

ANALYSIS OF ADMIXED ANIMALS USING INDIRECT HAPLOTYPE
INFORMATION FROM EXISTING TECHNOLOGIES

Chen-Ping Fu

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
2015

Approved by:

Leonard McMillan

Fernando Pardo-Manuel de Villena

Vladimir Jojic

Jan Prins

Wei Sun

Fei Zhou

© 2015
Chen-Ping Fu
ALL RIGHTS RESERVED

ABSTRACT

CHEN-PING FU: ANALYSIS OF ADMIXED ANIMALS USING INDIRECT
HAPLOTYPE INFORMATION FROM EXISTING TECHNOLOGIES.

(Under the direction of Leonard McMillan.)

The use of genotyping and sequencing technologies in genetic studies typically involves inspecting variants defined within a single reference genome. While this definition of genetic variation promotes a simple model of the genome that is easy to organize and analyze, it does not encompass the full breadth of variation possible between individuals. Fortunately, existing technologies capture information about genomic variation outside the original targeted variants. By incorporating these low-level signals, which classical methods generally regard as noise, we can make more accurate inferences about the relationship between admixed animals and their ancestral and parental strains. In this thesis, I use both genotyping microarrays and RNA sequencing data to demonstrate the utility of using signals from ancestral haplotype data to analyze admixed animals.

I introduce a novel method for designing a genotyping microarray that provides maximal information about ancestral haplotypes for the admixed population Collaborative Cross (CC). The result is the 78K-marker MegaMUGA array, which achieves high call rates and distinction power in a diverse set of mouse strains.

Using probe intensities from microarrays such as the MegaMUGA, I develop methods for founder haplotype inference as well as quantitative trait loci (QTL) mapping. I show that these intensity-based methods outperform traditional genotype call-based methods due to their ability to capture additional information about the local sequence,

which I confirm using high-throughput sequencing data within probe regions.

In addition to demonstrating my thesis with microarray intensity data, I also use RNA-seq read data from parental strains to estimate allele-specific expression (ASE) in the F1 offspring. By directly using parental read data as features in a regularized regression problem, I can achieve accurate estimations of the offspring's expressed gene transcripts and allele-specific expression levels, showing that no matter the data source, incorporating low-level signals directly from ancestral strains provides a more accurate template for analysis of admixed strains.

ACKNOWLEDGMENTS

Many thanks to my advisor Leonard McMillan, whose enthusiasm and offbeat yet brilliant ideas make this work possible. I am grateful for the time, advice, knowledge, and conversation he has generously shared with me.

I have been fortunate to collaborate with Fernando Pardo-Manuel de Villena, whose deep insights into both biological and non-biological topics amaze me. To my other committee members – Vladimir Jojic, Jan Prins, Wei Sun, and Fei Zou, thank you for the support and the insightful discussions that help enrich this research.

Thank you to Jeremy Wang, the first to read and revise this document. I’ve had a fantastic group of lab mates throughout the years – Catherine Welsh, Katy Kao, Matt Holt, and many others – thank you for making the lab such an enjoyable place to work.

Finally, love and gratitude to my parents for giving me the perfect childhood and never sending me to cram school, and much love to my husband Teddy who supports me in all but my craziest ideas.

TABLE OF CONTENTS

LIST OF TABLES	x
LIST OF FIGURES	xi
1 Introduction	1
1.1 Selection of Microarray Markers	3
1.2 Analysis of Microarray Markers	4
1.3 Analysis of High-Throughput Sequencing	5
1.4 Thesis Statement	5
1.5 Organization	6
2 Background	8
2.1 Genetic Structure and Nomenclature	8
2.1.1 DNA and Chromosome Structure	8
2.1.2 DNA to RNA	9
2.1.3 Genetic Recombination and Inheritance	9
2.2 Inbreeding to Achieve Genetic Reproducibility	11
2.3 Genetic Reference Populations	12
2.3.1 Collaborative Cross	13
2.3.2 Diversity Outbred	14
2.4 Tools for Analyzing Mouse Genomes	15
2.4.1 Mouse reference genome	15
2.4.2 Genotyping Microarrays	16
2.4.3 Next-Generation Sequencing	17
2.5 Conclusion	18

3	Microarray Design and Marker Selection	20
3.1	Introduction	20
3.2	Methods	23
3.2.1	Determining SNP marker spacing	24
3.2.2	Filtering available SNPs	25
3.2.3	Establishing a minimum marker window for informative SNPs . . .	27
3.2.4	Selecting maximally informative SNPs	29
3.2.5	Selecting non-SNP markers	30
3.3	Results	33
3.3.1	Haplotype Informativeness of Selected SNP Markers	33
3.3.2	The Final Manufactured MegaMUGA Array	34
3.3.3	Pseudoautosomal Region Markers on the MegaMUGA	41
3.4	Discussion	44
4	Ancestry Inference	47
4.1	Introduction	47
4.2	Materials and Methods	50
4.2.1	Materials	50
4.2.2	Algorithm overview	52
4.2.3	Creating reference clusters on MUGA	52
4.2.4	Creating reference clusters on MegaMUGA	53
4.2.5	Distance Model	54
4.2.6	Hidden Markov Model	57
4.2.7	Refining recombination breakpoints	58
4.2.8	Funnel constraints	58
4.3	Results	59
4.3.1	Reference intensity clusters	59

4.3.2	The role of off-target variants in intensity clusters	61
4.3.3	Ancestry inference comparisons using sequencing data	63
4.3.4	Other platforms and populations	68
4.4	Discussion	69
5	Mapping Quantitative Trait Loci	77
5.1	Introduction	77
5.2	Methods	79
5.2.1	Constructing distance matrices	80
5.2.2	Comparing distance matrices and significance	81
5.3	Results	85
5.3.1	Simulated data	86
5.3.2	Real data	87
5.3.3	Efficiency and memory	89
5.4	Discussion	89
6	Estimating Allele-Specific Expression using RNA-Seq Data	94
6.1	Introduction	94
6.2	Approach	97
6.2.1	Notation	99
6.2.2	Regression model	99
6.3	Methods	100
6.3.1	Simulated data	100
6.3.2	Real data	101
6.3.3	Selecting candidate transcripts	102
6.3.4	Coordinate descent	102
6.4	Results	104
6.4.1	Synthetic data results	104

6.4.2 Real data results	107
6.4.3 Speed and Memory	112
6.5 Discussion	113
7 Discussion and Conclusion	116
7.1 Microarray Design	117
7.2 Ancestry Inference	119
7.3 Quantitative Trait Loci Mapping	120
7.4 Estimating Allele-Specific Expression	123
7.5 Conclusion	124
BIBLIOGRAPHY	125

LIST OF TABLES

3.1	Mean number of distinguishable founder states in 5-SNP windows	34
3.2	Number of discordant markers between replicates on MegaMUGA	40
3.3	Informative markers between sister strains on MegaMUGA	41
4.1	Transitions between different states p and q	55
4.2	SNPs that disagree between the Distance Model (DM) vs. GAIN	66
4.3	Comparison of MUGA and MegaMUGA solutions to sequence data . . .	67
5.1	QTL positions for colitis phenotype	87
6.1	Notation	99
6.2	Dimensions and Results from Real Data	108
6.3	Comparisons to maternal contribution ratios found in [17]	111

LIST OF FIGURES

2.1	Genetic recombination through crossover	10
2.2	The founder strains of the CC and DO	13
2.3	Collaborative Cross breeding scheme	14
2.4	Illustration of a sample microarray marker	17
3.1	Flowchart for MegaMUGA CC SNP selection	24
3.2	Linkage map vs. Genomic position	25
3.3	Female-male MDA marker intensity ratio in the PAR	32
3.4	Number of distinguishable founder states using selected SNPs	35
3.5	Number of distinguishable founder states after discarding 10% of SNPs	36
3.6	Number of uniquely distinguishable founder states using random SNPs	36
3.7	Distribution of all MegaMUGA markers	38
3.8	All MegaMUGA markers colored by marker type	39
3.9	Illumina genotype calls on MegaMUGA PAR markers	42
3.10	PCA on FVB/NJ x (PWK/PhJ x CAST/EiJ) in the PAR	45
4.1	Intensity plots of four genotyping markers	49
4.2	Creating reference clusters in the MUGA	54
4.3	Intensity plots with replicates of CC founders highlighted in MUGA	60
4.4	The number of CC homozygous intensity clusters of MUGA markers	61
4.5	A MegaMUGA marker with 9 clusters within the CC founders and F1s	62
4.6	Distance to the closest variant among all documented CC SNPs	63

4.7	MUGA OTVs and their effects on intensity clusters	64
4.8	Two MUGA markers where OTVs result in unexpected intensities . . .	65
4.9	Intensity better resolves a breakpoint in sample OR1237m224	71
4.10	Erroneous genotypes result in errors in call-based ancestry inference . .	72
4.11	Comparison of ancestry inference results in MUGA and MegaMUGA .	73
4.12	Comparison between Distance Model and HMM solutions	74
4.13	The ancestry of a transgenic mouse	75
4.14	The ancestry of a transgenic mouse with non-CC backgrounds	76
5.1	Results from simulated data with varying window sizes	82
5.2	Intensity distances between pairs from 111 backcrossed samples	83
5.3	Intensity distances between pairs from 67 CC lines	84
5.4	Results on simulated QTLs created in 54 largely inbred CC samples . .	92
5.5	Full genome scans for QTLs affecting the albinism trait	93
5.6	My method applied to samples from Rogala et al. [55]	93
6.1	Pipeline for estimating allele-specific expression in F1 animals	98
6.2	Predicted versus actual expression levels from synthetic data	105
6.3	True positive rate vs. false positive rate for different values of λ	106
6.4	Stacked histogram of k-mers in the CASTxPWK k-mer profile	109
6.5	Histograms of estimated gene expression levels from real data	109
6.6	Histogram of the maternal contribution ratios of all expressed genes . .	111
6.7	Histogram of the maternal contribution ratios of X-chromosome genes .	112
7.1	Two MUGA markers that capture unexpected alleles beyond the CC .	121

Chapter 1 : Introduction

One main goal of genetics is to understand how differences in the genome relate to differences in individual traits. To achieve this goal, scientists often need to study large populations of individuals with diverse genetic backgrounds.

Genetic diversity is largely due to differences in the DNA sequences between individuals. Different individuals can carry different versions of genes, and these various versions are referred to as alleles. One common variation within genomes is a change at a single base in the DNA sequence known as a single-nucleotide polymorphism (SNP). Although there are other possible genetic variations that can span much longer regions of the genome, genetic studies have traditionally used SNPs as markers of local genetic variation, since they are common variants that are easy to detect. SNP-based studies measure SNPs between individuals at different locations within the genome, using a single standard genome as a sequence template.

Since individual genomes can vary significantly within the same species, this standard genome, called the reference genome, offers a convenient platform for annotating small and common genetic variants. This model of defining genetic variation in terms of SNPs in a reference genome has created many commonly used tools and techniques within the genetics community. However, individual genomes within the same species can differ greatly from each other and from the reference genome, which is obtained from one of very few members of the species. By using the reference genome—which is often unrelated to the experiment at hand—as the main template, large or novel variants which could be highly informative to downstream analyses may be ignored.

Fortunately, the tools for measuring SNPs often contain more information about

nearby regions that are frequently inherited together. Since these tools usually assess a SNP by using the sequence around it, variants close to a target SNP can cause subtle changes in signal that are usually viewed as noise. Yet since these subtle signal variations originate from actual sequence changes, they can convey valuable information about the region around the target SNP, allowing for finer discrimination between different genomes.

To obtain diverse individual genomes for reproducible experiments, scientists often use genetic reference populations with individuals that are bred from a set of known ancestors. The individuals in these populations have genomes composed of a mixture of known ancestral genomes, and they are therefore known as admixed individuals.

Admixed genomes are mosaics of their ancestral genomes because genetic information is passed from generation to generation through recombined segments of DNA. Therefore, genes are not inherited in the form of individual variants, but in contiguous subsequences from different ancestors. These segments of DNA that can be traced back to different ancestral genomes are called haplotypes. In practice, a haplotype can represent a set of variants within a region of the genome that are frequently inherited together.

In the analysis of admixed genomes, data is gathered from the admixed individuals as well as their ancestors. When ancestral genomes are measured with existing SNP-based tools, we can exploit the underlying subtle haplotype information to more accurately infer the relationship between the admixed individual and its ancestors. Using ancestral haplotype data gives a broader and more accurate view of the genome beyond single SNPs, and it provides a more direct means of comparison since the admixed animals are measured against the genomes from which they truly descended.

In this dissertation, I discuss improving the accuracy of analysis of admixed animals from genetic reference populations through the use of haplotype data from ancestral

genomes. I will first introduce the relevant biological terminology and background, then I will discuss the use of ancestral haplotypes in the context of genotyping array design, ancestry inference, quantitative trait loci mapping, and estimation of allele-specific expression.

1.1 Selection of Microarray Markers

Genotyping microarrays have been commonly used to determine the version of a gene, otherwise known as the allele, of individual samples at selected regions of the genome. The most common type of variation assessed with microarrays are SNPs. The classical use of a SNP genotyping microarray is, therefore, to sample different genomes at pre-specified points in their sequences, using probes that contain sequences adjacent to the SNPs of interest.

Since microarrays are often designed with the goal of probing SNPs within regions of functional consequence, many SNPs are selected based on annotations within known functional regions of interest. Another common selection criteria is that SNPs are spaced uniformly and their alleles segregate among a small set of strains, which are animals sharing the same ancestral genome sequences. While these selection techniques capture SNPs in functional regions and SNPs that are independently informative, they do not take into account the informativeness of SNPs within the scope of the entire genome. For instance, two nearby SNPs may be frequently inherited together, and having both these SNPs on an array does not provide much information beyond having just one. Since adjacent SNPs are frequently inherited together, neighboring SNPs should be considered during microarray design in order for the microarray to be locally informative. I present a method for microarray design that considers regions of several SNPs at once, so that nearby SNPs that are selected convey the optimal amount of information about the haplotype.

1.2 Analysis of Microarray Markers

In classical microarray studies, genotype alleles are determined across all markers using a set of samples. When a microarray is first manufactured, a group of individuals are selected to represent the amount of diversity expected within the majority of the population, and these individuals are the first to be genotyped using the microarray. Their hybridization intensities with the microarray probes then determine the thresholds at which each genotype allele is assigned. In the past, due to the high cost of genotyping a sample, each of these individuals is represented only once in the initial genotyping, and samples that may have minor differences are easily pooled together into the same genotype allele group. In a SNP genotyping microarray, three alleles, or variations, are typically defined: the wild type or ‘A’ allele, the mutant or ‘B’ allele, and the heterozygous allele ‘H’ which occurs when an individual has a different allele on each of two chromosomes. At each SNP marker, samples are then labeled as having one of the three variations, or having a “no call” allele ‘N’.

As the cost of genotyping microarrays has dropped, genotyping biological and technical replicates of the same strain has become increasingly feasible. Subtle differences between similar strains can be more easily distinguished now that scientists can afford to have several replicated genotyped samples of each animal. However, the microarray genotyping pipeline has largely remained the same, and typical genotyping microarrays still report only three possible variations for each marker, operating under the assumption that the target SNP is the only local variant across the population. In reality, the probe sequences around the target SNP can often contain variants within different strains, which then manifest as subtle variations in probe hybridization intensities. Therefore, SNP microarrays can often capture more alleles than the three possibilities that traditional methods assume.

1.3 Analysis of High-Throughput Sequencing

The advent of high-throughput sequencing technology has created a new set of challenges related to processing and understanding millions of short reads from a single genome or transcriptome. The standard technique for analyzing these short reads is to align them to the reference genome, allowing for a few mismatches within each aligned read. Since the possibilities for additional variations are endless, it is difficult to incorporate many SNPs or non-SNP variants within a read. This creates the problem of reference bias when some strains in the experiment are more similar to the reference and have more aligned reads, while other strains are dissimilar to the reference and have many unaligned or discarded reads. The substitution of known SNPs from each strain into the reference genome helps alleviate the problem of reference bias [32], but this technique still does not capture the full spectrum of variants and requires prior knowledge of SNPs from each strain and the modification of the reference genome for each ancestral strain.

However, when the ancestral strains of an admixed animal are also sequenced, the read data from the ancestors can be leveraged to provide a more accurate model for the admixed genome. By comparing the read data of the admixed strain directly to the read data of its ancestral strains, we can eliminate the need for annotated SNPs or reference alignment.

1.4 Thesis Statement

Low-level signals in existing tools capture more information about ancestral haplotypes than that examined in classical methods. By incorporating information that is generally regarded as noise, we can make more accurate inferences about the relationship between admixed animals and their ancestral strains. I present a novel approach

for the design of highly informative genotyping microarrays by maximizing distinguishable ancestral haplotypes. I then show that these microarrays are most informative when viewed on the haplotype scale in terms of probe intensities, as demonstrated by my methods for ancestry inference and quantitative trait loci (QTL) mapping. The use of ancestral haplotypes is further extended to RNA-seq data, where I show that direct use of ancestral data in place of the standard reference leads to more accurate allele-specific expression analysis.

1.5 Organization

This dissertation is organized into the following chapters:

- **Chapter 2** presents the relevant biological background of mouse genetics. This includes brief discussions of basic genetics and inheritance, genetic reference populations from which we obtained data, and the tools for genetic analysis used in this dissertation.
- **Chapter 3** presents the principles for designing a genotyping microarray that is maximally informative on the haplotype level, as well as the use of probe intensities to analyze non-SNP markers.
- **Chapter 4** presents methods for inferring ancestry using informative genotyping microarrays. These methods use microarray intensities, which I show contain more information about the underlying probe sequence than typical biallelic SNP genotypes.
- **Chapter 5** presents a method for mapping quantitative trait loci (QTL). As in Chapters 3 and 4, the method introduced for QTL mapping uses microarray intensities, which bypasses intermediate analysis steps and achieves more accurate results.

- **Chapter 6** moves beyond microarrays and applies the concept of using ancestor haplotype data to high-throughput RNA sequencing (RNA-seq). It presents a method for determining allele-specific expression (ASE) in F1 mice using RNA-seq data from parental strains.
- **Chapter 7** concludes this thesis and discusses potential areas for future research.

Chapter 2 : Background

2.1 Genetic Structure and Nomenclature

2.1.1 DNA and Chromosome Structure

The genes that we inherit from our ancestors reside in DNA. DNA is a molecule composed of a long sequence with four nucleotides as building blocks: adenine (A), cytosine (C), guanine (G), or thymine (T). Each DNA sequence has a complement sequence, so that the DNA consists of two strands, one called the forward strand, and the other called the reverse strand, contributing to the double helix structure of DNA. The nucleotides A and T on different strands bind to form a base pair, and C and G bind to form a base pair.

DNA molecules consist of long strands of base pairs and are packaged into units called chromosomes. Most mammals are diploid, meaning they have two copies of each chromosome, with one copy inherited from the mother and the other from the father. Humans have 23 pairs of chromosomes, while mice have 20 pairs. Both humans and mice have one pair of sex chromosomes known as the X and Y chromosomes, and individuals with two copies of the X chromosome are biologically female, while individuals with one copy of X and one of Y are biologically male. The 22 remaining chromosome pairs in human (19 in mice) are referred to as autosomes. In addition to these chromosome pairs, human and mouse cells have many copies of mitochondrial DNA, a small, circular chromosome that is inherited solely from the mother.

As a diploid mammal with a short reproductive cycle and small size, the house mouse – of the species *Mus musculus* – has long been used as a model organism for

genetic research. The length of the mouse genome is approximately 2.7 billion base pairs, which makes its size similar to the human genome, which is approximately 3.3 billion base pairs.

2.1.2 DNA to RNA

DNA acts as a stable template of biological guidelines; for these guidelines to become functional, DNA is copied through an intermediary called the RNA, which carries the genetic instructions within the organism. RNA is also composed of nucleotide bases, and different RNA molecules are made of nucleotide sequences copied from different regions of the DNA template. This copying of information from DNA to RNA is called transcription, and some of the transcribed RNA is then made into proteins, while others serve different functional roles within the cell. Different regions of the genome are transcribed into RNA in different cell types, and the amount of RNA transcribed varies by genetics, cell type, cell state, and environmental conditions. The expression of different genes are related to the abundance of RNA in different cells, so the abundance level of RNA is an active area of study in genetics.

2.1.3 Genetic Recombination and Inheritance

Genes are functional units in the DNA, and they vary in size from hundreds to millions of base pairs. Genetic diversity within a species can be attributed to variations in the DNA at the base pair level, which then create different functional variations of genes called alleles. One common type of sequence variation leading to different alleles is a single base change, called a single-nucleotide polymorphism (SNP). Deletion, insertion, or duplication of one or more base pairs are other possible allele sources, with some larger variations involving sequences upwards of thousands of base pairs.

Since different alleles arise due to rare mutations in the DNA sequence, each variable

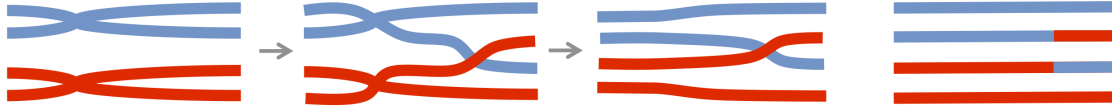


Figure 2.1: During the meiosis stage in sexual reproduction, the homolog chromosomes from the maternal and paternal sides pair together and crossover at some locus to produce four separate chromosomes, two with new combinations of alleles. One of these four chromosomes is then passed down to the offspring.

position within the genome typically only has two possible variants, as mutations very seldom occur at the same position along a sequence of billions of base pairs. These positions with two possible allele variants are called biallelic. Classically, one allele at a locus is called the reference or wild type allele, and the other allele is called the alternate or mutant allele.

As a consequence of being diploid, most mammals can be in one of three states at each locus. The two copies of a chromosome on a diploid animal can both contain the wild type allele, both contain the mutant allele, or one can contain the wild type and the other the mutant allele. Loci where both chromosomes have the same allele, either wild type or mutant, are called homozygous, and loci where the two chromosomes have different alleles are called heterozygous.

During meiosis, the process by which gametes are generated, the chromosomes undergo recombination. Recombination refers to the event where two homolog chromosomes (one inherited from the mother and one inherited from the father) pair together and crossover at some locus, so that chromosomes with new combinations of alleles are passed down to subsequent generations. This shuffling of gene alleles on each chromosome is the source of genetic diversity in subsequent generations.

Due to the interaction of a complex set of factors, including the accessibility of DNA packed into each chromosome, certain locations in the genome are more prone to crossover recombination events across an entire population, whereas other locations

tend to be more conserved. Regions with frequent crossover events are called recombination hotspots, and the genetic diversity of a population is often greatest in these regions due to the many different allele combinations.

2.2 Inbreeding to Achieve Genetic Reproducibility

When two parental genomes are very similar, recombination breakpoints are more difficult to detect since crossover events often do not create chromosomes with distinguishing alleles near the crossover breakpoint. By continuously breeding many generations of increasingly genetically similar individuals together, such as through father-daughter or sibling matings, scientists can create animals that have identical copies of each chromosome pair. This process is called inbreeding and results in inbred strains that are homozygous at each locus over the entire genome. Continued breeding within two members of the same line with identical chromosome pairs produces further generations with genomes identical to their parents. This intrafamily breeding scheme is known as inbreeding. Inbred animals are homozygous at every locus since they inherit the same alleles from both parents, effectively resulting in a loss of the heterozygous allele combination.

Inbreeding is a powerful technique for generating reproducible genomes within model organisms, significantly increasing the power of experiments that require a large number of samples with identical genomes. In the laboratory setting, animals are often inbred for many generations and maintained as inbred strains have identical genomes. On the other hand, animals in the wild typically have a mix of homozygous and heterozygous alleles, and they are referred to as outbred animals.

In this dissertation, I will discuss inbred populations whose genomes are assumed to be homozygous at every locus, as well as outbred populations whose genomes are heterozygous at some loci.

2.3 Genetic Reference Populations

To facilitate experiment design, geneticists often wish to know and control the genetic background of model organisms such as mice. The creation of genetic reference populations with specified breeding schemes enables scientists to manipulate the genetic makeup of such model organisms.

One simple and common type of genetic reference population is one consisting of different inbred strains, meaning they are homozygous at every locus, with the maternal and paternal chromosomes both containing the same allele. However, reference populations of standard inbred strains represent limited genetic diversity with unequal relatedness amongst individuals, and to obtain a large reference population with more reproducible strains that are also more genetically diverse, scientists often mix together different strains through various breeding schemes.

The simplest type of such a breeding scheme is the creation of F1 animals, or the first-generation offspring of two different inbred parents. Since the segregation of alleles between different allelic pairs are independent of one another during gamete formation, further breeding between different F1 animals and subsequent generations produces many more novel allele combinations than those present in the original founder strains [51]. After one or more generations of outbreeding to mix genomes, inbreeding commences to create genomes that can be replicated. The end result of such an inbreeding scheme is a population of inbred lines that are mosaics of their founders, and whose genomes can be replicated through continuous inbreeding. This ability to create reproducible genomes of many different founder mosaics greatly increases the power of controlled biological experiments, as well as decreases the need for genotyping each individual animal.

Since humans are outbred, many experiments use outbred animals to more accu-









Founder strain	Letter code	Color code	
A/J [1]	A	Yellow	
C57BL/6J [2]	B	Grey / Black	
129S1/SvImJ [3]	C	Pink	
NOD/ShiLtJ [4]	D	Blue	
NZO/H1LtJ [5]	E	Cyan	
CAST/EiJ [6]	F	Green	
PWK/PhJ [7]	G	Red	
WSB/EiJ	H	Purple	

Figure 2.2: The eight founder strains of the CC and DO, along with their letter codes and color designations.

rately model the human genome. Reproducible outbred populations can be created by making an F1 of two different inbred mosaics, creating countless combinations of outbred genomes that more closely model the human genome.

This dissertation makes use of two genetic reference populations of mice: the Collaborative Cross (CC), an inbred population, and the Diversity Outbred (DO), an outbred population, both of which are derived from the same set of eight inbred founders.

2.3.1 Collaborative Cross

The CC is an inbred mouse population with strains originating from a set of eight inbred founders – five classical inbred mouse strains A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, NZO/H1LtJ, and three wild-derived inbred mouse strains CAST/EiJ, PWK/PhJ, WSB/EiJ. [13, 67]. For ease of reference, each founder strain is assigned a letter code and a color (Figure 2.2), which will remain consistent throughout this dissertation. The eight founders were chosen to capture a high level of genetic diversity, representing on average 90% of known genetic variation across all 1-Mb intervals [53], and capturing 45 million segregating SNPs—four times those found in classical laboratory strains [81].

The CC lines were initiated in 2004 using a breeding scheme involving three genera-

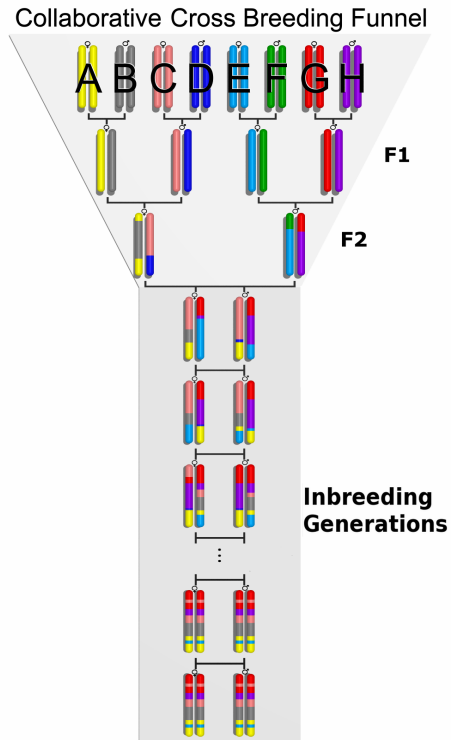


Figure 2.3: Collaborative Cross breeding scheme. Each funnel has an ordered list of eight founders which are crossed for three generations, then inbred for at least 20 generations to obtain recombinant inbred lines.

tions of outcrosses to incorporate all eight founder genomes, followed by many generations of inbreeding to generate reproducible genomes, as shown in Figure 2.3. Over one hundred inbred lines, each with a unique mosaic of the eight founders, are maintained through continued sibling matings.

By crossing two animals from different CC lines, scientists can create reproducible F1 animals with known but diverse genomes. Such a cross generated from two parental CC lines is known as a Recombinant Inbred Intercross (RIX).

2.3.2 Diversity Outbred

The DO mice originate from the same set of eight founders as the CC mice, and they were produced at the Jackson Laboratory by outbreeding 160 early lines from the

CC in such a way that maintained a balanced mixture of the founders [65, 15]. DO animals have a high degree of heterozygosity, and each DO mouse has a unique genome, so unlike a CC animal whose genome is consistent within its own line, or a CC RIX whose genome can be predicted through its parental lines, each DO animal’s genome mosaic can only be discovered through genotyping or sequencing the individual animal. The DO population therefore has a high resolution for genetic mapping, since their outbred genomes can capture any combination of the diversity present in the eight CC founders.

2.4 Tools for Analyzing Mouse Genomes

Since this dissertation analyzes the genomes of admixed mouse strains, I will discuss common tools for genetic analysis within the mouse community. The tools discussed here are not exclusive to mouse genetics; reference genomes are commonly used in both humans and model organisms, and genotyping microarrays and next-generation sequencing are the two most commonly used technologies for assessing the genomes of individual organisms.

2.4.1 Mouse reference genome

The genomes of individuals within the same species can vary greatly with SNPs, insertions, deletions, and large-scale genomic rearrangements. However, when conducting genetic experiments, it is often convenient to have a single representative genome to use as a reference for annotations. The mouse reference genome was assembled from samples of the classical inbred strain C57BL/6J [9]. Common genomic variants including SNPs and indels are catalogued in public databases such as Sanger Institute’s Mouse Genome Project [35], NCBI’s dbSNP [59], and Jackson Laboratory’s MGI [2], and the genomic positions of these variants are all catalogued in terms of chromosomes

and positions within the reference genome. Large-scale variations with sequences that do not occur in the reference strain but occur in other strains may have undefined reference positions for annotation purposes.

The analysis in this dissertation is done in NCBI37/mm9, or Build 37, assembled by NCBI and the Mouse Genome Sequencing Consortium [12]. The latest released reference genome assembly as of this writing is GRCm38, or Build 38, by the Genome Reference Consortium [21].

2.4.2 Genotyping Microarrays

Microarrays have long been used to sample the genome of an individual relative to a set of gene alleles, or genotypes [63, 42, 60]. The majority of genotyping microarrays detect SNPs through short DNA fragments, or probes, that target the complementary sequence in the sample's DNA. When a portion of the target DNA that is complementary to the probe sequence comes into contact with the probe, it will hybridize, or bind, to the probe. Most SNP markers contain copies of one probe sequence for each of the two common genotype alleles, which are commonly referred to as the 'A' allele and 'B' allele. As illustrated in Figure 2.4, when the target sample has the 'A' allele at the SNP locus, its DNA will hybridize to the 'A' allele probes, and similarly, sample DNA with the 'B' allele will hybridize to the 'B' allele probes. If a sample is heterozygous at the locus, with one chromosome containing the 'A' allele and the other containing 'B', its DNA will bind to both 'A' and 'B' allele probes.

Fluorescent markers are used to detect hybridization of probes to their target sequences. Since the fluorescence of a marker is usually measured over a unit of time, its intensity indicates the strength of the hybridization bond between the probe and its target DNA. Separate fluorescent markers are used for the 'A' and 'B' alleles, and the hybridization intensities are separately measured for the 'A' allele and the 'B' allele of

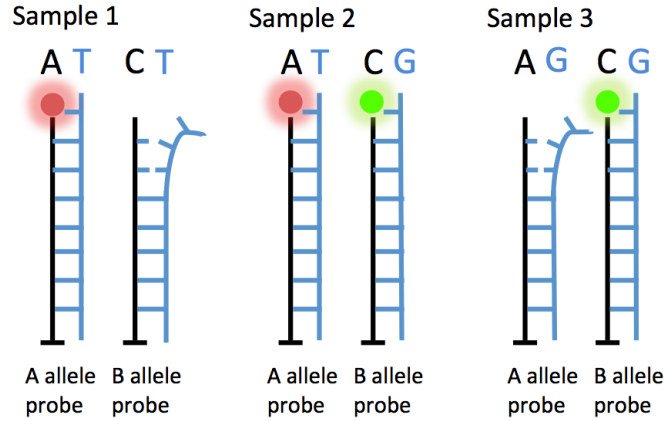


Figure 2.4: An example SNP marker in three different samples. Each SNP marker consists of a probe for the A allele and a probe for the B allele, where the probe sequence is the reverse complement sequence of the target DNA. In this case, the target SNP’s A allele has the “T” nucleotide, and the B allele has the “G” nucleotide. Sample 1 is homozygous with the A allele, so that the target DNA binds only to the A allele probe, and the marker fluoresces red. Sample 2 is heterozygous, meaning it has one copy of each the A and B alleles, and so its target DNA binds to both probes, and the marker fluoresces yellow from both the green and red fluorescence. Sample 3 is homozygous with the B allele, so that the target DNA binds only to the B allele probe, and the marker fluoresces green.

each SNP marker. Generally, samples with high A intensity and low ‘B’ intensity are assigned the homozygous ‘A’ allele genotype, samples with higher ‘B’ than ‘A’ intensity are assigned the homozygous ‘B’ genotype, and samples with both high ‘A’ and ‘B’ intensities are assigned the heterozygous, or ‘H’, genotype. Some samples may not contain the exact complementary sequences of the probe sequences, and may end up with low ‘A’ and ‘B’ intensities, resulting in a no call, or ‘N’ genotype.

2.4.3 Next-Generation Sequencing

While genotyping microarrays can contain many thousands, and even millions, of genome probes, it is currently not possible to probe all 2.7 billion base pairs in the mouse genome with a single microarray. However, next-generation sequencing technology has developed rapidly over the past decade as a scalable method for determining the precise

sequence of DNA molecules at the nucleotide level.

The main challenge that lies in current sequencing technology is that DNA sequences can only be read in fragments that are very short compared to the length of an entire chromosome. As a result, many overlapping read fragments are required to accurately determine the nucleotide at a locus. However, these overlapping short reads are often unevenly sampled from the genome based on the nucleotide content, physical shape of the DNA molecule, and other factors. The two main approaches for addressing the issue of abundant short reads are to perform *de novo* assembly of the genome, or to perform alignment of the short reads to a previously assembled reference genome. *De novo* assembly of short reads is a difficult problem, so read alignment is more commonly performed once several key genomic strains have been assembled. In the case of mouse genome sequencing, reads are typically aligned to the mouse reference genome.

Next-generation sequencing is used for sequencing both DNA and RNA, where RNA-seq typically produces fewer reads since only the transcriptome, instead of the entire genome, is read. The reference transcriptome refers to regions in the reference genome that are known to be transcribed, and most RNA-seq analysis relies on alignment of reads to the reference transcriptome.

2.5 Conclusion

The following chapters introduce methods for maximizing the amount of information that is available in existing genotyping microarray and next-generation sequencing technologies. The genetic reference populations of the CC and DO are sources of data on which I demonstrate many of my methods. Chapter 3 discusses how to design a highly informative microarray for admixed animals, and the inheritance of these admixed animals' genomes is explored in further depth in Chapter 4. Chapter 5 considers the association between genetics and physical traits using genotyping microarrays,

while Chapter 6 demonstrates a similar concept of maximizing available information in next-generation sequencing data.

Chapter 3 : Microarray Design and Marker Selection

3.1 Introduction

Genome-wide genotyping microarrays have been used for many years to characterize individual genomes. The design of a genotyping array involves selecting the genomic variants to be included on the array, taking into consideration the sequences immediately flanking the variants, and whether those sequences can be easily manufactured into working microarray probes. Although high-throughput sequencing is becoming more affordable today, for large experiments involving many model organism individuals, genotyping microarrays remain a more cost-effective option for assessing the content of genomes.

One of the first genome-wide genotyping arrays developed for the mouse was released in year 2000 and contained 2,848 SNP probes [42], with SNPs selected to segregate between eight common inbred strains. The SNPs were selected from known gene regions as well as random positions as annotated in [50], and the result was the most comprehensive SNP genotyping array for the mouse genome available at the time.

One bottleneck in the creation of high-density microarrays was that probe selection required known SNPs, so as SNP discovery improved, the achievable density of genotyping microarrays also increased. In 2006, the Wellcome Trust Institute developed an array for the mouse with 11,609 SNPs [60], which was the highest density array for any model organism at the time. The array was designed for a heterogeneous stock descended from eight inbred classical laboratory strains [71]. The criteria for selected SNPs in [60] was that they were polymorphic in at least some founders, and that they

were spaced roughly uniformly across the genome, regardless of known functionality. Any two SNPs closer than 50kb with the same strain distribution pattern (SDP) were discarded, since without any recombination event between them, both SNPs would contain the same information for any descendants.

Pruning neighboring SNPs containing the same haplotype information was one of the main considerations in the design of the Mouse Universal Genotyping Array (MUGA), a medium-density Illumina genotyping array with 7,851 markers designed for the Collaborative Cross (CC) population released in 2011 [75, 65]. Although another genotyping array with 13,000 SNPs was developed for the CC population at Oak Ridge National Laboratory [8], it uniquely identified all eight inbred CC founders at only 1,200 regions in the genome. The goal for the design of MUGA was to uniquely identify the eight CC founders in as many regions as possible, meaning the SNPs had to be selected to be highly informative between the CC founders on a haplotype scale, with the alleles of neighboring SNPs considered in the design. I participated in the design of the MUGA, and the markers were selected not only based on each SNP being polymorphic within the CC founders with a high minor allele frequency, but also based on the informativeness of each SNP with respect to its neighbors. We considered sliding windows of SNP markers, selecting SNPs so that a minimal number of markers could uniquely identify all eight founders in sliding windows across the genome. Therefore, the SNP markers on MUGA are highly informative when neighboring SNP markers are considered, and in most regions of the genome, all eight CC founders can be distinguished using sliding windows of 3-5 SNP markers [75, 65], making the haplotype information content on MUGA very high despite its low marker density.

The mouse genotyping arrays with the most dense set of SNP and non-SNP markers is the high-density genotyping Mouse Diversity Array (MDA) [80]. Released in 2009, the MDA was designed to contain 623,124 SNP markers. The selection criteria for

MDA SNPs included selecting those with high minor allele frequency in uniform bins across the genome, selecting SNPs that covered a high proportion of mouse phylogenetic trees, and also selecting singleton SNPs where only C57BL/6J is known to have the minor allele. Out of the 623,124 SNP markers manufactured on the MDA, nearly 7% of MDA SNP markers were determined to have unreliable genotype calls, bringing the total number of usable SNP markers to 581,672 [80]. Although the MDA includes SNP and non-SNP markers informative in many scenarios, due to the high cost of sample preparation and genotyping on the high-density Affymetrix array, it is not ideal for large experiments. Furthermore, there is no guarantee that adjacent SNPs have SDPs distinct from one another, or that the haplotypes of different strains can be determined using only a few consecutive SNPs. In fact, [75] shows that the mean number of CC founders that can be uniquely distinguished using a window of four consecutive SNPs is 1.76 in the MDA, compared to 5.01 in the MUGA, although the MDA boasts much higher marker density.

The need for a robust, low-cost, yet highly informative array motivates the design and creation of MegaMUGA, the successor of the MUGA that contains ten times as many markers for the same cost. While the design of MUGA maximized the number of uniquely identifiable inbred founders in each window of SNPs, doing so does not necessarily maximize the number of distinguishable heterozygous founder states. It is common for two different pairs of inbred founders, each with different alleles for a sequence of 4 or so SNPs, to produce two different F1s with the same sequence of alleles for these 4 or so SNPs. Therefore, the design of the MegaMUGA additionally optimizes for the detection of different heterozygous founder states as well as homozygous founder states, which makes the array idea of identifying founder haplotype in both the Collaborative Cross (CC) and Diversity Outbred (DO) populations. In addition to traditional SNP markers, MegaMUGA also includes additional non-SNP markers with invariant

probe sequences. Presented here in depth are the methods for selecting SNP markers that comprise the majority of MegaMUGA markers, which are SNP markers selected to differentiate between the CC and DO homozygous and heterozygous founders, along with non-SNP markers in the pseudoautosomal region (PAR) to demonstrate the design and use of invariant probe sequences.

3.2 Methods

MegaMUGA was designed as a medium-density microarray on the Illumina Infinium II platform with approximately ten times the density of the 7,854-marker Mouse Universal Genotyping Array (MUGA) [75]. A total of 80,000 markers were selected for the initial design of MegaMUGA, with 65,000 SNP markers designed to be informative for the Collaborative Cross (CC) and Diversity Outbred (DO) populations, 14,000 SNP markers chosen to be informative between wild mice, and the remaining non-SNP markers tracking structural variants or genetically engineered constructs. The wild mouse SNP markers and majority of non-SNP markers were selected by our genetics collaborators; the CC SNP markers, which comprised 81.25% of MegaMUGA markers, will be the main focus of this discussion. The design of non-SNP markers within the mouse pseudoautosomal region (PAR) is also presented here as an example of informative markers that were designed to distinguish between variants with only invariant probe sequences.

The overall procedure for the selection of CC SNP markers in the MegaMUGA is presented as a flowcart in Figure 3.1. The details of each operation are discussed in the following subsections.

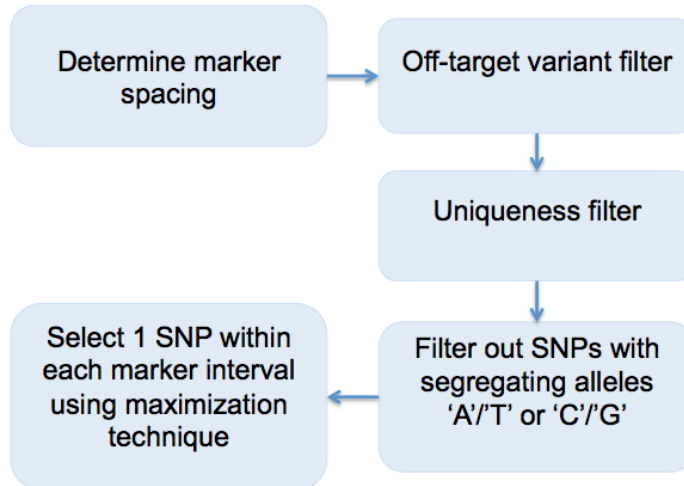


Figure 3.1: Flowchart overview for the selection of CC SNP markers on MegaMUGA.

3.2.1 Determining SNP marker spacing

The spacing of SNP markers in the MegaMUGA was based on a linkage map constructed from previously observed recombination frequencies within CC mouse genomes [75, 43]. This method ensured that marker density would be high in regions of frequent recombination and low in more non-recombinant regions of the genome, therefore effectively capturing the most information in the most diverse areas of the mouse genome.

The CC recombination map was determined using CC animals in the G2:F1 generation genotyped on the high-density Mouse Diversity Array (MDA) [43, 80]. Each chromosome was divided into sections at recombination breakpoints, with each section representing a conserved segment with no observed crossover events within the population. The linkage map based on crossover events was mapped onto genomic positions [75], as shown in Figure 3.2. The genome was then divided into 65,000 equal intervals on a recombination scale using this mapping, where the probability of a crossover event occurring at each interval was uniform across the genome. A SNP within each interval was then selected to be made into a SNP marker using the criteria below.

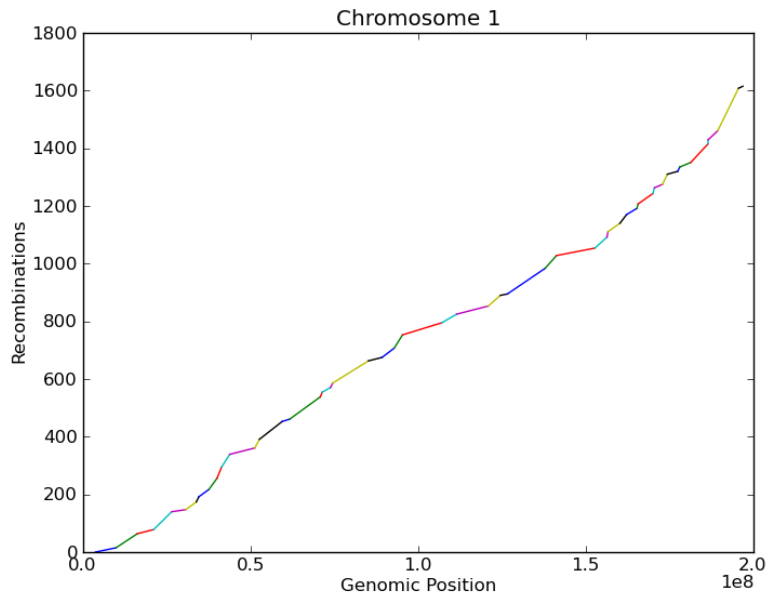


Figure 3.2: Linkage map vs. Genomic position on Chromosome 1, from Catherine Welsh [75] from recombinations observed in the G2:F1 generation [43]. Each unit on the y-axis represents one recombination segment, with genomic positions of its two endpoints plotted on the x-axis. Due to the large number of segments, the data was fitted with a piece-wise linear curve with 50 segments. To select intervals for SNP markers, the y-axis was sampled uniformly and the corresponding genomic positions were recorded.

Since telomeric regions are recombination-rich [43], the design of MegaMUGA includes 500 of the 65,000 markers in telomeric regions of all autosomes, spaced uniformly by genomic distance due to a lack of telomeric linkage map data.

3.2.2 Filtering available SNPs

For each interval selected to contain a SNP marker, I consider all high confidence SNPs documented by the Wellcome Trust Institute [35] in the set of possible SNPs for that marker. To ensure the manufactured probes work as intended, I filtered out SNPs with flanking sequences that might not result in properly functioning probes, as described below.

The MegaMUGA was designed for the Illumina Infinium II platform with 50 bp

probe sequences. Therefore, each SNP probe has a sequence including the SNP itself, along with one flanking 49 bp sequence from either upstream or downstream of the SNP. In order for the target DNA to fully hybridize to the probe sequence, no other SNPs or variants should be present within the 49 bp of the probe sequence. Unexpected variants within the probe sequence can lead to poor hybridization and low intensities, which is sometimes informative as discussed in later chapters, but which often results in the marker having no intensity signal across all samples. Therefore, I filtered out SNPs with any other high-confidence SNP or indel in all 17 mouse strains sequenced by the Wellcome Trust Institute [35] within both flanking 49 bp sequences. Those with off-target SNPs or indels in only one of the two flanking 49 bp sequences were still considered for marker production, but only the side with no known off-target variations was considered as a possible probe sequence.

In addition to the constraint of not containing off-target variants, the probe sequences were also filtered for uniqueness. If the 49 bp probe sequence occurs elsewhere in the reference genome, on either the primary or the reverse complement strand, the marker will hybridize with the target DNA in these genomic regions. This provides erroneous genotype information about the intended SNP, since the hybridization intensities may have been a result of hybridization with several different regions in the target genome. Early mouse genotyping arrays [42] observed this effect, noting that homologous sequences in the genome result in heterozygous genotype calls even in inbred strains.

To filter our set of potential SNP markers for uniqueness, I searched for each flanking 49 bp sequence in the Build 37 standard reference genome from NCBI [12], on both the primary and reverse complement strands. Any flanking sequence occurring in any other region of the reference genome was discarded as a potential probe sequence, and any SNP with two non-unique flanking sequences was discarded from the set of potential

SNPs.

Since Illumina’s fluorescent labels for the SNP nucleotides have only two colors (the ‘A’ and ‘T’ nucleotides share a color and the ‘C’ and ‘G’ nucleotides share a color), producing a marker targeted at a SNP where the two alleles have the same fluorescent color requires two separate probes. For instance, a SNP segregating between the nucleotides ‘A’ and ‘T’ would require an ‘A’/‘C’ probe and a ‘T’/‘C’ probe, and a sample with the ‘A’ allele would simply not hybridize to the second probe, a sample with the ‘T’ allele would not hybridize to the first probe, and no samples would hybridize to the ‘C’ allele on either probe. Due to this limitation that SNPs segregating between ‘A’/‘T’ or ‘C’/‘G’ would require two probes instead of one and therefore use up valuable space on the array, I also filtered out any SNPs that were documented to segregate between ‘A’/‘T’ or ‘C’/‘G’ alleles.

After filtering potential SNP probes for containing no off-target variants, uniqueness within the genome, and ‘A’/‘T’ or ‘C’/‘G’ alleles, I used the algorithm described below to select one SNP within each intended marker region to produce a maximally informative array.

3.2.3 Establishing a minimum marker window for informative SNPs

To produce an array of highly informative markers for mouse genotyping, adjacent markers should be able to differentiate between a diverse set of strains. The CC founders were used as representatives of diverse mouse strains since they were initially chosen for their genetic diversity [53]. In the MUGA, markers were selected to uniquely distinguish each of the 8 inbred CC founders using as few contiguous SNPs as possible. Theoretically, 3 continuous SNPs can represent 8 different inbred states, since each SNP has two possible homozygous genotypes, and $2^3 = 8$. In reality, due to similarity between strains and linkage disequilibrium, defined as the non-random association of

alleles at different loci within a population, 4 or more markers were typically required to represent all 8 inbred CC founders in the MUGA [75].

While the design of MUGA attempted to identify each of the 8 inbred CC founder states within each 4-marker window, one goal in designing the MegaMUGA array was to uniquely distinguish all 36 possible CC founder states using as few continuous SNP markers as possible. The 36 states include the 8 inbred CC founders (AA, BB, ... HH), as well as the 28 F1 states (AB, AC,... GH), representing heterozygous segments of the genome. Theoretically, 4 continuous SNP markers, each with 3 possible genotypes, should be sufficient to represent 81 different states, since $3^4 = 81$. However, since the 8 inbred founder states can only be represented with the 2 possible homozygous genotypes, and the 28 F1 founder states of the CC are not independent of each other and the 8 inbred founder states, a minimum of 5 continuous SNP markers are actually required to distinguish between all 36 possible CC founder states.

The following Python code enumerates all possible scenarios with 4 SNPs that uniquely distinguish all 8 inbred founders, and it shows that none of the possible scenarios can produce unique encodings for all possible 28 heterozygous founder states. Here, I use '0' and '1' to represent the two homozygous alleles, and '2' to represent the heterozygous allele at each marker.

```
import itertools
inbreeds = [a+b+c+d for a in '01' for b in '01' for c in '01' for d in '01']
successful = 0
for scenario in itertools.combinations(inbreeds, 8):
    hets = set()
    for pair in itertools.combinations(scenario, 2):
        het = ''.join([a if a==b else '2' for a, b in zip(pair[0], pair[1])])
        if het not in hets:
            hets.add(het)
        else:
            break
    successful += len(hets)==28
```

```
print successful
```

If we modify the code to enumerate all possible 5-SNP combinations that uniquely distinguish the 8 homozygous founder states, then there are 1,650,400 scenarios that can also uniquely distinguish the 28 heterozygous founder states, which establishes a 5-marker minimum for separating all 36 possible founder states.

3.2.4 Selecting maximally informative SNPs

Given the 5-marker minimum for distinguishing all possible founder states, I selected SNPs with alleles that maximize the number of uniquely identifiable founder states in sliding windows of 5 markers. Since the goal was to maximize the total founder states distinguishable by all sliding windows of 5 markers, this maximization was done using a dynamic programming-like algorithm on each chromosome.

Let $f(\mathbf{w})$ be the number of unique 5-SNP allele sequences present in the 36 founder states within the marker window w (note that $1 \leq f(\mathbf{w}) \leq 36$ for all \mathbf{w}). Let

$$TF(i, \mathbf{q}) = \sum_{\mathbf{w} \in \mathbf{q}} f(\mathbf{w}) \quad (3.1)$$

denote the total number of founder states distinguished by all 5-SNP sliding windows on a chromosome given a selected sequence of SNPs \mathbf{q} ending at position i . For each chromosome, the objective is to find the sequence of SNPs \mathbf{q} that maximizes $TF(n, \mathbf{q})$, where n is the total number of SNPs to be selected for the chromosome.

The recurrence given below considers all possible paths of SNPs within each chromosome, with the constraint of having a single SNP selected at each intended marker region. Given a current path with a score of $TF(i, \mathbf{q})$, the task of the recurrence is to determine the maximum score of $TF(i + 1, \mathbf{q} + s)$ for each possible SNP s at position $i + 1$. The optimal path of SNPs for the entire chromosome is then determined by

backtracking from the path with the maximum score for $TF(n, \mathbf{q})$

Given a current path of SNPs \mathbf{q} , for each possible SNP s at position $i + 1$, the recurrence for $TF(i + 1, \mathbf{q} + s)$ is

$$TF(i + 1, \mathbf{q} + s) = \max\{TF(i, \mathbf{p}) + f(\mathbf{p}[i - 3 :] + s) | \quad (3.2)$$

$$\forall \mathbf{p} \in Paths, \text{ where } \mathbf{p}[i - 3 :] = \mathbf{q}[i - 3 :]\},$$

where $\mathbf{p}[i - 3 :]$ and $\mathbf{q}[i - 3 :]$ denote the last 4 SNPs of the paths \mathbf{p} and \mathbf{q} . The score for adding SNP s is the maximum score obtainable from adding s to all current paths with the same last 4 SNPs as \mathbf{q} , including \mathbf{q} itself, since the number of unique CC allele sequences within the 5-SNP window ending at position $i + 1$ is independent of any SNP before position $i - 3$. When the scores for all paths to the end of the chromosome are calculated, we can backtrack from the path with the maximum score to find the optimal sequence of SNPs.

The number of paths grows exponentially since multiple SNPs are considered within each marker region. To limit our solution space, the number of paths is pruned at each step to preserve the top 100,000 possible paths with the highest scores. This pruning makes the problem computationally feasible, yet leaves enough possible paths to find a near-optimal solution.

3.2.5 Selecting non-SNP markers

In addition to SNP markers, non-SNP markers are also included in the MegaMUGA design, such as markers for mapping the pseudoautosomal region (PAR). The PAR refers to a homologous sequence between the X and Y chromosomes that facilitates pairing of the sex chromosomes during meiosis, and it is found at the distal end of both chromosomes in the mouse. Although the beginning of the PAR sequence varies between strains, most laboratory strains have the PAR occurring between 700-1130 kb

from the distal end of the X chromosome, starting as early as 166.0 Mb in [79].

I designed 20 invariant markers in the mouse PAR, where all strains are expected to have the same allele at the locus for target SNPs in typical SNP probes. We expect hybridization with the probe when the target DNA is present in the genotyped strain, and we expect hybridization intensity to increase when multiple copies of the target DNA is present. This is based on studies and observations in earlier arrays such as the MUGA and the MDA [20]. As shown in Figure 3.3, strains with higher copies of the probe sequence also have higher hybridization intensities. Therefore, invariant markers can detect the beginning of the PAR sequence in different strains, since males in each respective strain will have one copy of any unique sequence elsewhere in the X chromosome, but two copies of any PAR-specific sequence. Due to the fact that different factors such as the prevalence of ‘C’ and ‘G’ nucleotides in the probe sequence also affect hybridization intensity, making the baseline hybridization intensity different between different markers, it is easier to detect the PAR by observing the female to male hybridization intensity ratio, such as in Figure 3.3.

To select probe sequences for marker production in the PAR, I divided the X chromosome into 20 intervals after 166.0 Mb. Within each interval, the first unique 49 bp sequence with no off-target variants was selected as the probe sequence. The uniqueness criteria is the same as that for SNP markers, with the 49 bp sequence appearing nowhere else within the Build 37 reference genome. The off-target variants applied to any high-confidence SNPs or indels within the 49 bp probe sequence and the 1 bp ‘SNP’ position itself. The alleles at the ‘SNP’ position contained the invariant nucleotide at that position and another arbitrarily chosen nucleotide. Therefore, even though the PAR markers were designed for a SNP genotyping platform, we expect all strains to hybridize only in the direction of the reference allele and have zero intensity for the alternate allele.

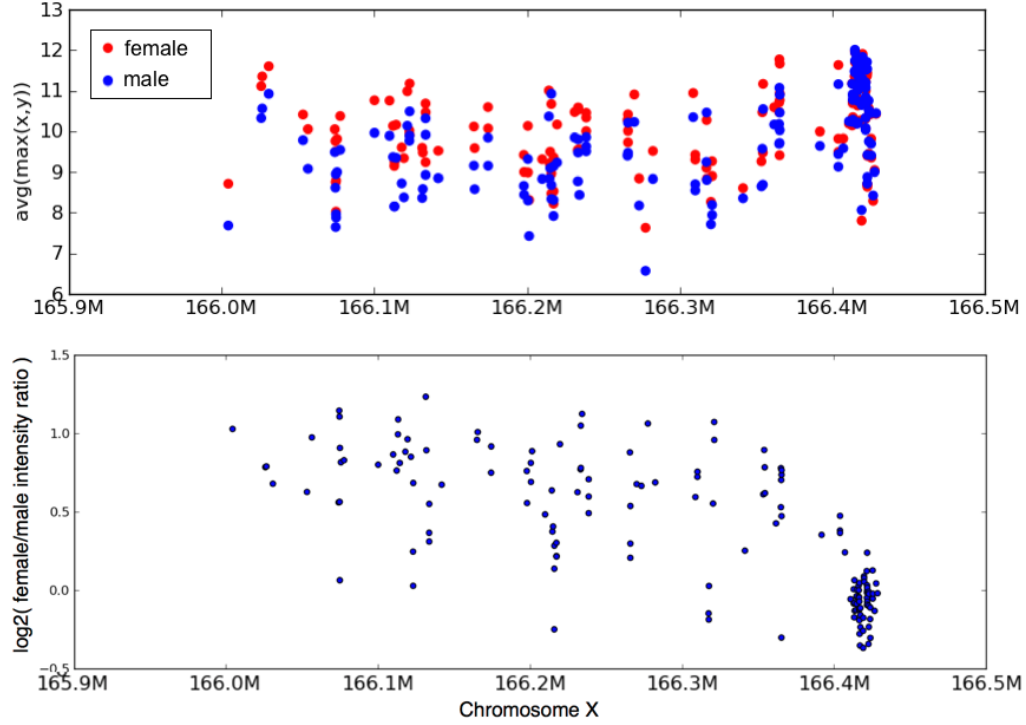


Figure 3.3: Female and male probe intensities in the pseudoautosomal region (PAR). I took the mean probe intensities of the hybridizing allele ($\max(x,y)$) from all females ($n=453$) and males ($n=1156$) genotyped on the Affymetrix-based Mouse Diversity Array (MDA). The top scatter plot shows female and male mean intensities for each probe, while the bottom plot shows the \log_2 of female/male intensity ratio. Prior to the PAR, females have two homologous copies of the sequence, while males have only one copy on the X chromosome. Once the PAR begins, both females and males have two homologous copies of the PAR sequence, since it is also present on the Y chromosome. In the first plot, we can see that females have higher hybridization intensities than males before 166.4 Mb, and roughly the same intensities after 166.4 Mb, which is where the PAR begins in most strains. From the bottom plot we can clearly see that the \log_2 of female/male intensity ratio drops to 0—indicating a 1:1 female/male intensity ratio—after 166.4 Mb. Although the female/male intensity ratio is less than 2 for non-PAR regions of the X chromosome, from the clear change in hybridization intensities at the distal end, we can still conclude that the copy number of a probe sequence is reflected in its hybridization intensity, leading us to design invariant markers within the PAR region to detect the beginning of the PAR in different strains.

3.3 Results

3.3.1 Haplotype Informativeness of Selected SNP Markers

To assess the effectiveness of our selection of maximally informative SNP markers, Figure 3.4 and Table 3.1 show the number of uniquely distinguishable homozygous and heterozygous founder states using SNP markers selected according to the presented algorithm, in sliding windows of 5 markers. In addition, I also plotted the number of distinguishable founder states using a random 90% of the selected SNPs (Figure 3.5), as well as using randomly selected CC SNPs (Figure 3.6). The results from the random selection of 90% of SNP markers shows that the number of uniquely identifiable founder states is robust even after discarding 10% of potentially nonperforming markers, while the results from the randomly selected SNPs serves as a model for traditional marker selection techniques, where SNP markers are chosen without consideration for the information content in nearby SNP markers. The randomly selected SNPs were first filtered for uniqueness and absence of off-target variants, which is a common filtering technique also done in the selection of MegaMUGA CC SNPs.

From Figure 3.4, we can see that the majority of 5-SNP windows can separate the 36 founders into 20 groups or more and separate the 8 inbred founders into 6 groups or more, with many windows achieving a separation power of more than 27 total founder states and 7 or 8 inbred founders. In contrast, the majority of randomly selected SNPs in Figure 3.6 can separate 15 or fewer founder states, or 5 and fewer inbred founder states in 5-SNP windows, with hardly any windows able to separate all 8 inbred founders. The high discrimination power of our SNPs selected with the maximization algorithm remains robust even when 10% of markers are randomly discarded in Figure 3.5, suggesting even a random sampling of the selected SNPs are highly informative on a haplotype scale. As shown in the next subsection, 63,637 CC SNP markers

Table 3.1: The mean number of distinguishable founder states in 5-SNP sliding windows

Chromosome	Mean # of distinct founders	Mean # of distinct homozygous founders
1	20.78	5.88
2	19.75	5.73
3	21.68	6.03
4	20.75	5.89
5	21.18	5.96
6	18.90	5.59
7	19.95	5.75
8	18.08	5.45
9	21.09	5.94
10	16.40	5.14
11	21.86	6.07
12	21.04	5.92
13	19.59	5.69
14	18.09	5.45
15	19.25	5.64
16	17.62	5.38
17	22.03	6.07
18	22.54	6.17
19	18.41	5.48
X	17.68	5.39

The middle column shows the mean number of distinguishable founder states, including both homozygous and heterozygous states, with the maximum possible distinguishable states being 36 (8 homozygous and 28 heterozygous) within each sliding window. The right column shows the mean number of distinguishable homozygous founder states, with the maximum possible states being 8.

are present in the manufactured MegaMUGA array, which is 97.90% of the 65,000 originally selected CC SNP markers, much higher than the 90% marker retention rate modeled in our analysis.

3.3.2 The Final Manufactured MegaMUGA Array

Of the 79,797 markers selected to be on the MegaMUGA, 77,808 markers were manufactured on the final array, achieving an extremely high conversion rate of 97.51%.

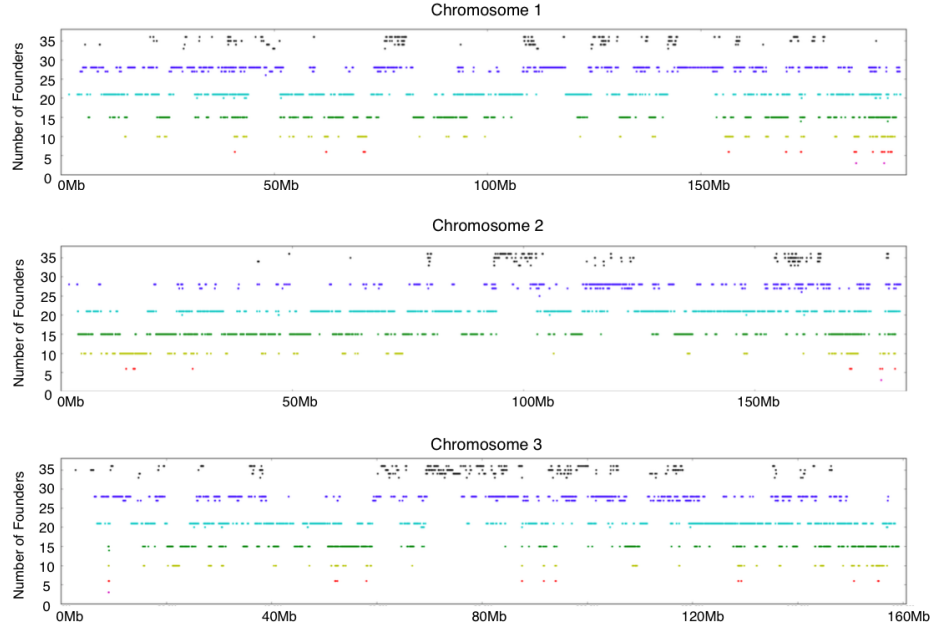


Figure 3.4: Number of uniquely distinguishable founder states for each 5-marker sliding window in the set of selected CC SNP markers, shown for Chromosomes 1-3. Each 5-marker window is plotted as a point, with the y-value corresponding to the number of founder states out of 36 which can be uniquely distinguished within that window. The colors correspond to the number of homozygous founder states out of 8 which can be uniquely distinguished, with the colors: {Black: 8, Blue: 7, Cyan: 6, Green: 5, Yellow: 4, Red: 3, Purple: 2}. Note for 7 distinct inbred founder allele patterns, at most 28 total founder states can be distinguished (7 homozygous and 21 heterozygous). Similarly, for 6, 5, 4, 3, and 2 distinct inbred founder allele patterns, at most 21, 15, 10, 6, and 3 total founder states can be distinguished.

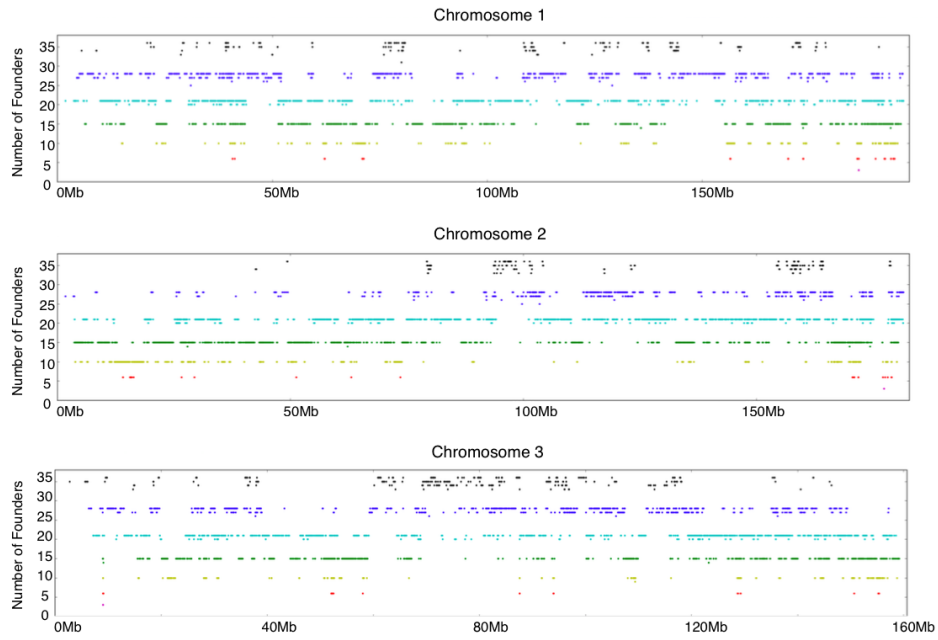


Figure 3.5: Number of uniquely distinguishable founder states for each 5-marker sliding window after discarding 10% of selected CC SNP markers.

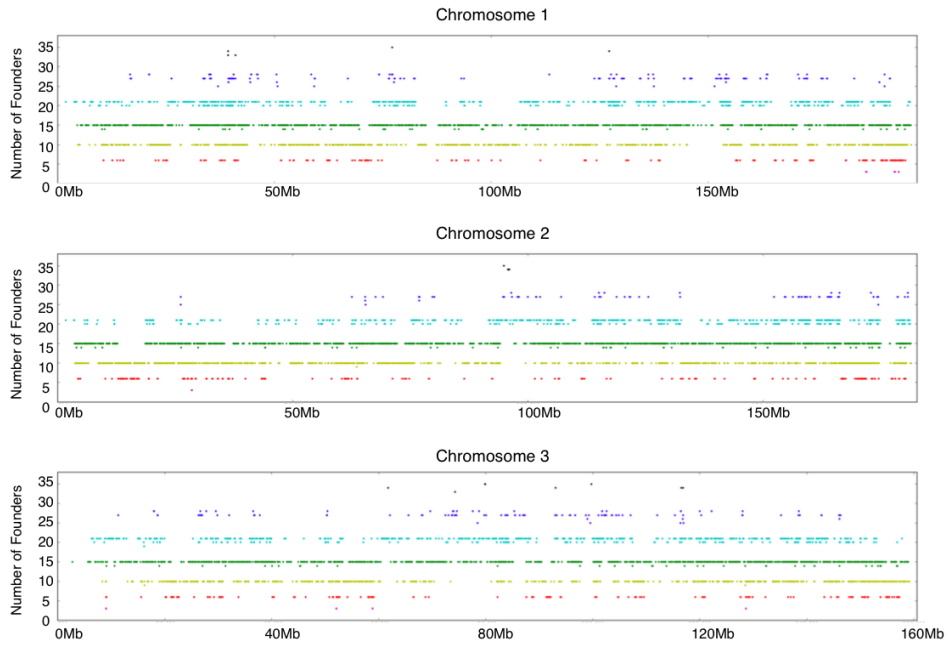


Figure 3.6: Number of uniquely distinguishable founder states for each 5-marker sliding window using randomly selected SNPs in each marker region.

The mean genomic spacing between markers on the autosomes and X chromosome was 35 Kb, and the distribution of all markers in the autosomes and sex chromosomes are shown in Figure 3.7. The markers that were manufactured include 63,637 CC SNP markers, 13,238 SNP markers from wild mice within the *Mus musculus* species, 1007 SNP markers from other species of *Mus*, including *Mus spretus* and *Mus caroli*, 149 SNP markers differentiating between two closely related substrains C57BL/6J and C57BL/6N, 17 markers in the PAR, 45 markers on the Y chromosome, 42 markers in the mitochondria, and 69 markers tracking genetically engineered constructs such as Cre and Luciferase. The distribution of each type of marker is shown in Figure 3.8.

I estimated the call rate, defined as the rate of calls which are not ‘N’, for MegaMUGA markers in 16 CC founder samples and 8 CC F1 samples. The mean number of ‘N’ calls in the CC founder samples was 2,540 with a standard deviation of 368.53. In CC F1 samples, the mean and standard deviation for ‘N’ calls were 2273.38 ± 45.32 . The overall ‘N’ call rate among these CC founder and F1 samples was 2451.71 ± 325.26 , or $3.15\% \pm 0.42\%$. This makes the MegaMUGA call rate 96.85%, in comparison to the call rate of 95% in MUGA [48] and the 93.35% of markers with reliable genotype calls in MDA [80]

In addition to assessing call rate, I estimated MegaMUGA’s genotyping error rate using biological replicates of pairs of the 8 CC founders genotyped on MegaMUGA. For each replicate pair, all markers where the two samples had differing genotype calls and neither sample had an ‘N’ call were included in the list of discordant markers. Markers where at least one sample from the pair had an ‘N’ call were labeled uninformative. Among the 8 replicates pairs, the mean number of markers that were inconsistent was 5.88, ranging from 1 in the C57BL/6J and NOD/ShiLtJ strains to 22 in the WSB/EiJ strain. The number of discordant calls between each replicate pair is summarized in Table 3.2. In addition to the 8 founder strains, I also report the number of discordant

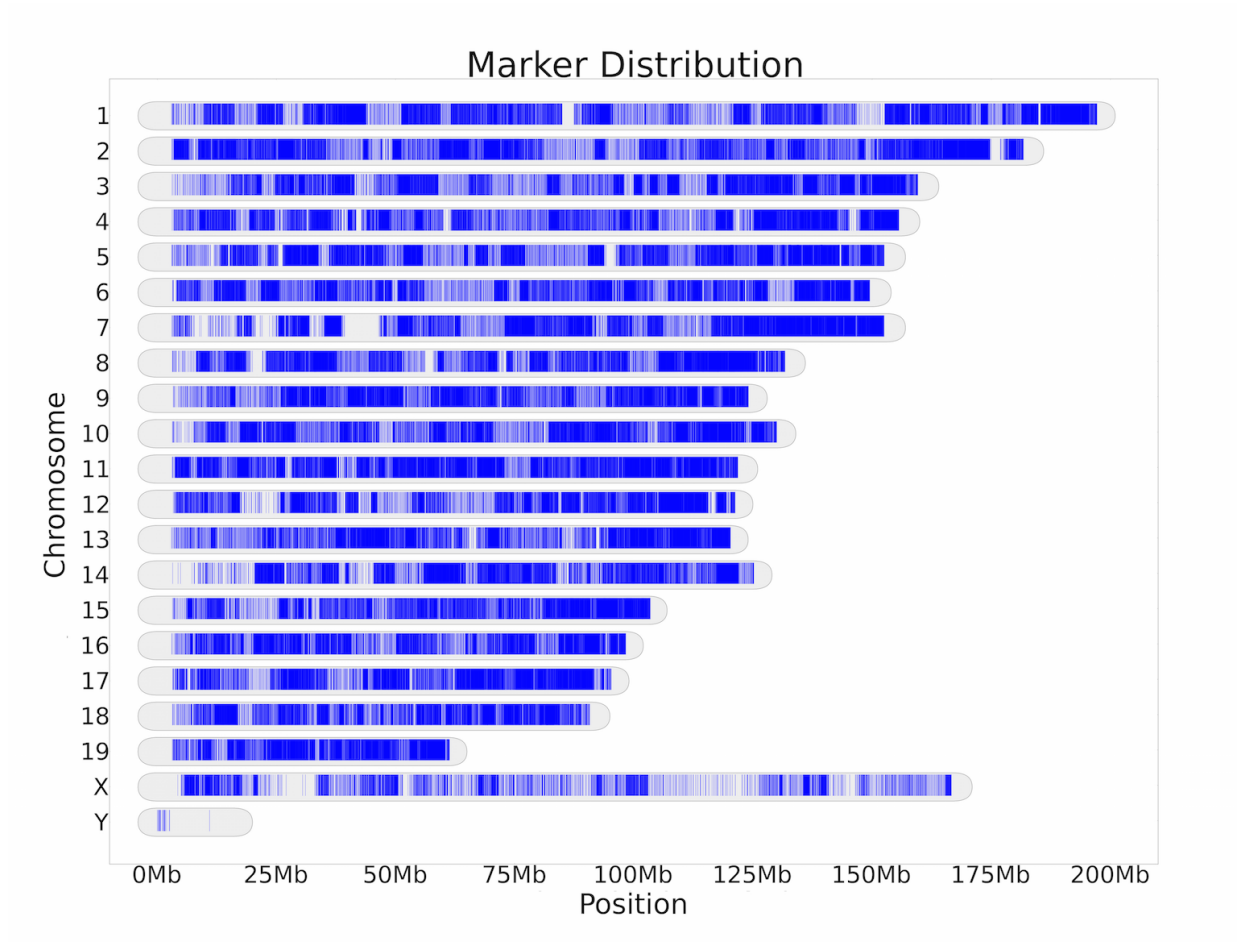


Figure 3.7: Distribution of all MegaMUGA markers in the autosomes and sex chromosomes. Different marker types and markers not in autosomes or the sex chromosomes are shown in Figure 3.8.

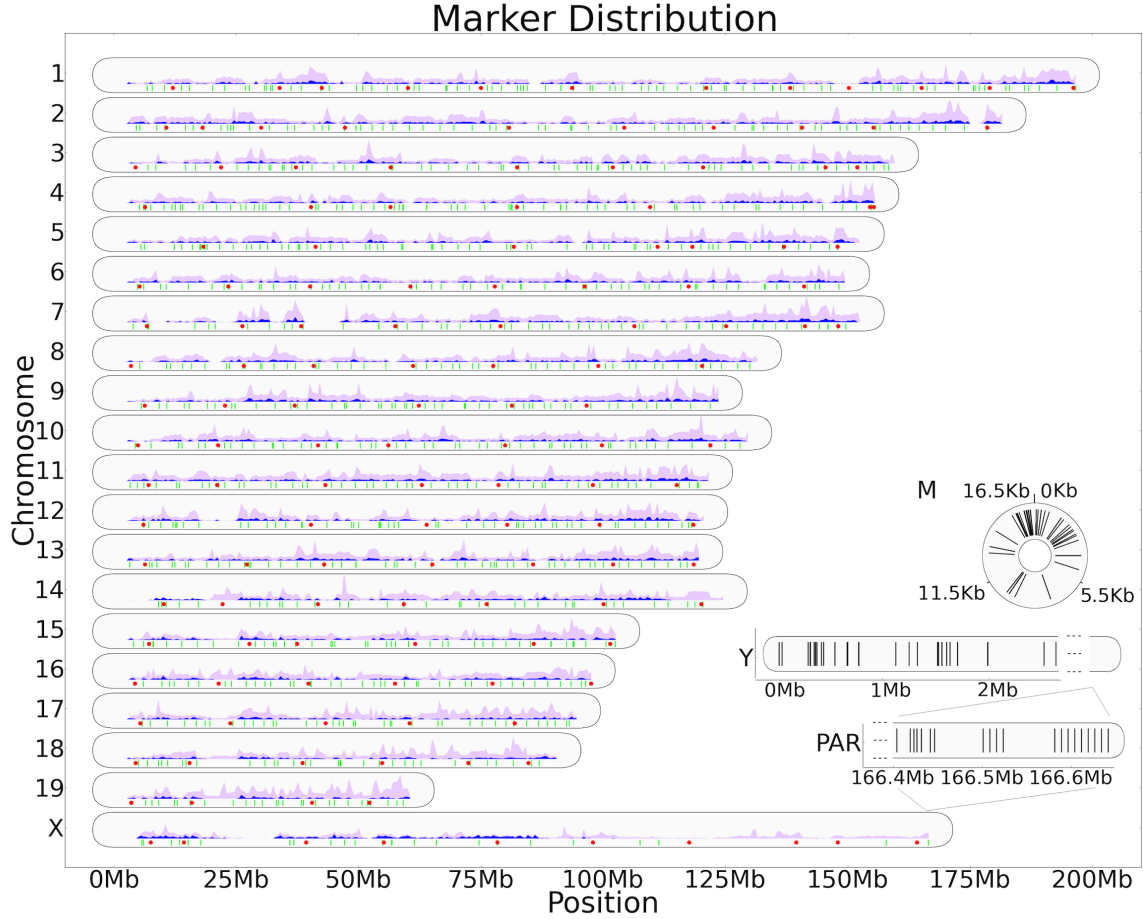


Figure 3.8: Manufactured MegaMUGA markers, colored by marker type. The light purple histogram shows the distribution of CC SNP markers (63,637 total), the blue histogram shows the distribution of SNP markers selected from wild mice from the *Mus musculus* species (13,238 total), the green ticks are SNP markers selected from mice outside the species *Mus musculus*, including those from the species *Mus spretus* and *Mus caroli* (1007 total), and the red dots are SNP markers differentiating two very similar sister strains, C57BL/6J and C57BL/6N (149 total). The 17 markers in the PAR, 45 markers on the Y chromosome, and 42 markers in the mitochondria are also shown. Some markers served as more than a single type, such as markers that distinguished between both CC SNPs and wild mice SNPs, so the totals add up over the 77,808 total marker count. Not shown in this plot are 69 markers that track genetically-engineered constructs that do not appear in the mouse genome.

Table 3.2: Number of discordant markers between replicates on MegaMUGA

Sample 1	Sample 2	Discordant	Uninformative
129S1/SvImJm35370	129S1/SvImJm1314	2	2420
A/Jm37621	A/Jm35593	5	2442
C57BL/6Jm1957	C57BL/6Jm38420	1	2286
CAST/EiJm0538	CAST/EiJm0042	5	3200
NOD/ShiLtJm0150	NOD/ShiLtJm1214	1	2417
NZO/HILtJm36511	NZO/HILtJm0591	2	2348
PWK/PhJm0175	PWK/PhJm1090	9	3152
WSB/EiJm0993	WSB/EiJm1345	22	3610
(B6 xPWK)F1m005	(B6 xPWK)F1m005-2	5	2338
(A/JxWSB)F1m0197	(A/JxWSB)F1m0197-2	10	2424
(CASTxNZO)F1m	(CASTxNZO)F1m-2	3	2423
(NODx129S1)F1m0030	(NODx129S1)F1m0030-2	7	2428
Total:		72	31488

The number of discordant markers include all markers where Sample 1 and Sample 2 did not share the same genotype call, excluding those where one had an ‘N’ genotype call. Markers where at least one of the pair had an ‘N’ genotype call were labeled as uninformative. The total number of discordant markers in these 12 pairs of biological and technical replicates was 72, which is a mean of 6 discordant markers per replicate pair, or an average of 0.008% of all informative markers.

calls between 4 pairs of technical replicates of CC F1 strains, to show MegaMUGA’s performance in highly heterozygous samples. The mean number of discordant calls between F1 technical replicates was 6.25, with a minimum of 3 and maximum of 10. Together, these 12 pairs of biological and technical replicates among CC founders and F1s yield a genotyping error rate of a mere 0.008 %, or 6 discordant markers per replicate pair. This extremely low error rate between CC replicate pairs means that informative markers in MegaMUGA have a reproducibility of over 99.99% in the CC population.

In addition to being highly reliable, MegaMUGA is also highly informative, even between genetically similar strains, or sister strains. The number of informative markers between seven pairs of sister strains is shown in Table 3.3. In order to avoid confusion

Table 3.3: Informative markers between sister strains on MegaMUGA

Strain 1	Strain 2	Informative markers
129S7	129S6/SvEvTac	626
C57BL/6J	C57BL/10J	1567
PWK/PhJ	PWD/PhJ	2560
DBA/2J	DBA/1J	3460
NOD/ShiLtJ	NOR/LtJ	3570
CBA/J	CBA/CaJ	4639
129S4/SvJaeJ	129S5/SvEvBrd	5019

of differing genotype calls on the sex chromosomes, all samples compared in Table 3.3 are male. Here, informative markers include all markers where the two sister strains have differing genotype calls, excluding any markers where at least one strain has an ‘N’ call. An analysis of sister strains that are even more genetically similar than the ones reported here is included in [18].

3.3.3 Pseudoautosomal Region Markers on the MegaMUGA

Out of the 20 invariant markers designed for the Pseudoautosomal Region (PAR), 17 were incorporated in the final array. Since the PAR markers were not selected to contain SNPs, genotype calls for the markers are non-informative, with most PAR markers having only Illumina genotype calls of one allele and ‘N’, as shown in Figure 3.9. However, since the hybridization intensities of the PAR markers vary according to the number of copies of the probe sequence, we can see clear variation between the intensities of different samples. Therefore, to assess the utility of these non-SNP invariant markers without using genotype calls, I examine the clustering of hybridization intensities of the samples.

The set of samples I used for assessing the PAR markers are progeny from crosses between FVB/NJ and (PWK/PhJ x CAST/EiJ) F1 animals. Since these samples are bred from FVB/NJ females and (PWK/PhJ x CAST/EiJ) males, males from this popu-

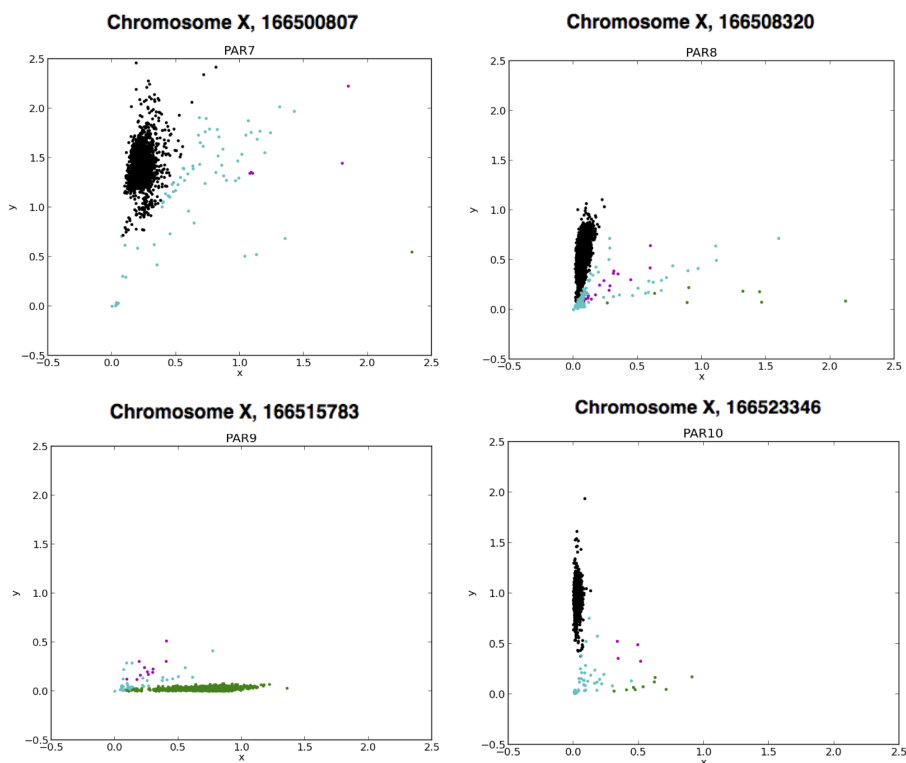


Figure 3.9: Four MegaMUGA PAR markers are shown here for all samples genotyped on MegaMUGA. Each point in each plot represents a single sample, colored by the genotype call provided by Illumina (Green: ‘A’ allele, Black: ‘B’ allele, Cyan: ‘N’, Magenta: ‘H’). The x and y axes are the hybridization intensities of the ‘A’ and ‘B’ alleles designed for each marker. Even though the markers were designed to not have a variant at the typical SNP position, Illumina still assumes a variant SNP and provides genotype calls for the marker. Since there is no actual SNP in the mouse genome, our samples hybridize only in one dimension, resulting in all the genotype calls for a marker being a single allele, ‘N’, or ‘H’. However, the hybridization intensities still differ between samples, which is why I use hybridization intensities for analysis of the PAR markers.

lation have Y chromosomes exclusively from CAST/EiJ and X chromosomes exclusively from FVB/NJ, while females are heterozygous between FVB/NJ and PWK/PhJ in the X chromosome.

I examine the hybridization intensities of male and female samples from this population in Figure 3.10. To observe the haplotype beyond single markers, I perform Principal Component Analysis (PCA) on all marker intensities within specified genomic regions, treating each n -marker window as a vector of length $2n$, which each marker provides an ‘A’ allele hybridization intensity and a ‘B’ allele hybridization intensity.

The PAR in CAST/EiJ begins around 166.0 Mb [79] on the X chromosome, which means males in our population are heterozygous between FVB/NJ and CAST/EiJ right after 166.0 Mb, since the PAR appears in both the X and Y chromosomes. This is easy to see in the first plot at the very beginning of the PAR shown in Figure 3.10, where the light gray male samples cluster together between CAST/EiJ (highlighted in green) and FVB/NJ (highlighted in yellow). The females are clustered with FVB/NJ and PWK/PhJ since they are heterozygous between the two.

Since the PAR is a recombination-rich region [79] where the X and Y chromosomes recombine, further into the PAR, crossover events become more common within our samples. After a crossover event in the PAR, male samples from our population become heterozygous between FVB/NJ and PWK/PhJ, while female samples become heterozygous between FVB/NJ and CAST/EiJ. This can be seen starting in the second and third windows, where a few female samples are drifting toward the male cluster, away from the FVB/NJ and PWK/PhJ heterozygous cluster. Similarly, male samples in these regions are drifting toward the female cluster, away from the FVB/NJ and CAST/EiJ heterozygous cluster. This continues throughout the PAR until sufficient crossover events have occurred and male and female samples are indistinguishable by intensity, and all samples no longer form distinct clusters.

This example shows that even without informative genotype calls, we can observe biologically significant phenomena using a traditional SNP genotyping array. Despite containing no informative SNP variants, these PAR markers have varying hybridization intensities that provide information that would otherwise have been unavailable with traditional genotype calls.

3.4 Discussion

Genotyping microarrays have long been a cost-effective method for assessing the genomes of experimental populations, and the opportunities for designing low-cost and high-information microarrays increase as manufacturing costs drop and reliable SNP and variant annotations become more widely available.

The methods for SNP selection presented in this chapter can be easily applied to any genetic reference population with founders that have well-documented SNPs and variants, or for any set of diverse strains representative of most sequence variations. For the CC and DO, MegaMUGA provides a highly informative platform that can distinguish between a high number of homozygous and heterozygous founder states. Due to the fact that some classical inbred strains are identical by state (IBS), or have identical haplotypes, in many regions of the genome, the number of distinct founder states in 5-SNP sliding windows can drop well below the theoretical maximum. However, with no variants distinguishing between strains that are IBS, even high-throughput sequencing data would not be able to separate all founder states in these regions, and I show that MegaMUGA remains highly informative even between sister strains that share large sequence similarity.

As the costs of both genotyping arrays and high throughput sequencing decrease, the challenge remains to design genotyping arrays so that their informativeness for certain applications can compete with that of sequencing. MegaMUGA achieves this

PCA on Marker Windows in the Pseudoautosomal Region

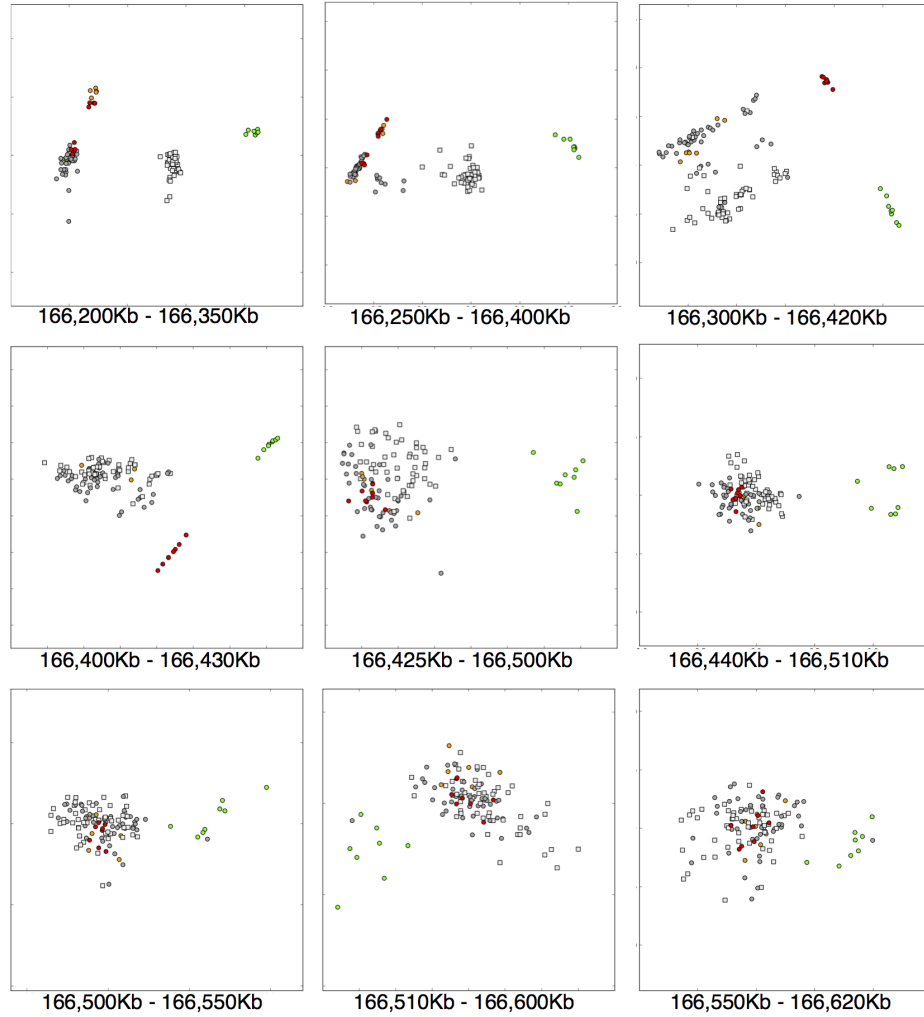


Figure 3.10: PCA on 96 FVB/NJ x (PWK/PhJ x CAST/EiJ) samples in the Pseudoautosomal Region. Replicates of FVB/NJ are highlighted in yellow, PWK/PhJ in red, and CAST/EiJ in green. The gray samples plotted are progeny of FVB/NJ females crossed with (PWK/PhJ x CAST/EiJ) F1 males. The dark gray circles are females, and the light gray squares are males. In each genomic region, I performed Principal Component Analysis (PCA) on the samples' hybridization intensities of all markers within that region. At the beginning of the PAR, the males are heterozygous between FVB/NJ and CAST/EiJ, and the females are heterozygous between FVB/NJ and PWK/PhJ. Frequent recombinations within the PAR eventually bring the male and female samples to have similar genomic backgrounds, as shown by the progression of male female samples blending together in intensity space further into the PAR.

goal in the application of ancestry inference in the CC, as shown by [75], where MegaMUGA achieves a concordance rate of over 98% with high-throughput sequencing data in three different CC samples. This concordance with sequencing data demonstrates the ability of MegaMUGA to distinguish between the different possible founder states in CC samples.

The analysis of invariant probes in the PAR demonstrates the utility of hybridization intensities in the case of non-informative genotype calls. In the MDA, off-target variations within the probe sequence have been shown to manifest as subtle but repeatable changes in the probe hybridization intensity patterns [20], which I discuss in the following chapter. Although MegaMUGA probes were selected to not contain off-target variations, these observations suggest that the intentional inclusion of off-target variants in probe sequences may create markers that can differentiate between more than two SNP alleles, enabling the design of even more informative genotyping arrays.

In the next chapter, I discuss ancestry inference, which is one important application for both the MUGA and MegaMUGA. The ability to assign ancestry relies on the informativeness of SNP markers on a haplotype scale, for which the MUGA and MegaMUGA are both optimized. In addition, the next chapter extends the discussion on using microarray hybridization intensities instead of genotype calls for maximizing information content.

Chapter 4 : Ancestry Inference

4.1 Introduction

In this chapter, I introduce methods for using hybridization intensities from microarrays such as MUGA and MegaMUGA to infer the ancestry of admixed animals [24]. Admixed animals have genomes that are mosaics of segments inherited from their ancestors. Mapping populations, in particular, consist of individuals with mixtures of haplotype segments derived from a set of known founders. 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. The ability to infer ancestry accurately not only enables linkage and quantitative trait loci mapping, but it also adds to our understanding of recombination.

Numerous methods have been proposed for inferring ancestry when given the genotypes of an individual and a set of ancestral haplotypes. Such methods generally use biallelic SNP data obtained from genotyping arrays or DNA sequencing as input. In humans, mapping ancestry is an essential step in admixture mapping, and methods such as HAPMIX [52], HAPAA [64], and LAMP [58] use HMM-based methods to infer the most likely ancestral blocks for each individual. While many methods require prior knowledge of the linkage disequilibrium landscape and use only unlinked markers, HAPMIX uses information from all neighboring markers and points out the amount of information lost by filtering linked markers. However, most of these methods accept genotypes from calling algorithms as ground truth and seldom discuss the impact 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 [49], a package for QTL mapping designed for outbred crosses. Methods for ancestry inference in recombinant inbred strains include two designed for the Collaborative Cross [44, 82], the same population with which I test my algorithms [13]. GAIN [44], 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 lower density arrays do not have the luxury of filtering out a significant percentage of SNPs and keeping only reliable genotype calls.

Moreover, even the best genotype calling algorithms often miscategorize markers with atypical hybridization intensity patterns [20]. In genotype calling, probe hybridization intensities are converted to one of four genotype calls (reference allele, alternate allele, heterozygous, or no call) via a classification algorithm. This is a difficult problem and genotype calling algorithms can generate questionable results when marker intensities deviate from the patterns seen in typical biallelic variants. Furthermore, many markers exhibit unusual intensity patterns due to polymorphisms in or around the target probe sequences [20]. Sometimes sequence variations within probes lead to a reduction in hybridization intensities, and other times they manifest as intensity patterns that can discriminate between more than two alleles (Figure 4.1). In either case, traditional genotype calling methods that assume biallelic SNPs do not correctly classify these markers. This results in a loss of information, or worse, incorrect calls.

I propose an algorithm for ancestry inference that does not require the conversion of hybridization intensities to discrete genotypes. The use of hybridization intensities from

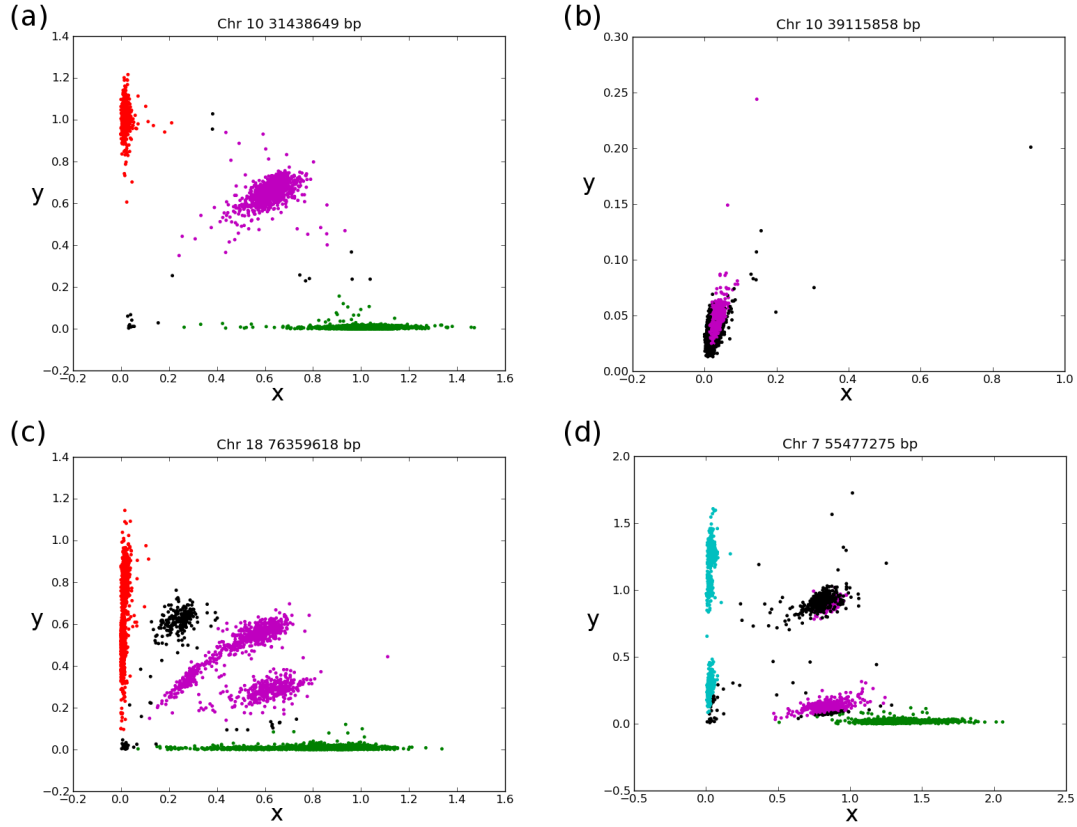


Figure 4.1: Intensity plots of four markers, colored by genotype calls obtained from Illumina’s GenomeStudio. The chromosome location of each marker in Build 37 [12] is shown on top of each panel. Each point represents a single MUGA sample with its reference probe intensity on the x-axis and its alternate 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. As shown in Figure 4.3, replicates of the same strains cluster together in these four markers, suggesting the atypical cluster patterns are driven by real sequence variations.

genotyping arrays is common in studies of copy number variation (CNV) [72]. I show that allele variations beside CNVs manifest as variations in hybridization intensities as well, and I try to implicitly capture these variants with the methods presented here.

At the advent of genotyping microarrays, samples were seldom replicated on the same microarray due to cost constraints. Now, we have the resources to genotype replicates of each strain, and all CC founders and F1 combinations have been genotyped multiple times on both the MUGA and MegaMUGA. This enables me to use hybridization intensities of CC founder and F1 replicates to produce cluster models for each ancestral haplotype, allowing the CC ancestral haplotypes to cluster into more groups than the three typical genotype alleles at each marker.

I do not filter any markers, allowing each marker to be potentially informative on low-density genotyping arrays. In this chapter, my methods are demonstrated on CC strains genotyped with the 7,854-marker Mouse Universal Genotyping Array [16]. Using available DNA sequencing data as ground truth, my algorithm compares favorably to GAIN, which is sensitive to incorrect genotype calls and loses information in atypical markers.

4.2 Materials and Methods

4.2.1 Materials

I implemented the following methods on the admixed population of the CC [13, 16], which is introduced in Chapter 2. I applied my methods to CC samples at various stages of the inbreeding process, ranging from four generations of inbreeding, which is near the peak of genetic diversity (with a large number of founder segments and significant heterozygosity), to 22 generations of inbreeding, where samples are expected to be nearly completely inbred [78]. For comparison purposes, the results in this chapter will focus on three CC samples that have been genotyped and full-genome sequenced.

To ascertain the founder contributions and level of inbreeding of CC lines, my colleagues and I designed the Mouse Universal Genotyping Array (MUGA), a 7,854-marker array based on the Illumina Infinium platform introduced in Chapter 3. Markers were selected to locally discriminate amongst the eight inbred founders, and despite being a low-density array, the MUGA provides highly informative markers that are well-suited for mapping genome ancestry [16]. My methods make use of normalized probe hybridization intensities returned by Illumina rather than genotype calls from Illumina’s GenomeStudio software [66].

To establish statistical distributions for each marker, each CC founder was genotyped on both the MUGA and MegaMUGA using a minimum of eight replicates, which were primarily biological replicates with a few technical replicates. We also genotyped at least two replicates of each of the 28 possible F1 combinations of the eight founders (ignoring the direction of the cross) for a total of 98 F1 samples on the MUGA and 112 F1 samples on the MegaMUGA. In total, I used 65 founder samples and 98 F1 samples to learn clustering models for the MUGA, and 64 founder samples and 112 F1 samples to learn clustering models for the MegaMUGA.

Of the 461 CC samples genotyped with the MUGA, we have high-throughput sequencing data for three. Each of these three samples has approximately 30X genomic coverage in the form of 100 bp, paired-end reads from an Illumina HiSeq 2000, with a mean fragment size of 300 bp. These data were aligned to a CC consensus genome using Bowtie 1.0 with the best-match criterion and allowance for 3 or fewer mismatches per read alignment [75]. The CC consensus genome was formed by substituting the majority allele among the 8 founders into the NCBI Genome Reference Consortium Build 37 mouse reference genome [12] at the high confidence SNP positions as determined by an early pre-release of the Wellcome Trust mouse genome sequencing effort [35]. I used this aligned sequence data to validate the accuracy of my ancestry inference in

the MUGA.

4.2.2 Algorithm overview

In contrast to genotype-based algorithms, I infer ancestor mosaics and probabilities from probe intensities to avoid the limitations of genotype calling introduced earlier. Using intensities from replicate samples, I construct a cluster model for each ancestor and find the set of ancestor intensities that best match the intensities of the target sample. This problem can be framed as both a distance optimization problem with penalties associated with making unnecessary transitions, as well as a classical hidden Markov model, both of which I will discuss.

Given n markers and m inbred ancestors, the distance model minimizes the cumulative distance in 2D probe intensities from the individual sample to each of the $m' = m + \binom{m}{2}$ two-founder haplotype combination states (m homozygous and $\binom{m}{2}$ heterozygous). Each of the m' states has a representative cluster of probe intensities per marker that is pooled from the available founder and F1 replicates, and transitions between different states are penalized via the addition of a transition penalty.

The hidden Markov model presents a different view of a very similar optimization problem, where instead of minimizing the overall intensity distance, the solution for the model maximizes the overall probability of observing the entire sequence of intensities across the genome. This is done by using emission probabilities based on the representative clusters of each of the m' states at each locus, as well as probabilities for transitioning between different states.

4.2.3 Creating reference clusters on MUGA

We have at least eight replicates of each CC founder on MUGA, as well as two or more replicates of each possible F1 combination. Each founder strain forms a re-

peatable 2D probe intensity cluster (Figure 4.2). To create reference clusters with increased statistical power, I pool together founders in common clusters (as determined by Hotelling’s T-square test with a p-value threshold of 0.001) and estimate each final cluster’s mean and covariance. In a second pass, I incorporate F1 samples. When the parental strains of the F1 map to a common cluster, I incorporate the F1 sample into the existing cluster model. When the parental strains of the F1 map to different clusters, I create a new heterozygous cluster model (Figure 4.2). I do not specify an expected number of alleles (clusters) prior to creating reference clusters, allowing for multiple homozygous alleles, each associated with one or more inbred founders. In extreme cases, a poorly performing marker might map all samples to a single cluster. The model handles this case transparently, whereas traditional genotype calling makes arbitrary calls that are likely erroneous.

4.2.4 Creating reference clusters on MegaMUGA

CC reference clusters on MegaMUGA were derived by Chia-yu Kao, as described in [34]. The algorithm used was a non-parametric probabilistic cluster model for each of the m' founder states. This improved cluster model does not rely on the assumption that clusters have Gaussian distributions, but instead interpolates the probabilities of a given intensity vector belonging to a specific founder using observed intensities of founder replicates. The method is similar to the method used for MUGA in that it relies on biological and technical replicates of the CC founders and F1s genotyped on MegaMUGA, and the homozygous founders are clustered in a first pass, with the F1s clustered to form heterozygous cluster models in a second pass. The results are stored in lookup tables for each i^{th} marker, indicating $P(f_i = q|x_i)$, the probability of the ancestor at the i^{th} marker coming from founder state q given the observed intensity vector x_i of the admixed sample.

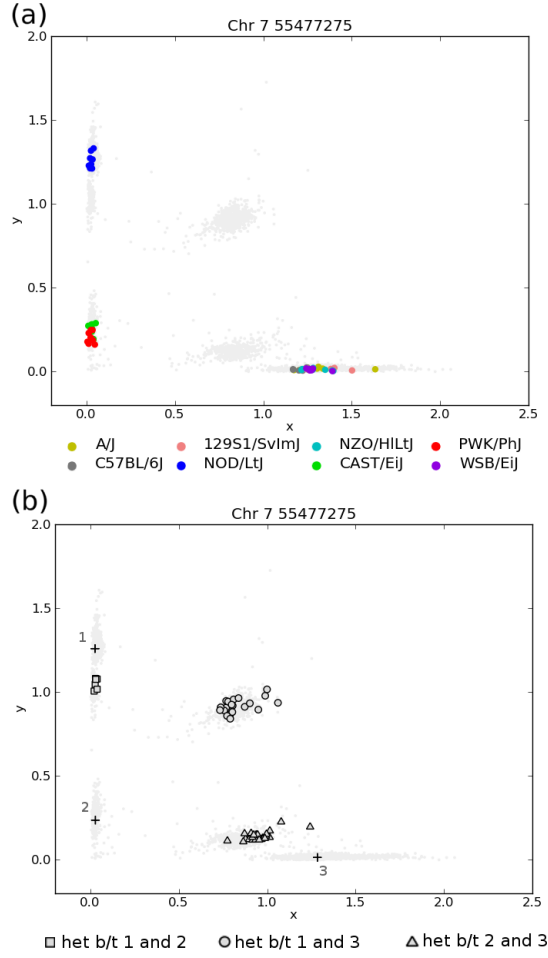







Figure 4.2: Creating reference clusters in the MUGA. (a) To create homozygous reference clusters, I first pool all inbred founders with overlapping clusters as determined by replicate samples. The SNP shown here has three homozygous clusters. (b) I then create heterozygous reference clusters by pooling F1 samples between founders in different clusters. This SNP has three heterozygous clusters. I also refine homozygous clusters by adding F1 samples between founders in the same homozygous clusters. The means of homozygous clusters 1, 2, and 3 are shown as crosses. Data points for all samples are shown in the background to provide context. They are not used in the cluster modeling.

4.2.5 Distance Model

The goal in the minimum distance model is to assign the set of most likely ancestor states $\{f_1, f_2, \dots, f_i, \dots, f_n\}$ for each marker from position 1 to n . The set of possible ancestor states F contains $m' = m + \binom{m}{2}$ possible haplotype combinations given m founders

Table 4.1: Transitions between different states p and q

p is hom	q is hom	p, q share 1 haplo	Graphical depiction	$penalty(p, q)$
yes	yes	no		mean D_M b/t hom clusters
yes/no	no/yes	yes		1.5* mean D_M b/t homo and het clusters
no	no	yes		1.5* mean D_M b/t different het clusters
yes/no	no/yes	no		5.0*mean D_M b/t hom and het clusters
no	no	no		5.0*mean D_M b/t het clusters

(with the exception of sex chromosomes on male samples, which only have m states). At marker i , each ancestor state $q \in F$ has a cluster model $cluster(q, i)$ with a stored mean and covariance. I define the distance at each marker i from the target sample to each ancestor state q as the Mahalanobis distance [46] of the sample's 2D probe intensities to $cluster(q, i)$. The goal is to find the set of ancestor intensities that best models the target sample's intensities across the genome without excessive transitioning. Hence, denoting the target sample's 2D intensity vector as x_i and the assigned ancestor as f_i at marker i , we wish to minimize

$$D_M(x_1, f_1) + \sum_{i=2}^n D_M(x_i, cluster(f_i, i)) + penalty(f_{i-1}, f_i), \quad (4.1)$$

where $D_M(x_i, cluster(f_i, i))$ is the Mahalanobis distance from the 2D point x_i to the reference cluster of f_i at position i , and $penalty(f_{i-1}, f_i)$ is the transition penalty from the assigned state at marker $i-1$ to the state at marker i , defined below.

I set up a dynamic program to find the path which minimizes $dist_{f_{i+1}=q}$, the total distance from the first marker to having the assigned founder q at marker $i + 1$. Given that the previously assigned founder state was p at position i , the main dynamic programming recurrence then becomes

$$\begin{aligned}
dist_{f_{i+1}=q|f_i=p} = & D_M(x_{i+1}, cluster(q, i + 1)) + penalty(p, q) \\
& + min\{dist_{f_i=p|f_0=r} | \forall r \in F\}.
\end{aligned} \tag{4.2}$$

Since the algorithm does not require knowledge of pedigree, transition penalties are based on observed differences in probe intensities between different founder states. Using the predetermined founder and F1 clusters, I calculate the mean Mahalanobis distance from homozygous clusters to other homozygous clusters, from heterozygous clusters to homozygous clusters, and from heterozygous clusters to other heterozygous clusters. Using these mean Mahalanobis distance values, I allow for transitions between homozygous states when I encounter a single SNP with typical Mahalanobis distance between two different homozygous clusters. To account for the fact that heterozygous clusters are typically closer to all other clusters, the penalty to transitioning to or from heterozygous states is equivalent to 1.5 times the typical Mahalanobis distance between heterozygous states and other states. Transitions that suggest two independent recombination events at the same locus (coincident transitions) are rare and are penalized more heavily in this model. I set this penalty to be five times the mean Mahalanobis distance between different states. The set of possible transitions between state p and state q , where $p \neq q$, are shown in Table 4.1. Transition penalties are symmetric, and there is no penalty value for staying in the same state, that is, $penalty(p, q) = penalty(q, p)$ and $penalty(p, p) = 0$.

For the CC dataset genotyped on MUGA, the penalty values are 0.082 between different homozygous states, 0.066 between heterozygous and compatible homozygous states, and 0.047 between compatible heterozygous states. Coincident transitions have penalty values of 0.22 and 0.16.

4.2.6 Hidden Markov Model

To obtain the probabilities of the admixed animal inheriting each locus of its genome from each ancestor, the above distance model can be extended to a hidden Markov model (HMM), similar to ones used in [82, 44, 49]. The hidden states in this case are $\{f_1, f_2 \dots f_i \dots f_n\}$, the true founders at each marker locus, and the observed outcomes are the admixed sample's probe intensities at each marker, or $\{x_1, x_2 \dots x_i \dots x_n\}$. The task of the HMM is then to learn the sequence of true founder states.

The clustering algorithms for both MUGA and MegaMUGA return $P(f_i = q|x_i)$ for each marker i and each founder state q , which is the probability of the sample having descended from founder q given its intensity vector x_i . The emission probability for each founder state $q \in F$ is $P(x_i|f_i = q)$, which can be calculated using Bayes' rule:

$$P(x_i|f_i = q) = \frac{P(f_i = q|x_i)P(x_i)}{P(f_i = q)} \quad (4.3)$$

We can assume that each ancestor is inherited with equal probability throughout the genome, and that every 2D intensity vector x_i is emitted with the same probability, so that $P(f_i = q) = \frac{1}{|F|} = \frac{1}{m'}, \forall q \in F$ and $P(x_i) = P(x_j), \forall i, j$. Given these assumptions, we can see that:

$$P(x_i|f_i = q) \propto P(f_i = q|x_i), \quad (4.4)$$

which means we can use $P(f_i = q|x_i)$ from the clustering algorithms as the emission probability as long as the outputs are scaled into probabilities that sum up to 1 at each marker locus.

The transition probabilities are estimated using previously observed recombinations in similar populations and are presented in [75]. To solve the HMM, I use the Viterbi algorithm to find the most likely founder at every locus, obtaining the length- n vector of most likely founders, $\{f_1, f_2, \dots f_i, \dots f_n\}$. I also run the Forward-backward algorithm

on the HMM to find the probability of the admixed genome belonging to every founder state at each locus, obtaining an $n \times m'$ matrix of results, where for each marker position i , we have the vector $\langle P(f_i = q_1 | \mathbf{x}_{1:n}), P(f_i = q_2 | \mathbf{x}_{1:n}), \dots, P(f_i = q_{m'} | \mathbf{x}_{1:n}) \rangle$ indicating the probability of the sample being descended from each founder given the entire sequence of observed sample intensities, where m' is the total number of homozygous and heterozygous founder states.

4.2.7 Refining recombination breakpoints

Once the sequence of most likely founders is obtained via the distance model solution or the Viterbi solution for the HMM, determining recombination breakpoints between founders that share similar or identical sequences near transitions remains a challenge. Although the inference algorithms will specify a transition between some pair of adjacent markers, I report the breakpoint as an interval of ambiguity where the true breakpoint falls. For a transition from ancestor states p to q , I start from the breakpoint locus given by the solution and extend the ambiguous interval both ways, stopping when I reach a left endpoint i where $D_M(x_i, \text{cluster}(p, i)) < D_M(x_i, \text{cluster}(q, i))$ for the distance model, or $P(f_i = p | x_i) > P(f_i = q | x_i)$ for the HMM. Similarly, we stop at a right endpoint j where $D_M(x_j, \text{cluster}(p, j)) > D_M(x_j, \text{cluster}(q, j))$, or $P(f_i = p | x_i) < P(f_i = q | x_i)$. These two endpoints indicate the markers where the target sample is noticeably more likely to be descended from one founder state over another, and the region in between these markers are assumed to be identical by state (IBS) between the assigned founder states p and q .

4.2.8 Funnel constraints

Assuming a funnel order of $ABCDEFGH$ for a CC strain, heterozygous combinations of the initial mating pairs AB , CD , EF , and GH cannot reappear in later

generations, since the genomic material passed from an F1 cross is carried on a single haplotype in all subsequent generations [7, 16]. When applying the algorithm to CC samples with available funnel information, I incorporate this constraint by removing these four prohibited founder states.

4.3 Results

4.3.1 Reference intensity clusters

I created reference clusters for 7,854 MUGA markers using a total of 65 CC founder samples and 98 CC F1 samples. The eight CC founders segregated into a single cluster for 1,104 markers. I observed the expected biallelic intensity clusters in 5,550 markers, with two homozygous clusters and a single heterozygous cluster among the CC reference samples. The remaining 1,200 markers exhibit three or more clusters among the eight inbred founders, with 1,021 exhibiting three homozygous clusters, and 179 exhibiting four or more (Figure 4.3). The maximum number of homozygous clusters I observed was six. The distribution of markers colored by the number of reference intensity clusters in each is shown in Figure 4.4.

Of the 6,750 informative markers in MUGA, there was a mean of 2.21 homozygous clusters per SNP, or 3.66 total (homozygous and heterozygous) clusters per SNP. Thus, using the reference clusters, each SNP provides more information than typical genotype calls with 2 homozygous and 1 heterozygous alleles. This is especially advantageous for low-density platforms such as the MUGA and allows us to break ties between similar founders and refine recombination breakpoints, as discussed below.

In MegaMUGA, the 75,132 SNP markers in autosomes and the X chromosome were clustered using the algorithm described in [34]. The eight CC founders appeared in a single cluster in 4,730 markers, making the number of informative SNP markers in the MegaMUGA 70,402. Of the 70,402 informative SNP markers, the mean number

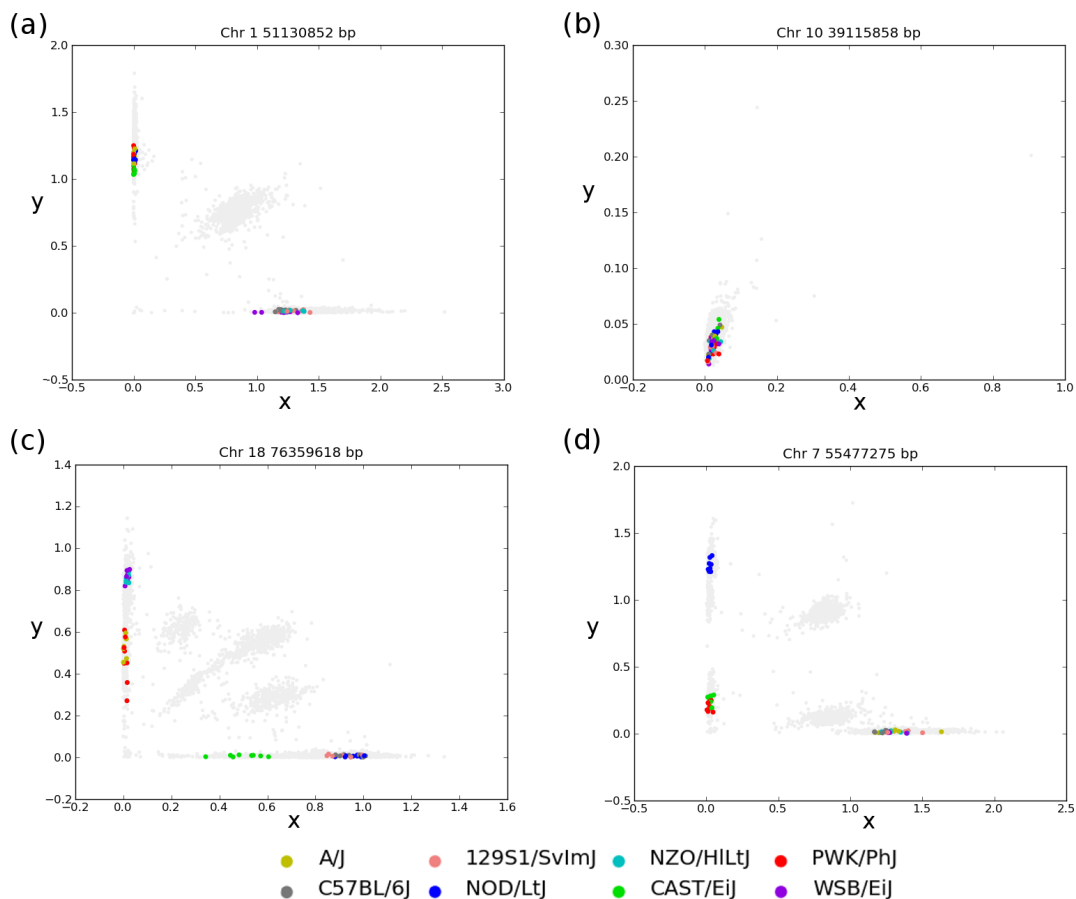


Figure 4.3: Intensity plots with replicates of CC founders highlighted and all other samples drawn as background, with the same four markers as Figure 4.1. Founders with overlapping clusters are pooled to create a single homozygous cluster. (a) A typical biallelic marker with the expected two homozygous clusters. (b) A poorly performing marker with a single cluster. (c) A marker with four homozygous clusters. (d) A marker with three homozygous clusters.

of clusters per SNP was 3.06, which is only slightly higher than the typically assumed three clusters per marker. This was potentially due to the intentional inclusion of non-SNP invariant markers with few clusters, as well as the successful elimination of Off-Target Variants in all SNP markers using high-confidence SNPs from the Sanger Institute. The highest number of clusters observed in the MegaMUGA was 9, and is shown in Figure 4.5

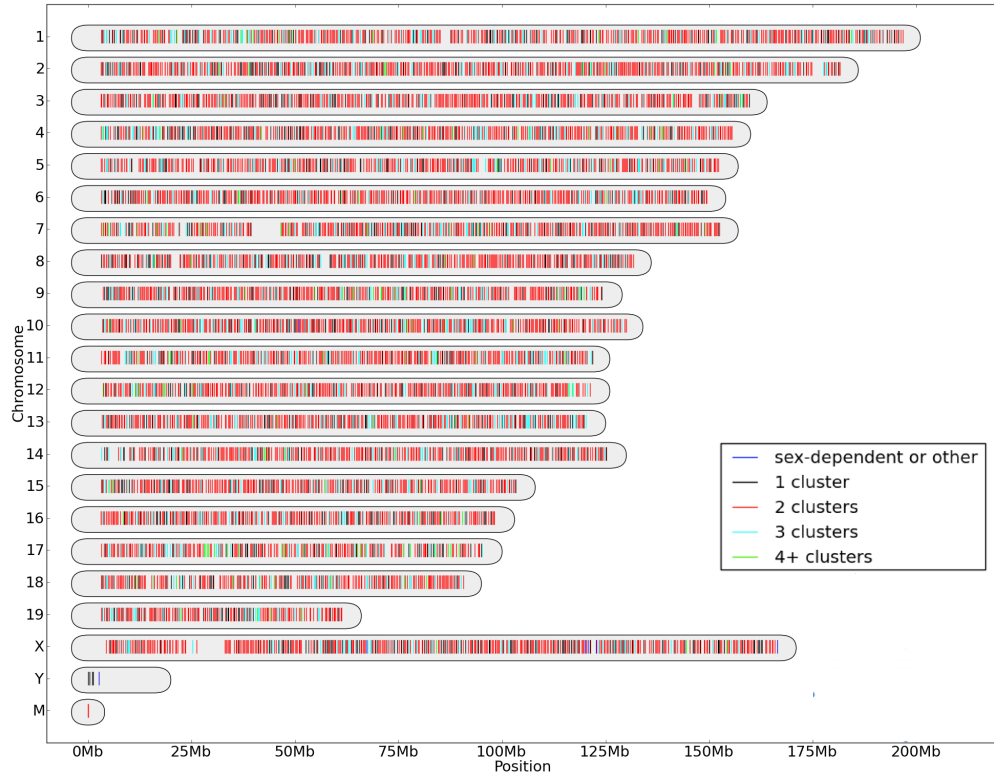


Figure 4.4: All MUGA markers colored by the number of CC reference homozygous intensity clusters found in each marker. The total number of multiallelic (three or more homozygous clusters) markers is 1,200, and the number of 1-cluster markers is 1,104. Both types of unexpected intensity variations are found more frequently in different regions of the genome, suggesting that distinct haplotypes spanning several markers can be captured in intensity clusters. The majority of markers (5,550) still have the traditional intensity pattern of two homozygous clusters.

4.3.2 The role of off-target variants in intensity clusters

To further investigate the probe sequence variations that lead to unexpected intensity clusters, I examined off-target variants (OTVs), defined as high-confidence SNPs and indels in CC founders annotated by Sanger Institute’s Mouse Genome Project [35] that occur in the 49 bp MUGA probe sequences. Although 2,342 OTVs were found in

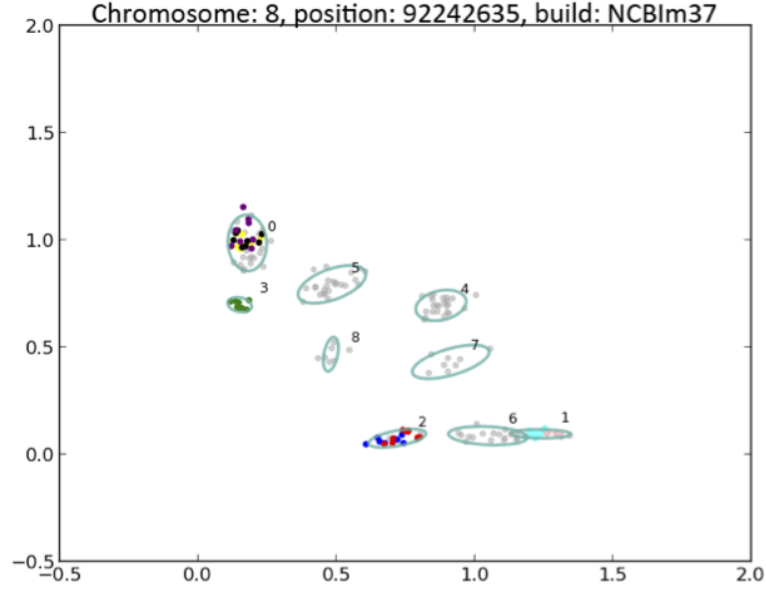


Figure 4.5: A MegaMUGA marker with 9 clusters within the CC founders and F1s, which is the highest number of clusters observed in the MegaMUGA. The inbred founders are highlighted with their respective color codes.

MUGA probes, only 1,474 resulted in observable intensity cluster differences. The observable effects on probe intensity seem to depend in part on the relative position of the OTV on the probe. As shown in Figure 4.7, OTVs that result in unexpected intensity patterns are distributed much closer to the target SNP than OTVs that do not affect intensity, with the majority of OTVs that affect intensity immediately adjacent to the target SNP. For reference, I also plotted the distance to the closest high-confidence SNP or indel within all CC founders for all high-confidence Sanger SNPs with other CC variants within 49 bps, and found that in most cases, the closest SNP position is immediately adjacent, as shown in Figure 4.6.

For each marker, I used the information on known OTVs to group the eight CC founders according to their SNP allele and probe sequences given by Sanger. These groupings from Sanger sequences were then compared to groupings by Illumina's consensus genotype calls for each founder, as well as by the CC intensity clusters. I found that the founders grouped by Illumina's genotype calls were in concordance with the

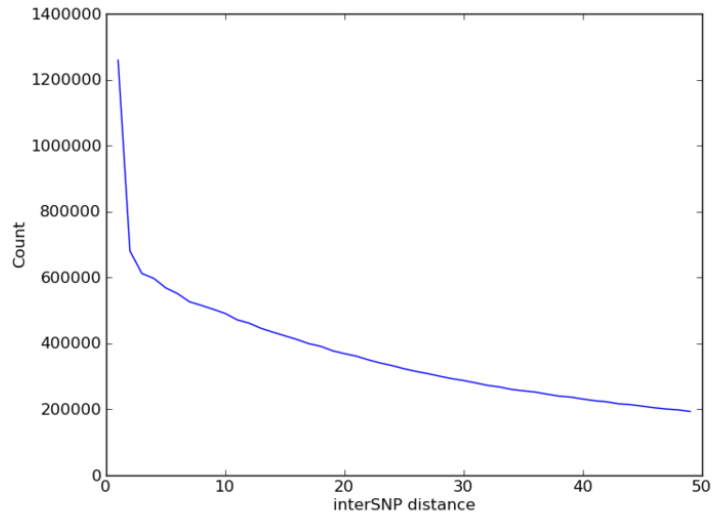


Figure 4.6: Distance to the closest variant among all high-confidence CC SNPs with variants within 49 bp, as documented by Sanger [35]. A total of 18,133,048 CC SNPs have other variants within 49 bp, which is 57.4% of all SNPs documented in the CC.

founders grouped by Sanger sequences in 69.36% of MUGA markers. In contrast, founders grouped by my intensity clusters were in concordance with founders grouped by Sanger in 81.15% of MUGA markers. This significantly higher concordance rate is explained when we examine individual markers such as the ones shown in Figure 4.8, where we see OTVs that further separate founders belonging to the same group of SNP alleles.

4.3.3 Ancestry inference comparisons using sequencing data

Using available sequencing data as ground truth, I compared the predictions of my distance model to those of GAIN, a genotype-based method optimized for animals with complex pedigrees such as CC animals [44]. GAIN uses knowledge of the breeding funnel and generations of inbreeding to approximate transition probabilities in a hidden Markov model. As with most genotype-based methods, GAIN infers heterozygous genotypes and requires genotypes from only the inbred founders. I used the consensus

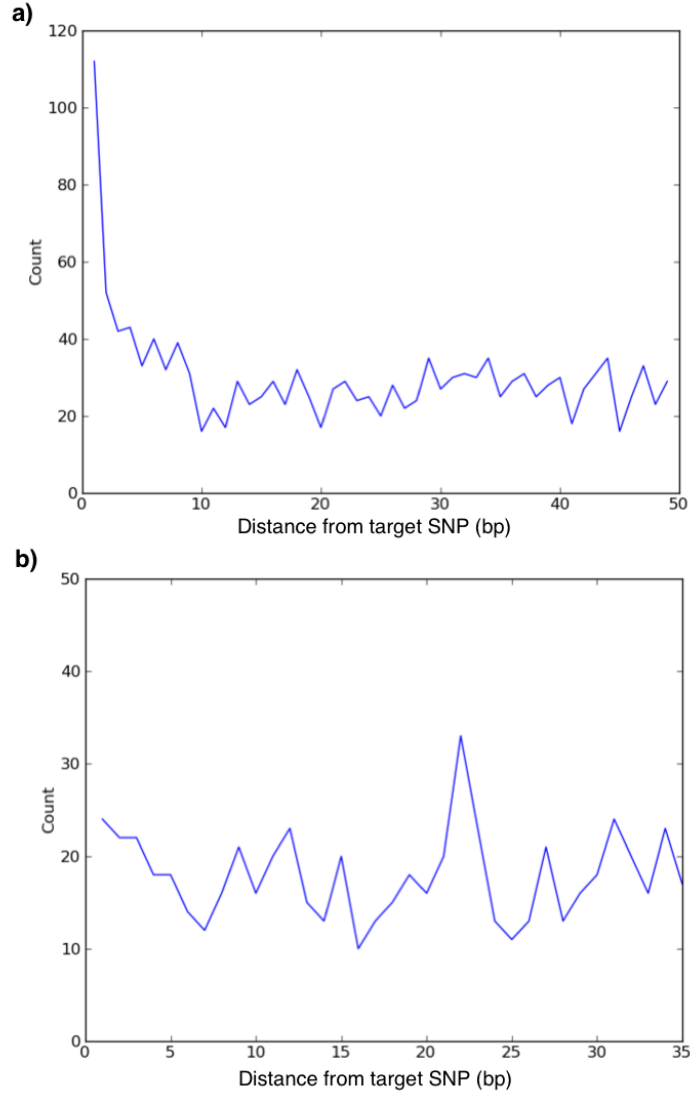


Figure 4.7: MUGA probe location of off-target variants (OTVs) and their effects on intensity clusters. (a) OTVs that effect marker intensity clustering. This shows a histogram of the 1474 OTVs on MUGA probes with unconventional intensity clusters, grouped by their distance to the target SNP. (b) OTVs that do not effect marker intensity clustering. This shows a histogram of the 868 OTVs on MUGA probes with typical biallelic intensity clusters, grouped by their distance to the target SNP. OTVs closer to the target SNP seem to have a more observable effect on probe intensities.

genotype calls given by Illumina’s GenomeStudio software [66] from all samples of each CC founder. Since GAIN requires that all founders be called a homozygous allele at each marker, I filtered the 7,854 MUGA markers by eliminating all markers where a

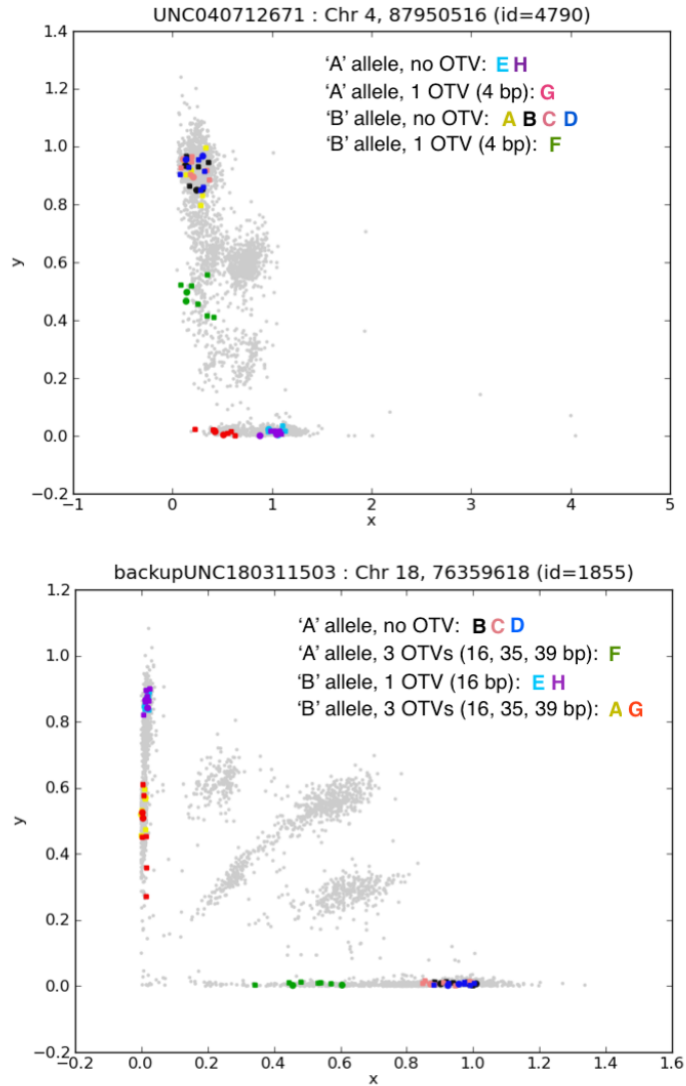


Figure 4.8: Two MUGA markers where off-target variants (OTVs) result in unexpected intensity clusters. The SNP alleles and OTVs for each markers are shown and are as determined by sequencing data from the Sanger Institute [35]. 'A' and 'B' alleles simply refer to the alleles at the target SNP hybridizing in the 'x' and 'y' intensity directions, respectively. Each OTV's position relative to the target SNP position is also noted.

CC founder's consensus call was 'H' or 'N,' as well as markers where all eight founders have the same call, leaving 5,782 markers. In comparison, my algorithm uses every marker. This includes 6,750 informative markers with more than one cluster, nearly 1000 markers more than the ones used by GAIN.

Table 4.2: SNPs that disagree between the Distance Model (DM) vs. GAIN

	Total SNPs disagreeing b/t DM and GAIN	SNPs where DM agrees w/ sequence	SNPS where GAIN agrees w/ sequence
OR867m532	33,026	24,092	8,934
OR1237m224	17,536	14,524	3,011
OR3067m352	38,621	23,095	15,526
Total	89,183	52,144 (69.2%)	27,471 (30.8%)

I ran my distance model and GAIN on three CC samples with DNA sequencing data available: OR867m532, OR1237m224, and OR3067m352. I examined the non-ambiguous regions where the two methods disagree and imputed high-confidence SNPs from the Wellcome Trust Sanger Institute [35] for these regions based on the inferred ancestries. I then estimated the true genotypes by examining the aligned reads at each SNP locus, considering only loci with a coverage of ten or more reads. Loci where the second most common nucleotide showed up with a frequency of more than 0.2 were declared heterozygous.

Of all high-confidence Sanger SNPs in regions where the two methods disagree, 69.2% of SNPs imputed using ancestor assignments from my method agree with the sequencing data, compared to 30.8% of SNPs imputed from GAIN that agree with sequencing (Table 6.3). With the assumption that the aligned sequencing data and Sanger Institute SNPs are correct, loci with imputed SNPs that differ from the sequencing data most likely result from erroneous ancestry assignments. As seen in the sample plot of chromosomes 3 and 5 on OR1237m532 (Figure 4.10), errors in GAIN are often driven by incorrect genotype calls, where a single miscalled genotype can result in an incorrect assignment. These incorrect genotype calls often occur in markers with intensity clusters that do not separate as well as typical biallelic intensity clusters, and the discretization from intensities to genotype calls in these cases easily lead to errors in algorithms relying on correct genotype calls.

Unlike genotype-based methods, the reference clusters here can make use of markers

Table 4.3: Comparison of MUGA and MegaMUGA solutions to sequence data

Sample	Number of segments			Concordance with HTS	
	HTS	MUGA	MegaMUGA	MUGA	MegaMUGA
OR867m532	117	108	118	95.56%	98.09 %
OR1237m224	116	102	116	95.97%	98.37 %
OR3067m352	112	102	113	96.76%	98.95 %

where ancestors have “H” or “N” calls, and they can discriminate between ancestors with the same genotype call but have different hybridization intensity patterns. For example, my algorithm defines a recombination breakpoint between 15,059,945 bp and 15,922,708 bp on chromosome 17, where the ancestor of OR1237m224 transitions from homozygous WSB/EiJ (purple in Figure 4.9) to homozygous CAST/EiJ (green). GAIN reports the recombination breakpoint to be between 14,675,894 bp and 18,347,703 bp, a region 2.8Mb larger than that reported by my algorithm. From the pileups of the aligned sequencing reads, we can refine the true breakpoint to the 5Kb region centered around 15,060,000 bp. My algorithm is able to more precisely discriminate the breakpoint region due to a marker with an ‘N’ genotype call and a marker with three homozygous clusters flanking the breakpoint. GAIN does not consider the marker immediately upstream of the true breakpoint at 15,059,945 bp since four of the eight CC founders are called ‘N’ at the locus, along with the target individual. However, WSB/EiJ and CAST/EiJ clearly segregate into separate clusters at the marker, with the target individual falling in WSB/EiJ’s reference cluster, which is recognized by my algorithm. The marker downstream of the true breakpoint at 15,922,708 bp has three homozygous reference clusters, with WSB/EiJ and CAST/EiJ sharing the same genotype calls but segregating into different clusters. GAIN is unable to differentiate between the two founders at that marker due to their shared genotype call, but my algorithm is able to assign the genomic region to CAST/EiJ’s reference cluster (Figure 4.9).

The hidden Markov model (HMM) is a logical extension of the distance model, and the optimization objectives for the two models are extremely similar. The additional information provided by the HMM forward-backward solution, compared to that of the shortest distance model, is illustrated in Figure 4.12.

In [76], my colleagues and I report comparisons between MUGA distance model, MegaMUGA distance model, and high-throughput sequencing data solutions for the ancestry in the three CC samples OR867m532, OR1237m224, OR3067m352. The results are summarized in Table 4.3 in terms of the number of detected founder segments and the percent of the genome with ancestry concordant with the sequencing solution. As expected, the higher-resolution MegaMUGA was able to detect more recombination segments, with higher concordance with the sequencing solution. A detailed example of results for the same sample genotyped on both MUGA and MegaMUGA using the distance model is shown in 4.11. In addition to reporting the most likely solution from the distance model, the matrix of probabilities provided by the forward-backward algorithm are provided in my online tool at <http://csbio.unc.edu/CCstatus/index.py?run=AvailableLines>,

4.3.4 Other platforms and populations

The presented algorithms work across different genotyping platforms. In addition to testing on the MUGA and MegaMUGA, I have also tested the algorithms on CC animals genotyped with the 600K-marker Mouse Diversity Array (MDA) [80], a high-density genotyping array on the Affymetrix platform. Since there are fewer replicates of CC founders and F1s genotyped on the MDA, instead of creating reference clusters and calculating Mahalanobis distances, I used other distance measures, such as 2D Euclidean and Manhattan distances, to calculate distance between the individual sample and each ancestor. In the case of F1 strains without available samples, I approximate the intensities of the F1 by taking the mean intensities of its two parental strains. This

approximation produces results similar to those of using real F1 samples.

Although I have focused the results on samples from the CC, my algorithms have been implemented on other populations that have been genotyped on MUGA and Mega-MUGA as well. I have also tested my algorithms on heterogeneous stocks such as the Diversity Outbred (DO) population being developed at The Jackson Laboratory [65], as well as transgenic, knockout, and knockin mice from the Mutant Mouse Regional Resource Center (MMRRC; <http://www.mmrrc.org>). Like the CC, these mice are derived from two or more ancestors. For an ancestor that is not a CC founder, I assign the closest CC founder-derived reference cluster at each marker and run the distance model with the most likely set of reference clusters representing the ancestor. Since the CC founders capture most genetic diversity in the mouse, the reference clusters I created using CC founders and F1s work well for modeling non-CC ancestors as well (Figures 4.13, 4.14). Online tools implementing my algorithms for CC animals on the MUGA and Mega-MUGA can be found at <http://csbio.unc.edu/CCstatus/index.py?run=NewFoundersMM>. Online tools for inferring ancestry in non-CC animals can be found at <http://www.csbio.unc.edu/MMR>.

4.4 Discussion

Existing methods for ancestry inference assume accurate genotype calls that model all variants within a marker. However, markers may capture multiallelic information due to off target polymorphisms in the target probe sequence, and I have observed a substantial number of markers in multiple genotyping platforms which consist of more than two homozygous intensity clusters. My ancestry inference methods cluster ancestors based on probe intensities and solves an optimization problem to find the most likely sequence of ancestors given the intensities of an admixed sample. By using probe intensities instead of discretized genotype calls, I obtain more information from multiallelic markers and markers with many ‘N’ calls, and eliminate errors due to

incorrect genotype calls in markers with atypical intensity patterns.

Since microarrays sample the genome at only selected points, there is a fundamental limit on the resolution of detectable ancestral segments. However, I show that the inferred ancestry in both MUGA and MegaMUGA have high concordance with solutions obtained from high-throughput sequencing data, with MegaMUGA in particular capturing most haplotype segments. Although markers in MegaMUGA were spaced according to recombination frequency from a linkage map, in arrays where inter-marker distances vary greatly on a recombination scale, using the same transition probabilities between all adjacent markers may not be ideal, and parameterized transition probabilities dependent on inter-marker distances from a linkage map may lead to even more accurate solutions.

I have demonstrated that probe hybridization intensities provide valuable information that is often lost after genotype calling. Although some perceive intensities as noisy data, intensity-based ancestry inference produces good results in even the low-density MUGA, eliminating noise originating from incorrect genotype calls. Furthermore, I am able to specify recombination regions more precisely due to additional information from intensities. Intensity-based methods can be used to solve many other problems that traditionally rely on discretized genotype calls, such as the problem of quantitative trait loci (QTL) mapping, which is presented in the next chapter. Using intensity-based methods provides more direct comparison of admixed animals and their ancestors, whereas using genotyping calls assume the admixed animal and its ancestors are all similar to the reference sequence used in the microarray marker selection and probe design. In cases where discretized genotype calls are desired, genotype calling algorithms that allow for an arbitrary number of alleles per marker, such as one presented in [34], could lead to more accurate results than would traditional biallelic calls.

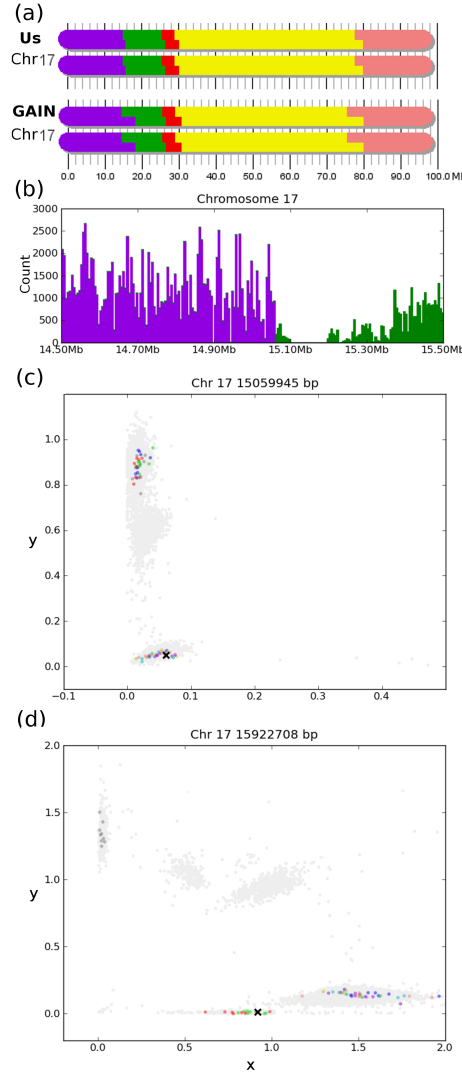


Figure 4.9: Intensity information better resolves a recombination breakpoint on chromosome 17 of sample OR1237m224. (a) Assigned ancestry from my distance model (top) and GAIN (bottom). Both algorithms show a recombination breakpoint between WSB/EiJ (purple) and CAST/EiJ (green) around 15Mb. The distance model shows the region containing the breakpoint as 15,059,945 - 15,922,708 bp, while GAIN shows the region as 14,675,894 - 18,347,703 bp. (b) Sequencing data pinpoints the breakpoint to a 5Kb region centered around 15.06 Mb. Here, I show the number of SNPs from aligned reads which are informative between WSB/EiJ and CAST/EiJ, colored by the SNP's allele. (c) The marker immediately upstream of the true breakpoint. CC founders are highlighted and OR1237m224 is marked as "x." This marker was filtered by GAIN due to the high number of 'N' calls among CC founders, but the sample falls within the cluster with the WSB/EiJ allele. (d) A marker downstream of the true breakpoint. WSB/EiJ and CAST/EiJ share the same genotype call at the marker, so GAIN cannot discriminate between the two. However, WSB/EiJ and CAST/EiJ fall in different reference clusters, so we can accurately assign the sample to the cluster containing CAST/EiJ.

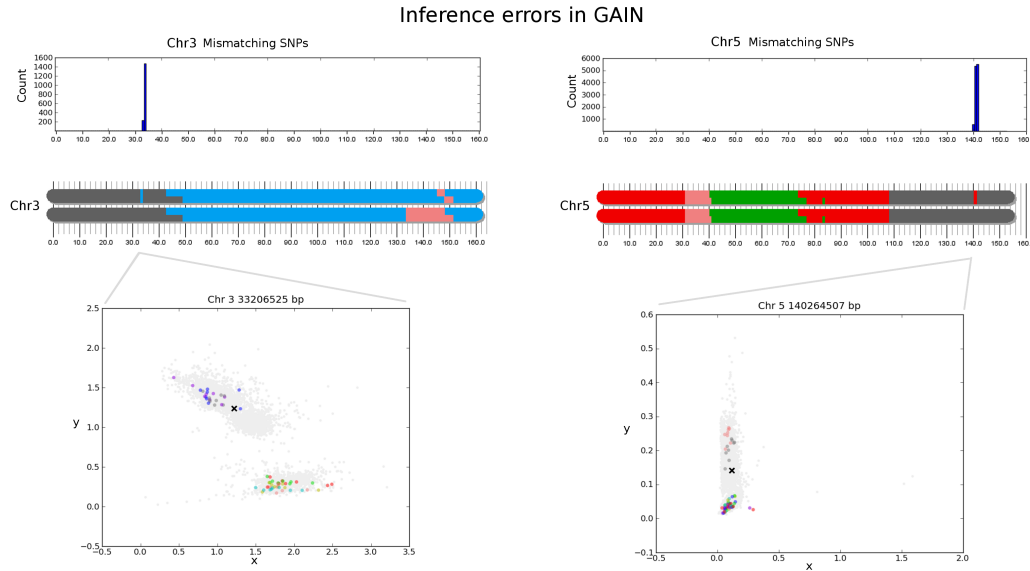


Figure 4.10: Errors in GAIN are often due to questionable genotype calls. Here I show results from GAIN on chromosomes 3 (left) and 5 (right) of sample OR1237m224. The histograms on the top show the SNPs with alleles imputed from GAIN that differ from sequence data, out of all SNPs in regions where my assignments differ from GAIN's. This suggests the small heterozygous segments assigned by GAIN on both chromosomes are erroneous. GAIN's ancestry assignment is depicted in the middle, and the bottom plots show SNPs where the sample is called 'H'. In both chromosomes, the errors occur in regions of markers where the sample is called 'H' yet has an intensity vector close to the correct homozygous cluster (dark gray). CC inbred founders are highlighted in the intensity plots, and the intensity of OR1237m224 is marked "x."

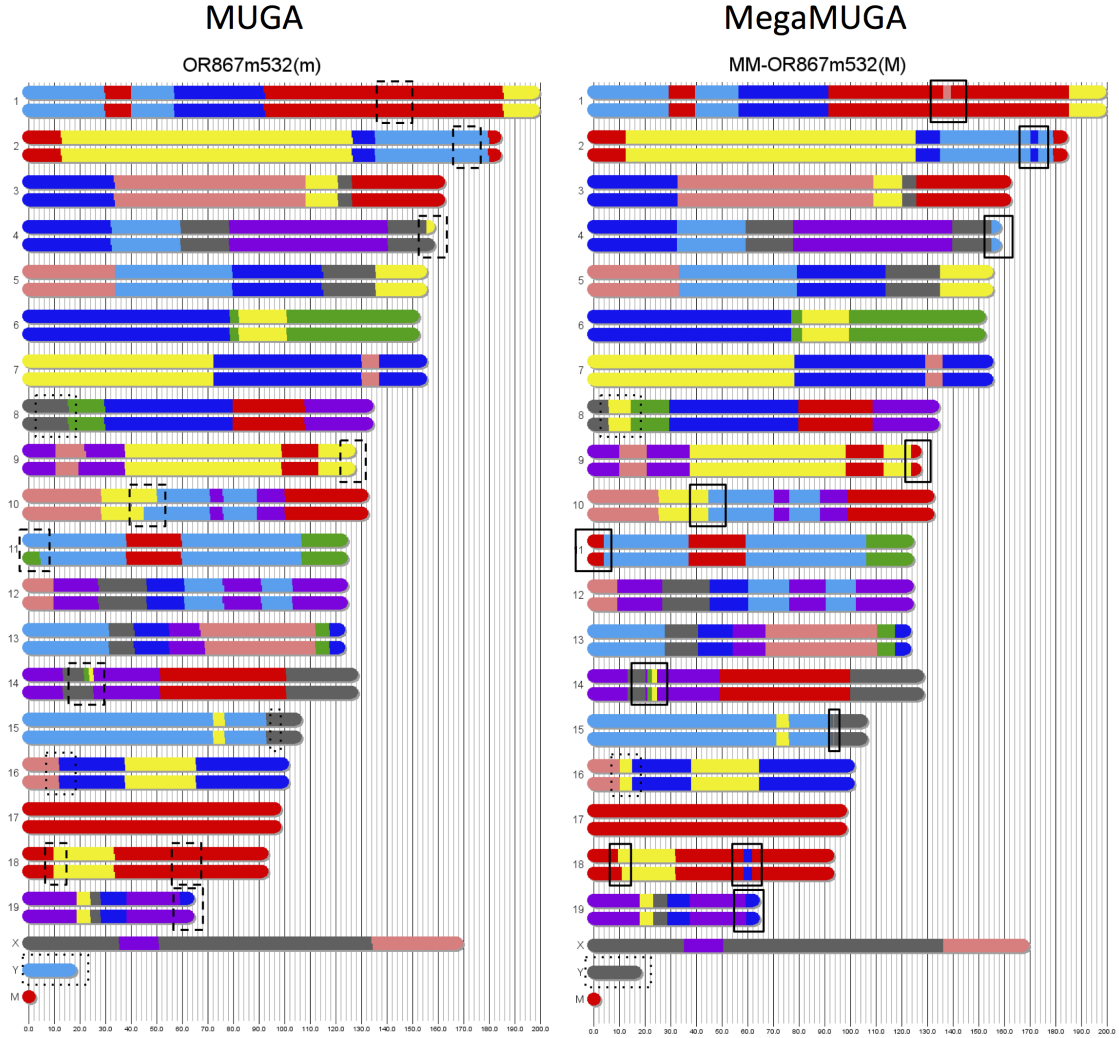


Figure 4.11: Comparison of ancestry inference results on sample OR867m532 using the distance model in MUGA and MegaMUGA. Segments that differ between the two results are outlined with boxes. Segments that have been confirmed using sequencing data from [78] are boxed with a solid lines, while segments that are in discordance with sequencing data are boxed in dashed lines. Segments where sequencing data is inconclusive, where no reads are informative between two or more founders, are boxed in light dotted lines. Due to its higher density, MegaMUGA more accurately captures small recombination segments, especially small heterozygous segments such as the ones on Chromosomes 1 and 18.

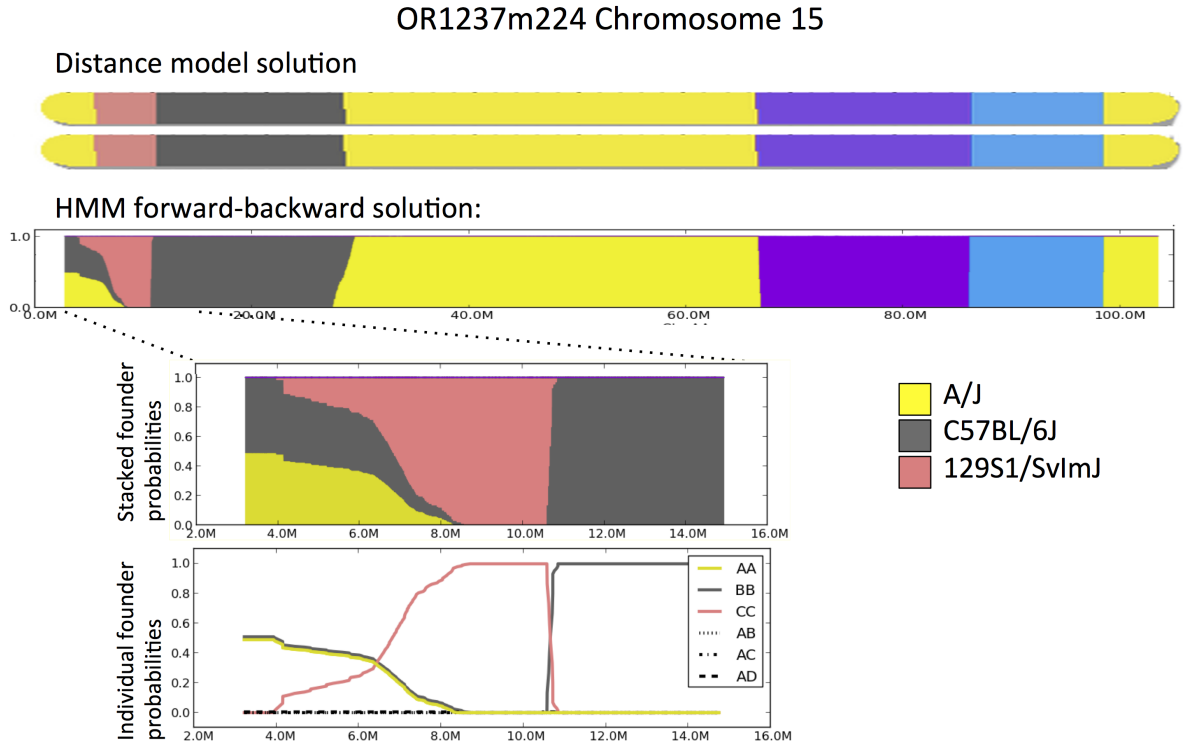


Figure 4.12: Comparison between Distance Model and HMM forward-backward solutions for Chromosome 15 of sample OR1237m224. Sequencing data [78] shows no informative reads between A/J and C57BL/6J until 8.5 Mb, and no informative reads between A/J, C57BL/6J, and 129S1/SvImJ between 4.5 Mb and 8.5 Mb. Informative reads have 129S1/SvImJ alleles starting at 8.5 Mb. While the shortest distance model selects a single founder in regions of multiple probable founders, the forward-backward solution from the HMM include probabilities for each possible founder, reflecting the regions of identity by state (IBS) where some founders have identical sequences.

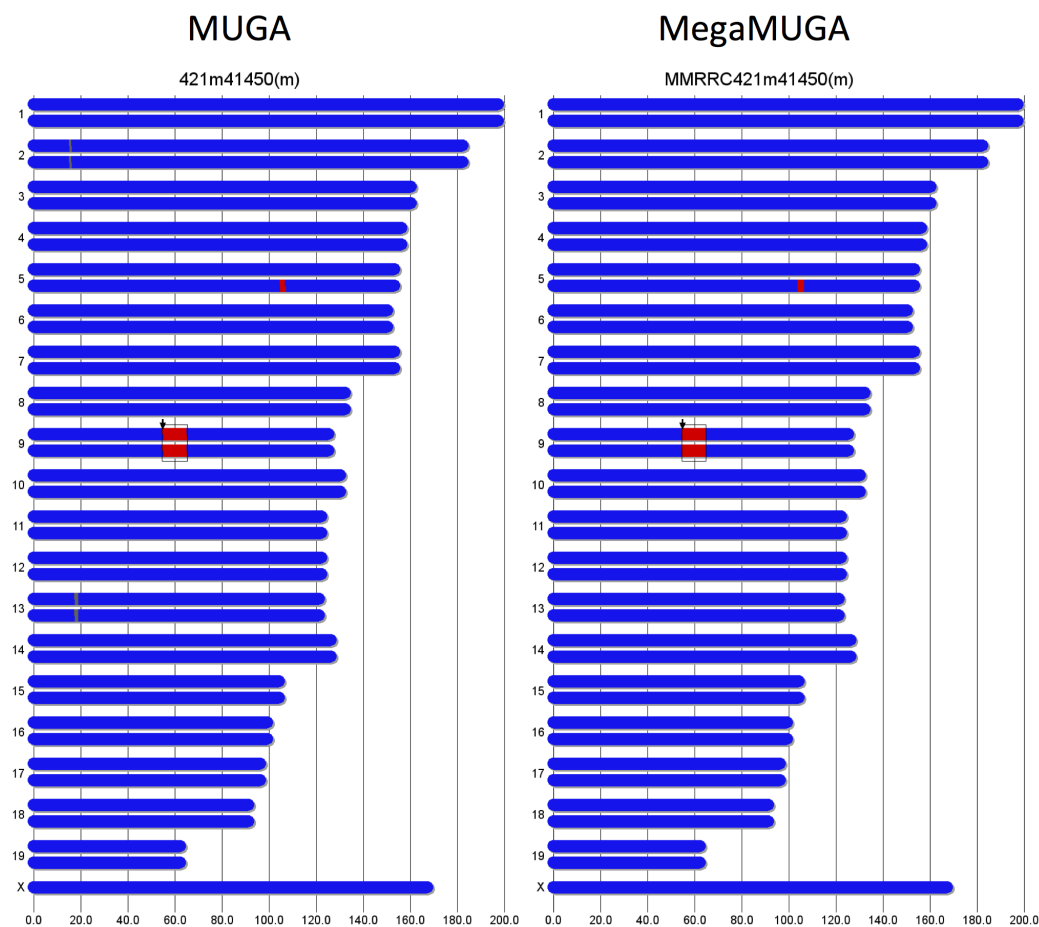


Figure 4.13: The ancestry of a transgenic mouse from the Mutant Mouse Regional Resource Center. This strain is bred on a C57BL/6J (blue) background, with a target mutation on Chromosome 9 at 54.8 Mb (denoted by the arrow) carried by an ES cell line derived from 129S7/SvEvBrd (red). My distance model in both MUGA and MegaMUGA finds the region contributed by 129S7/SvEvBrd (in box), which can be useful in predicting which SNPs are in linkage disequilibrium with the target allele, such as regions surrounding the target gene and the small heterozygous region in Chromosome 5.



Figure 4.14: The ancestry of a transgenic mouse with non-CC background strains from the Mutant Mouse Regional Resource Center. This strain is heterozygous on a BALB/cJ (blue) background, with a target mutation on Chromosome 8 at 8.6 Mb (denoted by the arrow) carried by an ES cell line derived from 129S7/SvEvBrd (red). Even though neither background strain are CC founders, the founder reference clusters are still informative for ancestry inference.

Chapter 5 : Mapping Quantitative Trait Loci

5.1 Introduction

This chapter describes a method for mapping quantitative trait loci (QTL) using hybridization intensities of genotyping microarrays [22]. Quantitative trait loci (QTL) mapping attempts to identify genomic loci that are associated with a given phenotype value within a population. Many methods have been developed over the years, the simplest and most straightforward of which is single-marker analysis of variance (ANOVA), where samples are grouped by their genotype calls ('A', 'B', or 'H') at each marker, and the phenotype values of samples in the two groups are compared for a significant difference. Markers where samples in different genotype groups have significantly different phenotype values are considered possible QTLs. However, until recently, genotyping arrays had few markers, and it was difficult to capture QTLs using ANOVA tests with sparse markers. Lander and Bolstein [37] developed the maximum likelihood-based interval mapping to consider loci with missing genotypes. Although interval mapping gives complete information at all desired loci, it is computationally intensive, and efficient approximations to interval mapping were then developed, such as the regression-based method proposed by Haley and Knott [30]. As medium and high-density arrays have become more affordable, interval mapping has become less of a necessity as markers cover the genome at a much finer scale.

Many packages have been offered for QTL mapping that accommodate different algorithms and different mapping populations. R/qtl [3] by Broman et al. provides a QTL mapping environment with the option of implementing different methods and

processing different types of experimental crosses. QTLRel [6] by Cheng et al. enables modeling of genetic relatedness and the incorporation of covariates. Most packages also provide the option to incorporate hidden Markov Models (HMM) to create genetic maps or ancestry mosaics that correct for missing genotypes and genotyping errors.

Despite the wealth of methods that exist for QTL mapping, the overwhelming majority of methods rely on the assumption that genotype calls represent all relevant genetic information, while in practice the conversion of microarray intensities into genotype calls not only introduces genotyping errors, but oftentimes also results in loss of valuable information, as I showed in Chapter 4.

In this chapter, I propose a novel method for mapping QTLs directly using genotype intensities. There exists one other QTL mapping method that explicitly makes use of genotype intensities, which is DOQTL by Gatti et al. [25]. This method uses genotype intensities to infer local genomic ancestry of all samples in the mapping population. Instead of being converted to genotype calls, the intensities are converted to founder origin probabilities at each marker locus. These intermediaries of founder probabilities are then correlated with the phenotype values. BAGPIPE by Valdar et al [71] provides a similar framework that takes as input founder probabilities calculated from either genotype calls or intensities. The HMM algorithm used in DOQTL is similar to those used in genotype call-based ancestry inference methods such as HAPPY by Mott et al. [49], but the approach is computationally intensive, requires a population with known ancestors and ancestor genotype intensity data, and also defines genotype-phenotype correlation via a derived intermediate result. Our approach eliminates the need for all intermediate steps such as calling genotypes via clustering algorithms, computing a genetic map, or inferring ancestry via HMMs, and instead focuses on the direct relationship between genotype intensity and phenotype values.

Since there is no intuitive way to apply a traditional regression model on two-

dimensional genotype intensities and phenotype values without making assumptions on the distribution and shape of the intensity data, I instead examine the relationship between genotype distances and phenotype distances between samples. Our proposed method is similar to the Mantel test by Nathan Mantel [47], which has long been used in ecological studies to ascertain the correlation between different distance metrics, such the geographical distance versus genetic distance between pairs of different species. I use a similar approach to assess the correlation between genetic distance in terms of Euclidean intensity difference and phenotype value difference between pairs of samples. To reduce error due to a single unreliable marker, I calculate the genetic distance matrices across windows of several markers. I then estimate the significance on a genome-wide level using permutation testing, which is often used for establishing significance levels in QTL mapping [14].

5.2 Methods

I chose to map QTLs directly with genotype intensities in order to preserve information that may be lost or misrepresented in the conversion to discrete genotype calls, as discussed in Chapter 4.

One challenge in mapping phenotype values directly to genotype intensities is the lack of an inherent ordering in two-dimensional intensity values. In traditional QTL mapping, only three classes of genotypes are considered: ‘A’, ‘B’, and ‘H’. In additive QTL models, samples with genotype ‘H’ are assumed to have phenotype values between those of samples with genotypes ‘A’ and ‘B’, and in dominance QTL models, no ordering of the genotype calls is necessary since there are only two possible classes. This makes it straight-forward to find a linear relationship between phenotype values and genotype calls. However, when we consider the intensities without assumptions on the number or distribution of clusters, it is unclear how to directly map a two-dimensional xy-intensity

value to a one-dimensional phenotype value. Intuitively, in the QTL region, samples that have the similar phenotype values should be more genetically similar than samples with different phenotype values. Therefore, instead of searching for a direct mapping of intensity and phenotype values for each sample, I chose to search for regions where genotype intensity distances correlate with the phenotype differences between all pairs of samples. Figures 5.2 and 5.3 demonstrate that pairwise genotype intensity distances vary according to pairwise phenotype differences for the albinism phenotype in two different mouse populations.

5.2.1 Constructing distance matrices

Given n samples and m markers, I construct matrices representing the distance between pairs of samples in both the phenotype space and the genotype intensity space. Our inputs are \mathbf{z} , the length n phenotype vector with quantitative trait values of each sample, and \mathbf{X} , the $n \times 2m$ matrix of xy intensities of each sample at each marker.

The pairwise distance matrix for phenotypes, \mathbf{A} , has $n \times n$ entries, where a_{ij} , the entry in row i column j , is calculated as the difference between the quantitative trait values of sample i and sample j :

$$a_{ij} = |z_i - z_j| \quad (5.1)$$

For binary traits such as albinism, the phenotypes are encoded as 1 or 0, though any two different values could be used if the correlation measure used is scale-invariant.

An $n \times n$ pairwise distance matrix for genotype intensities, \mathbf{B} , is calculated for each sliding window of k markers, where $k \in \{1, 2 \dots m\}$. The sliding window approach here ensures that information about the underlying haplotype is captured. Each entry b_{ij} in matrix \mathbf{B} is the Euclidean distance between the vector of intensity values of sample

i and the vector of intensity values of sample j :

$$b_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2, \quad (5.2)$$

where \mathbf{x}_i is a vector of length $2k$ containing the x and y genotyping intensities of sample i at each marker within the sliding window, and where \mathbf{x}_j is a vector of the same length containing the intensities of sample j . Given m markers examined in sliding windows of size k , the number of \mathbf{B} matrices constructed would be $m - k + 1$.

I used $k = 5$ in subsequent analyses, as 5 was the minimum number of consecutive markers required for the MegaMUGA to fully distinguish the highly diverse CC homozygous and heterozygous founder states, as discussed in Chapter 3. I tested my method on values of k ranging between 1 and 25 and found that the results are highly consistent between different values of k , as shown in Figure 5.1.

5.2.2 Comparing distance matrices and significance

Since Euclidean distance matrices are symmetric and all diagonal entries are zero, I examine only the entries above the diagonal when comparing distance matrices. At each sliding window of k markers, I treat both distance matrices \mathbf{A} and \mathbf{B} as vectors of length $n(n - 1)/2$. I then calculate Pearson's correlation coefficient r between the flattened matrices to use as the QTL mapping statistic. There is no need for normalization of the phenotype values and genotype intensities since Pearson's r is scale invariant, and since r can be negative if there is an inverse relationship, I report $|r|$ to represent the strength of genotype-phenotype correlation.

The entries in distance matrices are non-independent, since $n - 1$ entries have to be changed at once every time an original data point changes. Permutation tests have commonly been used to correct for this non-independence. Most notably, the Mantel test, developed by Nathan Mantel [47] for use in the field of ecology for assessing simi-

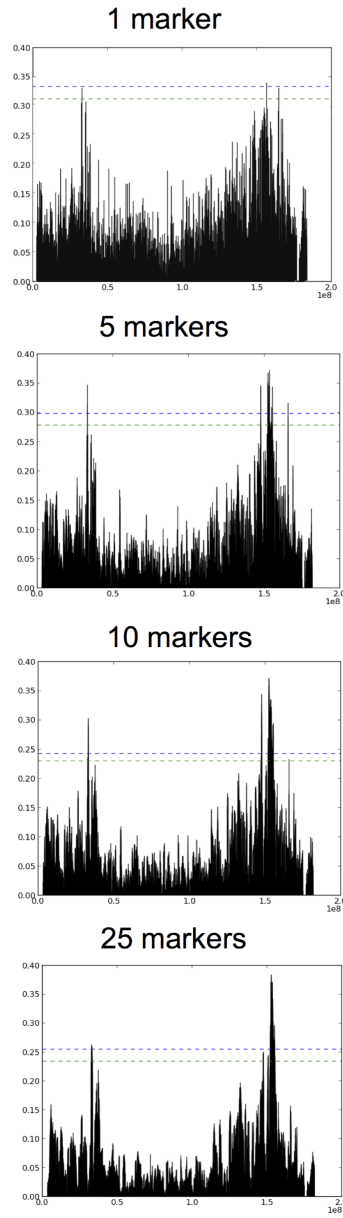


Figure 5.1: Results from simulated data with varying marker window sizes k , showing that the significant peaks remain largely consistent with window size. The $|r|$ metric is on the y-axis and genomic location is on the x-axis. The simulated data used here is the same as the data used in Figure 5.4.

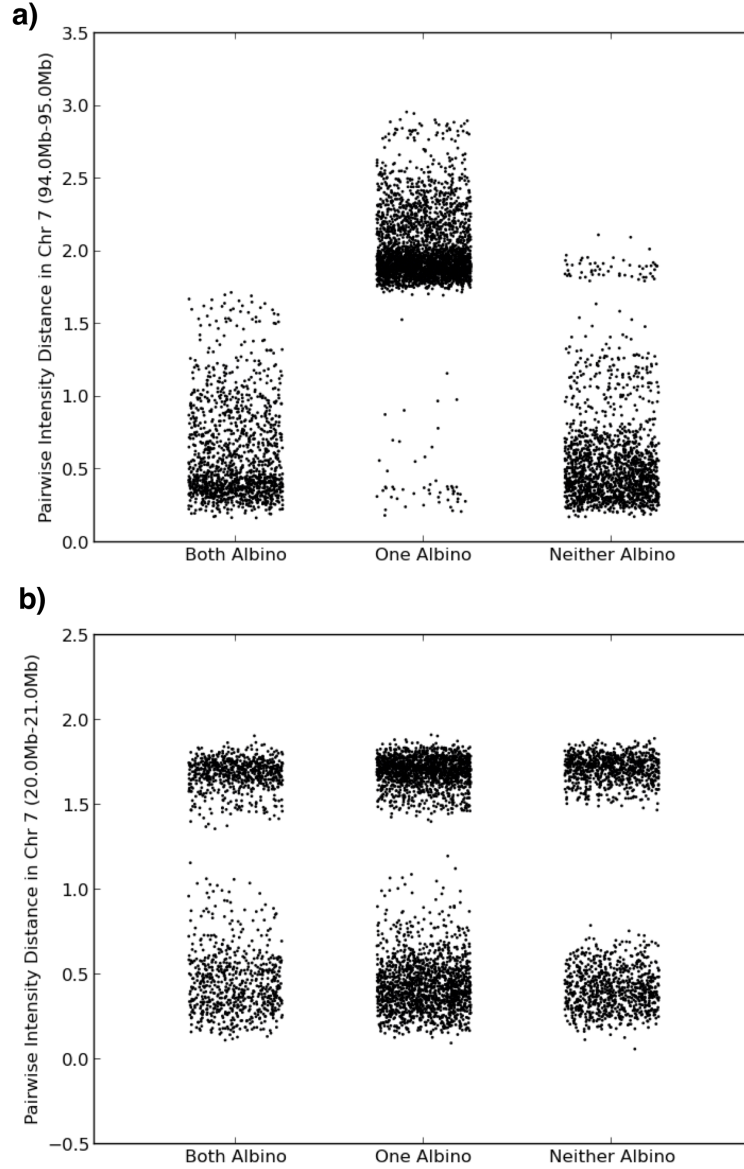


Figure 5.2: a) Intensity distances in the QTL region between pairs of samples grouped by whether the pair shared the same trait for albinism. The intensity vector of each sample includes all markers in Chromosome 7: 94 Mb - 95 Mb, and pairwise distance is calculated as Euclidean distance between intensity vectors of sample pairs. Tyrosinase (*Tyr*), the causal gene, is located on Chromosome 7: 94.58 Mb - 94.64 Mb. The samples used are the 111 MegaMUGA samples used in [55] and presented in the results section. At each locus, all samples are either homozygous CC011/Unc or heterozygous between CC011/Unc and C57BL/6J. b) Intensity distances between sample pairs plotted in Chromosome 7: 20 Mb - 21 Mb, a region of the genome not known to be associated with albinism. Unlike the first plot, the intensity distances have the same distribution in all three groups.

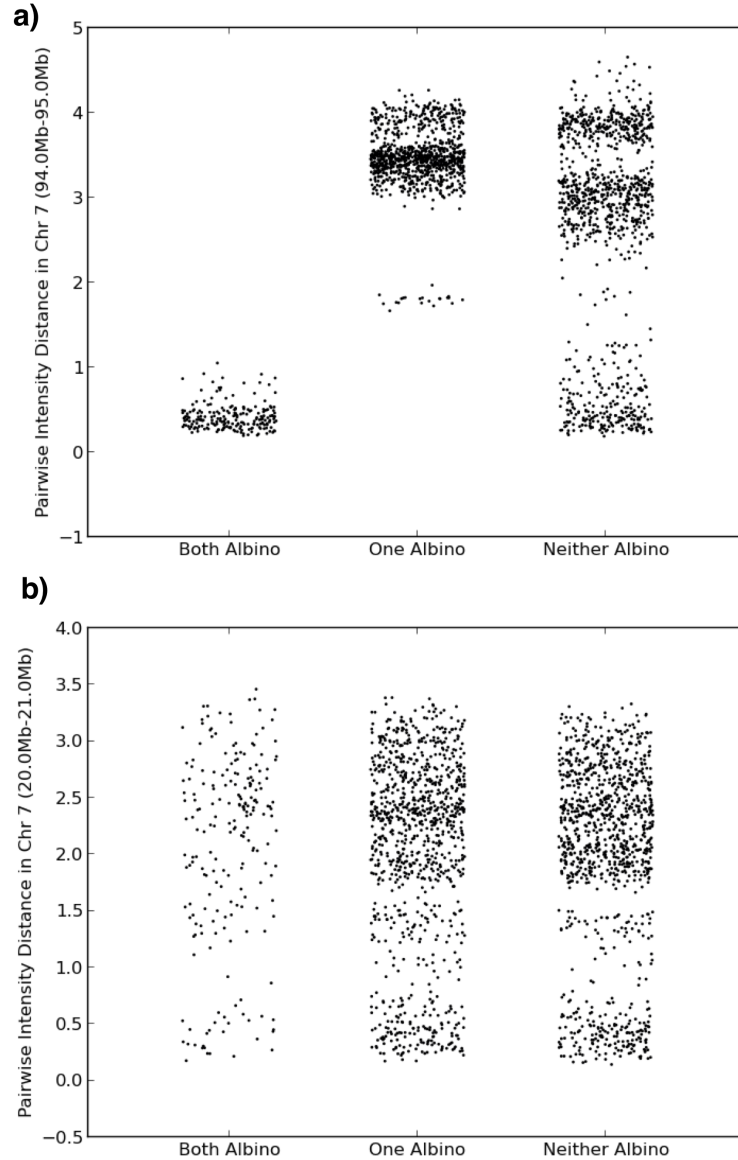


Figure 5.3: a) Intensity distances in the QTL region between pairs of samples grouped by whether the pair shared the same trait for albinism. This is the same as Figure 5.2, except the samples used are 67 samples from CC lines presented in the results section. Unlike the samples used in Figure 5.2, the CC samples have more than two possible haplotypes in each region, resulting in more intensity distance variation in pairs where neither are albino. Although the group where neither are albino has a wide variation in intensity distance, the distribution of intensity distances between the other two groups is still significantly different from that of an unrelated region in the genome. b) Intensity distances between CC sample pairs plotted in Chromosome 7: 20 Mb - 21 Mb, a region of the genome not known to be associated with albinism.

larity between two different distance measures, involves taking Pearson’s r between two distance matrices, permuting the rows and columns of one distance matrix many times, and reporting the significance of the r obtained from the original matrices compared to the permuted matrices.

Since the rows and columns of these matrices are samples, this is equivalent to permuting the phenotype vector \mathbf{z} and recalculating r . For a p-value threshold of 0.05, I perform 1000 permutation tests at each window of markers. For each permutation iteration, I permute the n phenotype values stored in vector \mathbf{z} , and recalculate Pearson’s r with the new \mathbf{A} matrix and original \mathbf{B} matrices across the genome. To assess the significance of genome-wide QTL peaks, for each permutation, I select the maximum $|r|$ value across the entire genome to represent the highest peak achievable by the null hypothesis. I then select the $|r|$ value at the top 5 percentile among all permutations as an estimate for the p-value threshold of 0.05. Any peaks of the $|r|$ values provided by the original \mathbf{z} vector that are over the p-value threshold are then significant and represent putative QTLs.

5.3 Results

I implemented my methods on mouse strains genotyped on the 77,808-marker MegaMUGA. Hybridization intensities were provided by Illumina, with each sample having an xy-intensity at each marker. These intensities are converted into discrete genotype calls of ‘A’, ‘B’, and ‘H’ by Illumina’s GenomeStudio for typical use. For my method, I directly use the xy-intensities as input, and the Illumina genotype calls are ignored.

I tested my method on real data from two different populations. The first mapping population consisted of one sample genotyped on MegaMUGA from each of 67 completed lines from the CC [13], a recombinant inbred population. All 67 lines were phenotyped for coat color, and I chose to map the albino phenotype, with 21 albino

and 46 non-albino lines.

The second mapping population was a subset of samples used in mapping spontaneous colitis by Rogala et al., 2014 [55]. These samples were generated by backcrossing a single male mouse from the CC line CC011/Unc to F1 females which are hybrid between CC011/Unc and C57BL/6J. At each locus, these samples are either homozygous CC011/Unc or heterozygous between CC011/Unc and C57BL/6J. Rogala et al. phenotyped all samples for traits associated with colitis and genotyped 111 backcrossed samples on the MegaMUGA. I selected to use the 111 MegaMUGA samples to map the total colitis score described by Rogala et al. in [55], and I used the same 1059 MegaMUGA markers selected by Rogala et al. that are reliably informative between the strains CC011/Unc and C57BL/6J. In addition, since these 111 samples' coat colors were also phenotyped by Rogala et al., with 53 albino and 58 non-albino samples, I also mapped the albino coat color phenotype using all markers on the MegaMUGA.

5.3.1 Simulated data

Using one sample from each of 54 inbred distributable CC lines as the mapping population, I used two simulated QTLs on mouse chromosome 2 at 145.73 Mb and 30.50 Mb. The majority of variance is explained by the QTL at 145.73 Mb, while the remaining variance can be explained by the QTL on 30.5 Mb. My method detected both QTL at a p-value threshold of 0.01, with significant intervals at 145.80 Mb - 147.40 Mb and 30.65 Mb - 30.71 Mb, and the two highest peaks at 147.39 Mb and 30.69 Mb (Fig. 5.4). Although my method also produced a significant peaks around 137 Mb and 157 Mb, the highest two peaks on the chromosome recovered the main and secondary simulated QTLs.

For comparison, I ran single marker scans using two different methods provided in R/qtl [3]: interval mapping and Haley-Knott regression, as well as a scan using

Table 5.1: QTL positions for colitis phenotype

Chr	Position (Mb) - from [55]	Position (Mb) - my method
12	94.8-112.3 (peak 110.8)	90.1-121.3 (peak 112.9)
14	60.3-94.3 (peak 64.5)	60.3-78.2 (peak 64.5)

BAGPIPE [71], using a matrix of founder probabilities from the forward-backward solution of my hidden Markov model introduced in Chapter 4. I filtered the marker set from MegaMUGA to include only informative markers with no markers containing duplicate information, reducing the set of 5334 markers on chromosome 2 to 2717 markers. Both genotype-based methods had the highest LOD score peaks around 157 Mb, 11 Mb downstream of the simulated QTL. The second highest peak in both genotype-based methods was at 159.83 Mb, and the third highest peak fell within the correct interval of 145.86 Mb - 147.42 Mb. The secondary simulated QTL at 30.50 Mb was not recovered using either method with a p-value threshold of 0.05. BAGPIPE recovered the correct peaks, with large regions of significant LOD scores surrounding the peaks. Although accurate, BAGPIPE’s main drawback is that it requires preprocessing in the form of ancestry inference to obtain founder probabilities.

5.3.2 Real data

Using 67 mouse samples from inbred CC lines, I mapped the Mendelian trait of albinism based on observed coat color. I found a strong peak with $p < 0.01$ to chromosome 7 in the region containing the gene tyrosinase (*Tyr*), which is known to recessively cause albinism when mutated in both humans and mice [36, 1]. The location for *Tyr* in the NCBI Build 37 mouse reference genome is on chromosome 7: 94.58 Mb - 94.64 Mb [12], and the strongest peak found was located in the chromosome 7 region 94.58 Mb - 94.90 Mb. The larger region of 87 Mb -107 Mb on chromosome 7 also had peaks with $p < 0.01$. The results of this scan is shown in Fig 5.5 a).

To evaluate the method on more complex traits, I used 111 samples genotyped on MegaMUGA from a study on spontaneous colitis by Rogala et al [55]. The samples they used were CC011/Unc x C57BL/6J hybrids backcrossed to CC011/Unc. Rogala et al. selected to use 1059 MegaMUGA markers that are maximally informative between CC011/Unc and C57BL/6J, and they defined a quantitative phenotype called “colitis score” based on seven colitis-related phenotypes. They used Haley-Knott regression in R/qtl and found two QTLs with $p < 0.05$, one on chromosome 12: 94.8 Mb - 112.3 Mb (peak at 110.8 Mb) and another on chromosome 14: 60.3 Mb - 94.3 Mb (peak at 64.5 Mb).

Using the colitis score as phenotype and the 1059 MegaMUGA markers from [55], I was able to locate the same two QTLs, with intervals of $p < 0.05$ at chromosome 12: 90.1 Mb - 121.3 Mb (peak at 112.9 Mb) and chromosome 14: 60.3 Mb - 78.2 Mb (peak at 64.5 Mb). The location of these QTLs are summarized in Table 5.1. Rogala et al. also reported on two QTL regions, one on chromosome 1 (3.6 Mb - 197.2 Mb) and another on chromosome 8 (67.2 Mb - 79.8 Mb), with peaks not even reaching the $p = 0.1$ LOD score threshold. My method produced a peak on chromosome 1 at 72.9 Mb with significance over the $p = 0.1$ threshold, and also has an insignificant peak at the distal end of chromosome 8. The full genome scan using this method is shown in Fig 5.6.

The CC population used to map albinism was largely inbred, so I also mapped albinism in the backcross population from Rogala et al. to test my method’s use in more heterozygous populations. Since only two ancestor types were involved in the breeding of the backcrossed samples, their haplotypes were inherited in much larger intervals than those of the 67 CC samples, resulting in larger peak intervals. I obtained a strong peak within chromosome 7: 56.89 Mb - 142.02 Mb that was well over the $p = 0.01$ threshold, with the highest peak around 88.89 Mb - 98.31 Mb, which includes

the (*Tyr*) locus. The results of this scan is shown in Fig 5.5 b).

5.3.3 Efficiency and memory

On a Macbook Air with a single 1.3 GHz processor and 4 GB RAM, my method required 5.5 minutes to run 1000 permutations on a single chromosome with approximately 3000 markers. This was comparable to R/qtl's single marker scan using interval mapping, which required seven minutes to run with the same data and machine. The number of pairwise distances calculated scales as $O(n^2)$, where n is the number of samples within the mapping population, so my method has time complexity $O(mn^2)$, where m is the number of markers.

5.4 Discussion

In this chapter, I introduced a novel intensity-based QTL mapping method that is straightforward and does not require intermediate processing such as genotype calling, clustering, or ancestry inference. I tested my method on simulated and real data and found that it works well on Mendelian and non-Mendelian traits, as well as different mapping populations. Unlike traditional methods that are constrained to modeling only additive effects (where heterozygous samples have phenotype values between the two homozygous phenotype values) or only dominance effects (where heterozygous samples have similar phenotype values as one of the two homozygous phenotype values), my method can detect QTLs with either additive or dominance effects since I do not consider how to order genotype values to reflect the ordering of phenotype values. In addition, my method can detect QTLs with atypical ordering of phenotype values, such as in the case of overdominance, where heterozygous samples have a high phenotype value and the two homozygous groups have low phenotype values. The pairwise distance-based measure used here inherently considers all possible models that tradi-

tionally would require running different regression models to evaluate.

Another advantage of this method is the incorporation of information on copy number variations, deletions, and other off-target variations that manifest as atypical genotype intensity patterns. In traditional QTL mapping, these unexpected genetic variations are either not detected, converted to ‘N’ calls and ignored, or assigned incorrect genotypes. Even though I only tested my method on samples genotyped on the MegaMUGA, I have observed the phenomenon of off-target variations causing atypical intensity patterns in several different microarrays [20, 24] on Affymetrix and Illumina platforms, which suggests my intensity-based method is applicable to different microarrays on different platforms.

I have observed highly consistent intensity clustering of technical and biological replicates across different batches in all our microarrays. Nevertheless, I calculate the intensity distance matrices in sliding windows of five markers, partly to incorporate information about the haplotype, and partly so that in the rare event of a single outlier marker with unreliable or contaminated intensities, the solution would be resistant to the outlier marker.

The proposed method represents a new way to approach the use of microarray data for QTL mapping, with ample room for extension and exploration. For instance, the method is easily extensible to two-locus scans for detecting QTLs that act in pairs by adding the first QTL from a preliminary scan to the marker windows considered in the secondary scan. I believe the incorporation of covariates such as sex, environmental factors, or population structure is also possible using a variation of the partial Mantel test [62].

I obtained results that are comparable to those obtained from traditional methods on real data, and my results compare favorably to traditional QTL mapping methods in recovering secondary QTLs from simulated data. Many tasks that traditionally

take genotype calls as input can be modified to use genotype intensities, and that in doing so, error and uncertainty due to the grouping of continuous intensities into a predetermined small number of discrete genotype calls can be avoided. I showed in the previous chapter that my intensity-based method for ancestry inference captures more information and is less prone to error than genotype-based methods [24]. The use of intensity-based methods for QTL mapping also eliminates the need for intermediate steps, such as genotype imputation or ancestry inference, that are necessary for handling missing genotypes due to ‘N’ calls or modeling the possibility of genotype errors. The use of microarray intensity data offers the promise of decreasing computational costs and leads to higher accuracy and precision, and the applications of using genotype intensities instead of discrete genotype calls offer broad areas for future research.

This chapter, along with the Chapters 3 and 4, demonstrate the use of hybridization intensities, a source of information not usually included in existing SNP microarray analyses, to maximize information context about ancestral haplotypes. This concept of using ancestral haplotype data that is inherently available in existing technologies is explored further in Chapter 6, which discusses the estimation of allele-specific expression using RNA-seq reads from parental strains as a template for their F1 offspring’s RNA-seq reads.

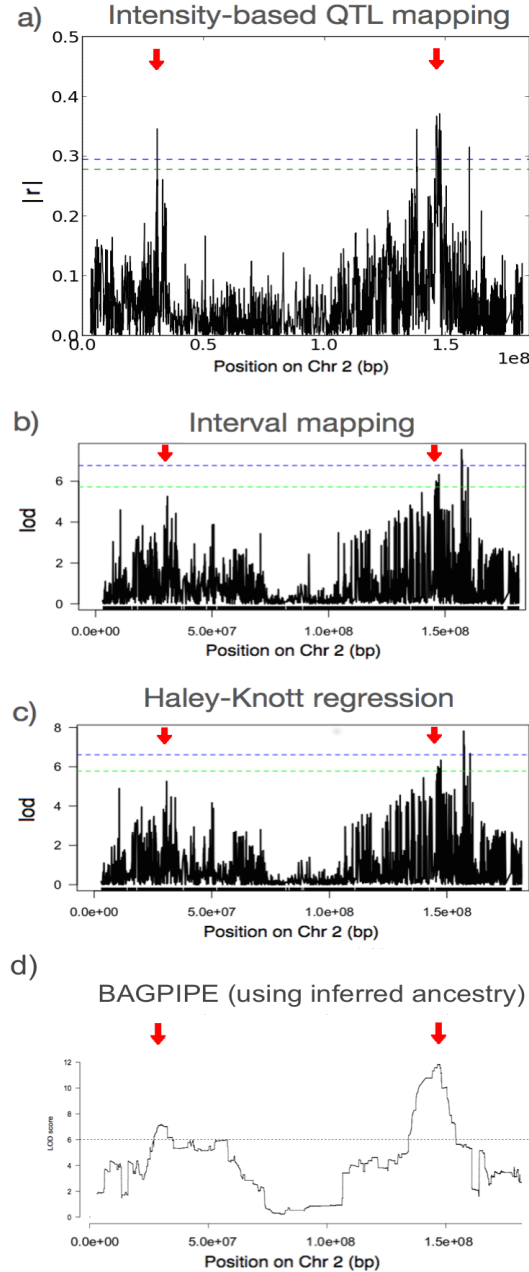


Figure 5.4: Results on simulated QTLs created in 54 largely inbred CC samples. The two QTLs were selected to be on mouse chromosome 2 at 145.73 Mb and 30.50 Mb. The blue line is $p=0.01$ and the green line is $p=0.05$. **a)** My intensity-based method recovered the two QTLs (in red arrows) in the two highest peaks at 147.39 Mb and 30.69 Mb. **b)** Interval mapping in R/qtl produced the highest peak at 157.39 Mb, with a significant peak in the correct interval of 145.86 Mb - 147.42 Mb. The secondary QTL at 30.50 Mb was not recovered. **c)** Haley-Knott regression in R/qtl produced the highest peak at 157.12 Mb, with a significant peak in the correct interval of 145.86 Mb - 147.42 Mb. The secondary QTL at 30.50 Mb was not recovered. **d)** BAGPIPE [71], which uses the founder probabilities from the forward-backward algorithm discussed in Chapter 4. The highest peaks are in the correct intervals, yet the region of significance is large due to haplotype blocks within the inferred ancestry. The dotted line here is $p=0.05$

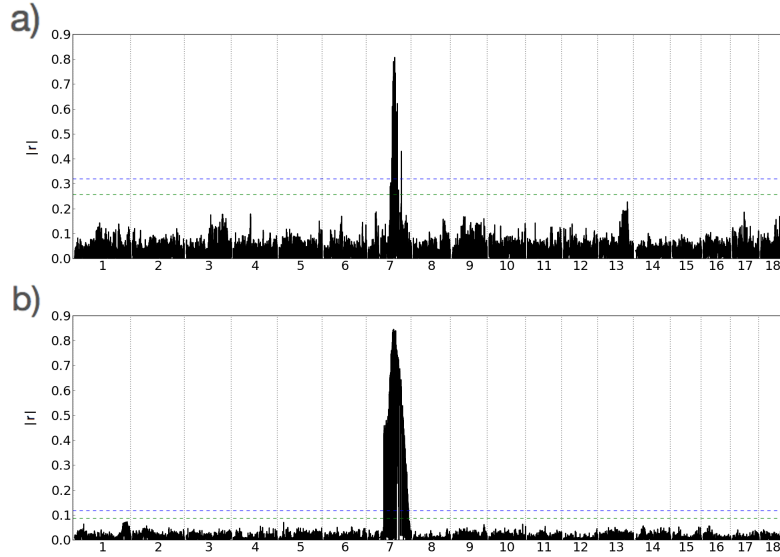


Figure 5.5: Full genome scans for QTLs affecting the albinism trait. Blue dotted lines are $p=0.01$ and green dotted lines are $p=0.05$. The known causal gene tyrosinase (*Tyr*) is located on chromosome 7: 94.58 Mb - 94.64 Mb in NCBI Build 37 of the mouse genome. I performed two separate scans using two different populations: a) 67 CC samples which were all from different inbred lines. The strongest peak is on chromosome 7: 94.58 Mb - 94.90 Mb. b) 111 backcrossed samples used in [55], generated with the crosses (CC011/Unc x C57BL/6J) x C57BL/6J. The strongest peak is on chromosome 7: 88.89 Mb - 98.31 Mb. Note the longer significant QTL interval in b), which is due to a longer non-recombinant interval within the backcrosses, which underwent fewer generations of crossing.

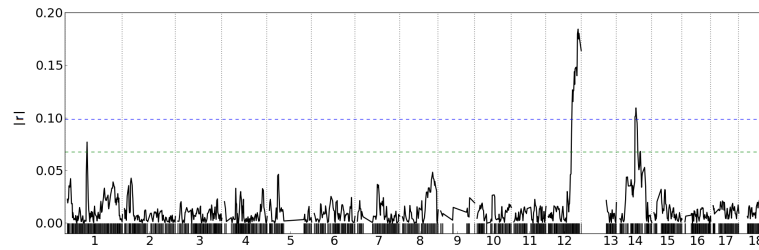


Figure 5.6: My method applied to the samples and phenotype values from Rogala et al. [55]. The phenotype mapped here is “total colitis score” as defined in [55]. The two significant peaks I found matched the two significant peaks reported by Rogala et al. in Figure 6a of their paper using R/qtl. The exact positions of these peaks are reported in Table 5.1

Chapter 6 : Estimating Allele-Specific Expression using RNA-Seq Data

6.1 Introduction

While the previous chapters discussed methods for exploiting ancestor haplotype information from genotyping microarrays by using probe intensities, similar approaches can also be applied to the reads of high-throughput sequencing data from both ancestors and admixed animals. In this chapter, I describe methods for assessing the contribution of maternal versus paternal alleles in an F1 sample using short sequencing read data from RNA-seq [23].

Recent advances in high-throughput RNA-seq technology have enabled the generation of massive amounts of data for investigation of the transcriptome. While this offers exciting potential for studying known gene transcripts and discovering new ones, it also necessitates new bioinformatic tools that can efficiently and accurately analyze such data.

Current RNA-seq techniques generate short reads from RNA sequences at high coverage, and the main challenge in RNA-seq analysis lies in reconstructing transcripts and estimating their relative abundances from millions of short (35-250 bp) read sequences. A common approach is to first map short reads onto a reference genome, and then estimate the abundance in each annotated gene region. Such reference-alignment methods include TopHat [68], Cufflinks [70] and Scripture [29], which use algorithms such as the Burrows-Wheeler transform [4] to achieve fast read alignment. These methods are well established in the RNA-seq community and there exist many auxiliary tools [68] [69] for downstream analysis.

However, aligning reads to a reference genome has some disadvantages. First, read alignment assumes samples are genetically similar to a reference genome, and as a result, samples that deviate significantly from this reference frequently have a large portion of unmapped reads. This is particularly a problem when samples are a mix of founder genomes that individually and regionally deviate from the reference. This leads to what is known as “reference bias,” which is a bias that impacts abundance estimates and favors mapping reads from samples more similar to the reference genome. Second, alignment methods typically cannot resolve the origin of reads that map to multiple locations in the genome, resulting in reads being arbitrarily mapped or discarded from analysis. Suggested workarounds to the first problem of reference bias involve creating new genome sequences, typically by incorporating known variants, to use in place of the reference genome for read alignment [57, 32]. However, this requires prior knowledge of genomic variants in the targeted RNA-seq sample, which is sometimes difficult and expensive to obtain, since it generally requires additional DNA sequencing of all founder genomes that potentially contributed to the sample.

Another class of methods perform *de novo* assembly of transcriptomes using *De Bruijn* graphs of k-mers from reads [26] [54]. These methods enable reconstruction of the transcriptome in species for which no reference genomic sequence is available. While these methods offer the possibility of novel transcript discovery, their *de novo* nature makes it difficult to map assembled subsequences back to known annotated transcripts. Furthermore, estimation of transcript expression levels in these methods is not straightforward and generally involves alignment of assembled contigs to a reference genome [26] [54], which reintroduces the possibility for reference bias.

Expression level estimation is particularly difficult for outbred diploid organisms, since each expressed transcript may contain two different alleles, one from each parental haplotype.

Allele-specific expression (ASE) is an approach to determine the contribution of each parent or ancestral strain. In some transcripts, one allele is preferentially expressed over another, resulting in what is known as allelic imbalance. It is often biologically interesting to identify genes and transcripts exhibiting allelic imbalance through ASE, as well as estimate the relative expression levels of the maternal and paternal alleles [27] [74]. Prior to the introduction of RNA-seq, ASE studies often relied on microarray technology. Although microarrays are able to identify genes exhibiting ASE, they generally examine a small number of genes, with expression level estimates in highly relative terms [45] [56]. The abundance of data from RNA-seq not only enables large-scale ASE studies incorporating the entire transcriptome, but it also provides an estimate of overall gene-expression levels.

Current RNA-seq-based methods for analyzing ASE rely on reference transcriptome alignment [57] [61], requiring prior knowledge of genomic variants between the strains of interest, which is again subject to reference bias. Reference bias is particularly problematic in ASE analysis, since it can falsely enhance relative expression in one parental strain over another.

In the case where RNA-seq data of all three members of a mother-father-child trio are available, we can utilize the RNA-seq data from the parental strains and eliminate the need for prior knowledge of their genomic variants. In this chapter, I examine ASE in F1 mouse hybrids, which are first-generation offspring of two distinct homozygous parental strains. I separately construct maternal and paternal versions of transcripts using RNA-seq reads from the parental strains and annotated reference transcripts, creating a set of candidate transcript sequences the F1 hybrid could express. I then estimate the expression level of each candidate transcript in the F1 hybrid using a modified lasso regression model [31]. Lasso regression has been proposed by Li et al. [40] in the context of RNA-seq isoform expression level estimation, but not in the context

of estimating ASE without reference alignment. I choose to use lasso regularization since it prefers a sparse solution for parameter settings (i.e. it drives most parameters to zero). This models the expectation that only a small subset of known transcripts are actually expressed in a given tissue sample. I modify the lasso penalty slightly to prefer assigning higher expression levels of the F1 child's transcripts with subsequences that appear frequently in its parents' RNA-seq reads, thus assuming that most highly expressed genes in the parents should also be highly expressed in the child.

I tested these methods on synthetic RNA-seq data from the wild-derived mouse strains CAST/EiJ and PWK/PhJ, along with F1 offspring CASTxPWK, with CAST/EiJ as the maternal strain and PWK/PhJ as the paternal strain. I also tested on real RNA-seq data from a CAST/EiJ, PWK/PhJ, CASTxPWK trio and a CAST/EiJ, WSB/EiJ, CASTxWSB trio, both using CAST/EiJ as the maternal strain. The CAST/EiJ, PWK/PhJ, and WSB/EiJ mouse strains are isogenic, and all three have well-annotated genomes [35] that differ significantly from each other and from the mouse reference sequence [81], which is largely based on the C57BL/6J strain (NCBI37 [12]). CAST/EiJ and PWK/PhJ each have a high variation rate of approximately one variant per 130 bp with respect to the reference genome, and a slightly higher rate with respect to each other, while WSB/EiJ is more similar to the reference genome with approximately one variant per 375 bp. The genetic distance between these three strains make them ideal candidates for studying ASE, since we expect a large percentage of reads to contain distinguishing variants.

6.2 Approach

In this section, we discuss the parameters and assumptions of our proposed model and the underlying optimization problem.

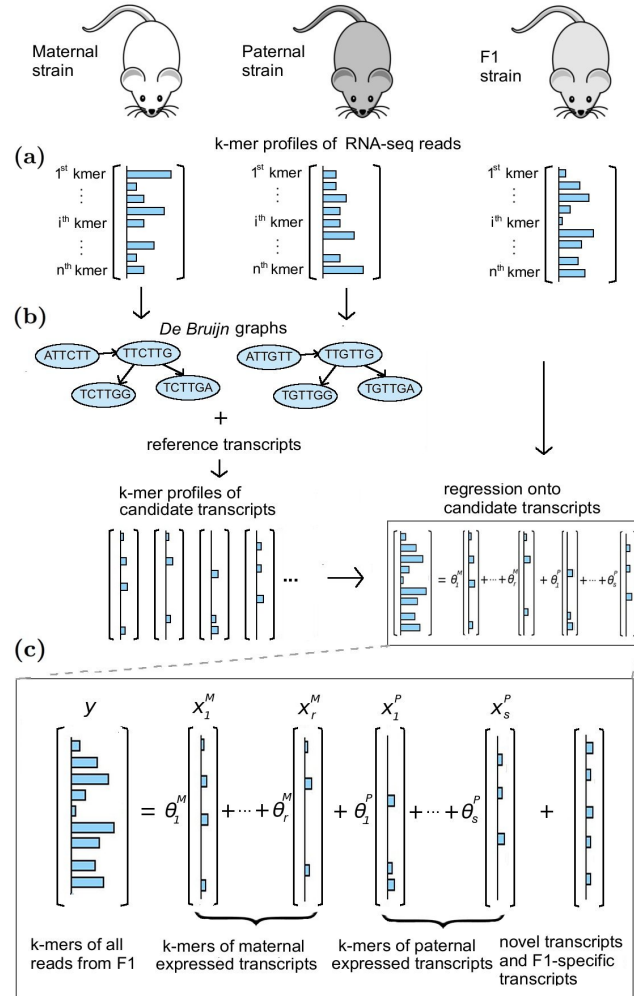


Figure 6.1: Our pipeline for estimating allele-specific expression in F1 animals. **(a)** k-mer profiles are created for the maternal, paternal, and F1 strains, using all available RNA-seq reads from one sample of each strain. Each k-mer is also saved as its reverse complement, since we do not know the directionality of the read. **(b)** *De Bruijn* graphs are created for the maternal and paternal samples. Using annotated reference transcripts and the parental *De Bruijn* graphs, we select candidate transcripts which incorporate parental alleles from the *De Bruijn* graphs. **(c)** The k-mer profile of the F1 sample, y , is then regressed onto the candidate parental transcripts, $\{x_1^M, x_2^M, \dots, x_r^M\} \cup \{x_1^P, x_2^P, \dots, x_s^P\}$, and we estimate the expression level θ of each candidate transcript.

Table 6.1: Notation

\mathbf{y}	F1 k-mer profile. An $n \times 1$ vector where y_i indicates the number of times the i^{th} k-mer appears in the F1 sample
$\mathbf{z}^M, \mathbf{z}^P$	maternal and paternal k-mer profiles
\mathbf{X}^M	set of k-mer profiles of candidate transcripts from \mathbf{z}^M
\mathbf{X}^P	set of k-mer profiles of candidate transcripts from \mathbf{z}^P
\mathbf{X}	an $n \times m$ matrix equal to $[\mathbf{X}^M \cup \mathbf{X}^P]$, where n is number of k-mers and m is number of transcripts
\mathbf{x}_j	k-mer profile of the j^{th} candidate transcript
$x_{i,j}$	number of times the i^{th} k-mer occurs in the j^{th} candidate transcript
θ_j	estimated expression level for the j^{th} candidate transcript

6.2.1 Notation

Table (6.1) includes a description of the variables used in this paper. The primary genomic feature used in my analysis is the frequency of k-length substrings, called k-mers, from a given set of reads. Thus, each read of length n contributes counts for $n - k + 1$ k-mers. We denote the k-mer profiles of maternal candidate transcripts, $\mathbf{X}^M = \{\mathbf{x}_1^M, \mathbf{x}_2^M, \dots, \mathbf{x}_r^M\}$, and the k-mer profiles of paternal candidate transcripts, $\mathbf{X}^P = \{\mathbf{x}_1^P, \mathbf{x}_2^P, \dots, \mathbf{x}_s^P\}$, jointly as $\mathbf{X} = \mathbf{X}^M \cup \mathbf{X}^P$, a matrix representing the k-mer profiles of all candidate transcripts. Each candidate transcript k-mer profile is labeled as originating from the maternal k-mer profile, the paternal k-mer profile, or both if there are no differentiating variants between the parental k-mer profiles.

6.2.2 Regression model

We propose a modified lasso penalized regression model for estimating the abundance of each candidate transcript, with the assumption that the F1's k-mer profile \mathbf{y} can be expressed as a linear combination of its expressed transcripts $\mathbf{X} =$

$\{\mathbf{x}_1, \dots, \mathbf{x}_j, \dots, \mathbf{x}_m\}$ multiplied by their relative expression levels θ_j :

$$\mathbf{y} = \sum_{j=1}^m \theta_j \mathbf{x}_j. \quad (6.1)$$

To filter out non-expressed transcripts and prevent overfitting, each candidate transcript is penalized by an l_1 -norm, parameterized by the regularization parameter λ and the inverse of w_j , where

$$w_j = \text{median} \begin{cases} \{z_i^M/x_{i,j}, \forall x_{i,j} > 0\}, & \mathbf{x}_j \in \mathbf{X}^M \\ \{z_i^P/x_{i,j}, \forall x_{i,j} > 0\}, & \mathbf{x}_j \in \mathbf{X}^P \\ \{(z_i^M + z_i^P)/x_{i,j}, \forall x_{i,j} > 0\}, & \mathbf{x}_j \in \mathbf{X}^P \cap \mathbf{X}^M \end{cases} \quad (6.2)$$

Therefore, transcripts that are expressed at a high level in the parental samples are subject to a lesser penalty than those that are expressed at lower levels or not seen at all in the F1's parents. Our objective function then becomes

$$\begin{aligned} \underset{\theta}{\operatorname{argmin}} \quad & \frac{1}{2} \sum_{i=1}^n (y_i - \sum_{j=1}^m \theta_j x_{i,j})^2 + \lambda \sum_{j=1}^m \frac{\theta_j}{w_j} \\ \text{subject to} \quad & \theta_j \geq 0, \forall j, \end{aligned} \quad (6.3)$$

with each θ_j constrained to be nonnegative since they represent transcript expression levels.

6.3 Methods

6.3.1 Simulated data

I used the Flux Simulator [28] to create simulated reads from the CAST/EiJ, PWK/PhJ, and CASTxPWK mouse genomes. I chose these two parental strains

because they are well-annotated strains that differ significantly from the reference strain C57BL/6J and from each other. The transcript sequences for CAST/EiJ and PWK/PhJ were created using Cufflinks' gffread utility [70] with genomes from the Sanger Institute [35] and transcript annotation from the Ensembl Genome Database [10]. The positions from the reference transcript annotation files were updated with positions to the CAST/EiJ and PWK/PhJ genomes using MODtools [33].

I simulated 10,000,000 100bp paired-end reads from both the CAST/EiJ and the PWK/PhJ genomes to represent reads from a maternal CAST/EiJ genome and a paternal PWK/PhJ genome. I specified the same set of 1000 transcripts with a positive number of expressed RNA molecules in both genomes. In addition, I merged two sets of 5,000,000 separately simulated reads from both CAST/EiJ and PWK/PhJ to create a simulated F1 fastq file. From the merged CAST/EiJ and PWK/PhJ versions of transcript sequences, the Flux Simulator output 1156 unique transcripts sequences where at least 95% of the sequence is covered by reads, and I define this set of 1156 transcript sequences, representing 626 reference transcripts, as the truly expressed transcripts.

6.3.2 Real data

RNA from whole-brain tissues (excluding cerebellum) was extracted from 5 samples (CAST/EiJ female, PWK/PhJ male, WSB/EiJ male, CASTxPWK male and CASTxWSB female) using the Illumina TruSeq RNA Sample Preparation Kit v2. The barcoded cDNA from each sample was multiplexed across four lanes and sequenced on an Illumina HiSeq 2000 to generate 100 bp paired-end reads (2x100). This resulted in $2 \times 71,291,857$ reads for the CAST/EiJ sample, $2 \times 49,877,124$ reads for the PWK/PhJ sample, $2 \times 62,712,206$ reads for the WSB/EiJ sample, $2 \times 77,773,220$ reads for the CASTxPWK hybrid sample, and $2 \times 57,386,133$ reads for the CASTxWSB hybrid sample. Note that the selected samples were not true biological trios, but genetically

equivalent. I also used the same female CAST/EiJ sample as the maternal model for both F1 hybrids.

6.3.3 Selecting candidate transcripts

I used a greedy approach for selecting candidate transcript sequences from the *De Bruijn* graphs of each parental k-mer profile. The k-mer size used for this and subsequent analyses was 32 bp. For each of the 93,006 reference transcripts provided by Ensembl [10], I match the reference transcript sequence to a path of k-mers in the *De Bruijn* graph, allowing for a maximum number of 5 mismatches within a sliding window of 25 bp, which is a sensible choice except in the case of unusually dense SNPs or indels. In the case of mismatches, I replace the reference sequence with the sequence in the parental *De Bruijn* graph, thus creating updated candidate transcript sequences which reflect variants in the parental strains. When more than 5 mismatches occur in 25 bp, I continue along the transcript until another 25 bp subsequence is found in the graph. If more than 80% of a transcript's k-mers are found in the *De Bruijn* graph, I consider it a candidate transcript. The k-mer profiles of the selected candidate transcript sequences are then used as features in this regularized regression model.

6.3.4 Coordinate descent

To optimize the objective function Eq. (6.3), I update θ_j using coordinate descent:

$$\theta_j = \frac{\max(\sum_{i=1}^n y_i^{(-j)} x_{i,j} - \frac{\lambda}{w_j}, 0)}{\|x_j\|_2^2}, \text{ where} \quad (6.4)$$

$$y_i^{(-j)} = y_i - \sum_{k \neq j} \theta_k x_{i,k}.$$

Due to the high dimensional nature of the data (in real data, the number of k-mers, n , is approximately 5×10^7 , and the number of candidate transcripts, m , is

approximately 2×10^4), updating each θ_j on every iteration becomes inefficient. I therefore adapt the coordinate descent with a refined sweep algorithm as described by Li and Osher [41], where I greedily select to update only the θ_j that changes the most on every iteration. To save on computation per iteration, we can let $\beta_j = \sum_{i=1}^n y_i^{(-j)} x_{i,j}$ and precompute the matrix product $\mathbf{X}^T \mathbf{y}$, so that β can be updated at every iteration using only addition and a scalar-vector multiplication. The algorithm is described in Eq. (6.5), and proof of its convergence is provided by Li and Osher [41].

Initialize:

$$\begin{aligned}\theta^0 &= \mathbf{0} \\ \beta^0 &= \mathbf{X}^T \mathbf{y} \\ \gamma &= \text{diag}(\|\mathbf{x}_j\|_2^2) - \mathbf{X}^T \mathbf{X}\end{aligned}$$

Iterate until convergence:

$$\begin{aligned}\theta^* &= \frac{\max(\beta - \frac{\lambda}{\mathbf{w}}, 0)}{\|\mathbf{x}_j\|_2^2} \\ j &= \text{argmax}|\theta^* - \theta^k|\end{aligned}\tag{6.5}$$

Updates:

$$\begin{aligned}\theta_j^{k+1} &= \theta_j^* \\ \beta^{k+1} &= \beta^k + \gamma_{j,:}(\theta_j^* - \theta_j^k) \\ \beta_j^{k+1} &= \beta_j^k\end{aligned}$$

The coordinate descent algorithm terminates when the minimization objective Eq. (6.3) decreases by less than a threshold of 0.001 per iteration. For computational efficiency, the value of the objective function Eq. (6.3) is evaluated per τ iterations, where $\tau = 10^4$ initially. I decrease τ as the objective increases, until $\tau = 1$ for the final iterations.

This saves significant computation time since the computation of the objective function contains a matrix multiplication and the regular updates do not, and the convergence of the algorithm is not affected as the updates are still being performed per iteration.

The lasso regularization parameter λ is chosen via 4-fold cross validation. It is important to note that the value of λ depends on the mean observed values for w_j , so different values of λ could be chosen for each trio.

6.4 Results

I analyzed a synthetic data set to ascertain the sensitivity and specificity of my estimation framework. I then applied my technique to two real data sets and evaluated them based on their ability to recapitulate known biological properties.

6.4.1 Synthetic data results

In the synthetic F1 sample, the Flux Simulator generated 1156 unique transcript sequences from both the maternal and paternal haplotypes with positive expression levels, representing 626 reference transcripts. I identified 4517 candidate parental transcript sequences from all reference mouse transcripts annotated by Ensembl, 1055 of which were truly expressed, representing 598 out of 626 truly expressed reference transcripts.

I selected the lasso regularization parameter λ to be 500 using 4-fold cross validation. I took $\theta_j = 0$ to indicate transcript j was not expressed and calculated the sensitivity and specificity of my method in identifying which transcripts were expressed. For the chosen value of λ , I found the sensitivity to be 0.9553 (598/626) and the specificity to be 0.9880 (91278/92385).

Of the correctly identified expressed transcripts, the true and estimated expression levels had a Pearson correlation coefficient of 0.85, indicating high positive correlation, as shown in Fig. (6.2). To allow for comparison of relative expression levels, I normalized

both true and predicted expression levels to have a mean value of 1 across all expressed transcripts. The mean absolute error between true and predicted expression levels was 0.3128 for the chosen value of λ . True positive rates, false positive rates, and mean absolute error of predicted expression levels for different values of λ are summarized in Fig. (6.3).

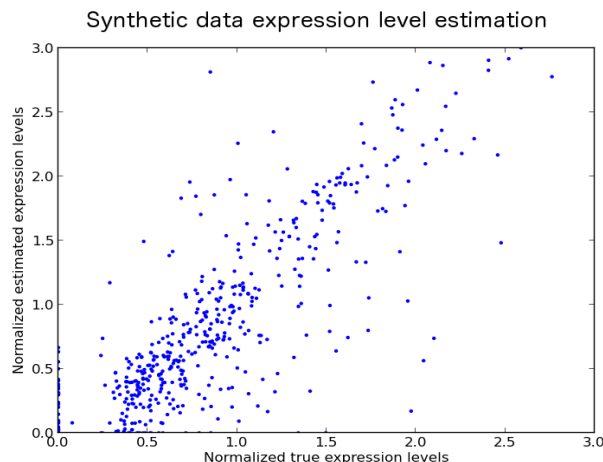


Figure 6.2: Predicted versus actual expression levels from synthetic data, with $\lambda = 500$. Expression levels were normalized to have a mean value of 1. The Pearson correlation coefficient is 0.85 among the 1055 correctly identified expressed transcript sequences.

Among the 598 correctly identified expressed transcripts, 544 had differentiable paternal and maternal candidate sequences. Of these, 141 exhibited ASE, as defined by having a maternal contribution ratio (maternal expression level divided by total expression level) outside the range $[0.4, 0.6]$. My model correctly identified 109 transcripts exhibiting ASE and correctly rejected 293 transcripts not exhibiting ASE, achieving a sensitivity of 0.77 and specificity of 0.73.

I compared my results with Trinity [26], since its *de novo* assembly methods are able to separate maternal and paternal versions of transcripts better than reference alignment-based methods.

To assemble candidate transcripts from the maternal and paternal strains, I ran Trinity with its default parameters on the synthetic maternal CAST/EiJ and paternal

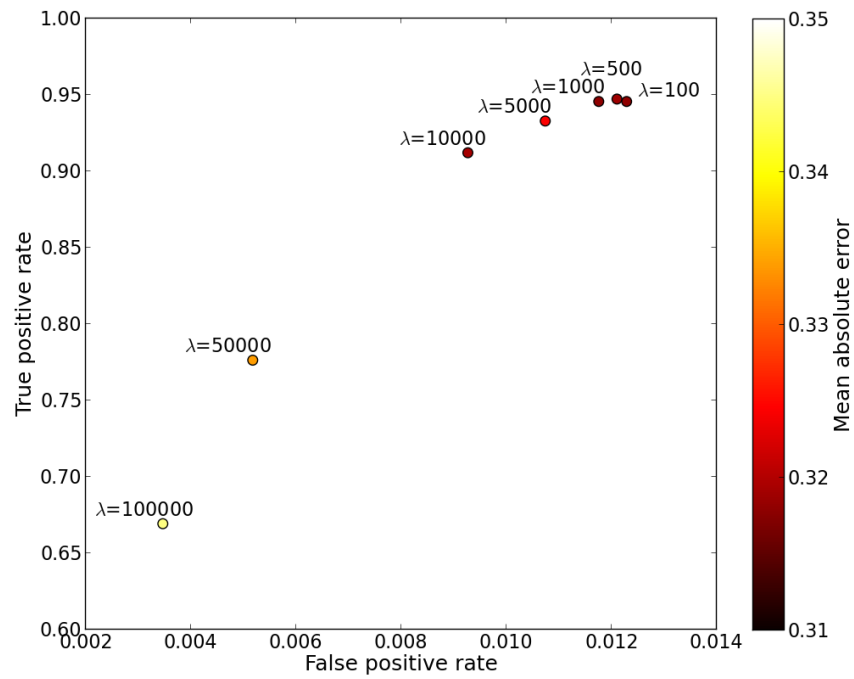


Figure 6.3: True positive rate vs. false positive rate for different values of λ . Each point is colored by the mean absolute error between normalized true and estimated expression levels for all transcripts correctly classified as expressed.

PWK/PhJ samples. Per Trinity’s downstream analysis guidelines, I then aligned reads from the synthetic F1 sample to the assembled parental transcript sequences using Bowtie [38] then estimated expression levels using RSEM [39].

Trinity assembled 4215 transcript sequences from both parental strains. Following their guidelines to eliminate false positives, I retained 3336 transcript sequences representing at least 1% of the per-component expression level. I used a criterion of Levenshtein distance less than 10% of the true transcript length to match annotated transcripts to the *de novo* transcripts sequences reported by Trinity. With this criterion, only 110 out of 626 truly expressed transcripts were present in the set of expressed transcripts found by Trinity. In this set, the mean Levenshtein distance from each true transcript sequence to the Trinity sequences was 0.12% of the true transcript length, with the maximum distance being 2.6% of the true transcript length, suggesting the matching criterion of 10% Levenshtein distance was generous.

Out of the 110 assembled transcripts correctly identified, 81 had nonzero expression levels, making the sensitivity for baseline expression detection 0.13. However, of the 81 correctly identified transcripts, the Trinity-Bowtie-RSEM pipeline produced a high correlation of 0.88 between true and estimated expression levels.

Of the 81 expressed transcripts correctly identified by Trinity, 63 originated from reference transcripts with ASE. Trinity correctly identified 20 true positives and 16 true negatives, with a sensitivity of 0.32 and specificity of 0.89.

6.4.2 Real data results

I applied the methods to a male CASTxPWK F1 sample and a female CASTxWSB F1 sample. I first created *De Bruijn* graphs for a CAST/EiJ female, a PWK/EiJ male, and a WSB/EiJ male, representing the parental *De Bruijn* graphs of the two F1 samples. To eliminate erroneous reads in each strain, I filtered k-mers appearing fewer than five times. Using Algorithm 2, I selected 15,287 candidate transcripts from the

Table 6.2: Dimensions and Results from Real Data

	CASTxPWK	CASTxWSB
k-mers in merged trio k-mer profile	118,100,824	118,383,117
k-mers in candidate transcripts	42,688,910	52,715,089
k-mers in estimated expressed transcripts	42,482,315	52,162,586
candidate transcripts	23,585	29,155
estimated expressed transcripts	17,118	20,596
candidate genes	7,393	8,532
estimated expressed genes	7,148	8,242
expressed genes with isoforms from both parents	4,065	5,183

CAST/EiJ *De Bruijn* graph, 9,852 candidate transcripts from the PWK/EiJ graph, and 16,023 candidate transcripts from the WSB/EiJ graph. For each F1 sample, transcript sequences without differentiating variants between the two parental strains were merged into a single candidate transcript. This resulted in 23,585 candidate transcripts for CASTxPWK and 29,155 candidate transcripts for CASTxWSB, representing 7,393 and 8,532 candidate genes, respectively.

The CAST/EiJ, PWK/EiJ and CASTxPWK trio had a merged k-mer profile of 118,100,824 k-mers, 42,688,910 (36.1%) of which appeared in the candidate transcripts. Similarly, the CAST/EiJ, WSB/EiJ and CASTxWSB trio had a merged k-mer profile of 118,383,117 k-mers, 52,715,089 (44.5%) of which appeared in its set of candidate transcripts. I verified most the k-mers in the F1 samples not appearing in candidate transcripts have few occurrences. The k-mers with high profiles which do not appear in candidate transcripts were mostly due to poly(A) tails, transcripts with dense variants in the parental strains, or transcripts expressed by the F1 strain but not the parents, as shown in Fig. (6.4)

Using the penalty parameter $\lambda = 10^4$ for both F1 samples, my methods found 17,118 non-zero θ values in the CASTxPWK sample and 20,596 non-zero θ values in the CASTxWSB sample, corresponding to as many estimated expressed transcripts. This

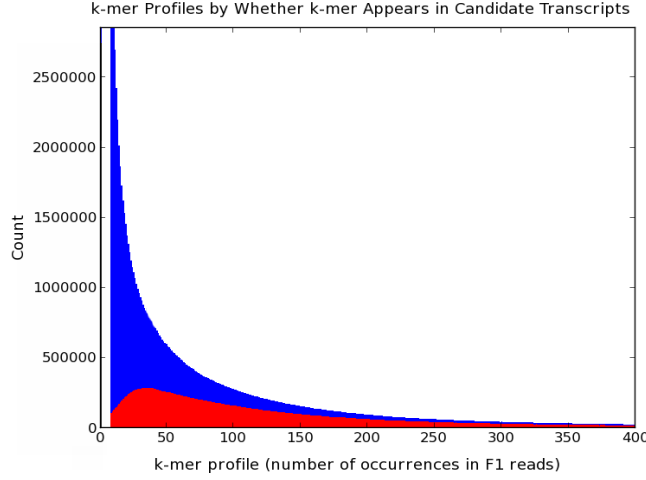


Figure 6.4: Stacked histogram of k-mers in the real CASTxPWK k-mer profile, sorted by the number of times each k-mer appears in the F1 reads. K-mers appearing in candidate transcripts are in red, and k-mers not appearing in candidate transcripts are in blue. The majority of k-mers not appearing in candidate transcripts have low number of occurrences, suggesting they are from lowly expressing genes or erroneous reads.

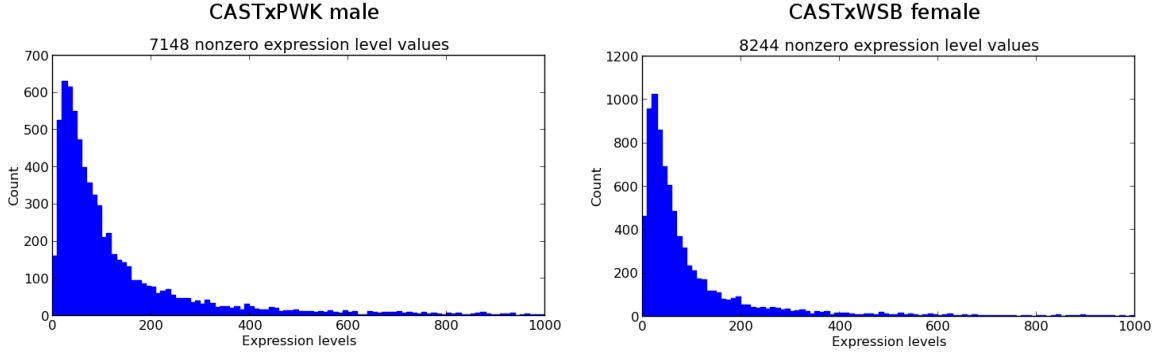


Figure 6.5: Histograms of estimated gene expression levels from real data that are non-zero. CASTxPWK has a median expression level of 75.97, with a max expression level of 5383.75. CASTxWSB has a median expression level of 52.38, with a max expression level of 9897.35.

represented 7,148 of 7,393 and 8,242 of 8,532 estimated expressed genes, respectively. These results are summarized in Table (6.2). I estimated the expression level of each gene by summing the θ values for all expressed isoforms, both maternal and paternal, of each gene. The distributions of non-zero expression levels are shown in Fig. (6.5).

To assess the ability to estimate ASE, I looked at the maternal contribution ratio of all expressed genes with candidate isoforms from both parents and differentiating variants between the two parents. Maternal contribution ratio of a gene is defined as the ratio of the expression levels from all maternal isoforms to the expression levels from both paternal and maternal isoforms of the gene. The distribution of maternal contribution ratios for both F1 samples is shown in Fig. (6.6). The median maternal contribution ratio for both the male CASTxPWK sample and the female CASTxWSB sample is around 0.5, as expected. In the male CASTxPWK sample, a higher number of genes are maternally expressed, which is expected since genes on the X chromosome and mitochondria in males are inherited exclusively from their mothers. I verified several genes that are known to exhibit ASE [27] [74] [17] as having high maternal contribution ratios if maternally expressed and low maternal contribution ratios if paternally expressed.

Table 6.3 shows comparisons of the maternal contribution ratios found in ten genes known to exhibit ASE, using both my method and the method used in [17], which is described in detail in [32]. The pipeline described in a recent study using this same dataset [32] addresses reference bias by modifying the reference sequence with all known variants in the parental strains, a process that is shown to result in more aligned reads and more reads with assigned origins. Although this modification of an alignment-based method greatly improves traditional single-reference pipelines, it requires previously annotated variants in the parental strains, a requirement that many experiments do not meet. My method does not require any known annotations, and the maternal contribution ratios are highly consistent when compared to those found in [32].

In addition, I examined the maternal contribution ratios of all expressed genes on the X chromosome with candidate isoforms from both parents and differentiating variants

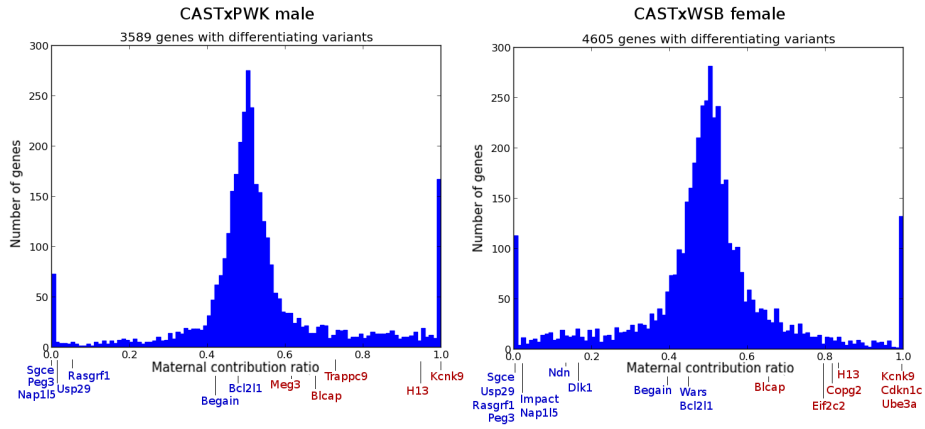


Figure 6.6: Histogram of the maternal contribution ratios of all expressed genes with candidate isoforms from both parental strains and containing differentiating variants between the parental strains. On the bottom of each plot, several genes known to be maternally expressed in literature are highlighted in red, and several genes known to be paternally expressed are highlighted in blue.

Table 6.3: Comparisons to maternal contribution ratios (MCR) found in [17], which uses methods from [32]

	CASTxPWK		CASTxWSB	
Gene	MCR from [17]	MCR from my pipeline	MCR from [17]	MCR from my pipeline
Kcnk9	0.94	1.0	0.98	1.0
H13	0.87	0.95	0.78	0.82
Blcap	0.63	0.66	0.64	0.64
cl2l1	0.46	0.47	0.45	0.43
Begain	0.39	0.42	0.38	0.39
Nap1l5	0.07	0.00	0.14	0.02
Rasgrf1	0.05	0.05	0.04	0.00
Usp29	0.00	0.01	0.00	0.00
Peg3	0.06	0.00	0.01	0.00
Sgce	0.00	0.00	0.00	0.00

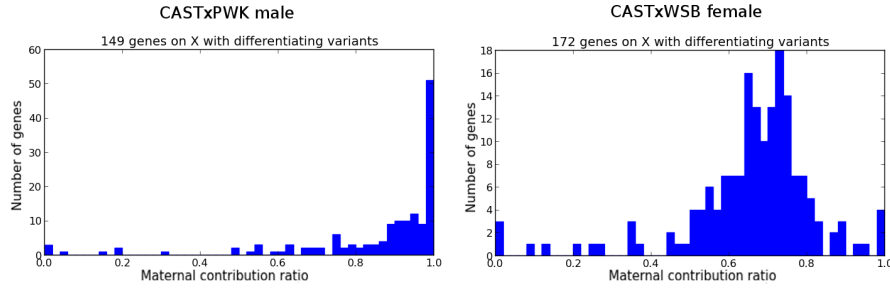


Figure 6.7: Histogram of the maternal contribution ratio of all expressed genes on the X chromosome with candidate isoforms from both parental strains and containing differentiating variants between the parental strains. In the male CASTxPWK sample, the median maternal contribution ratio is 0.94. In the female CASTxWSB sample, the median maternal contribution ratio is 0.68. Both are in the expected range of maternal contribution ratio of X-chromosome genes in male and female animals, respectively.

between the parents. In the male CASTxPWK sample, we expect all genes on the X chromosome to be maternally expressed, since its X chromosome is inherited from the maternal strain. In the female CASTxWSB sample, we expect most genes on X to be expressed with a 0.6-0.7 maternal contribution ratio due to a known maternal bias in X inactivation [73] [5]. As expected, I found the median maternal contribution ratio to be 0.94 in the male CASTxPWK sample and 0.68 in the female CASTxWSB sample. The distributions of maternal contribution ratios of genes on the X chromosome are plotted in Fig. (6.7).

6.4.3 Speed and Memory

I ran the methods on a single 1600 MHz processor on a machine with 32 GB RAM. The *De Bruijn* graphs of the samples are approximately 1GB after combining reverse complements and filtering low-count k-mers. Inferring candidate transcripts takes approximately 2-3 hours per parental strain, and the coordinate descent algorithm converges after approximately 1 to 3 million iterations, which takes around 1-2 hours on this specific machine. I was able to take advantage of the sparseness of the candidate transcript k-mer profile matrix \mathbf{X} by storing it as a sparse matrix using the Scipy.sparse

package.

6.5 Discussion

I have developed methods to estimate expression levels for maternal and paternal versions of transcripts from RNA-seq trio data. The need for such methods arose when I realized that although we have RNA-seq data of biological trios and wish to analyze ASE of F1 strains, current methods, both alignment-based and *de novo*, do not include standard pipelines that take advantage of available RNA-seq data from parental strains. My model is able to exploit the information from the maternal and paternal RNA-seq reads and build candidate transcripts that accurately reflect the F1 strain's transcriptome, and it does so without requiring annotated variants of the parental strains. My proposed methods still rely on the existence of an annotated reference transcriptome, which is refined to make it more consistent with the observed data. The use of annotated transcripts also serves as a strong prior towards biologically meaningful solutions.

The proposed methods performed well when compared to a Trinity-Bowtie-RSEM pipeline, which incorporates a state-of-the-art *de novo* assembler and aligner. I was able to achieve high sensitivity and specificity (0.9553 and 0.9883) in detecting baseline expression of transcripts. Of the correctly identified expressed transcripts, I was also able to correctly identify more transcripts exhibiting ASE, with a sensitivity of 0.77, compared to Trinity's low ASE sensitivity of 0.32. The pipeline I used with Trinity also made use of parental RNA-seq data, since I separately assembled transcript sequences from maternal and paternal reads, then aligned the F1 reads to the entire set of assembled transcript sequences. However, Trinity still had a low sensitivity of 0.13 for determining baseline expression, since the main challenge I faced using Trinity was mapping the assembled sequences back to known reference transcripts.

When compared with a modified alignment-based method that requires known variants of the parental strains [32], my method yields consistent estimates of the maternal contribution ratio.

The dimensionality of the data can be large. In the real data, we have approximately 5×10^7 k-mers after filtering and tens of thousands of candidate transcripts. Despite the high dimensionality of the k-mer space and transcripts space, I was able to use a refined coordinate descent algorithm to efficiently perform lasso regression. Although not implemented, we could also decrease the k-mer space without affecting the solution by merging overlapping k-mers into contigs of variable lengths greater than k .

Since the candidate transcripts are generated from annotated reference transcripts, my methods do not currently assemble novel transcript sequences. However, it is possible to model the k-mer profiles of all novel transcripts as the residual of the linear regression, and *de novo* assembly of the residual k-mers could then generate sequences of novel transcripts. Another limitation of my model lies in its inability to detect genes exhibiting overdominance, where the expression level is high in the F1 animal but nonexistent in the parental strains. This could be remedied by also selecting candidate transcripts from the F1 *De Bruijn* graph itself as additional features. The strength of my methods lies in the ability to determine ASE directly from RNA-seq data in diploid trios without prior knowledge of genomic variation in the parental genomes. This straightforward regression approach is tolerant of imbalanced read counts in different samples, as demonstrated by the reasonable maternal contribution ratio distribution in the male CASTxPWK F1 sample (Fig 6.6), despite the CAST/EiJ read count being nearly 1.5 times as high as the PWK/EiJ read count. These methods could even be extended to ascertain ASE in any animal that is a hybrid of two or more isogenic ancestral genomes, such as the recombinant inbred strains often used as genetic reference panels. It can also be applied to outbred samples, such as human trios, at a subset

of genes where the parents are homozygous with different alleles. For other genes, the parental origin cannot be established, but the total abundance can still be estimated.

Chapter 7 : Discussion and Conclusion

In this thesis, I presented methods for analyzing admixed animals using information that can be inferred by relating measurements from a sample to the same measurements made on its ancestors. I have shown that these methods compare favorably to traditional methods, which rely on categorical measures, such as biallelic SNPs annotated in the reference genome, as the main source of information. The quantitative measures used for admixed animals and their ancestors presented in this thesis come from two main sources – genotyping microarrays and RNA-seq reads – which are common data sources for assessing the genomes of individual organisms. The innovative use of microarray probe intensities and unaligned RNA-seq reads, and the increase in accuracy they provide over genotype calls and reference-aligned reads, demonstrate the amount of ancestral haplotype information that is available in these widely-used platforms. However, this information about ancestral haplotypes is often inaccessible by traditional methods, which examine only point variations within the reference genome, and my results suggest that the traditional use of genotyping microarrays and next-generation sequencing data results in loss of information, leading to incomplete or incorrect conclusions. Although this subtle information about the haplotype is sometimes regarded as noise, it originates from actual sequence variation and should be included for accurate analysis.

Many of the methods and resources I presented in this thesis have been used by our collaborators to aid and further genetics studies using the mouse genetic reference population Collaborative Cross (CC), among other samples [16, 17, 65, 77, 55, 19]. I first discuss in Chapter 3 the design of a genotyping microarray that is informative

for assessing ancestor haplotypes, which differs from traditional array design that only considers the informativeness of single SNPs independently. The MegaMUGA genotyping platform I introduce has been widely used by the mouse genetics community, with 7179 samples, both CC and non-CC, genotyped to date. The ancestry inference method described in Chapter 4 has been run on all genotyped samples in both MUGA and MegaMUGA, and our collaborators have used the information to facilitate the continued inbreeding of CC lines. Furthermore, the ancestor haplotype reconstructions of CC mice contributed to the discovery of the *R2d2* – responder to meiotic drive 2 – gene, a “selfish” gene with certain founder alleles that are inherited more frequently than others [19]. I have also developed several other intensity-based analysis tools for the MDA, MUGA, and MegaMUGA, including those that determine the karyotype of sex chromosomes and detect copy number variations. These tools are available online at <http://compgen.unc.edu/Tools> and <http://csbio.unc.edu/CCstatus>. Additionally, I extended the ancestry inference algorithm to detect and verify genetically engineered constructs in non-CC mutant mice from the Mutant Mouse Regional Resource Centers (MMRRC), using both the MUGA and MegaMUGA genotyping platforms. This tool is available online at <http://csbio.unc.edu/MMRRC>. The quantitative trait loci (QTL) mapping algorithm introduced in Chapter 5 is currently implemented for the MegaMUGA, and it is also available online at <http://csbio.unc.edu/CCstatus/index.py?run=mapIntenQTL>.

In the sections below, I summarize my results and possible directions for future research.

7.1 Microarray Design

In Chapter 3, I discussed the design of the genotyping microarray MegaMUGA. Unlike previously designed genotyping arrays, MegaMUGA takes into account nearby

variants of each SNP marker. By considering SNPs in sliding windows, I was able to design an array that is maximally informative for the 36 possible founder haplotype states in the CC and DO populations. In addition to selecting SNP markers that could distinguish between a high number of CC founders, I also included non-SNP invariant markers in the pseudoautosomal region (PAR). MegaMUGA was manufactured on the Illumina Infinium II and had a very high marker conversion rate of 97.51% with 77,808 final markers. The markers in the PAR demonstrate that even without a variable SNP, hybridization intensities can be informative for invariant probe sequences.

The techniques I introduced for maximizing information content in consecutive markers can be modified to optimize different aspects of genetic information. For instance, similar methods were used in the most recent array designed for the CC and DO populations, called GigaMUGA. GigaMUGA is a recently-released 143,259-marker Illumina genotyping array which I helped to develop with several colleagues. While the design of both MUGA and MegaMUGA optimized for distinguishing the CC founder states uniformly across the genome, GigaMUGA was designed to better localize recombinations between CC founder states. Therefore, the markers in GigaMUGA were selected to immediately flank known recombination hotspots, and the maximization of detectable ancestral haplotypes was performed locally around each hotspot, so that the ability to recognize recombinations between CC founder states would be optimal.

Given the distinctive hybridization intensity patterns generated by samples with off-target variants within SNP probe sequences reported in Chapter 4, we can consider the design of SNP markers with intentional off-target variants within their probe sequences. Theoretically, these SNP markers would each exhibit more than the typical three clusters ('A', 'B', 'H') in hybridization intensities. This would enable the design of genotyping arrays that contain significantly more information per marker than traditional arrays, since each selected SNP marker would potentially differentiate between

more than three possible alleles. However, such an array would require more sophisticated genotyping calling algorithms that allow more than two homozygous alleles, such as the one introduced in Chapter 4 and [34].

7.2 Ancestry Inference

In Chapter 4, I discussed a method I developed for inferring the ancestry of admixed animals using hybridization intensities from the genotyping microarrays MUGA and MegaMUGA [24]. I show that the intensity-based method achieves a higher accuracy and agreement with high-throughput sequencing data than GAIN, a method which uses discretized genotype calls [44]. Due to the prevalence of non-informative and erroneous genotype calls from the Illumina platform, the intensity-based method is able to better refine recombination breakpoints and handle suboptimal markers.

The clustering of CC founders used in my ancestry inference algorithm reveals many SNP markers with more than the three expected ‘A’, ‘B’, and ‘H’ allele clusters, as well as some SNP markers with no clear clustering yet appear to have arbitrarily defined genotype calls. I show that these unexpected probe intensity patterns can be caused by off-target variants or deletions in the probe sequence, and we expect some patterns are also due to copy number variations or homologous sequences elsewhere in the genome [20].

Using probe intensity clusters, we can redefine genotype calls to include more than two homozygous and one heterozygous alleles, as done in [34]. However, methods that can handle an arbitrary number of genotype calls at each SNP would be necessary to analyze such data, yet many traditional methods for array and sequence analysis rely exclusively on the binary encoding of ‘0’ and ‘1’ to represent the two homozygous alleles. The development of methods which allow for more than biallelic SNPs would be a crucial downstream area of research for genotyping arrays with multiallelic markers

and genotype calling algorithms with multiallelic classes.

One concern with defining more than biallelic genotype calls with intensity clusters is the selection of the set of samples used for initial cluster. The methods I present rely on CC founders and F1s as the set of control samples used for clustering. While the CC founders are highly diverse, other non-CC strains may have completely different alleles in certain regions, as shown in Figure 7.1. However, our arrays were designed using probe sequences and SNPs from the CC, and only a small handful of non-CC genotype clusters have been discovered in MUGA and MegaMUGA.

Although the maximization method for ancestry inference I present here uses fixed transition probabilities based on previous observations, the transition probabilities between different founder states can be learned from a large population of samples so that transitioning between SNP pairs with frequent crossover events is more probable than transitioning between SNPs with few recombinations in between. The drawback to this method would be more difficult detection of rare recombinations.

As data from next-generation sequencing data becomes more widely available, we can consider methods for ancestry inference using DNA-seq reads. Such a method that relies on reference alignment of DNA-seq reads from the admixed animal and its ancestors is discussed in [76], but its current speed and cost constraints still make ancestry inference in microarrays the more practical approach by far.

7.3 Quantitative Trait Loci Mapping

One of the important applications of creating genome mosaics from inferred ancestry is the mapping of Quantitative Trait Loci (QTL). QTL mapping is the task of associating physical traits or diseases with regions in the genome, in order to find genes of interest affecting the observed trait. Since microarray probe intensities have the ability to elucidate haplotype information, in Chapter 5, I developed an intensity-

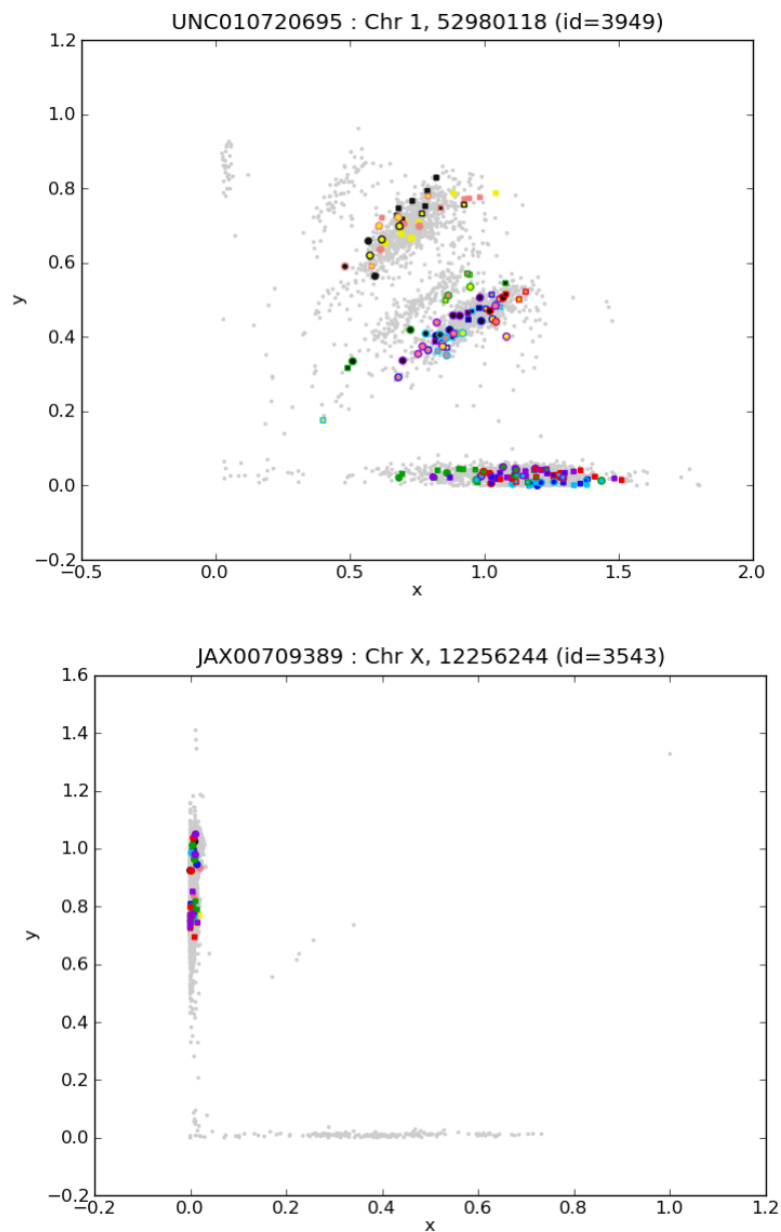


Figure 7.1: Two MUGA markers that capture unexpected alleles in strains beyond the CC founders. Here, all samples genotyped in MUGA are plotted with their hybridization intensities, with the CC founders and their F1s highlighted.

based method that bypasses the need for assigning genotypes, but instead uses sliding windows of probe intensities to identify genomic regions correlated with physical traits [22]. This creates a more direct association between haplotype and phenotype, and it also avoids the potential issues that arise when the samples used for clustering do not represent the full spectrum of possible alleles, as in the case of the markers shown in Figure 7.1.

My intensity-based method for QTL mapping uses the correlation between genotype intensity and phenotype distance matrices, and I show that the results capture the same QTLs as traditional genotype-based methods in real data. In synthetic data, my method captures true QTL peaks more accurately and with higher significance than traditional genotype-based methods, and with more specificity than ancestor haplotype-based methods.

Although I have shown that CC lines exhibit very low levels of non-local linkage disequilibrium [16], other populations used for QTL mapping may have significant population structure. The intensity-based method can be easily extended to handle such cases by comparing pairwise intensity vectors on a genome-wide or chromosome-wide scale.

Physical traits are sometimes affected by more than one gene, and these genes can be in epistasis, meaning the physical outcome depends on the combination of the two gene alleles together. Given n markers, epistasis can only be modeled with $O(n^2)$ algorithms with varying degrees of pruning. The intensity-based method I present only examines one genomic region at a time, but once a candidate gene region is found, it can be treated as a covariate in a second iteration of mapping, enabling the discovery of gene pairs in epistasis where at least one gene is strongly correlated with the phenotype.

7.4 Estimating Allele-Specific Expression

In Chapter 6, I used quantitative counts of ancestral subsequences (k-mers) from RNA-seq reads to ascertain allele-specific expression (ASE) in F1 animals [23]. My method uses the reads from the two parental strains as a template for regression of the F1 strain's reads, with the added constraint of a lasso penalty to filter out non-expressed transcripts. The results in synthetic data show the sensitivity and specificity of my method in both the detection of expressed transcripts and the prediction of expression levels when compared with existing methods. The results in real data verify known gene transcripts that exhibit ASE in the autosomes and the X chromosome.

The method I present assumes that the F1 is descended from inbred parents. Outbred populations such as humans have parents that are heterozygous, and for the F1 animal, each parental transcript should come from one of two possibilities from two different grandparents' genome. Therefore, in outbred animals, ancestry inference of the admixed animal can be first performed to select the appropriate parental reads to be used in each region. Otherwise, the algorithm would require two separate transcript sequences from each parent, with a total of four candidate transcript sequences per gene transcript. This can be obtained from reads from each parent, but would require minor modifications of the algorithm.

Although the ancestors used in my examples are the parental strains of F1 animals, it is trivial to imagine extending my method to admixed animals with more than two founders. This can be done after ancestry inference so only two possible ancestor haplotypes are considered in each region of the genome, or all possible ancestors can be used as features for reads from the admixed sample to be regressed against. When performed using reads from all possible ancestors as features, it is even possible to infer ancestry in conjunction with allele-specific expression. This is an area of active research,

specifically with applications to DO animals, which are outbred samples between the eight CC founders [11].

7.5 Conclusion

The methods and subsequent analyses presented in this thesis show the power of using traditional genotyping and sequencing platforms in innovative ways to obtain maximum information. When analyzing admixed animals, data from ancestral genomes serve as a better template for comparative analysis than single-point variation data on a single reference genome. Although this thesis focuses mainly on the use of microarray and RNA-seq data, the idea of using ancestor data as a standard for comparing admixed animals can be extended to other genetic platforms, such as DNA-seq data. A method for ancestry inference with DNA-seq read data aligned to a CC consensus genome is presented in [76], and similar methods for both ancestry inference and quantitative trait loci mapping that require no reference alignment, such as ancestry inference with RNA-seq reads [11], are potential topics for future research.

BIBLIOGRAPHY

- [1] David E Barton, Byoung S Kwon, and Uta Francke. Human tyrosinase gene, mapped to chromosome 11 (q14? q21), defines second region of homology with mouse chromosome 7. *Genomics*, 3(1):17–24, 1988.
- [2] Judith A Blake, Joel E Richardson, Carol J Bult, Jim A Kadin, Janan T Eppig, et al. Mgd: the mouse genome database. *Nucleic acids research*, 31(1):193–195, 2003.
- [3] Karl W Broman, Hao Wu, Śaunak Sen, and Gary A Churchill. R/qtl: Qtl mapping in experimental crosses. *Bioinformatics*, 19(7):889–890, 2003.
- [4] Michael Burrows and David J Wheeler. *A block-sorting lossless data compression algorithm*. Citeseer, 1994.
- [5] Lisa Helbling Chadwick, Lisa M Pertz, Karl W Broman, Marisa S Bartolomei, and Huntington F Willard. Genetic control of x chromosome inactivation in mice: definition of the xce candidate interval. *Genetics*, 173(4):2103–2110, 2006.
- [6] Riyan Cheng, Mark Abney, Abraham A Palmer, and Andrew D Skol. Qtlrel: an r package for genome-wide association studies in which relatedness is a concern. *BMC genetics*, 12(1):66, 2011.
- [7] E.J. Chesler, D.R. Miller, L.R. Branstetter, L.D. Galloway, B.L. Jackson, V.M. Philip, B.H. Voy, C.T. Culiati, D.W. Threadgill, R.W. Williams, et al. The collaborative cross at oak ridge national laboratory: developing a powerful resource for systems genetics. *Mammalian Genome*, 19(6):382–389, 2008.
- [8] Elissa J Chesler, Darla R Miller, Lisa R Branstetter, Leslie D Galloway, Barbara L Jackson, Vivek M Philip, Brynn H Voy, Cymbeline T Culiati, David W Threadgill, Robert W Williams, et al. The collaborative cross at oak ridge national laboratory: developing a powerful resource for systems genetics. *Mammalian Genome*, 19(6):382–389, 2008.
- [9] Asif T Chinwalla, Lisa L Cook, Kimberly D Delehaunty, Ginger A Fewell, Lucinda A Fulton, Robert S Fulton, Tina A Graves, LaDeana W Hillier, Elaine R Mardis, John D McPherson, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.
- [10] A.T. Chinwalla, L.L. Cook, K.D. Delehaunty, G.A. Fewell, L.A. Fulton, R.S. Fulton, T.A. Graves, L.D.W. Hillier, E.R. Mardis, J.D. McPherson, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.
- [11] Kwangbom Choi. Populase: Population model for improved estimation of allele-specific expression. <https://github.com/jax-cgd/populase>, January 2015.

- [12] Deanna M Church, Leo Goodstadt, LaDeana W Hillier, Michael C Zody, Steve Goldstein, Xinwe She, Carol J Bult, Richa Agarwala, Joshua L Cherry, Michael DiCuccio, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS biology*, 7(5):e1000112, 2009.
- [13] 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, et al. The collaborative cross, a community resource for the genetic analysis of complex traits. *Nature Genetics*, 36(11):1133–1137, 2004.
- [14] Gary A Churchill and Rebecca W Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–971, 1994.
- [15] Gary A Churchill, Daniel M Gatti, Steven C Munger, and Karen L Svenson. The diversity outbred mouse population. *Mammalian Genome*, 23(9-10):713–718, 2012.
- [16] Collaborative Cross Consortium. The genome architecture of the collaborative cross mouse genetic reference population. *Genetics*, 190:389–401, 2012.
- [17] James J Crowley, Vasyl Zhabotynsky, Wei Sun, Shunping Huang, Isa Kemal Pakatci, Yunjung Kim, Jeremy R Wang, Andrew P Morgan, John D Calaway, David L Aylor, Zaining Yun, Timothy A Bell, Ryan J Buus, Mark E Calaway, John P Didion, Terry J Gooch, Stephanie D Hansen, Nashiya N Robinson, Ginger D Shaw, Jason S Spence, Corey R Quackenbush, Cordelia J Barrick, Randal J Nonneman, Kyungsu Kim, James Xenakis, Yuying Xie, William Valdar, Alan B Lenarcic, Wei Wang, Catherine E Welsh, Chen-Ping Fu, Zhaojun Zhang, James Holt, Zhishan Guo, David W Threadgill, Lisa M Tarantino, Darla R Miller, Fei Zou, Leonard McMillan, Patrick F Sullivan, and Fernando Pardo-Manuel de Villena. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature genetics*, advance online publication:–, 03 2015.
- [18] John P Didion, Ryan J Buus, Zohreh Naghashfar, David W Threadgill, Herbert C Morse, and Fernando Pardo-Manuel de Villena. Snp array profiling of mouse cell lines identifies their strains of origin and reveals cross-contamination and widespread aneuploidy. *BMC genomics*, 15(1):847, 2014.
- [19] John P Didion, Andrew P Morgan, Amelia M-F Clayshulte, Rachel C McMullan, Liran Yadgary, Petko M Petkov, Timothy A Bell, Daniel M Gatti, James J Crowley, Kunjie Hua, et al. A multi-megabase copy number gain causes maternal transmission ratio distortion on mouse chromosome 2. *PLoS genetics*, 11(2):e1004850–e1004850, 2015.
- [20] John P Didion, Hyuna Yang, Keith Sheppard, Chen-Ping Fu, Leonard McMillan, Fernando PM de Villena, and Gary A Churchill. Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC genomics*, 13(1):34, 2012.

- [21] Paul Flicek, Ikhlaq Ahmed, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, et al. Ensembl 2013. *Nucleic acids research*, page gks1236, 2012.
- [22] Chen-Ping Fu, Fernando Pardo-Manuel de Villena, and Leonard McMillan. Quantitative trait loci mapping with microarray marker intensities. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 472–478. ACM, 2014.
- [23] Chen-Ping Fu, Vladimir Jovic, and Leonard McMillan. An alignment-free regression approach for estimating allele-specific expression using rna-seq data. In *Research in Computational Molecular Biology*, pages 69–84. Springer, 2014.
- [24] Chen-Ping Fu, Catherine E Welsh, Fernando Pardo-Manuel de Villena, and Leonard McMillan. Inferring ancestry in admixed populations using microarray probe intensities. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 105–112. ACM, 2012.
- [25] Daniel M Gatti. Doqtl: Qtl mapping for diversity outbred mice. <http://cgd.jax.org/apps/doqtl/DOQTL.shtml>, 2014.
- [26] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, et al. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, 2011.
- [27] Christopher Gregg, Jiangwen Zhang, Brandon Weissbourd, Shujun Luo, Gary P Schroth, David Haig, and Catherine Dulac. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *science*, 329(5992):643–648, 2010.
- [28] Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. Modelling and simulating generic rna-seq experiments with the flux simulator. *Nucleic acids research*, 40(20):10073–10083, 2012.
- [29] M. Guttman, M. Garber, J.Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M.J. Koziol, A. Gnirke, C. Nusbaum, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincnas. *Nature biotechnology*, 28(5):503–510, 2010.
- [30] Chris S Haley and Sarah A Knott. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69(4):315–324, 1992.
- [31] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

- [32] Shunping Huang, James Holt, Chia-Yu Kao, Leonard McMillan, and Wei Wang. A novel multi-alignment pipeline for high-throughput sequencing data. *Database*, 2014:bau057, 2014.
- [33] Shunping Huang, Chia-Yu Kao, Leonard McMillan, and Wei Wang. Transforming genomes using mod files with applications. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. ACM, 2013.
- [34] Chia-Yu Kao, Chen-Ping Fu, and Leonard McMillan. Instantgenotype: a non-parametric model for genotype inference using microarray probe intensities. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 147–154. ACM, 2014.
- [35] T.M. Keane, L. Goodstadt, P. Danecek, M.A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294, 2011.
- [36] Byoung S Kwon, Asifa K Haq, Seymour H Pomerantz, and Ruth Halaban. Isolation and sequence of a cdna clone for human tyrosinase that maps at the mouse c-albino locus. *Proceedings of the National Academy of Sciences*, 84(21):7473–7477, 1987.
- [37] Eric S Lander and David Botstein. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185–199, 1989.
- [38] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, et al. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.
- [39] B. Li and C.N. Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.
- [40] W. Li, J. Feng, and T. Jiang. Isolasso: a lasso regression approach to rna-seq based transcriptome assembly. *Journal of Computational Biology*, 18(11):1693–1707, 2011.
- [41] Y. Li and S. Osher. Coordinate descent optimization for l-1 minimization with application to compressed sensing; a greedy algorithm. *Inverse Probl. Imaging*, 3(3):487–503, 2009.
- [42] Kerstin Lindblad-Toh, Ellen Winchester, Mark J Daly, David G Wang, Joel N Hirschhorn, Jean-Philippe Lavolette, Kristin Ardlie, David E Reich, Elizabeth Robinson, Pamela Sklar, et al. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature genetics*, 24(4):381–386, 2000.
- [43] Eric Yi Liu, Andrew P Morgan, Elissa J Chesler, Wei Wang, Gary A Churchill, and Fernando 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, 2014.

- [44] 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, 2010.
- [45] Ruijie Liu, Ana-Teresa Maia, Roslin Russell, Carlos Caldas, Bruce A Ponder, and Matthew E Ritchie. Allele-specific expression analysis methods for high-density snp microarray data. *Bioinformatics*, 28(8):1102–1108, 2012.
- [46] P.C. Mahalanobis. On the generalized distance in statistics. In *Proceedings of the National Institute of Sciences of India*, volume 2, pages 49–55. New Delhi, 1936.
- [47] Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220, 1967.
- [48] Darla R Miller. The mouse universal genotyping array is available. <http://atlas.uthsc.edu/pipermail/ctc/2011-February/000003.html>, February 2011.
- [49] Richard Mott, Christopher J Talbot, Maria G Turri, Allan C Collins, and Jonathan Flint. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proceedings of the National Academy of Sciences*, 97(23):12649–12654, 2000.
- [50] Chad Nusbaum, Donna K Slonim, Katrina L Harris, Bruce W Birren, Robert G Steen, Lincoln D Stein, Joyce Miller, William F Dietrich, Robert Nahf, Victoria Wang, et al. A yac-based physical map of the mouse genome. *Nature genetics*, 22(4):388–393, 1999.
- [51] Luanne L Peters, Raymond F Robledo, Carol J Bult, Gary A Churchill, Beverly J Paigen, and Karen L Svenson. The mouse as a model for human biology: a resource guide for complex trait analysis. *Nature Reviews Genetics*, 8(1):58–69, 2007.
- [52] 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, 2009.
- [53] 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):473–481, 2007.
- [54] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S.D. Jackman, K. Mungall, S. Lee, H.M. Okada, J.Q. Qian, et al. De novo assembly and analysis of rna-seq data. *Nature methods*, 7(11):909–912, 2010.
- [55] Allison R Rogala, Andrew P Morgan, Alexis M Christensen, Terry J Gooch, Timothy A Bell, Darla R Miller, Virginia L Godfrey, and Fernando Pardo-Manuel de Villena. The collaborative cross as a resource for modeling human disease:

- Cc011/unc, a new mouse model for spontaneous colitis. *Mammalian Genome*, pages 1–14, 2014.
- [56] James Ronald, Joshua M Akey, Jacqueline Whittle, Erin N Smith, Gael Yvert, and Leonid Kruglyak. Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Research*, 15(2):284–291, 2005.
 - [57] J. Rozowsky, A. Abyzov, J. Wang, P. Alves, D. Raha, A. Harmanci, J. Leng, R. Bjornson, Y. Kong, N. Kitabayashi, et al. Alleleseq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology*, 7(1), 2011.
 - [58] S. Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin. Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, 82(2):290–303, 2008.
 - [59] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.
 - [60] Sagiv Shifman, Jordana Tzenova Bell, Richard R Copley, Martin S Taylor, Robert W Williams, Richard Mott, and Jonathan Flint. A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS biology*, 4(12):e395, 2006.
 - [61] D.A. Skelly, M. Johansson, J. Madeoy, J. Wakefield, and J.M. Akey. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome research*, 21(10):1728–1737, 2011.
 - [62] Peter E Smouse, Jeffrey C Long, and Robert R Sokal. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic zoology*, pages 627–632, 1986.
 - [63] Frank J Steemers, Weihua Chang, Grace Lee, David L Barker, Richard Shen, and Kevin L Gunderson. Whole-genome genotyping with the single-base extension assay. *Nature methods*, 3(1):31–33, 2006.
 - [64] A. Sundquist, E. Fratkin, C.B. Do, and S. Batzoglou. Effect of genetic divergence in identifying ancestral origin using HapAa. *Genome Research*, 18(4):676–682, 2008.
 - [65] 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, 2012.
 - [66] Illumina Technote. Infinium genotyping data analysis, 2010.

- [67] David W Threadgill and Gary A Churchill. Ten years of the collaborative cross. *G3: Genes— Genomes— Genetics*, 2(5):153–156, 2012.
- [68] C. Trapnell, L. Pachter, and S.L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.
- [69] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, H. Pimentel, S.L. Salzberg, J.L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *nature protocols*, 7(3):562–578, 2012.
- [70] C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. Van Baren, S.L. Salzberg, B.J. Wold, and L. Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.
- [71] William Valdar, Leah C Solberg, Dominique Gauguier, Stephanie Burnett, Paul Klenerman, William O Cookson, Martin S Taylor, J Nicholas P Rawlins, Richard Mott, and Jonathan Flint. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature genetics*, 38(8):879–887, 2006.
- [72] Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan FA Grant, Hakon Hakonarson, and Maja Bucan. Penncnv: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome research*, 17(11):1665–1674, 2007.
- [73] Xu Wang, Paul D Soloway, Andrew G Clark, et al. Paternally biased x inactivation in mouse neonatal brain. *Genome Biol*, 11(7):R79, 2010.
- [74] Xu Wang, Qi Sun, Sean D McGrath, Elaine R Mardis, Paul D Soloway, and Andrew G Clark. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PloS one*, 3(12):e3839, 2008.
- [75] Catherine E. Welsh. *Computational Tools to Aid the Design and Development of a Genetic Reference Population*. PhD thesis, University of North Carolina at Chapel Hill, December 2014.
- [76] Catherine E Welsh, Chen-Ping Fu, Fernando Pardo-Manuel de Villena, and Leonard McMillan. Fine-scale recombination mapping of high-throughput sequence data. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, page 585. ACM, 2013.
- [77] Catherine E Welsh, Darla R Miller, Kenneth F Manly, Jeremy Wang, Leonard McMillan, Grant Morahan, Richard Mott, Fuad A Iraqi, David W Threadgill, and Fernando Pardo-Manuel de Villena. Status and access to the collaborative cross population. *Mammalian Genome*, 23(9-10):706–712, 2012.

- [78] 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):191–198, 2012.
- [79] Michael A White, Akihiro Ikeda, and Bret A Payseur. A pronounced evolutionary shift of the pseudoautosomal region boundary in house mice. *Mammalian Genome*, 23(7-8):454–466, 2012.
- [80] 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):663–666, 2009.
- [81] 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, et al. Subspecific origin and haplotype diversity in the laboratory mouse. *Nature genetics*, 43(7):648–655, 2011.
- [82] Q. Zhang, W. Wang, L. McMillan, J. Prins, F. Pardo-Manuel de Villena, and D. Threadgill. Genotype sequence segmentation: Handling constraints and noise. *Algorithms in Bioinformatics*, pages 271–283, 2008.