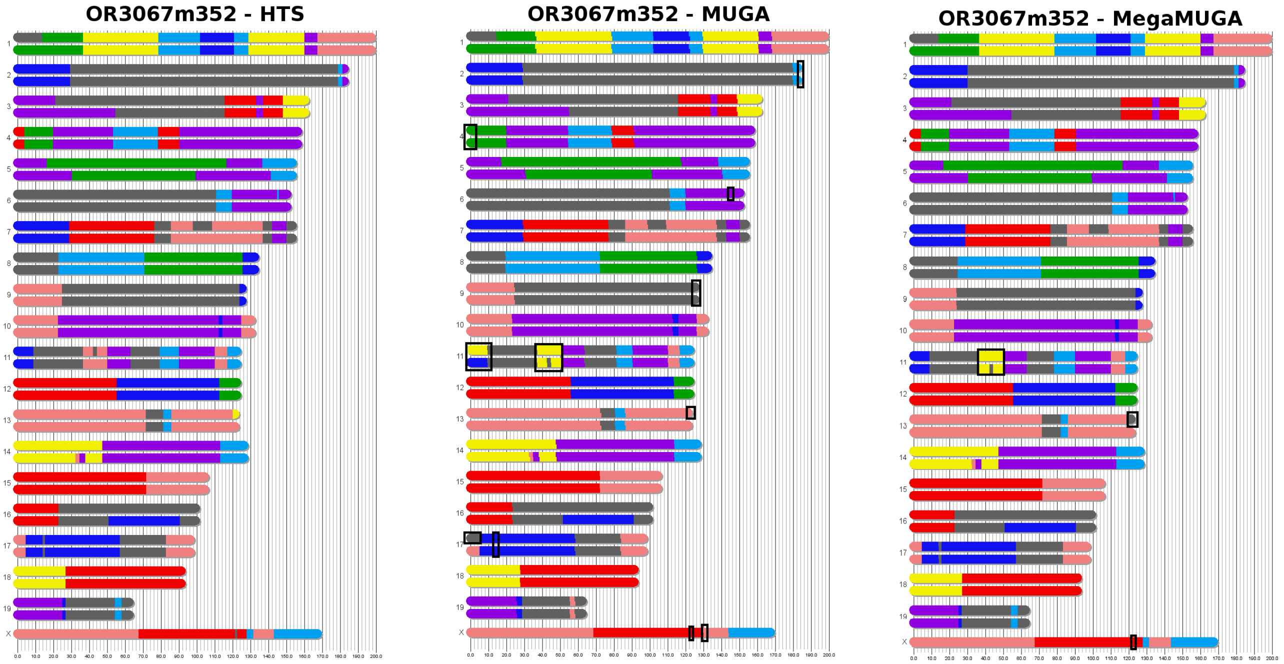


Figure 6.10: Comparison of HTS full coverage solutions with MUGA and MegaMUGA solutions for OR3067m352. The HTS solution is shown first, followed by the MUGA solutions and then MegaMUGA solution and black boxes are drawn on the MUGA and MegaMUGA solutions to depict differences between each and the HTS solution. One difference is that both MegaMUGA and MUGA mislabeled the pink founder (129S1/SvImJ) as yellow (A/J) on chromosome 11. For this sample, MUGA missed 7 recombinations, had 0 false positives, and chose a different founder three times. MegaMUGA missed 1 interval (on Chromosome X), had 0 false positives, and chose a different founder twice.

6.4.3 Read Coverage Analysis

The most significant variable influencing cost in HTS is the read coverage. In order to use HTS as a cost-effective alternative to genotyping arrays in the future, one needs to determine the necessary read coverage to compute haplotype reconstructions that are, at a minimum, equivalent in resolution to algorithms based on a fixed marker set. The resolution of array-based methods is a function of marker density, genetic state, and the informativeness of each marker. MUGA was designed to be able to resolve recombinations to within 1Mb on average when the sample was nearly inbred. MegaMUGA was designed to resolve recombinations to within 160Kb for samples that are highly heterozygous. To determine the necessary read coverage, I sampled the reads at various coverage levels, such that if I wanted 2x coverage, I used about 1/15th of the available reads. In this way, I sampled the genome at 0.25x, 0.5x, 1x, 4x, and 16x. Since I randomly decided which reads to keep, each experiment was run 10 times with a different random seed and the resulting solutions are used in this analysis. For coverage levels of 1x and above, I used the same size bins (1Kb) as the full coverage solution. However, in order to maintain a similar level of evidence per bin at the lower coverage levels, I used 2Kb bins for the 0.5x coverage and 4Kb bins for the 0.25x coverage.

I sampled the HTS reads at various coverage levels (16x, 4x, 1x, 0.5x, and 0.25x) to ascertain the level of accuracy of the haplotype reconstructions and the recombination breakpoints at each level. Since reads were chosen randomly, I repeated each coverage level 10 times. I compared each of the 10 solutions to the full coverage solution to determine the number of times recombinations were found, missed or when new recombinations not in the full coverage solution were created (false positives). For all true recombinations, I calculated the average distance from the recombination breakpoints of the low coverage solutions to the full coverage solution, and also noted the maximum distance between the full coverage recombination location and the low coverage solutions. A synopsis of these comparisons can be seen in Table 6.2. For comparison, similar statistics for the genotyping solutions are also shown in Table 6.2. In addition, Figures

Sample	#Recombs	#FP	#Missing Recombs	Avg. Distance to HTS	Max. Distance to HTS
OR867m532	95	-	-	-	-
16.0x	93	0	2	14471.51	824000
4.0x	93	0.2	2.2	19088.47	824000
1.0x	92.6	1.8	4.2	34485.99	869000
0.5x	79.6	9.6	25	46993.45	855000
0.25x	66	18	47	66466.53	998000
MUGA	86	1	8	820717.70	3832590
MegaMUGA	95	2	2	161699.68	1748837
OR1237m224	95	-	-	-	-
16.0x	92.4	0.4	3	7623.21	558000
4.0x	92.6	0.8	3.2	14482.41	558000
1.0x	91.3	3.1	6.8	35263.74	815000
0.5x	78.5	12.1	28.6	49397.94	982000
0.25x	77.3	10.3	28	51197.37	993000
MUGA	81	0	11	827496.10	3204258
MegaMUGA	93	2	2	252832.99	2263917
OR3067m352	90	-	-	-	-
16.0x	88	0.3	2.3	152.93	19000
4.0x	88.2	0.8	2.6	2765.18	753000
1.0x	87.6	4.1	6.5	26921.33	919000
0.5x	77.4	8.5	21.1	45009.11	933000
0.25x	76.6	9.9	23.3	48702.63	990000
MUGA	81	0	7	870420.44	3575568
MegaMUGA	90	0	2	320968.16	3562834

Table 6.2: Statistics for various coverage levels of sequencing and genotyping data for the three CC samples. I show here the average number of recombinations found among the 10 runs at each coverage level, as well as the average number of false positive recombinations (recombinations found that did not occur in the full-coverage solution), and the average number of missing recombinations (recombinations that occurred in the full-coverage solution that were not present in the lower coverage solution). I also show the average distance and the maximum distance from the lower coverage solution to the full-coverage solution for the recombinations that was found. For the genotyping solutions (MUGA and MegaMUGA), I include the actual statistics from the single run done on each platform.

6.11, 6.12, and 6.13 show the 3 full-coverage solutions compared to one of their 4x and 0.25x coverage solutions. At 4x coverage, most solutions were very similar to the 30x baseline and were between 99.8% and 99.9% concordant with the full coverage solution. At 0.25x coverage though, the solutions varied more dramatically depending on whether or not the randomly selected reads fell over enough informative SNPs for a founder pair in a particular region. These solutions ranged from 49.1% to 99.4% concordant with the full coverage solution. In Figures 6.11, 6.12, and 6.13 I have shown a 99.1% concordant 0.25x solution for sample OR867m532, a 95.6% concordant 0.25x solution for sample OR1237m224, and a 76.8% concordant 0.25x solution for OR3067m352. Note that the majority of the disconcordant solutions include a heterozygous state rather than the expected homozygous state chosen by the full coverage HMM. Calls of heterozygous states with only a single observation tend to be 50% correct, in that they always have one founder that matches the correct homozygous state solution. This issue could be addressed in low coverage cases by considering the degree of inbreeding when establishing the emission probabilities. I have shown that it is possible to accurately reconstruct founder mosaics using HTS data at relatively low coverage levels. In order to maintain the ability to distinguish between homozygous and heterozygous founder-pair states, we found that 1x coverage was sufficient. Below this level of coverage, the results were highly variable depending if reads were available at the informative SNPs within the recombination breakpoint areas. Solutions at 16x were very consistent, with the majority of the solutions choosing the exact same bins at which to transition for 87.2% of found recombinations. As I lowered the coverage level, the 10 solutions at each coverage level became more inconsistent, as shown in Figure 6.14, although they still maintained relatively concordant solutions with the full coverage HTS solution. As shown in Table 6.2, less recombinations are found and of the recombinations found, the distance from the HTS solution grows. The largest difference in coverage levels comes between 1x and 0.5x, where I start to lose informative SNPs since only about half of the SNPs will have reads, and the bin size is also doubled in order to maintain a similar level of evidence.

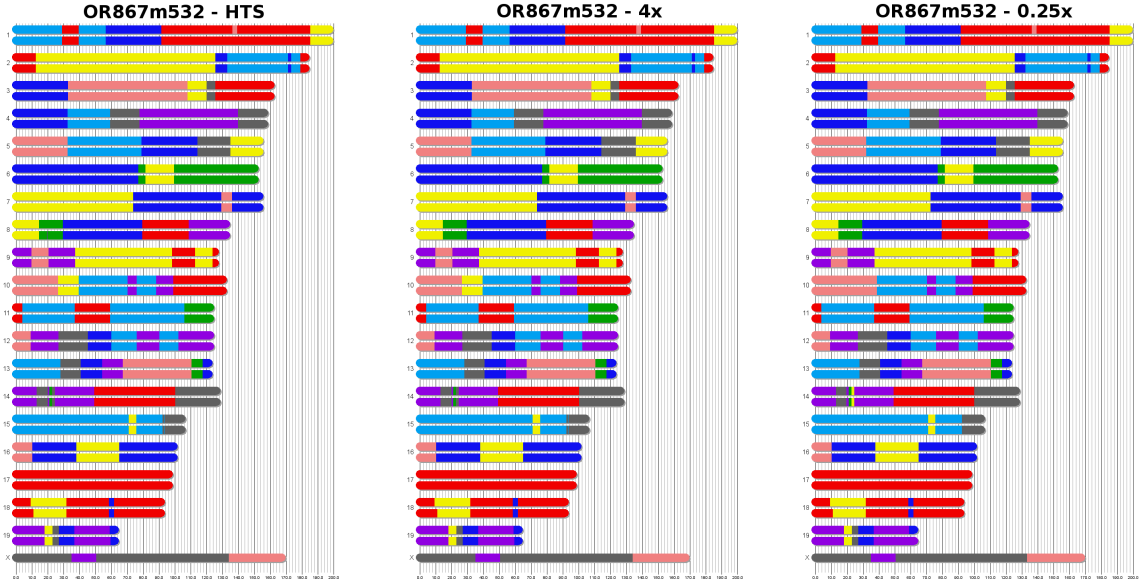


Figure 6.11: Comparison of HTS full coverage solutions for OR867m532 with 4x and 0.25x coverage solutions. At 4x coverage, the solution is 99.9% concordant with the full coverage solution. The 0.25x solutions shown here is 99.1% concordant with the full coverage solution.

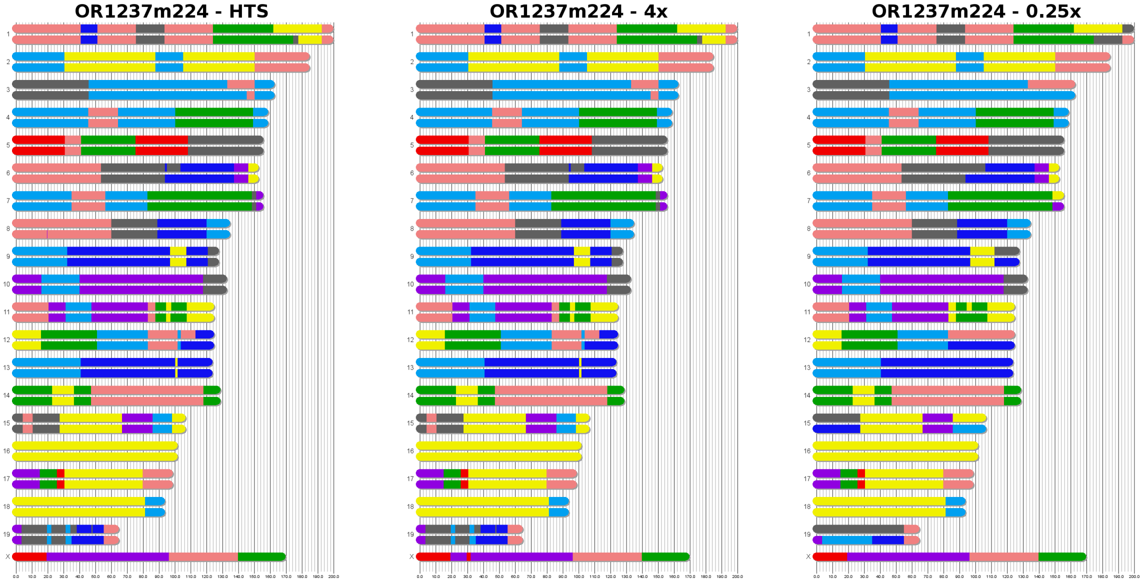


Figure 6.12: Comparison of HTS full coverage solutions for OR1237m224 with 4x and 0.25x coverage solutions. At 4x coverage, the solution is 99.8% concordant with the full coverage solution. The 0.25x solutions shown here is 95.6% concordant with the full coverage solution.

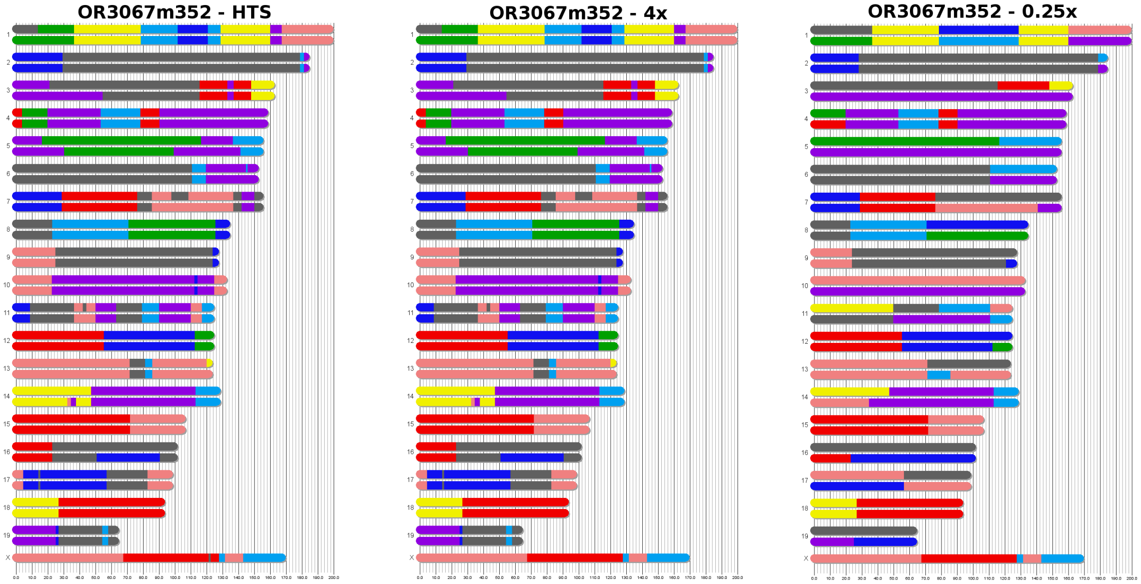


Figure 6.13: Comparison of HTS full coverage solutions for OR3067m352 with 4x and 0.25x coverage solutions. At 4x coverage, the solution is 99.9% concordant with the full coverage solution. At 0.25x coverage, the solution were more variable, depending on which reads were selected. The 0.25x solutions shown here is only 76.8% concordant with the full coverage solution. In the 0.25x solution shown, it can be seen that the majority of the disconcordance is from a heterozygous state being selected rather than the correct homozygous state. In each of these cases, the 0.25x solution is 50% correct, in that the heterozygous state selected includes the correct homozygous founder.

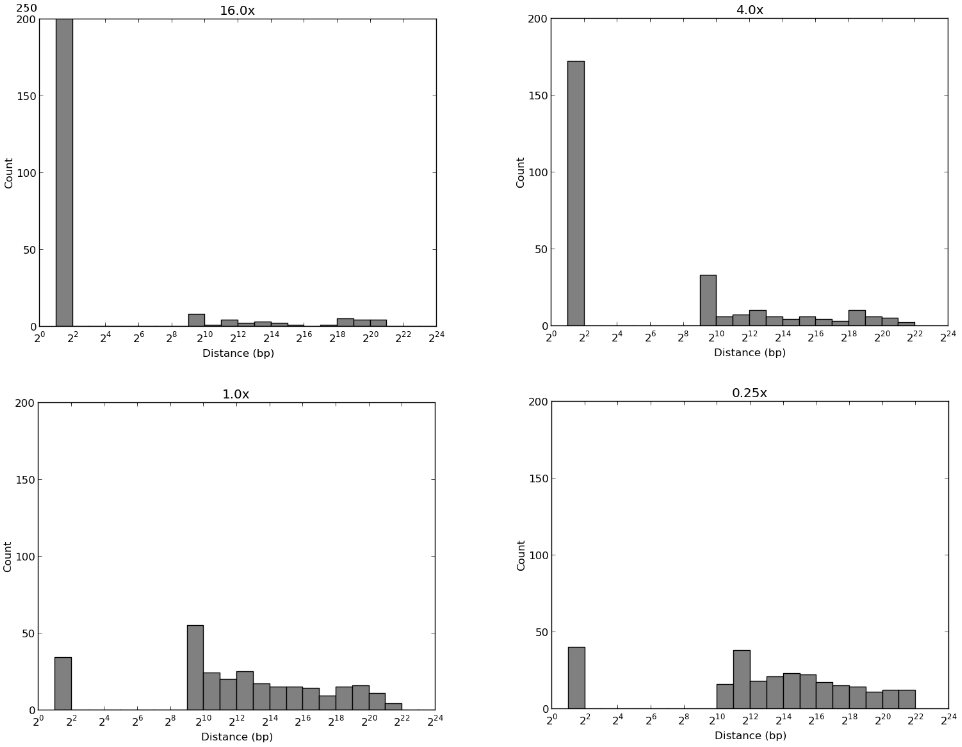


Figure 6.14: Histograms of the delta between the maximum position and minimum position found at each recombination among the 10 runs at each coverage level. Coverage levels of 16x, 4x, 1x, and 0.25x are shown. One can see that at 16x coverage, almost all 10 solutions were identical, while at 1x coverage, the solutions tended to be more divergent, although the majority still found transitions within 1-4 bins of each other.

6.5 Conclusion

By developing a method for computing founder mosaics from HTS data, users of the CC, DO and other mouse populations will be able to easily transition from genotyping arrays to HTS. This means that in the near future when HTS becomes price competitive with genotyping, the current pipelines for marker-assisted inbreeding[60], detection of residual heterozygosity and other tools for the CC lines[61] will be able to transition from using genotyping arrays to HTS. I have shown that even at relatively low coverage levels of 1x, the founder mosaics are just as reliable, if not more accurate than the current genotyping platform algorithms. This is caused by the ability to see almost all informative SNPs for each founder-pair genome-wide, rather than a pre-selected subset of SNPs.

The accuracy with which one can resolve recombination breakpoints in HTS data depends on both the density of reads and the genetic diversity of the genomes on either side of the breakpoint. I have attempted to address both of these factors by combining an HMM data driven model with a refine process that is based entirely on the known genetic differences between a given founder pair. In this setting the HMM is responsible for finding a rough estimate of the breakpoint location, but more importantly it is responsible for identifying the founders on either side of the breakpoint. I can then refine the location of the breakpoint using informative SNPs down to the limits of the sequence diversity.

One of the limitations of my algorithm was that I was only able to refine the location of the breakpoint to within the distance between consecutive informative SNPs. One of my early design decisions was to filter out any SNP that had a no-call (N) genotype for any of the eight founders. Therefore, it is possible that I could have further refined some regions using one of these filtered SNPs if the SNP was informative (and not an N call) between the two founders on either side of the breakpoint. Another drawback to my approach is that it relies heavily on the sequence alignment process being done correctly. If the consensus genome to which the HTS data was aligned was not accurate, it could affect the results of the algorithm.

CHAPTER7: CONCLUSION AND FUTURE WORK

In this thesis, I have developed a breeding simulator suitable for evaluating various breeding schemes. Through simulations, I was able to test breeding schemes and breeder metrics to decrease the number of generations it takes to achieve inbred strains as well as engineer user-specified mice more efficiently. I have also designed two full-genome genotyping platforms and proven their effectiveness in tracking residual heterozygosity as well as achieving accurate founder assignments in fixed genomic regions. These two components allow the results of these genotyping platforms to act as input into the simulator as well as a number of other useful online tools, so that live mice data can be utilized to make breeding decisions in a timely fashion. I also explored the use of high-throughput sequencing (HTS) as an alternative to genotyping and have developed tools to create accurate haplotype reconstructions from HTS data for when this technology becomes more cost effective than microarrays.

In the following sections, I summarize my results and describe avenues for future investigation.

7.1 Theoretical Marker-Assisted Techniques

Through simulations, I developed several alternatives to random sib-matings to dramatically accelerate the creation of RILs by as much as 16 generations. These include the judicious use of parental backcrossing and the selection of mating pairs based on genotypes from genome-wide SNPs. Both of these techniques, when applied after the point of peak diversity is reached, result in a negligible reduction in the number of segments. I also propose an advanced intercross variant in which MAI is applied during the early generations to increase the number of haplotype segments for better mapping resolution.

7.1.1 Future Work

While the results of my simulations for standard two-way and eight-way RILs were comparable to those of Broman [8], the recombination model of the simulator could still be improved. My simulator currently uses a statistical model of recombination similar to that used by Broman, but instead of using the typical centimorgan measurement for each chromosome, it represents the length of each chromosome in base pairs and equates the randomly chosen recombination locations to the base pair location in the genome. This allows the simulator to more easily accept input from live mouse genotyping data, but it creates a bit of uncertainty since the exact centimorgan to base pair mapping is not known. Therefore, a better recombination model would be to use the actual recombination map of the mouse. This would allow the simulator to be more accurate in randomly determining recombination spots when simulating breedings and lead to simulation results that are even more similar to what is seen in the live mouse populations.

7.2 Tools for Analysis of Live Mice

To implement the MAI techniques described in Chapter 3 on live mice a number of tools are necessary. The first tool that was needed was an inexpensive platform for interrogating mouse genotypes. Therefore, I codesigned MUGA, a 7,851 SNP genotyping array. This array was used to genotype CC animals for about 2 years until it was possible to design a new array with 10x more SNPs for the same price. This new array is called MegaMUGA and contains 77,808 SNP markers genome-wide. Using either MUGA or MegaMUGA to learn about the underlying genomes of the mice in the CC, I then built a series of tools on top of these genotyping platforms to facilitate the implementation of the MAI techniques. These tools were built to ensure quality control, do analysis of the CC mice, and select breeders throughout the inbreeding process. Through the use of these tools, researchers and mouse room techs have been able to make informed decisions regarding the CC population throughout the inbreeding process.

7.2.1 Future Work

Since MegaMUGA was designed a few years ago, it is now possible to create a 3rd generation array that has about 10x more SNPs that MegaMUGA for the same price per sample. By putting all working SNPs from MegaMUGA on the new array and filling in the additional markers with SNPs selected similarly to those on MegaMUGA, a 3rd generation genotyping array will be even more powerful and informative in determining the underlying genomic structure of the mouse samples. However, as with the development of MegaMUGA, all the tools that have previously been built to work specifically with MUGA or with MegaMUGA will need to be modified to work with the new array. This will only affect tools that do not use haplotype reconstructions as their underlying data structure though, since those tools will continue to work with the new array seamlessly.

While most of the tools mentioned were built with particular goals in mind, future tools that would be very helpful to have would be the ability to upload and import genotyping array results online rather than needing to do this rather time intensive process offline. Another great tool would allow for the automatic selection of lines that have reached the required homozygosity thresholds to be considered “Available Lines”. These “Available Lines” are those lines that are currently being distributed online to other labs for research purposes.

7.3 High-Throughput Sequencing Data

As most of the tools mentioned in this thesis rely heavily on the use of haplotype reconstructions as input, it was essential to both verify the accuracy of those reconstructions as well as determine a way to create haplotype reconstructions using HTS data rather than genotyping array data. I was able to verify that the results from the genotyping array data were very accurate with MUGA achieving about 95%-97% concordance with the full coverage HTS data and MegaMUGA achieving 98%-99% concordance with the full coverage HTS data. The HTS data used in this experiment was 30x coverage, meaning that each genomic coordinate is covered by an average of 30 reads.

Since all HTS data is genome-wide, the cost is not determined by the number of SNPs, but rather by the coverage level. Therefore, I compared different coverage levels (16x, 4x, 1x, 0.5x, and 0.25x) and I determined that even at relatively low coverage levels of 1x, the haplotype reconstructions produced are just as reliable, if not more accurate than the current genotyping platform algorithms. This is caused by the ability to see almost all informative SNPs for each founder-pair genome-wide, rather than a pre-selected subset of SNPs.

7.3.1 Future Work

High-throughput sequencing (HTS) data holds a wealth of information and its use has become prevalent in many genome-wide studies. It has been suggested that HTS may enable us to detect gene conversions and de novo mutations that were previously undetectable by genotying microarrays. Gene conversions appear as two nearby recombinations, as if they were a tiny double recombination. Finding gene conversions is very difficult when pooling read data as they tend to be very small (100bp-3000bp) and the rate at which they occur is currently unknown. Using the full coverage HMM solutions, I plan to explore each bin for evidence of informative SNPs for some founder pair similar to how I refine recombination breakpoints. The primary difference being that when refining breakpoints the founder pair is given by the HMM solution. In the case of gene conversion all combinations would have to be explored while controlling for noise. I plan to take advantage of the observation that gene conversions tend to fall near recombinations, and in particular are found primarily in recombination hotspots. By looking at both the recombination regions in the HTS solutions as well as those regions of the genome known to be hotspots in mouse[42, 43, 10], I can test at what coverage level HTS allows for the discovery of gene conversions among multi-parental crosses.

As I described in Chapter 6, I used HTS data from three of our CC animals to determine the accuracy of founder calls in the haplotype reconstructions from microarray data. The sequencing data allows one to better see the exact recombination breakpoints as well as better determine “blind” spots among our eight founders. Sequencing data is a great resource for finding hot and cold spots of recombination, areas of IBD, and it allows us to utilize non-linear genomes as not every strain has the exact same genomic length. Therefore, in the future, I plan to use this sequencing data to build a better recombination model for my simulator.

Once I’ve improved the simulation model and proven its effectiveness using speed congenics as an example, I will further test it by simulating more complex genomic engineering. One of these more complex problems could include fixing multiple genes, possibly from more than one founder strain, to varied backgrounds. Other problems will use the recombinant inbred cross (RIX) mice that are developed from the CC inbred strains. These RIX mice will be developed to model outbred human populations. This is done by selecting a random ordering of the CC inbreds and mating them in a circular fashion, such that each resultant strain is the product of a 2-way cross between different CC mice. Selecting panels of RIX to achieve certain gene combinations or maximizing the diversity outside of a particular gene location are some examples of problems scientists may want to solve to create the best strains for their experiments. Also choosing the ordering of the CC inbreds for creating the RIX lines has some interesting computational implications. As this panel will be used as a mapping population, the goal is to maximize the mapping resolution of it by minimizing the number of shared recombination breakpoints among the breeding pairs. This resource will create a number of interesting problems, many of which can be solved using simulations.

APPENDIXA: SEQUENCE SIMILARITY MAPS

These plots depict the sequence similarity between CC founder pairs. Each map shows the percentage of 1000 base pair bins within a 100 kilobase window for which there is at least 1 informative SNP distinguishing the founder pair. Because CC founders fall into two categories, classical lab strains and wild-derived lab strains, there are three possible categories of founder-pair combinations. WW, shown in blue, occurs when both founders are wild-derived. These founder-pairs typically have low sequence similarity and many informative variants as seen by the relative density of the blue plot. LW, shown in green, occurs when one founder is wild-derived and the other is a classical lab strain. These founder-pairs also typically have many informative SNPs, but less than WW pairs. LL, shown in red, is an example where both founders are classical lab strains. These combinations typically have significantly more sequence similarity. As can be seen by the areas in dark gray, there are many 1000 bp bins with few to no informative SNPs. Regions shown in dark gray on these sequence similarity maps indicate areas of the genome that will be difficult to detect recombinations between these founder-pairs

As many different algorithms are used to create founder mosaics of the extant CC lines, it is necessary to validate the results and determine if a particular region that is segregating between two founders is an area where recombinations can be closely resolved. If your recombination of choice falls into a dark gray area of the sequence similarity map between your founder-pair, you should proceed with caution, as this area has very few segregating SNPs for your founder-pair of choice.

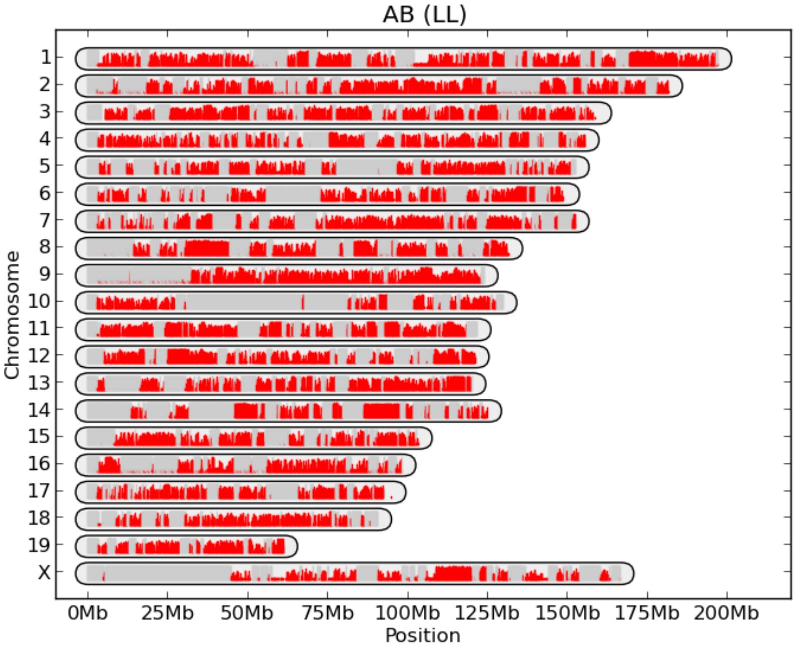


Figure A.1: Sequence similarity map for CC founders A (A/J) and B (C57BL/6J).

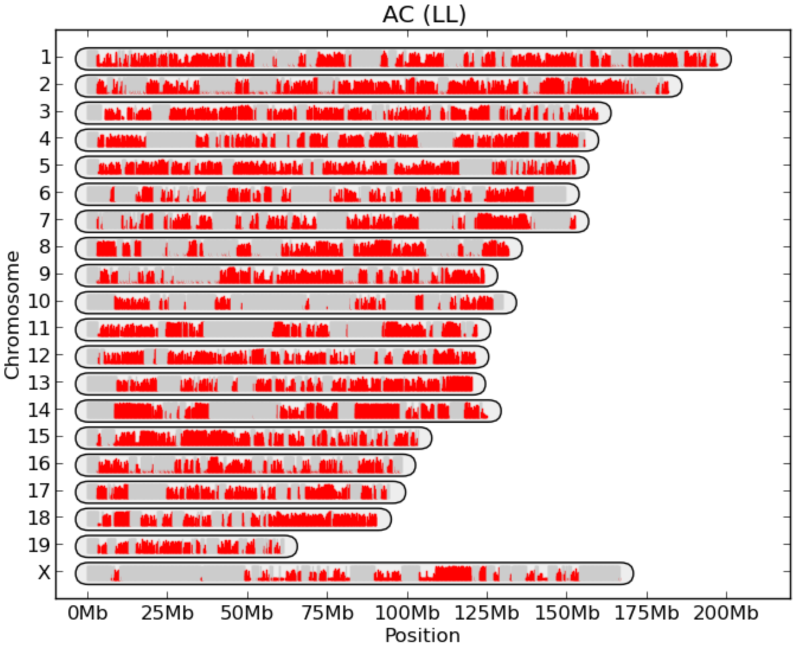


Figure A.2: Sequence similarity map for CC founders A (A/J) and C(129S1/SvImJ).

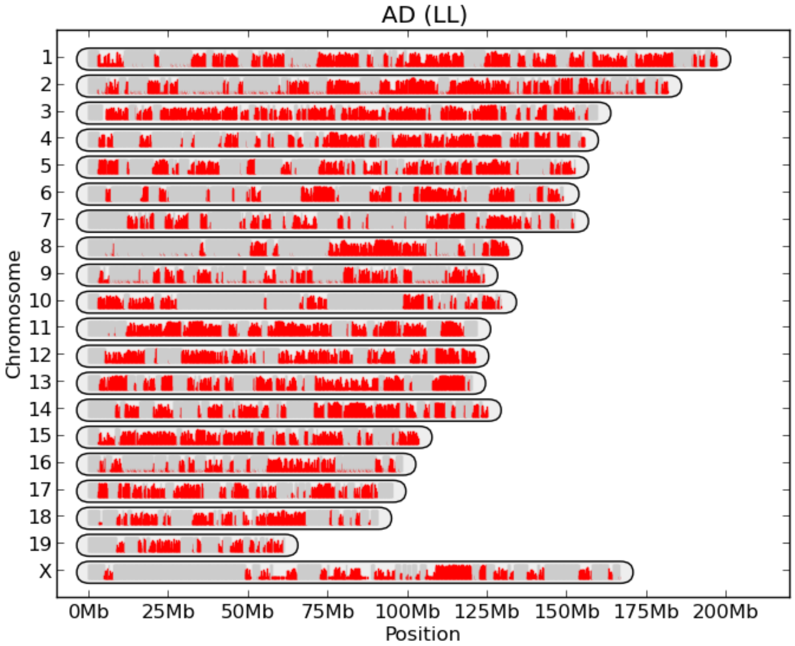


Figure A.3: Sequence similarity map for CC founders A (A/J) and D (NOD/ShiLtJ).

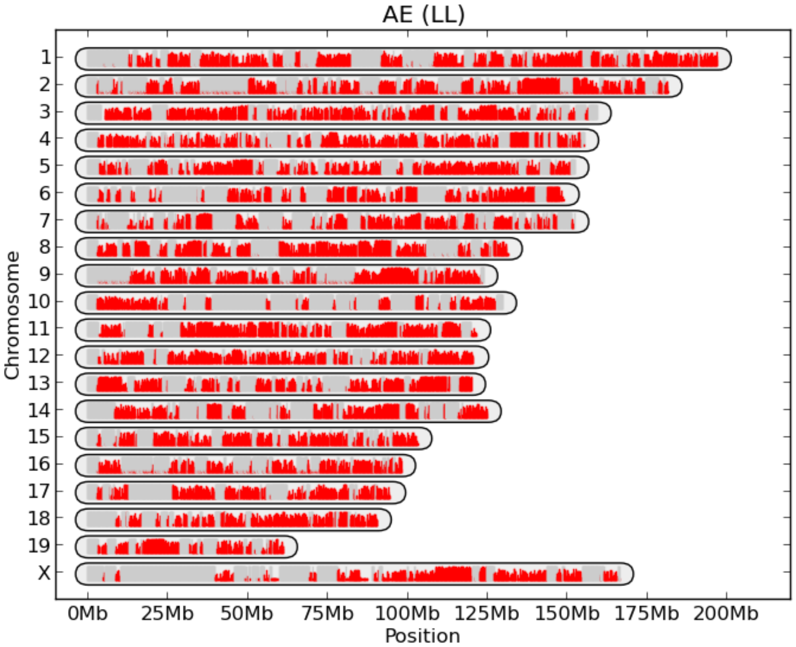


Figure A.4: Sequence similarity map for CC founders A (A/J) and E (NZO/HlLtJ).

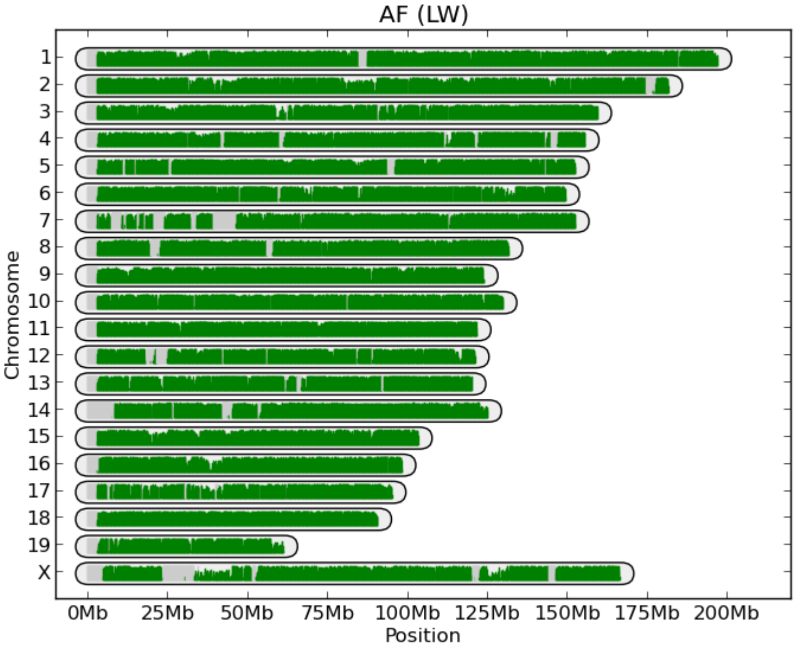


Figure A.5: Sequence similarity map for CC founders A (A/J) and F (CAST/EiJ).

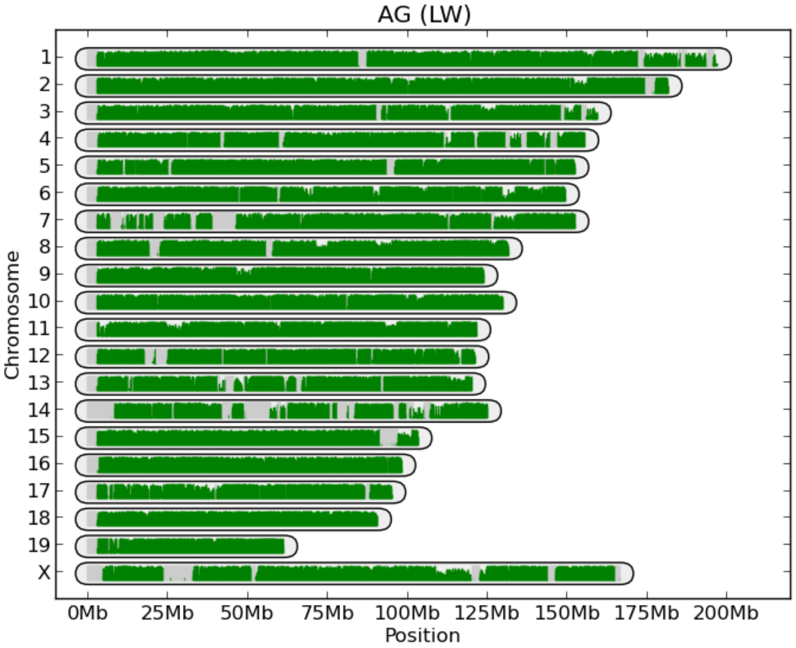


Figure A.6: Sequence similarity map for CC founders A (A/J) and G (PWK/PhJ).

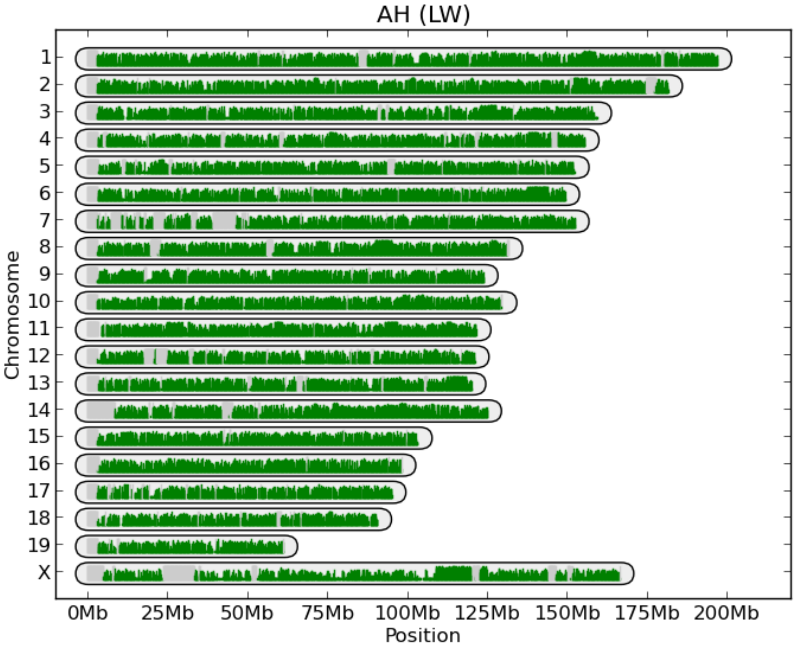


Figure A.7: Sequence similarity map for CC founders A (A/J) and H (WSB/EiJ).

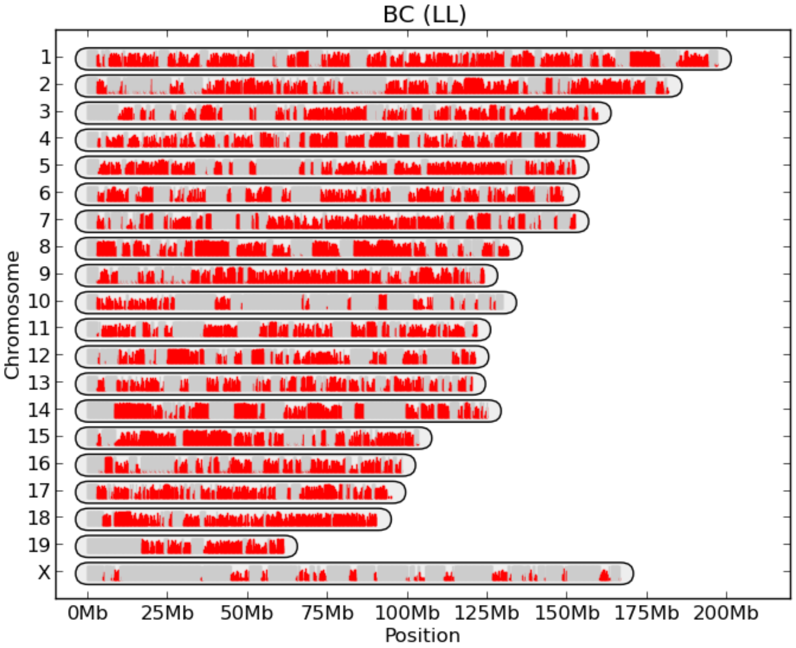


Figure A.8: Sequence similarity map for CC founders B (C57BL/6J) and C (129S1/SvImJ).

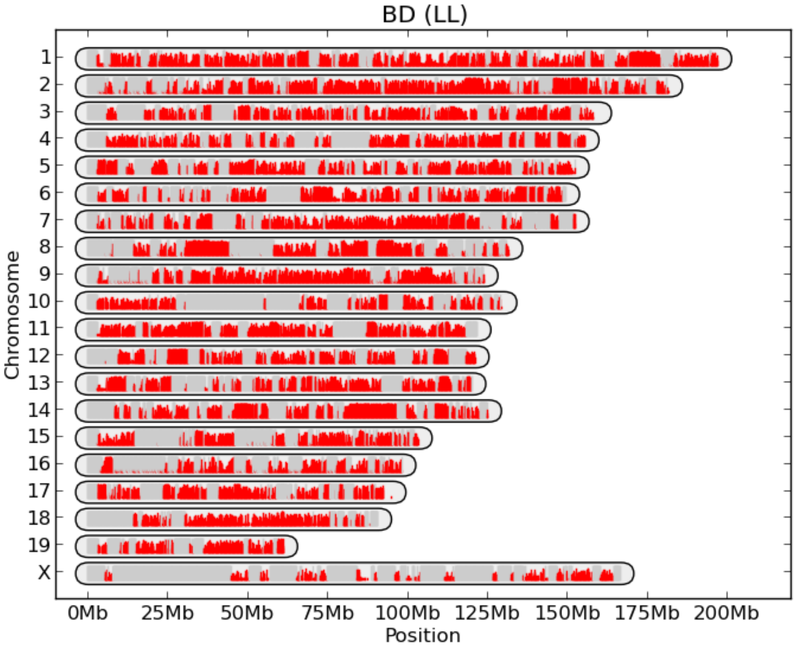


Figure A.9: Sequence similarity map for CC founders B (C57BL/6J) and D (NOD/ShiLtJ).

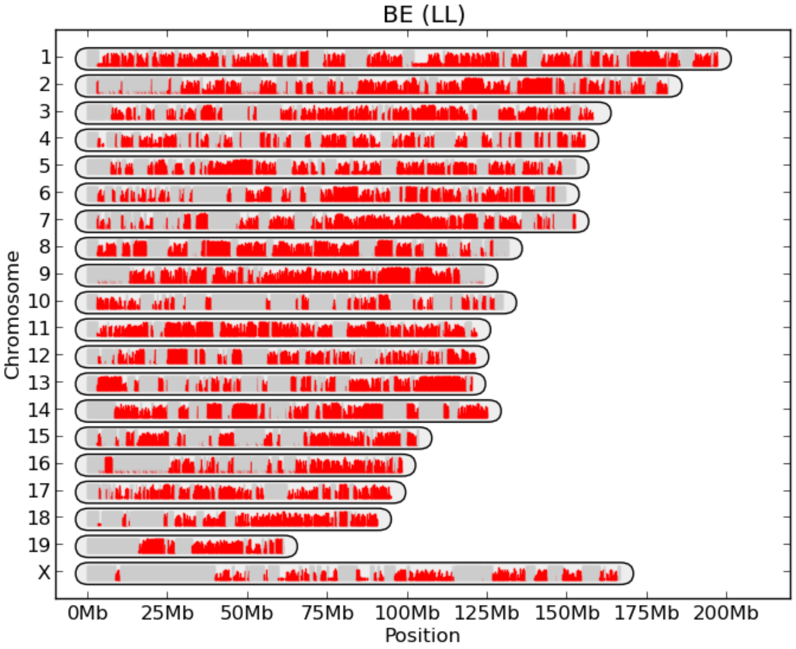


Figure A.10: Sequence similarity map for CC founders B (C57BL/6J) and E (NZO/HlLtJ).

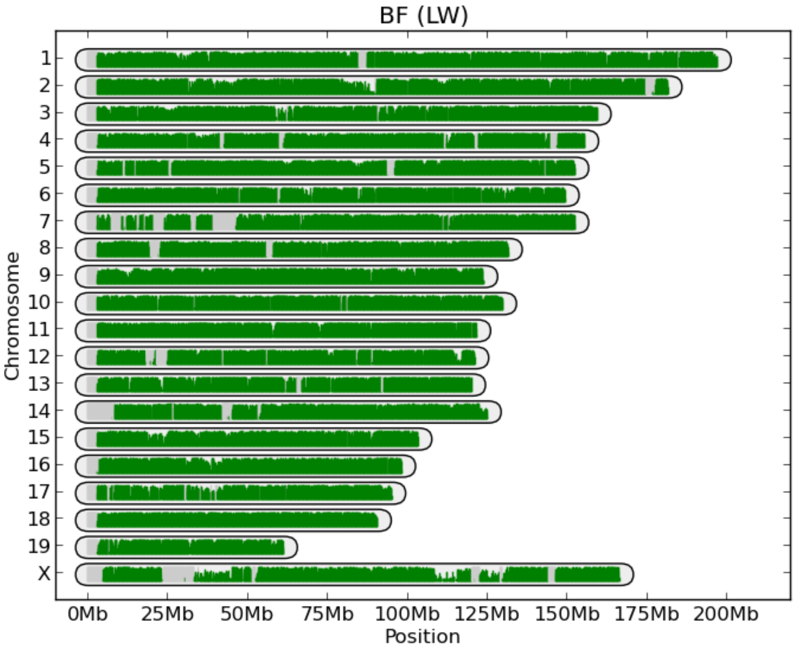


Figure A.11: Sequence similarity map for CC founders B (C57BL/6J) and F (CAST/EiJ).

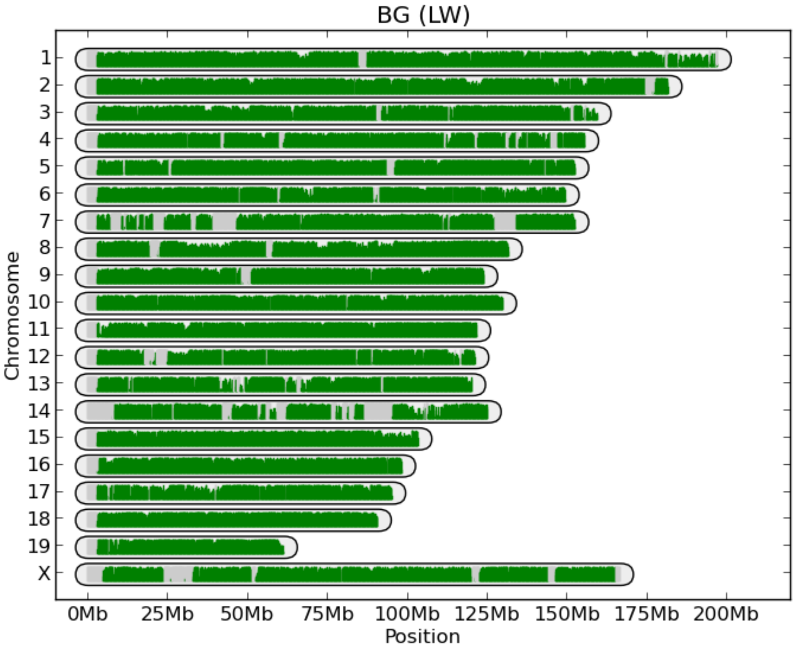


Figure A.12: Sequence similarity map for CC founders B (C57BL/6J) and G (PWK/PhJ).

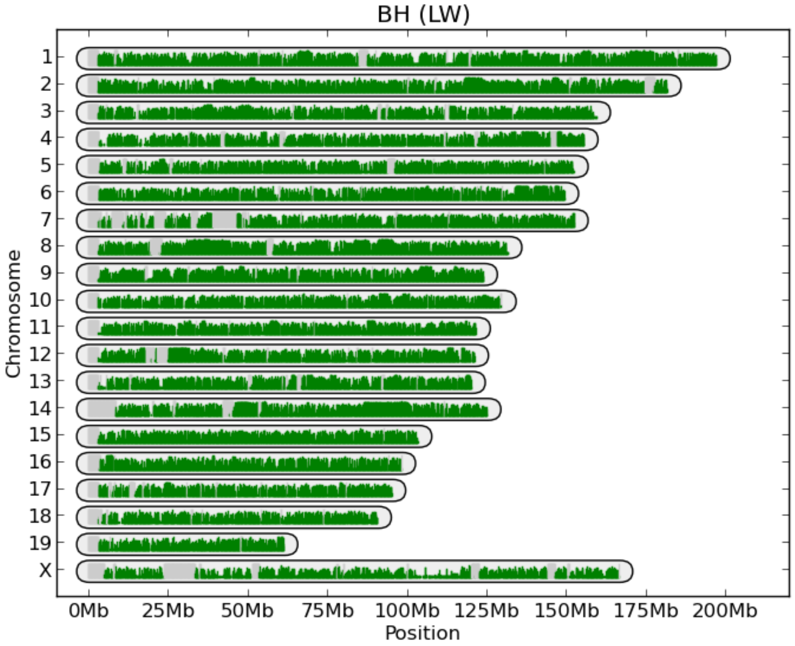


Figure A.13: Sequence similarity map for CC founders B (C57BL/6J) and H (WSB/EiJ).

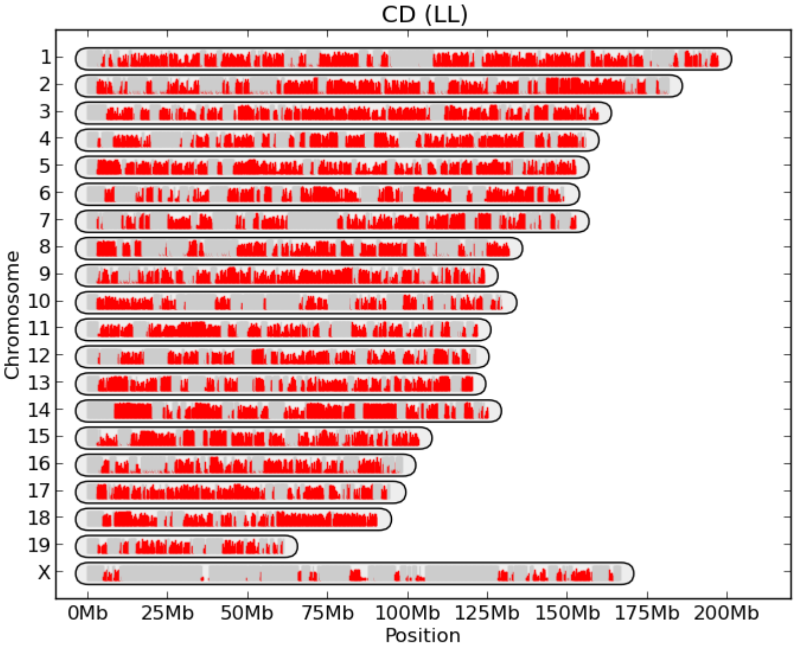


Figure A.14: Sequence similarity map for CC founders C (129S1/SvImJ) and D (NOD/ShiLtJ).

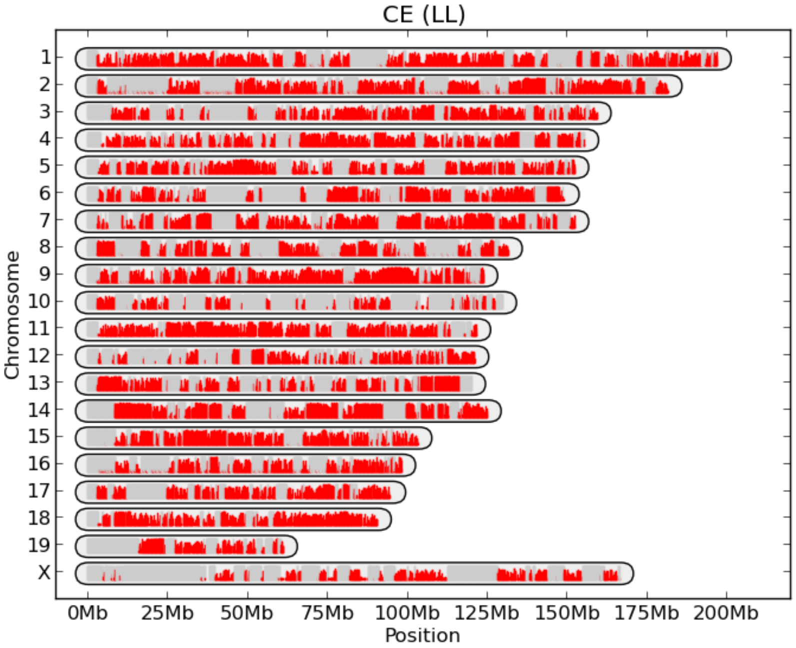


Figure A.15: Sequence similarity map for CC founders C (129S1/SvImJ) and E (NZO/HlLtJ).

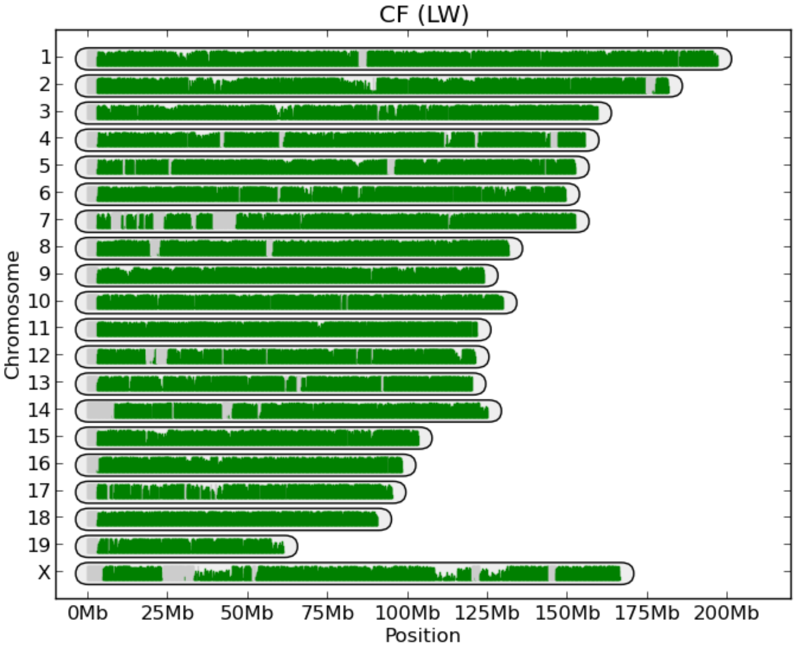


Figure A.16: Sequence similarity map for CC founders C (129S1/SvImJ) and F (CAST/EiJ).

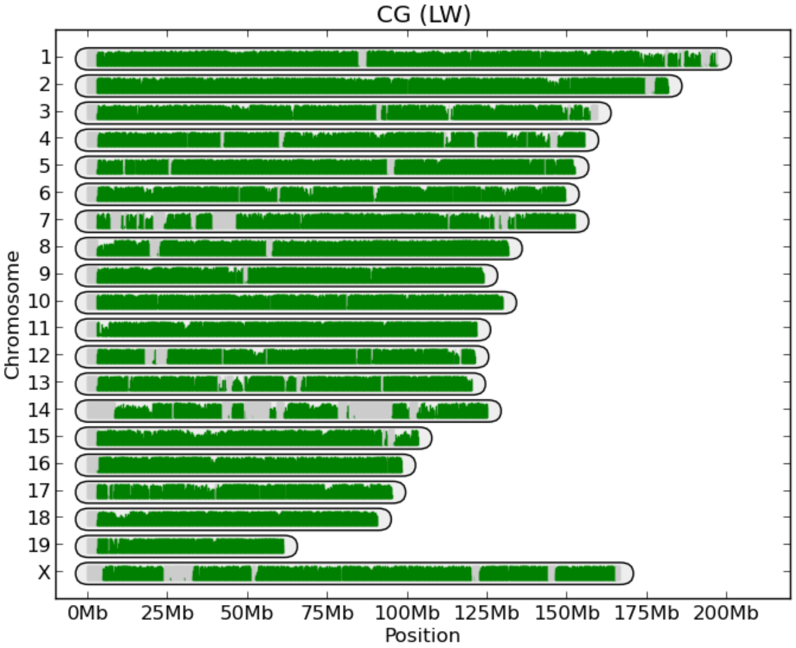


Figure A.17: Sequence similarity map for CC founders C (129S1/SvImJ) and G (PWK/PhJ).

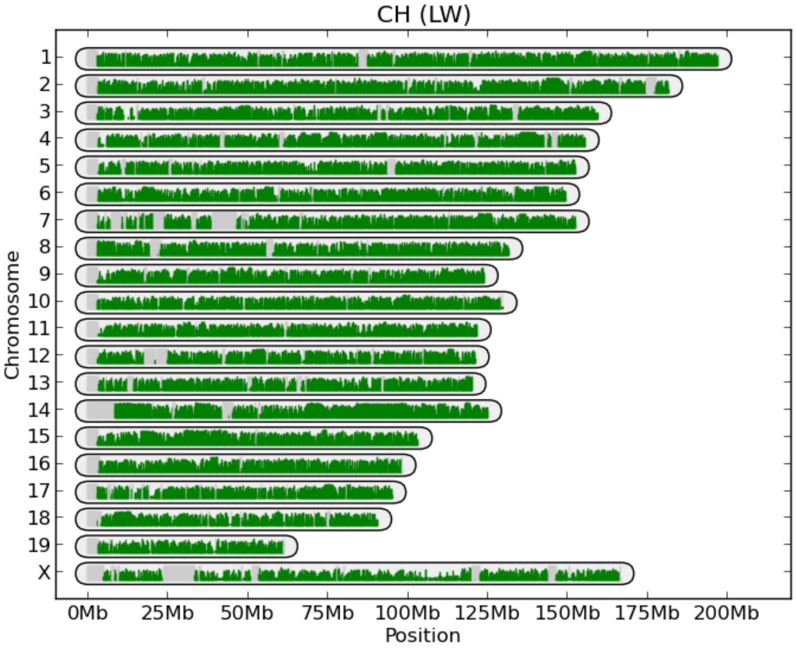


Figure A.18: Sequence similarity map for CC founders C (129S1/SvImJ) and H (WSB/EiJ).

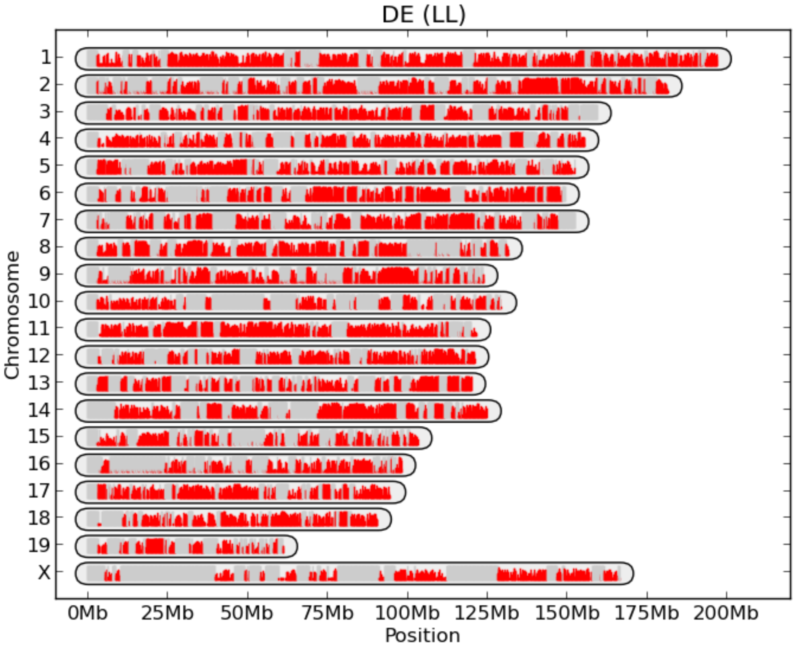


Figure A.19: Sequence similarity map for CC founders D (NOD/ShiLtJ) and E (NZO/HlLtJ).

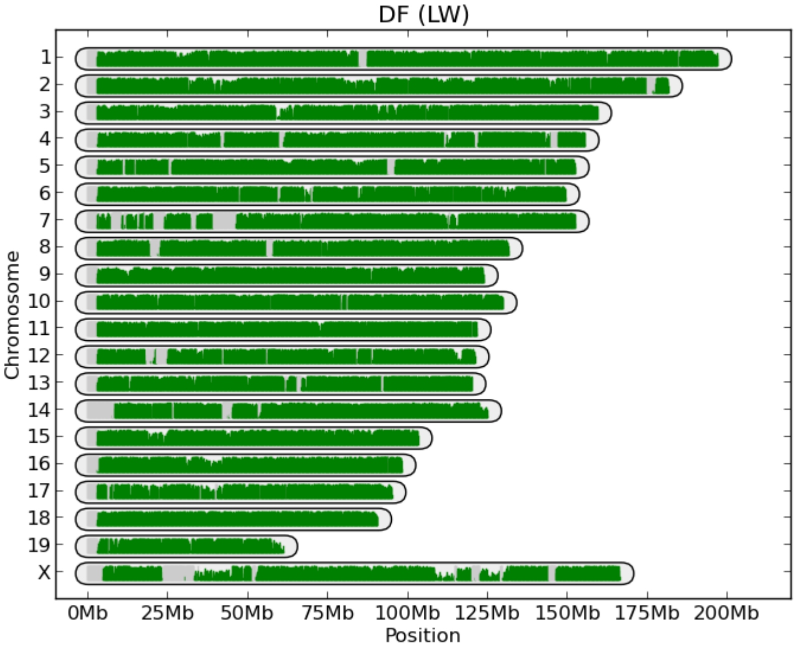


Figure A.20: Sequence similarity map for CC founders D (NOD/ShiLtJ) and F (CAST/EiJ).

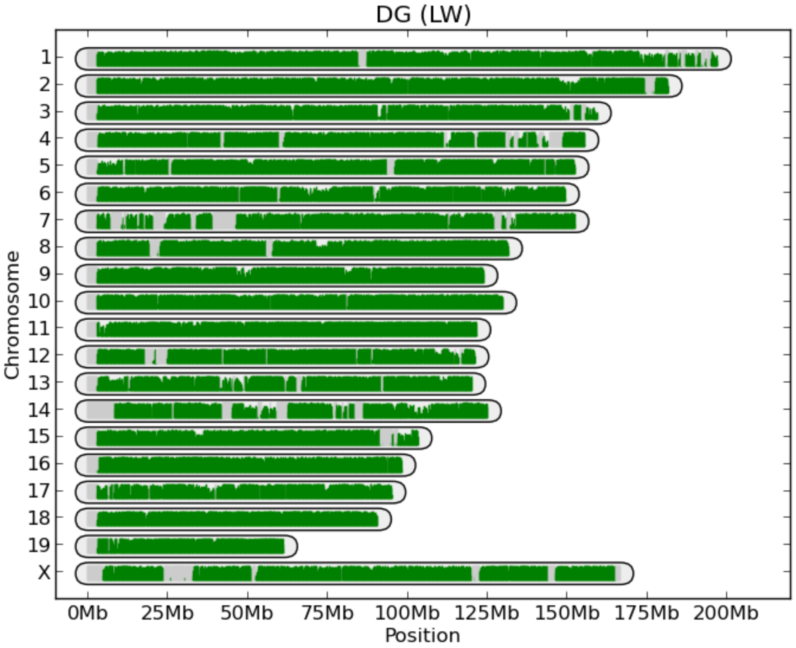


Figure A.21: Sequence similarity map for CC founders D (NOD/ShiLtJ) and G (PWK/PhJ).

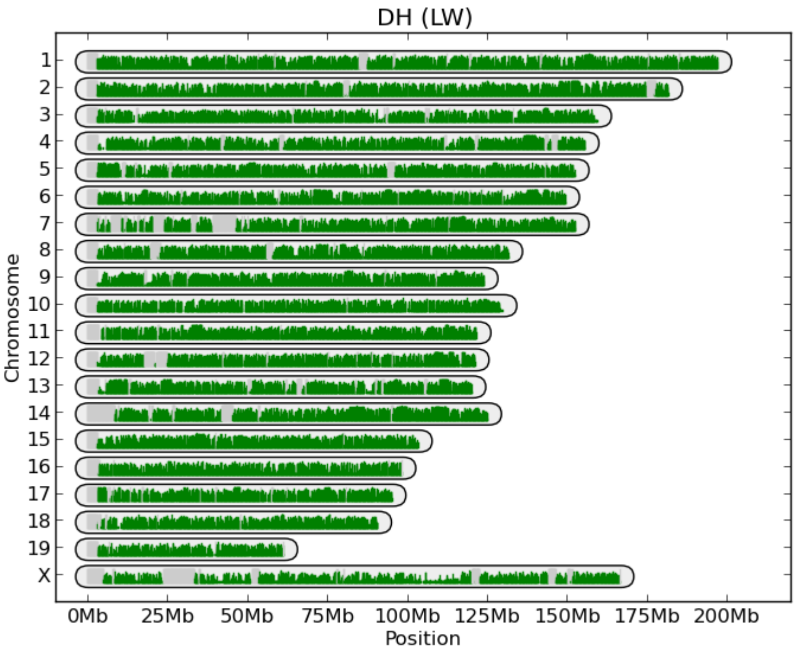


Figure A.22: Sequence similarity map for CC founders D (NOD/ShiLtJ) and H (WSB/EiJ).

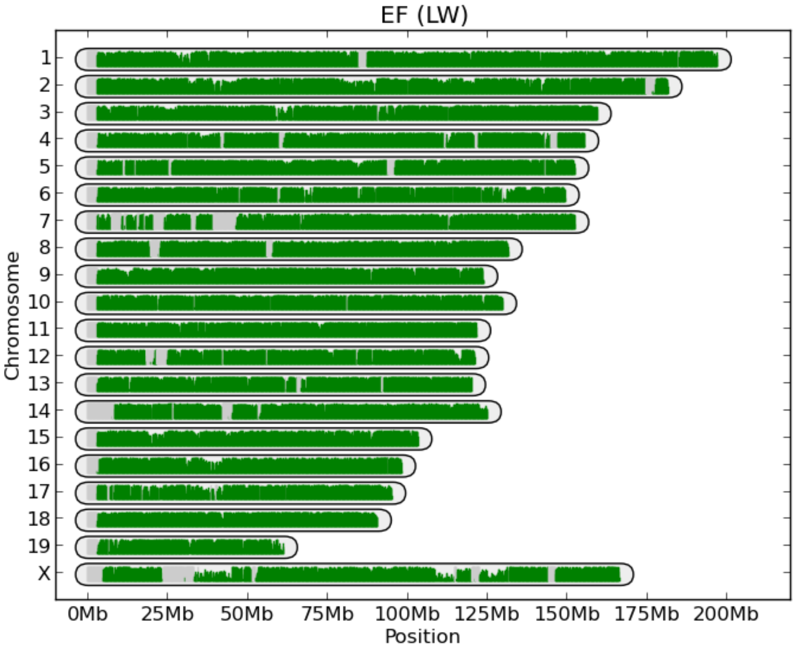


Figure A.23: Sequence similarity map for CC founders E (NZO/HlLtJ) and F (CAST/EiJ).

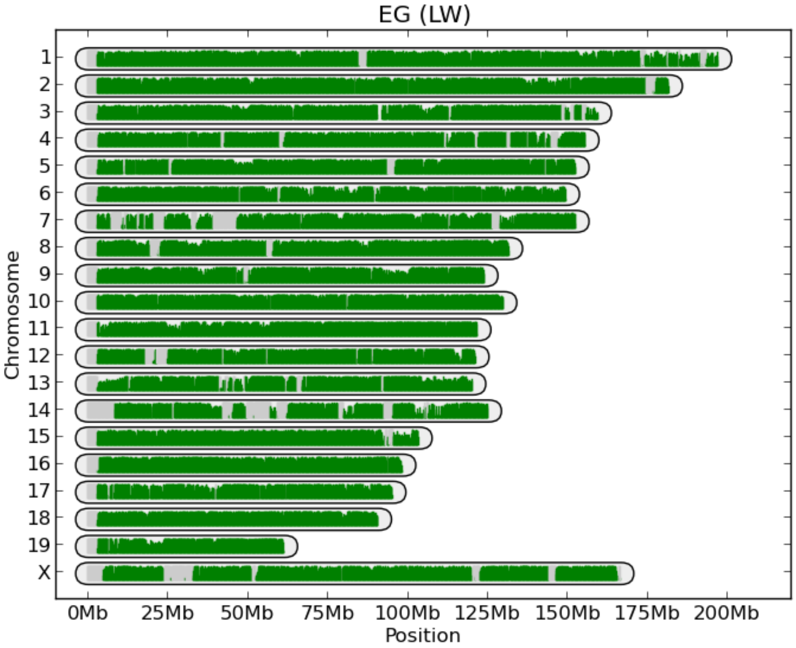


Figure A.24: Sequence similarity map for CC founders E (NZO/HlLtJ) and G (PWK/PhJ).

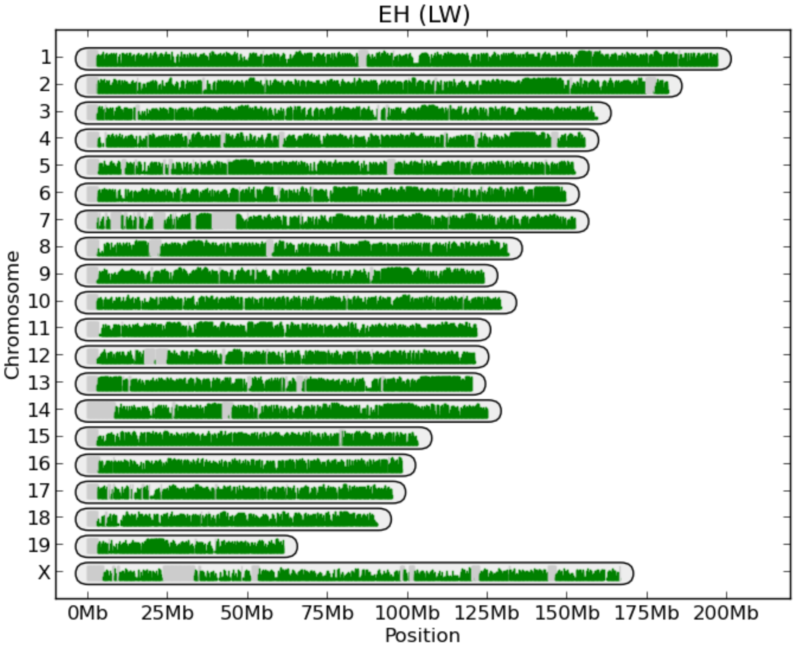


Figure A.25: Sequence similarity map for CC founders E (NZO/HlLtJ) and H (WSB/EiJ).

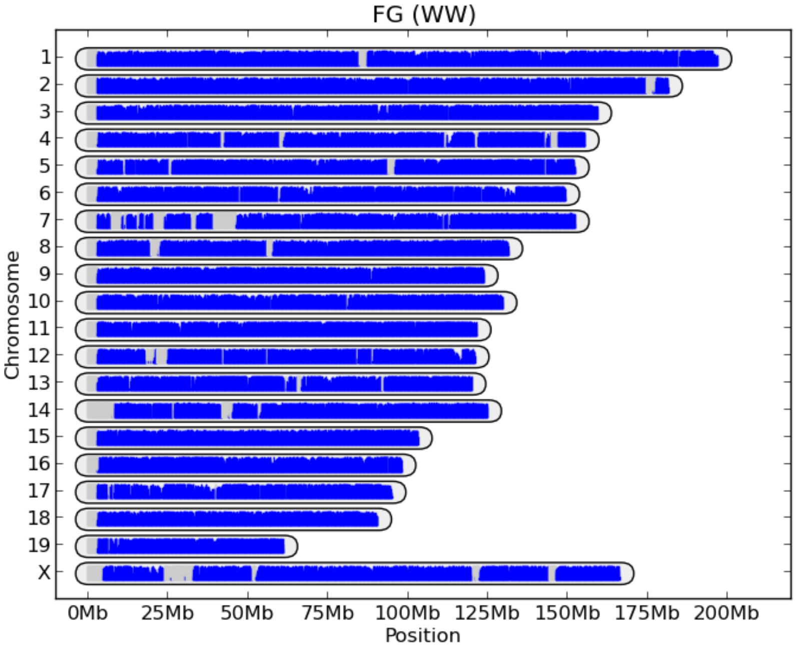


Figure A.26: Sequence similarity map for CC founders F (CAST/EiJ) and G (PWK/PhJ).

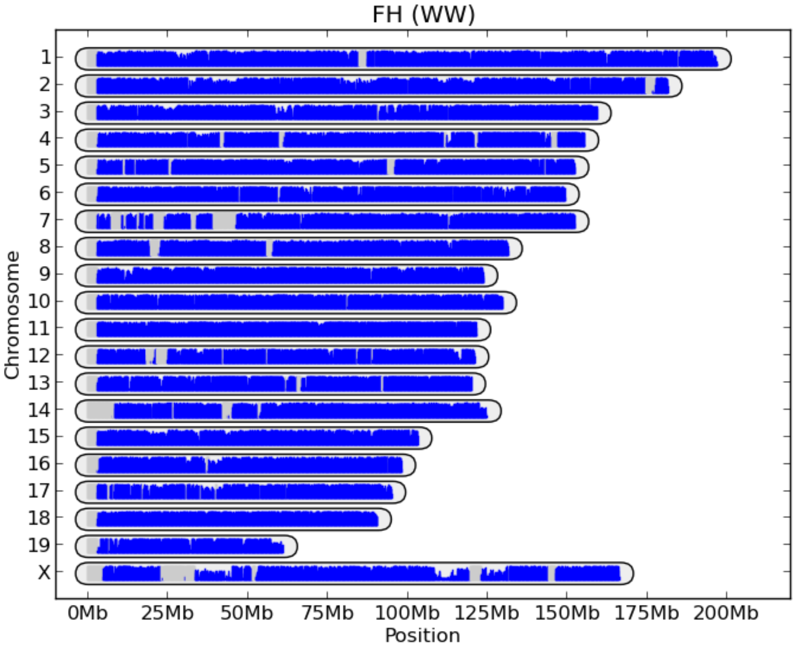


Figure A.27: Sequence similarity map for CC founders F (CAST/EiJ) and H (WSB/EiJ).

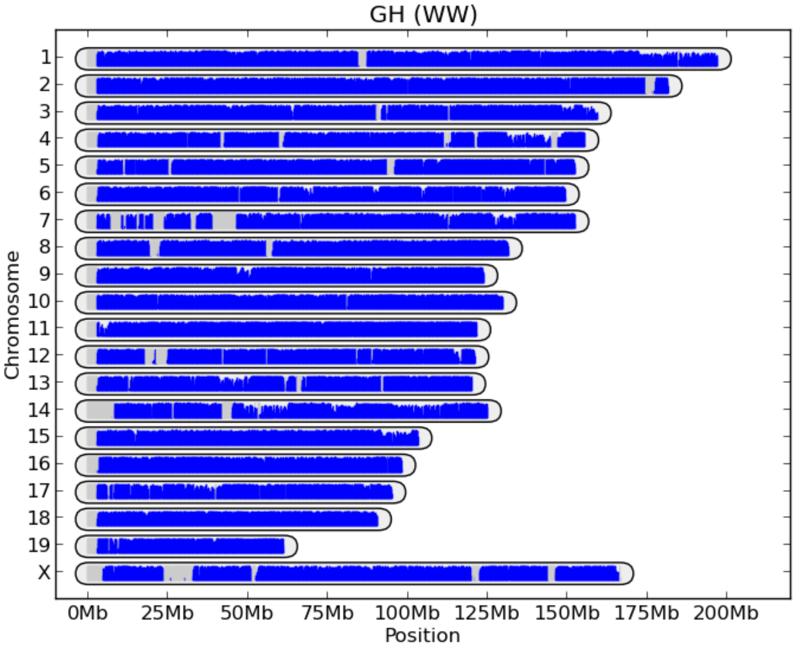


Figure A.28: Sequence similarity map for CC founders G (PWK/PhJ) and H (WSB/EiJ).

BIBLIOGRAPHY

[1] N. J. Armstrong, T. C. Brodnicki, and T. P. Speed. Mind the gap: analysis of marker-assisted breeding strategies for inbred mouse strains. *Mammalian Genome*, 17(4):273–87, Apr. 2006.

[2] D. L. Aylor, W. Valdar, W. Foulds-Mathes, R. J. Buus, R. a. Verdugo, R. S. Baric, M. T. Ferris, J. a. Frelinger, M. Heise, M. B. Frieman, L. E. Gralinski, T. a. Bell, J. D. Didion, K. Hua, D. L. Nehrenberg, C. L. Powell, J. Steigerwalt, Y. Xie, S. N. P. Kelada, F. S. Collins, I. V. Yang, D. a. Schwartz, L. a. Branstetter, E. J. Chesler, D. R. Miller, J. Spence, E. Y. Liu, L. McMillan, A. Sarkar, J. Wang, W. Wang, Q. Zhang, K. W. Broman, R. Korstanje, C. Durrant, R. Mott, F. a. Iraqi, D. Pomp, D. Threadgill, F. P.-M. de Villena, and G. a. Churchill. Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome Research*, 21(8):1213–22, Aug. 2011.

[3] J. F. Ayroles, M. A. Carbone, E. A. Stone, K. W. Jordan, R. F. Lyman, M. M. Magwire, S. M. Rollmann, L. H. Duncan, F. Lawrence, R. R. H. Anholt, and T. F. C. Mackay. Systems genetics of complex traits in Drosophila melanogaster. *Nature Genetics*, 41(3):299–307, Mar. 2009.

[4] D. W. Bailey. Recombinant-inbred strains. An aid to finding identity, linkage, and function of histocompatibility and other genes. *Transplantation*, 11(3):325–327, 1971.

[5] T. M. Beissinger, C. N. Hirsch, R. S. Sekhon, J. M. Foerster, J. M. Johnson, G. Muttoni, B. Vaillancourt, C. R. Buell, S. M. Kaeppler, and N. de Leon. Marker Density and Read-Depth for Genotyping Populations Using Genotyping-by-Sequencing. *Genetics*, 193:1073–1081, Feb. 2013.

[6] B. J. Bennett, C. R. Farber, L. Orozco, H. M. Kang, A. Ghazalpour, N. Siemers, M. Neubauer, I. Neuhaus, R. Yordanova, B. Guan, A. Truong, W.-p. Yang, A. He, P. Kayne, P. Gargalovic, T. Kirchgessner, C. Pan, L. W. Castellani, E. Kostem, N. Furlotte, T. a. Drake, E. Eskin, and A. J. Lusis. A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome research*, 20(2):281–90, Feb. 2010.

[7] P. Boddhireddy, J.-L. Jannink, and J. C. Nelson. Selective Advance for Accelerated Development of Recombinant Inbred QTL Mapping Populations. *Crop Science*, 49(4):1284, 2009.

[8] K. W. Broman. The genomes of recombinant inbred lines. *Genetics*, 169(2):1133–46, Mar. 2005.

[9] K. W. Broman. Genotype probabilities at intermediate generations in the construction of recombinant inbred lines. *Genetics*, 190(2):403–12, Feb. 2012.

[10] H. Brunschwig, L. Levi, E. Ben-David, R. W. Williams, B. Yakir, and S. Shifman. Fine-Scale Maps of Recombination Rates and Hotspots in the Mouse Genome. *Genetics*, 191(3):757–764, July 2012.

[11] E. S. Buckler, J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown, C. Browne, E. Ersoz, S. Flint-Garcia, A. Garcia, J. C. Glaubitz, M. M. Goodman, C. Harjes, K. Guill, D. E. Kroon, S. Larsson, N. K. Lepak, H. Li, S. E. Mitchell, G. Pressoir, J. A. Peiffer, M. O. Rosas, T. R. Rocheford, M. C. Romay, S. Romero, S. Salvo, H. S. Villeda, H. Sofia da Silva, Q. Sun, F. Tian, N. Upadyayula, D. Ware, H. Yates, J. Yu, Z. Zhang, S. Kresovich, and M. D. McMullen. The Genetic Architecture of Maize Flowering Time. *Science*, 325(5941):714–718, Aug. 2009.

[12] J. D. Calaway, A. B. Lenarcic, J. P. Didion, J. R. Wang, J. B. Searle, L. McMillan, W. Valdar, and F. Pardo-Manuel de Villena. Genetic Architecture of Skewed X Inactivation in the Laboratory Mouse. *PLoS Genetics*, 9(10):e1003853, Oct. 2013.

[13] E. J. Chesler, L. Lu, J. Wang, R. W. Williams, and K. F. Manly. WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. *Nature Neuroscience*, 7(5):485–486, May 2004.

[14] E. J. Chesler, D. R. Miller, L. R. Branstetter, L. D. Galloway, B. L. Jackson, V. M. Philip, B. H. Voy, C. T. Culiat, D. W. Threadgill, R. W. Williams, G. a. Churchill, D. K. Johnson, and K. F. Manly. The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. *Mammalian Genome*, 19(6):382–9, June 2008.

[15] D. M. Church, L. Goodstadt, L. W. Hillier, M. C. Zody, S. Goldstein, X. She, C. J. Bult, R. Agarwala, J. L. Cherry, M. DiCuccio, W. Hlavina, Y. Kapustin, P. Meric, D. Maglott, Z. Birtle, A. C. Marques, T. Graves, S. Zhou, B. Teague, K. Potamousis, C. Churas, M. Place, J. Herschleb, R. Runnheim, D. Forrest, J. Amos-Landgraf, D. C. Schwartz, Z. Cheng, K. Lindblad-Toh, E. E. Eichler, and C. P. Ponting. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biology*, 7(5):e1000112, May 2009.

[16] G. a. Churchill, D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie, J. Beatty, W. D. Beavis, J. K. Belknap, B. Bennett, W. Berrettini, A. Bleich, M. Bogue, K. W. Broman, K. J. Buck, E. Buckler, M. Burmeister, E. J. Chesler, J. M. Cheverud, S. Clapcote, M. N. Cook, R. D. Cox, J. C. Crabbe, W. E. Crusio, A. Darvasi, C. F. Deschepper, R. W. Doerge, C. R. Farber, J. Forejt, D. Gaile, S. J. Garlow, H. Geiger, H. Gershenfeld, T. Gordon, J. Gu, W. Gu, G. de Haan, N. L. Hayes, C. Heller, H. Himmelbauer, R. Hitzemann, K. Hunter, H.-C. Hsu, F. a. Iraqi, B. Ivandic, H. J. Jacob, R. C. Jansen, K. J. Jepsen, D. K. Johnson, T. E. Johnson, G. Kempermann, C. Kendziorski, M. Kotb, R. F. Kooy, B. Llamas, F. Lammert, J.-M. Lassalle, P. R. Lowenstein, L. Lu, A. Lusis, K. F. Manly, R. Marcucio, D. Matthews, J. F. Medrano, D. R. Miller, G. Mittleman, B. a. Mock, J. S. Mogil, X. Montagutelli, G. Morahan, D. G. Morris, R. Mott, J. H. Nadeau, H. Nagase, R. S. Nowakowski, B. F. O'Hara, A. V. Osadchuk, G. P. Page, B. Paigen, K. Paigen, A. a. Palmer, H.-J. Pan, L. Peltonen-Palotie, J. Peirce, D. Pomp, M. Pravenec, D. R. Prows, Z. Qi, R. H. Reeves, J. Roder, G. D. Rosen, E. E. Schadt, L. C. Schalkwyk, Z. Seltzer, K. Shimomura, S. Shou, M. J. Sillanpää, L. D. Siracusa, H.-W. Snoeck, J. L. Spearow, K. Svenson, L. M. Tarantino, D. Threadgill, L. a. Toth, W. Valdar, F. P.-M. de Villena, C. Warden, S. Whatley, R. W. Williams, T. Wiltshire,

N. Yi, D. Zhang, M. Zhang, and F. Zou. The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genetics*, 36(11):1133–7, Nov. 2004.

[17] C. C. Consortium. The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics*, 190(2):389–401, Feb. 2012.

[18] J. F. Crow. Haldane, Bailey, Taylor and Recombinant-Inbred Lines. *Genetics*, 176(2):729–732, June 2007.

[19] F. A. Cubillos, E. Billi, E. Zorgo, L. Parts, P. Fargier, S. Omholt, A. Blomberg, J. Warringer, E. J. Louis, and G. Liti. Assessing the complex architecture of polygenic traits in diverged yeast populations. *Molecular Ecology*, 20(7):1401–1413, 2011.

[20] A. Darvasi and S. M. Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics*, 141(3):1199–1207, 1995.

[21] P. Démant and A. A. M. Hart. Recombinant congenic strains A new tool for analyzing genetic traits determined by more than one gene. *Immunogenetics*, 24(6):416–422, 1986.

[22] E. J. Eisen. *The Mouse in Animal Genetics and Breeding Research*. Imperial College Press, London, 2005.

[23] L. Flaherty and V. Bolivar. Congenic and Consomic Strains. In *Neurobehavioral Genetics*, pages 115–127. CRC Press, Aug. 2006.

[24] K. A. Frazer, E. Eskin, H. M. Kang, M. A. Bogue, D. A. Hinds, E. J. Beilharz, R. V. Gupta, J. Montgomery, M. M. Morenzoni, G. B. Nilsen, C. L. Pethiyagoda, L. L. Stuve, F. M. Johnson, M. J. Daly, C. M. Wade, and D. R. Cox. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*, 448(7157):1050–1053, Aug. 2007.

[25] C.-P. Fu, C. E. Welsh, F. P.-M. de Villena, and L. McMillan. Inferring ancestry in admixed populations using microarray probe intensities. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, BCB '12, pages 105–112, New York, NY, USA, 2012. ACM.

[26] X. Gan, O. Stegle, J. Behr, J. G. Steffen, P. Drewe, K. L. Hildebrand, R. Lyngsoe, S. J. Schultheiss, E. J. Osborne, V. T. Sreedharan, A. Kahles, R. Bohnert, G. Jean, P. Derwent, P. Kersey, E. J. Belfield, N. P. Harberd, E. Kemen, C. Toomajian, P. X. Kover, R. M. Clark, G. Ratsch, and R. Mott. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*, 477(7365):419–423, Sept. 2011.

[27] G. Gibson and T. F. C. Mackay. Enabling population and quantitative genomics. *Genetical Research*, 80(1):1–6, Aug. 2002.

[28] C. C. Hudgins, R. T. Steinberg, D. M. Klinman, M. J. Reeves, and A. D. Steinberg. Studies of consomic mice bearing the Y chromosome of the BXSB mouse. *The Journal of Immunology*, 134(6):3849–3854, June 1985.

[29] H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics*, 178(3):1709–1723, Mar. 2008.

[30] T. M. Keane, L. Goodstadt, P. Danecek, M. a. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, N. a. Furlotte, E. Eskin, C. Nellå ker, H. Whitley, J. Cleak, D. Janowitz, P. Hernandez-Pliego, A. Edwards, T. G. Belgard, P. L. Oliver, R. E. McIntyre, A. Bhomra, J. Nicod, X. Gan, W. Yuan, L. van der Weyden, C. a. Steward, S. Bala, J. Stalker, R. Mott, R. Durbin, I. J. Jackson, A. Czechanski, J. A. Guerra-Assunção, L. R. Donahue, L. G. Reinholdt, B. a. Payseur, C. P. Ponting, E. Birney, J. Flint, and D. J. Adams. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–94, Sept. 2011.

[31] P. X. Kover, W. Valdar, J. Trakalo, N. Scarcelli, I. M. Ehrenreich, M. D. Purugganan, C. Durrant, and R. Mott. A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in *Arabidopsis thaliana*. *PLoS Genetics*, 5(7):e1000551, July 2009.

[32] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, Apr. 2012.

[33] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, Jan. 2009.

[34] K. Lindblad-Toh, E. Winchester, M. J. Daly, D. G. Wang, J. N. Hirschhorn, J. P. Laviolette, K. Ardlie, D. E. Reich, E. Robinson, P. Sklar, N. Shah, D. Thomas, J. B. Fan, T. Gingeras, J. Warrington, N. Patil, T. J. Hudson, and E. S. Lander. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genetics*, 24(4):381–6, Apr. 2000.

[35] E. Y. Liu, A. P. Morgan, E. J. Chesler, W. Wang, G. A. Churchill, and F. Pardo-Manuel de Villena. High-Resolution Sex-Specific Linkage Maps of the Mouse Reveal Polarized Distribution of Crossovers in Male Germline. *Genetics*, 197(1):91–106, May 2014.

[36] E. Y. Liu, Q. Zhang, L. McMillan, F. P.-M. de Villena, and W. Wang. Efficient genome ancestry inference in complex pedigrees with inbreeding. *Bioinformatics*, 26(12):i199–i207, June 2010.

[37] P. Markel, P. Shu, C. Ebeling, and G. Carlson. Theoretical and empirical issues for marker-assisted breeding of congenic mouse strains. *Nature*, 17(3):280–284, 1997.

[38] R. Mott, C. J. Talbot, M. G. Turri, a. C. Collins, and J. Flint. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(23):12649–54, Nov. 2000.

[39] J. H. Nadeau, J. B. Singer, A. Matin, and E. S. Lander. Analysing complex genetic traits with chromosome substitution strains. *Nature Genetics*, 24(3):221–225, Mar. 2000.

[40] K. Paigen. One Hundred Years of Mouse Genetics: An Intellectual History. I. The Classical Period (19021980). *Genetics*, 163(3):1–7, Nov. 2003.

[41] K. Paigen and J. T. Eppig. A mouse phenome project. *Mammalian Genome*, 11(9):715–717, 2000.

[42] K. Paigen, J. P. Szatkiewicz, K. Sawyer, N. Leahy, E. D. Parvanov, S. H. S. Ng, J. H. Graber, K. W. Broman, and P. M. Petkov. The Recombinational Anatomy of a Mouse Chromosome. *PLoS Genetics*, 4(7):e1000119, July 2008.

[43] E. D. Parvanov, P. M. Petkov, and K. Paigen. Prdm9 Controls Activation of Mammalian Recombination Hotspots. *Science*, 327(5967):835, Feb. 2010.

[44] M.-J. a. Paulo, M. Boer, X. Huang, M. Koornneef, and F. Eeuwijk. A mixed model QTL analysis for a complex cross population consisting of a half diallel of two-way hybrids in *Arabidopsis thaliana*: analysis of simulated data. *Euphytica*, 161(1-2):107–114, Feb. 2008.

[45] J. E. Pool, I. Hellmann, J. D. Jensen, and R. Nielsen. Population genetic inference from genomic sequence variation. *Genome Research*, 20(3):291–300, Mar. 2010.

[46] A. L. Price, A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, 5(6):e1000519, June 2009.

[47] D. T. Qi Zhang, Wei Wang, Leonard Mcmillan, Fernando Pardo-Manuel De Villena. Inferring genome-wide mosaic structure. *Proceedings of the Pacific Symposium on Biocomputing*, pages 150–161, 2009.

[48] D. T. Qi Zhang, Wei Wang, Leonard McMillan, Jan Prins, Fernando Pardo-Manuel de Villena. Genotype sequence segmentation: Handling constraints and noise. *Proceedings of the Workshop on Algorithms in Bioinformatics*, 2008.

[49] A. Roberts, F. Pardo-Manuel de Villena, W. Wang, L. McMillan, and D. W. Threadgill. The polymorphism architecture of mouse genetic resources elucidated using genome-wide resequencing data: implications for QTL discovery and systems genetics. *Mammalian Genome*, 18(6-7):473–81, July 2007.

[50] T. B. Sackton, R. J. Kulathinal, C. M. Bergman, A. R. Quinlan, E. B. Dopman, M. Carneiro, G. T. Marth, D. L. Hartl, and A. G. Clark. Population genomic inferences from sparse high-throughput sequencing of two populations of Drosophila melanogaster. *Genome Biology and Evolution*, 1:449–65, Jan. 2009.

[51] S. Sankararaman and S. Sridhar. Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, 82(2):290–303, 2008.

[52] S. Shifman, J. T. Bell, R. R. Copley, M. S. Taylor, R. W. Williams, R. Mott, and J. Flint. A High-Resolution Single Nucleotide Polymorphism Genetic Map of the Mouse Genome. *PLoS Biology*, 4(12):e395, Nov. 2006.

[53] L. M. Silver. *Mouse Genetics*. Oxford University Press, Oxford, UK, 1995.

[54] A. Sundquist, E. Fratkin, C. Do, and S. Batzoglou. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Research*, 18(4):676–82, Apr. 2008.

[55] K. L. Svenson, D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng, E. J. Chesler, A. A. Palmer, L. McMillan, and G. A. Churchill. High-Resolution Genetic Mapping Using the Mouse Diversity Outbred Population . *Genetics*, 190(2):437–447, Feb. 2012.

[56] B. A. Taylor, H. Meier, and D. D. Myers. Host-Gene Control of C-Type RNA Tumor Virus: Inheritance of the Group-Specific Antigen of Murine Leukemia Virus. *Proceedings of the National Academy of Sciences*, 68(12):3190–3194, Dec. 1971.

[57] R. W. Threadgill, D. W.; Hunter, K. W; Williams. Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mammalian Genome*, 13:175–178, 2002.

[58] W. Valdar, J. Flint, and R. Mott. Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. *Genetics*, 172(3):1783–97, Mar. 2006.

[59] J. Wang, R. Williams, and K. Manly. WebQTL. *Neuroinformatics*, 1(4):299–308, 2003.

[60] C. E. Welsh and L. McMillan. Accelerating the inbreeding of multi-parental recombinant inbred lines generated by sibling matings. *G3: Genes— Genomes— Genetics*, 2(2):191–8, Feb. 2012.

[61] C. E. Welsh, D. R. Miller, K. F. Manly, J. Wang, L. McMillan, G. Morahan, R. Mott, F. a. Iraqi, D. W. Threadgill, and F. P.-M. de Villena. Status and access to the Collaborative Cross population. *Mammalian Genome*, July 2012.

[62] B. Yalcin, J. Nicod, A. Bhomra, S. Davidson, J. Cleak, L. Farinelli, M. Ø sterås, A. Whitley, W. Yuan, X. Gan, M. Goodson, P. Klenerman, A. Satpathy, D. Mathis, C. Benoist, D. J. Adams, R. Mott, and J. Flint. Commercially available outbred mice for genome-wide association studies. *PLoS Genetics*, 6(9), Sept. 2010.

[63] B. Yalcin, K. Wong, A. Agam, M. Goodson, T. M. Keane, X. Gan, C. Nellaker, L. Goodstadt, J. Nicod, A. Bhomra, P. Hernandez-Pliego, H. Whitley, J. Cleak, R. Dutton, D. Janowitz, R. Mott, D. J. Adams, and J. Flint. Sequence-based characterization of structural variation in the mouse genome. *Nature*, 477(7364):326–329, Sept. 2011.

[64] H. Yang, Y. Ding, L. N. Hutchins, J. Szatkiewicz, T. a. Bell, B. J. Paigen, J. H. Graber, F. P.-M. de Villena, and G. a. Churchill. A customized and versatile high-density genotyping array for the mouse. *Nature Methods*, 6(9):663–6, Sept. 2009.

[65] H. Yang, J. R. Wang, J. P. Didion, R. J. Buus, T. a. Bell, C. E. Welsh, F. Bonhomme, A. H.-T. Yu, M. W. Nachman, J. Pialek, P. Tucker, P. Boursot, L. McMillan, G. a. Churchill, and

F. P.-M. de Villena. Subspecific origin and haplotype diversity in the laboratory mouse. *Nature Genetics*, 43(7):648–55, July 2011.