COMPUTATIONAL ANALYSIS OF VARIABILITY OF RECOMBINATION RATES IN MICE

by

VISHNUKUMAR GALIGEKERE NAGABHUSHANA

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT ARLINGTON
August 2008

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to all those people who have helped with this research project. I sincerely thank my advisor, Dr. Nikola Stojanovic for providing an opportunity to work with him as a research student. I would also like to thank him for his continual support and valuable guidance throughout the course of the project, without which it would not have materialized. I also thank Dr. Elena de la Casa-Esperon from the UTA Biology Department for providing an opportunity to work on the computational aspects of her ambitious project. I would like to thank her for all the support and patience in providing all the biological inputs much needed for the work, and also for serving on my committee. I thank Dr. J.C. Loredo-Osti for providing all the inputs pertaining to the statistical analysis. I also thank Mr. David Levine for serving on my committee and for being a very helpful guide. I would like to take this opportunity to thank all the members of my family, without whose support and patience this endeavor would have been unfruitful. I thank my friends for their valuable support and help throughout the course of the project.

June 25, 2008

ABSTRACT

COMPUTATIONAL ANALYSIS OF VARIABILITY

OF RECOMBINATION RATES IN MICE

Vishnukumar Galigekere Nagabhushana (M.S.)

The University of Texas at Arlington, 2008

Supervising Professor: Nikola Stojanovic

Rapid advances in engineering and technology have fueled research in areas such as molecular biology and genetics to a new level and thus have given birth to multidisciplinary fields like bioinformatics and computational biology. These fields involve the application of one's expertise in one or more areas including computer science, information science, statistics, chemistry and mathematics to solve problems related to biology in general and molecular biology and genetics in particular. It is a challenging research area which requires one to grasp diverse fields at almost equal rigor and at the same time provides new insights into the fascinating field of biology.

Recombination is a natural phenomenon which occurs in germ cells during meiotic division in higher organisms. Much about meiotic recombination is already known, however, the genetic mechanism of its regulation remains elusive. Our research team was interested in studying the genetic control of this phenomenon by analyzing the variability of rates of recombination between several strains of inbred mice. To meet that goal we have assembled a

large amount of data and performed computational analysis of linkage maps of different mice strains.

Genome data are available from various public databases which are complex and geographically dispersed. These databases tend to evolve rapidly, due to the fast generation and inclusion of new or more accurate genomic data. In addition, the interfaces to these databases are usually designed to provide information in a way most suitable for viewing by human investigators, and are hence limited in functionality. These issues add up to make the data collection and preprocessing a rather complex problem, alas one which rarely features prominently in any scientific report.

In this thesis we discuss two different approaches to estimate the variability of recombination rates in mice. We present efficient computational tools we have designed to gather data from public databases and discuss the preprocessing specific to the requirements dictated by the nature of our analysis. We also provide a description of the software tools we have used for the analysis of recombination rates. Thereafter, we discuss the algorithms, scripts and programs we have designed to arrive at many intermediate results required for the statistical analysis. After we have delivered the results described in this thesis, it is now up to our biology collaborators to use them for the study of the mechanisms driving the recombination in mice, the original motivation for this work. In that sense, this project is still much of a work in progress, and we expect that we, from the computer science side, have indeed paved the way for a major discovery in life sciences.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

There are millions of living organisms known to inhabit our planet, from the smallest and simplest bacteria through the most complex mammals including ourselves. Cells are the building blocks of all the organisms known thus far [1]. They are hence ubiquitously referred to as the structural and functional units of all living organisms. Cells are classified as prokaryotic or eukaryotic. The primary difference between the two is in that prokaryotic cells have no membrane encapsulating the nucleus (or even a nucleus as a unit), where as the eukaryotic cells do.

The nucleus of a eukaryotic cell houses almost all the genetic information in the form of two long chains of molecules running anti-parallel to each other. These chains of complex polymers coil around each other to form a double helical structure called the Deoxyribonucleic Acid (DNA). The basic units constituting the DNA are called nucleotides, and the entire polymer is made of a combination of only four nucleotides called Adenine, Cytosine, Guanine and Thymine. The DNA of higher eukaryotes is further organized in one or more linear structures called chromosomes. Most mammalian cells are diploid, having two copies of each chromosome, one from the mother and the other form the father. For example, the human DNA is divided into twenty three pairs of chromosomes and the mouse DNA is organized into twenty pairs of chromosomes. The chromosomes are long chains of nucleotides, compressed and packaged by proteins.

Numerous genes are spread across each chromosome. A gene is defined as the functional and physical unit of heredity passed from parent to offspring. Genes are segments of DNA, and the majority of genes contain the information necessary for making a specific protein. All the genes which occur in one copy of a chromosome also occur in the corresponding positions in the other copy Alleles are defined as variants of a gene (or any other genomic

feature) at a particular locus on a chromosome. Different alleles produce variations in inherited traits, such as hair color or blood type, and their combination is called the organism's genotype. Functional alleles can be dominant or recessive. In an individual, the dominant alleles are primarily responsible for the trait, also known as the individual's phenotype.

One of the reasons for life to have flourished so far is in the ability of cells to divide to form new cells. The original cells are called parent cells and the newly formed ones are called daughter cells. Mitosis is a type of eukaryotic cell division where the parent cell divides into two daughter cells. In this type of division, the parent nucleus first divides into two daughter nuclei, each of them having the same copy of the duplicated genome form the parent cell. After the division of the nucleus, the cell divides into two daughter cells such that each daughter cell carries the complete nuclear genome of the parent. Another type of cell division, meiosis, is defined as two successive nuclear divisions, with corresponding cell divisions which produce haploid gametes having only one half of the genetic material of the original cell. Meiotic cell division occurs in the cells of the reproductive system of sexually reproducing eukaryotes, also called germ cells. During the meiotic cell division, a diploid (two copies of each chromosome) genome replicates only once to form four haploid (one copy of each chromosome) cells. In consequence, a meiotic division results in four haploid cells called gametes. New diploid cells are formed only when these gametes are subjected to fertilization. During the first phase of meiosis genetic material is exchanged between homologous chromosomes resulting in unique genetic content in each of the haploids. This phenomenon is called recombination. The combination of parental alleles at different sites of a chromosome is called a haplotype. Haplotypes are used for the identification of recombination events between two or more loci.

DNA sequencing is used to determine the exact layout of four nucleotides in DNA fragments [2]. Complete genomes of many organisms have been assembled by sequencing fragments and then stitching together the sequences of many different fragments [3]. The rapid advances in the fields of engineering and computer science has led to the development of

various high-throughput technologies, capable of generating more accurate genetic sequences at a much faster rate and at the same time reducing the costs dramatically [4]. These technologies have been steadily fueling genetic research, taking it to a new level. In the process, they have given birth to cross-disciplinary areas such as bioinformatics and computational biology.

Bioinformatics is an interdisciplinary field between computer science, mathematics, statistics, biochemistry, molecular biology and genetics. Within computer science the field incorporates concepts from various sub-fields, such as algorithm theory, artificial intelligence, data mining and database theory [5]. Some typical bioinformatics tasks include analyzing the DNA sequence using a specific algorithm, finding out the similarity between two sequences, predicting genes, developing visualization tools, and many more. For these purposes many tools have been developed, such as BLAST [6] and BLAT [24].You have reached the chapter sections of the template. Under Format, the Chapter text's Font settings are 10-point font size, Regular font style and Arial font with no Effects boxes checked. Under Format, the Chapter text's Paragraph settings are Alignment set to Justified, and Line Spacing set to Double.

1.1 Meiotic Recombination

Meiotic recombination is the swapping of DNA fragments between homologous chromosomes during meiosis, as depicted in Figure 1.1. The paternal chromatid and the maternal chromatid are represented in blue and red respectively. Genetic material is swapped between the two due to a phenomenon called chromosomal crossover, resulting in mixed chromosomes. 'A', 'B' and 'C' are genes on the paternal chromosome and 'a', 'b' and 'c' are the corresponding genes on the maternal chromosome. From Figure 1.1, it is evident that a crossover can result in hybrid chromosomes with a combination of genes from both paternal and maternal chromosomes.

Recombination in eukaryotes occurs during the first phase (prophase-1) of the meiosis. When the four chromatids (single chromosome) are bundled tightly during prophase-1,

chromosomal crossover can happen and genetic material can be exchanged. The point at which the DNA material is exchanged and which can be seen under a microscope is called chiasma.

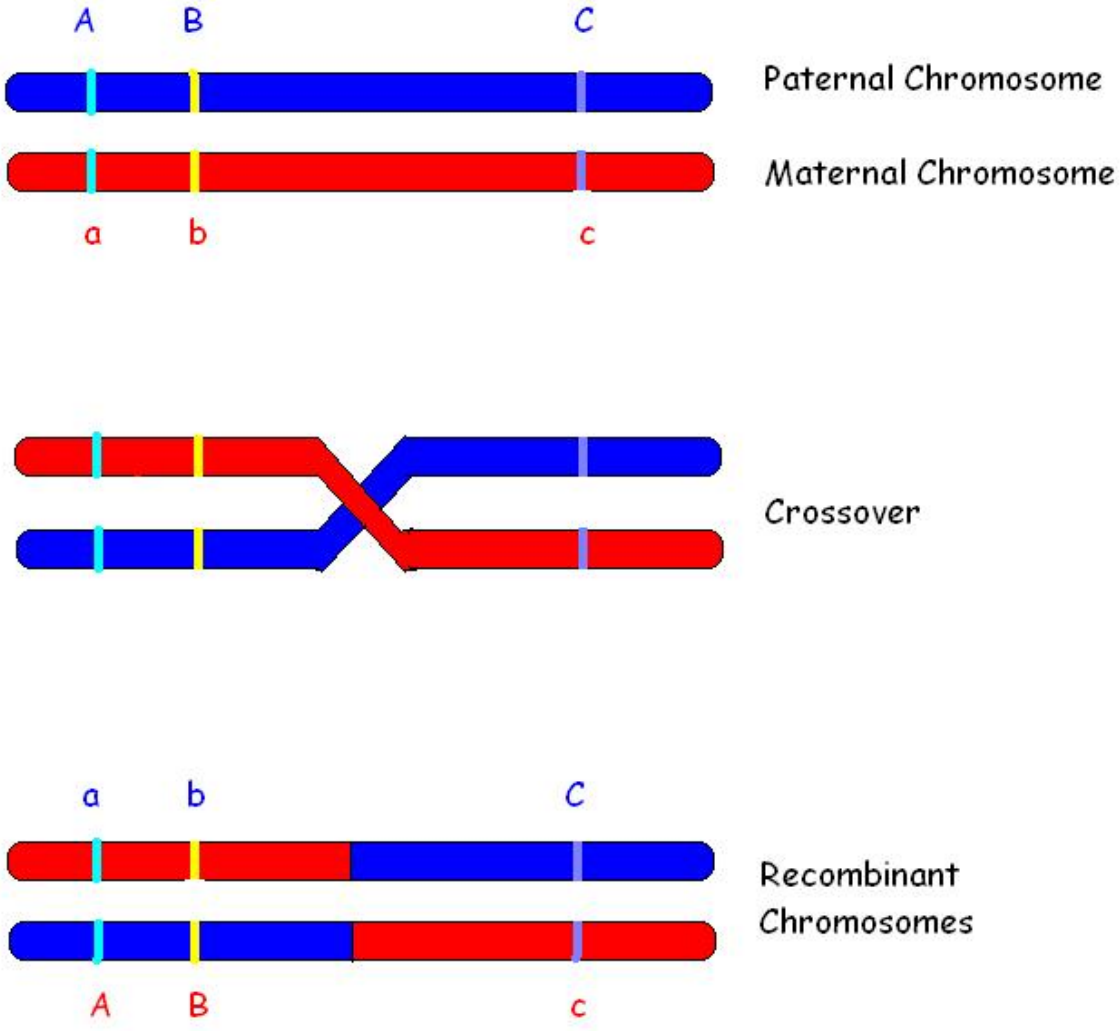


Figure 1.1 Genetic Recombination.

Figure 1.2 depicts a schematic of a crossover. In diploid cells each chromosome occurs in pairs and in our representation we show only one such pair, for the sake of simplicity. The ovals represent the nucleus, and red and blue lines represent the chromosomes. ‘A’ and ‘B’ are the alleles of two genes on the blue chromosome and, ‘a’ and ‘b’ are the corresponding alleles on the red chromosome. The crossover is illustrated by the following steps:

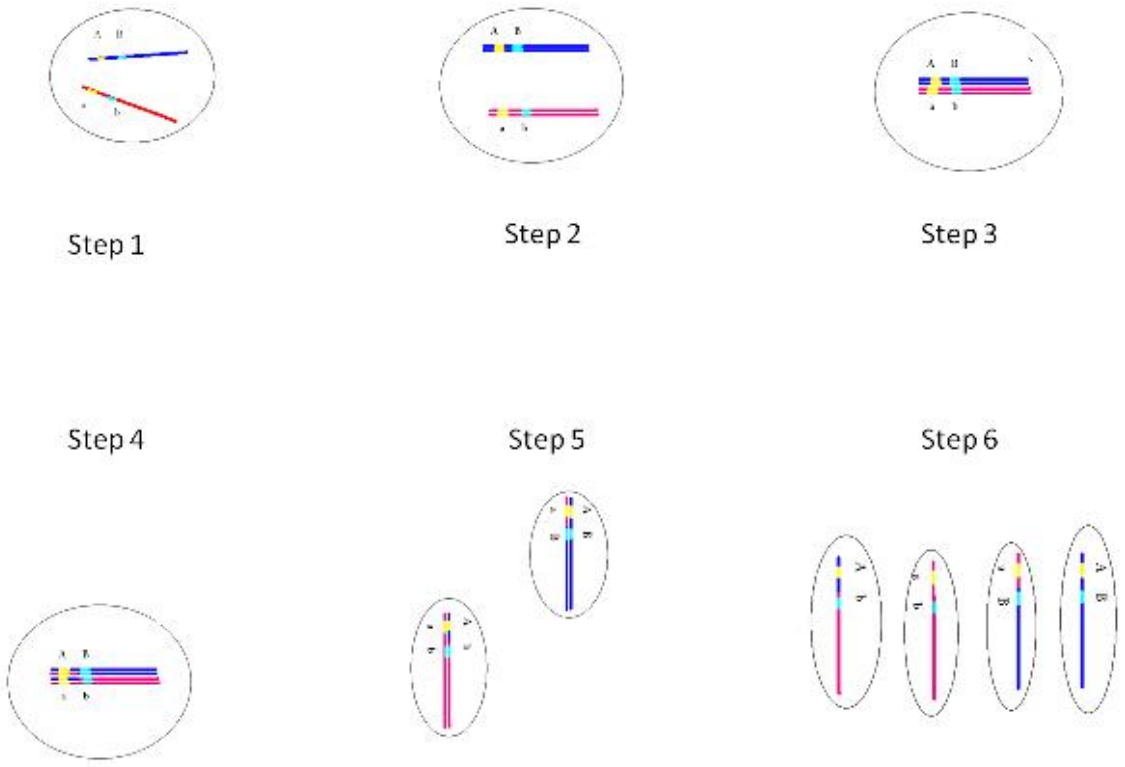


Figure 1.2 Schematic Representation of Meiotic Recombination.

Step 1: Shows the two copies of a particular chromosome in the nucleus.

Step 2: This step is called DNA synthesis. During this step each chromosome duplicates itself and the copies are attached at the centromere. After the duplication the copies are called sister chromatids.

Step 3: During this step, called synapsis, each pair finds its respective homologous chromatid pair and bundle up together. These bundles are called tetrads.

Step 4: At this step each of these tetrads line up in the middle of the nucleus. Chromosomal crossovers usually occur during this step.

Step 5: The nucleus separates into two nuclei segregating homologous chromosomes. This is the first of the two divisions occurring during meiosis. In this example, after one single crossover each nucleus will have a recombinant chromatid and a parental chromatid.

Step 6: Between step 5 and step 6 the nuclei and the sister chromatids separate again, forming four haploids.

Meiotic recombination is fundamental for the viability of a species, as the exchange of genetic material between homologous chromosomes increases the genetic diversity of the next generation. Much about meiotic recombination is already known, however, the genetic mechanism of its regulation remains elusive. In particular, the recombination rates have been found to be different in different regions of the chromosome. If recombination loci were randomly distributed on a chromosome, the distances on a genetic map (measuring the recombination percentage in a population) would be proportional to the distances on the corresponding physical map. Experiments related to this distribution were applied to the mouse genome maps, and the distribution of features provided evidence for heterogeneity in recombination [15]. Moreover, it is well established that the recombination rate is significantly higher near the chromosome ends, the telomeres, than near the centromere in mice [16]. In some experiments where a large number of offspring were genotyped (checked for the alleles at loci of interest) it was found that recombination events tended to cluster in small regions of a few kilobases in size. The regions where such clusters are formed are called recombination hotspots [17] [18]. The recombination rate is also known to be dramatically different between the sexes in mice, as the average rate of recombination observed in males was 50-85% of that observed in females [19]. Based on these results we can assume that the distribution of recombination is not stochastic, but that there exists a genetic control of meiotic recombination levels.

Our group has thus undertaken a research project to identify the genetic control of the levels of meiotic recombination in mice, and our collaborators have already identified two alleles of an X-chromosome linked gene and observed that changes in their expression affected the recombination frequencies [20]. However, the project described in this thesis concerns only the

computational support of this work, including the data collection, preprocessing and data analysis.

Our first step was to identify the recombination levels in as many crosses of different strains of mice (Mus musculus) as possible. Once we establish the levels for each of the considered crosses, our goal is to compare the individual distributions in an effort to evaluate the variation of recombination levels between the crosses involving different inbred strains.

1.2 Recombination Fraction

Recombination fraction gives us the probability of a crossover between two loci on the same chromosome, as a function of distance. If the loci are far apart, then the probability of observing a crossover between them, which is the recombination fraction, is high.

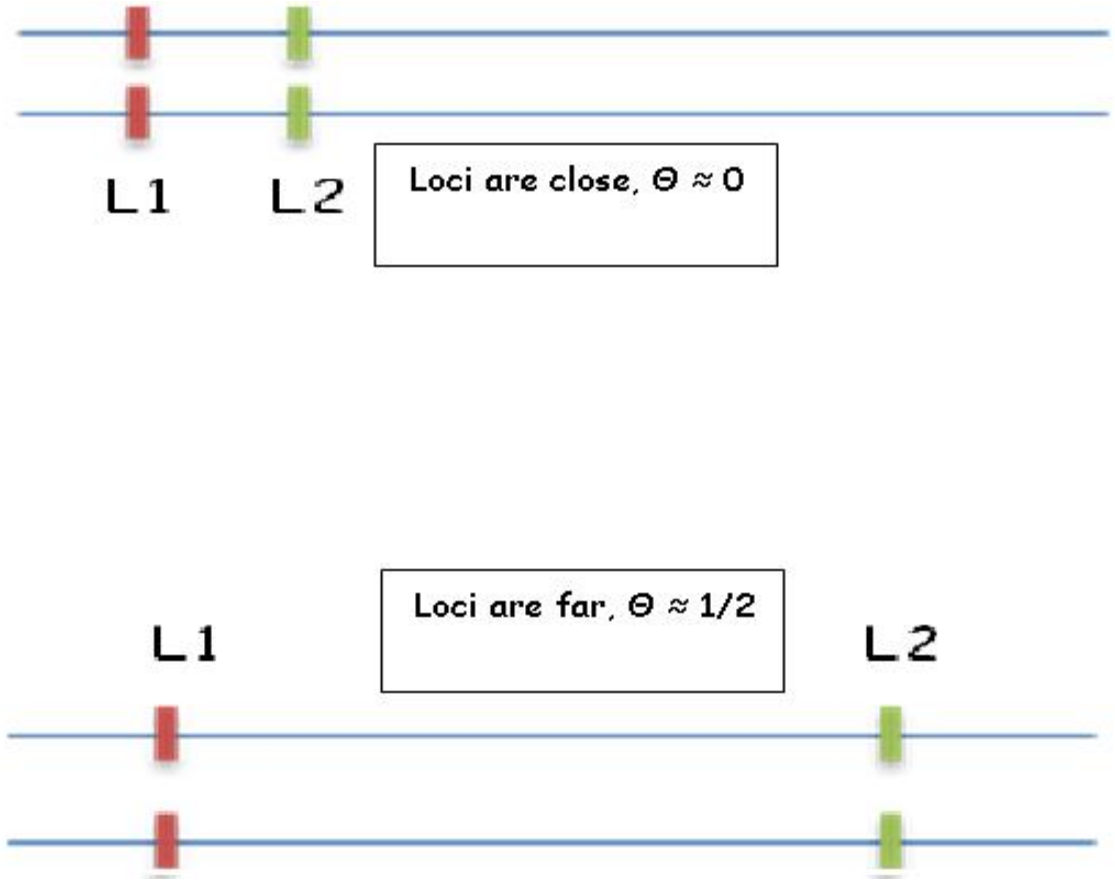


Figure 1.3 Recombination Fraction, [0, 1/2].

As described above, the end result of meiosis is four gametes, generally laid out as two recombinant chromosomes and the other two parental. Regardless of the number of crossovers and the distance, the maximum value of the recombination fraction between two loci is 0.50. Hence, Recombination fraction is bound within [0, 1/2].

Figure 1.3 shows the recombination fraction as a function of distance. The red and green stubs represent two loci (L1, L2) on a pair of homologous chromosomes. In the top part the two loci are close, so the recombination fraction is almost zero, but at the bottom of the figure the loci are far apart and the probability of observing a crossover is maximal, which is 0.50.

Recombination always starts with a crossover. Given two loci, L1 and L2, on a chromosome it is possible to observe the recombination between them only if there had been an odd number of crossovers (in an absence of additional information, such as more loci to observe between L1 and L2), as shown in Figure 1.4. This is because if there were an even number of crossovers the alleles at the two loci will be from the original chromatid again. Hence, we have to modify our calculation of the recombination fraction to reflect the fact that the probability of observing recombination in an interval equals the probability of an odd number of crossovers in that interval.

1.3 Physical Distance and Genetic Distance

Physical distance measures the actual length of the DNA as the number of nucleotides, also called bases (once the nucleotides in DNA pair with the correspondents on the other strand, they are often referred as base pairs). Considering the size of the DNA, a base is a rather small unit, and hence physical distances are usually measured in kilo-bases (one thousand bases) or even mega-bases (one million bases).

Genetic distance, on the other hand, is a more complex method of quantifying the genome. It is defined as the expected number of crossovers occurring on a single chromosome between two loci. We shall represent it with ‘d’. Genetic distance is measured in centi-Morgans

(cM), where, 1 cM is the distance between two loci where the crossover rate is 1%. It is mportant to note that genetic distance is not equal to physical distance.

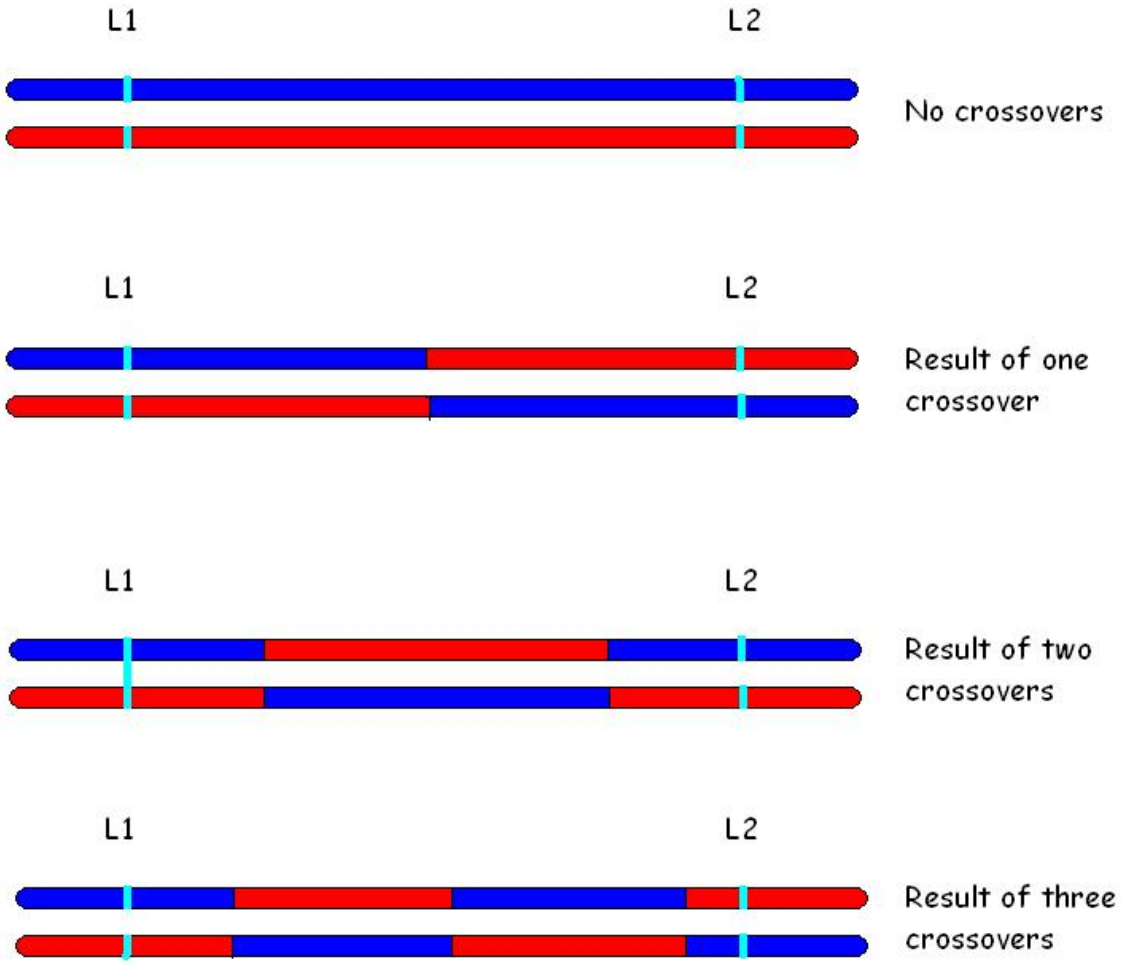


Figure 1.4 Only Odd Number of Crossovers Can be Identified.

1.4 Genetic Markers

A genetic marker is defined as a gene or other identifiable section of DNA whose inheritance can be followed [http://genomics.energy.gov/]. Genetic markers are chosen according to different criteria, depending on their intended use and technological constraints. As an example, Figure 1.5 (Source: http://www.informatics.jax.org/mgihome/nomen/gene.shtml) shows a sample distribution of markers over a small segment from 8 cM to 10 cM of chromosome 1. The numbers represent genetic distances between the markers in centi-

Morgans, and marker names are shown on the right, in blue color. A key feature of mouse nomenclature is the laboratory code, which usually consists of three to four letters, which identifies a particular institute, laboratory, or investigator who produced the DNA marker.

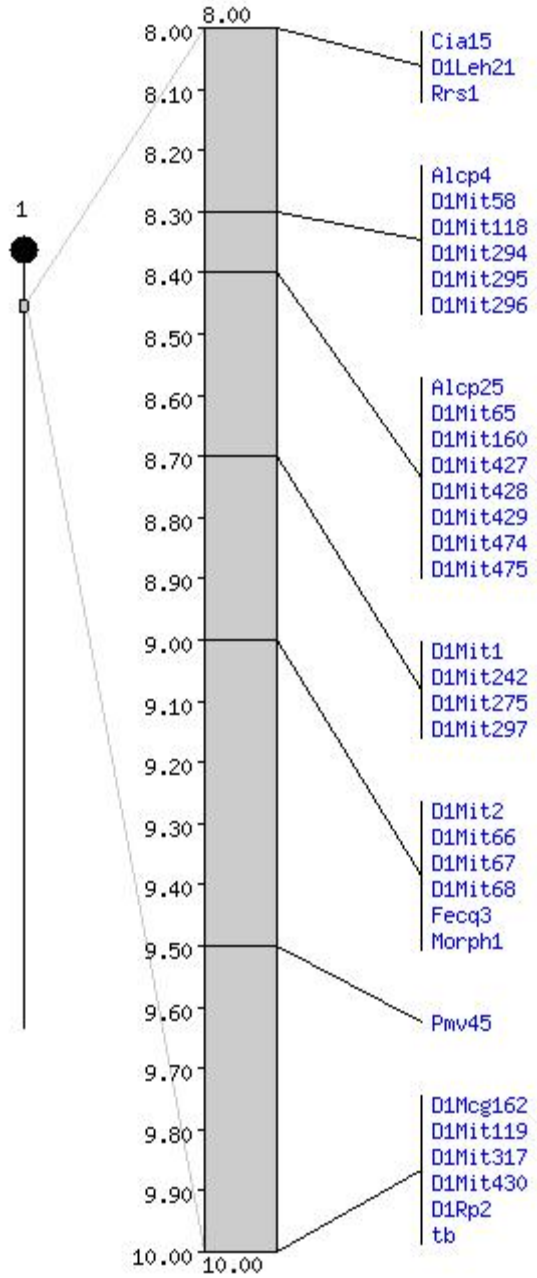


Figure 1.5 Magnified View of the Distribution of Markers on a Small Segment of Mouse Chromosome 1.

For example, in marker D1Mit58, ‘D’ stands for DNA segment, ‘1’ for Chromosome 1, ‘Mit’ stands for the lab in which it is obtained, Massachusetts Institute of Technology in this case, and 58 is a number identifying this particular marker.

Some of the common types of markers are:

- Restriction Fragment Length Polymorphism (RFLP): These markers are detected by a gel electrophoresis separation of DNA fragments obtained by digesting DNA from different individuals with a restriction enzyme [11].
- Short Tandem Repeats or Microsatellites (STR): These markers are detected by identifying highly polymorphic short repeats. For instance, $(ca)_n$ would stand for cacacacacacaca... with "ca" repeated *n* times.
- Single Nucleotide Polymorphisms (SNP): This technique relies on detecting a single different nucleotide from the sequence (such as G/A variation in "ATCGTAC" as one allele, and "ATCATAC" as the other).

1.5 Crossover Interference

Crossover interference is defined as the non random placement of crossovers, relative to one another, along chromosomes in meiosis [9]. It is a natural mechanism which prevents crossovers form occurring in close proximity to one another. The ratio between the number of double crossovers expected without interference and the number actually observed in the presence of interference is termed as the index of interference [25]. Moreover, interference regulates crossovers across the chromosome such that each pair of chromosomes has at least one crossover and multiple crossovers are evenly spaced [26]. Although the existence of a mechanism which affects the chances of crossovers occurring in the vicinity of each other is known, its molecular basis is presently not understood.

1.6 Genetic Linkage and Linkage Maps

If two genes are physically close on a chromosome then they usually segregate together, so the probability of observing particular combinations of their alleles is very high.

Such genes are said to be linked, and groups of such genes are called linked genes. However, if two genes are far apart, then the probability of them being separated due to recombination is high. A recombination taking place between such genes results in a hybrid chromosome. Linkage is measured in terms of the recombination frequency, which is the ratio of the number of recombinants to the total number of progeny and its unit of measurement is centi-Morgan, as described above.

Linkage maps are generated by counting the number of offspring which receive either parental or recombinant allele combinations from a parent who carries different alleles at two or more loci. Such map is also known as a recombination map, sometimes wrongly termed as genetic map (a genetic map usually includes chromosomal details as well as a physical map [11]). Mapping is a critical tool for many different areas of biological and medical research.

CHAPTER 2

RESOURCES AND TECHNIQUES

2.1 Backcross

A backcross is defined as a cross between a hybrid of two pure breeds mated with one of the parental strains. It provides a mechanism for the simplest form of linkage analysis, as only one parent is heterozygous (i.e. has different alleles on the chromosome pair) at each locus, and the other parent is homozygous (identical alleles on the chromosome pair) at these same loci.

Alleles can be of two types, dominant and recessive. A study of a backcross can be simplified (and made cheaper) if the right combination of homozygous and heterozygous parents is selected. The homozygous parent should have a pair of recessive alleles and the heterozygous parent should have a dominant allele and a recessive allele. In this scenario, we can be sure that if the offspring has the dominant allele then it has been transmitted from the heterozygous parent. The other case implies that the recessive allele has been transmitted from the heterozygous parent.

The backcross greatly simplifies the interpretation of genetic data because it allows one to jump directly from the genotypes of offspring to the frequencies with which different meiotic products have been formed by the heterozygous parent [11]. All mapping panels we have used for our research are backcross panels collected from various sources.

2.2 Mapping Panel Data

A screenshot of a part of mouse chromosome 1 DNA mapping panel data is shown in Figure 2.1 (Source: http://www.informatics.jax.org/). This panel, JAX (BSB), has been obtained from the Jackson Laboratory [http://www.jax.org/] in Bar Harbor, Maine. This particular panel has been generated by backcrossing a first filial generation (F1) of a cross between C57BL/6J

mouse strain and *M. spretus*, with a pure breed C57BL/6J mouse. Technically this backcross is represented as (C57BL/6J x M. spretus) F1 x C57BL/6J.

Marker	Animal #											
	1111112222 CDEFGHABCD	2222333333 EFGHABCDEF	3344444444 GHABCDEFGH	5555555566 ABCDEFGHAB	6666667777 CDEFGHABCD		7777888888 EFGHABCDEF	8899999999 GHABCDEFGH	1111111111 0000000011 ABCDEFGHAB	1111111111 1111112222 CDEFGHABCD	1111 2222 EFGH	
D1Mit427	bsbbbbbbbs	bbssbsbbsb	sbbbbsbssbb	ssssbbssss	sbbbbsbbbs	D1Mit427	sssbsbsbbs	sssbsbbsss	sbbsssbssb	ssbbbssssb	bbsb	D1Mit427
0/94						0/94						0/94
D1Mit475	bsbbbbbbbs	bbssbsbbsb	sbbbbsbssbb	ssssbbssss	sbbbbsbbbs	D1Mit475	sssbsbsbbs	sssbsbbsss	sbbsssbssb	ssbbbssssb	bbsb	D1Mit475
1/94				x		1/94						1/94
D1Mit58	bsbbbbbbbs	bbssbsbbsb	sbbbbsbssbb	ssbsbbssss	sbbbbsbbbs	D1Mit58	sssbsbsbbs	sssbsbbsss	sbbsssbssb	ssbbbssssb	bbsb	D1Mit58
0/94						0/94						0/94
D1Mit65	bsbbbbbbbs	bbssbsbbsb	sbbbbsbssbb	ssbsbbssss	sbbbbsbbbs	D1Mit65	sssbsbsbbs	sssbsbbsss	sbbsssbssb	ssbbbssssb	bbsb	D1Mit65
0/94						0/94						0/94
D1Mit167	bsbbbbbbbs	bbssbsbbsb	sbbbbsbssbb	ssbsbbssss	sbbbbsbbbs	D1Mit167	sssbsbsbbs	sssbsbbsss	sbbsssbssb	ssbbbssssb	bbsb	D1Mit167
0/94						0/94						0/94
D1Mit296	bsbbbbbbbs	bbssbsbbsb	sbbbbsbssbb	ssbsbbssss	sbbbbsbbbs	D1Mit296	sssbsbsbbs	sssbsbbsss	sbbsssbssb	ssbbbssssb	bbsb	D1Mit296
0/94						0/94						0/94
D1Mit316	bsbbbbbbbs	bbssbsbbsb	sbbbbsbssbb	ssbsbbssss	sbbbbsbbbs	D1Mit316	sssbsbsbbs	sssbsbbsss	sbbsssbssb	ssbbbssssb	bbsb	D1Mit316
0/94						0/94						0/94
D1Mit428	bsbbbbbbbs	bbssbsbbsb	sbbbbsbssbb	ssbsbbssss	sbbbbsbbbs	D1Mit428	sssbsbsbbs	sssbsbbsss	sbbsssbssb	ssbbbssssb	bbsb	D1Mit428
0/94						0/94						0/94
D1Mit429	bsbbbbbbbs	bbssbsbbsb	sbbbbsbssbb	ssbsbbssss	sbbbbsbbbs	D1Mit429	sssbsbsbbs	sssbsbbsss	sbbsssbssb	ssbbbssssb	bbsb	D1Mit429
3/94					x x	3/94			x			3/94
D1Mit2	bsbbbbbbbs	bbssbsbbsb	sbbbbsbssbb	ssbsbbssss	bbbsbsbbbs	D1Mit2	sssbsbsbbs	sssbsbbsss	bbbsssbssb	ssbbbssssb	bbsb	D1Mit2
0/94						0/94						0/94
D1Mit64	bsbbbbbbbs	bbssbsbbsb	sbbbbsbssbb	ssbsbbssss	bbbsbsbbbs	D1Mit64	sssbsbsbbs	sssbsbbsss	bbbsssbssb	ssbbbssssb	bbsb	D1Mit64
0/94						0/94						0/94
D1Mit67	bsbbbbbbbs	bbssbsbbsb	sbbbbsbssbb	ssbsbbssss	bbbsbsbbbs	D1Mit67	sssbsbsbbs	sssbsbbsss	bbbsssbssb	ssbbbssssb	bbsb	D1Mit67
3/94	x			xx		3/94						3/94
D1Mit430	ssbbbbbbbs	bbssbsbbsb	sbbbbsbssbb	bbbsbbssss	bbbsbsbbbs	D1Mit430	sssbsbsbbs	sssbsbbsss	bbbsssbssb	ssbbbssssb	bbsb	D1Mit430
5/94	x x			x		5/94		x x				5/94
D1Mit168	sssbsbbbbs	bbssbsbbsb	sbbbbsbssbb	bbbsbbbsss	bbbsbsbbbs	D1Mit168	sssbsbsbbs	sssbssbsbs	bbbsssbssb	ssbbbssssb	bbsb	D1Mit168
1/94			x			1/94						1/94
D1Mit298	sssbsbbbbs	bbssbsbbsb	sbbbbsbbsbb	bbbsbbbsss	bbbsbsbbbs	D1Mit298	sssbsbsbbs	sssbssbsbs	bbbsssbssb	ssbbbssssb	bbsb	D1Mit298
1/94						1/94	x					1/94
D1Mit4	sssbsbbbbs	bbssbsbbsb	sbbbbsbbsbb	bbbsbbbsss	bbbsbsbbbs	D1Mit4	sssssbsbbs	sssbssbsbs	bbbsssbssb	ssbbbssssb	bbsb	D1Mit4
1/94						1/94					x	1/94
D1Mit363	sssbsbbbbs	bbssbsbbsb	sbbbbsbbsbb	bbbsbbbsss	bbbsbsbbbs	D1Mit363	sssssbsbbs	sssbssbsbs	bbbsssbssb	ssbbbssssb	sbsb	D1Mit363
0/94						0/94						0/94
D1Mit526	sssbsbbbbs	bbssbsbbsb	sbbbbsbbsbb	bbbsbbbsss	bbbsbsbbbs	D1Mit526	sssssbsbbs	sssbssbsbs	bbbsssbssb	ssbbbssssb	sbsb	D1Mit526
1/94	x					1/94						1/94
D1Mit70	sssssbbbbs	bbssbsbbsb	sbbbbsbbsbb	bbbsbbbsss	bbbsbsbbbs	D1Mit70	sssssbsbbs	sssbssbsbs	bbbsssbssb	ssbbbssssb	sbsb	D1Mit70
0/94						0/94						0/94
D1Mit411	sssssbbbbs	bbssbsbbsb	sbbbbsbbsbb	bbbsbbbsss	bbbsbsbbbs	D1Mit411	sssssbsbbs	sssbssbsbs	bbbsssbssb	ssbbbssssb	sbsb	D1Mit411
0/94						0/94						0/94

Figure 2.1 Screenshot of a Mapping Panel.

This panel has been constructed with genotypes of 94 animals. The ‘s’ and ‘b’ genotypes indicate that the allele is from M. spretus and C57BL/6J animals, respectively, and ‘.’ indicates a missing genotype. The genotypes are arranged in a matrix format with the animal identification names as the column headers and the genetic markers as the row headers. The marker names are repeated in the center and at the end for better visualization of the panel. In order to explain the matrix fields let us consider the first row, animal number ‘1c’. This row

indicates that at the loci of the first twelve markers, from D1Mit427 to D1Mit67, the offspring ‘1c’ has inherited the ‘b’ allele form the C57BL/6J parent. Between the 13th and 14th marker locus a crossover has occurred, and hence the genotypes from the 14th marker locus onwards is an allele from (C57BL/6J x M. spretus) F1 parent. The 13th and 14th markers, D1Mit67 and D1Mit430, are the markers which flank the crossover site and they are hence called flanking markers. Also, there is a "3/94" note between these two markers, indicating that 3 out of 94 animals have experienced a recombination in this particular DNA segment. Moreover, if there were to be another crossover down the line, then the genotype would get switched back from ‘s’ to ‘b’. The rest of the panel is read in the same fashion. In the next chapter we shall discuss all the panels we have used for our research.

2.3 Mapping Functions

It is intuitive to imagine the recombination events as independent of each other. A direct consequence of this assumption would be to expect a linear relationship between the genetic distance and recombination frequency, in fact a one-to-one correspondence. This would imply that the recombination fraction should be equal to the mapping distance. However, this is true only for genetic distances less than 10cM. For distances greater than 10 cM, it has been found that the relationship between genetic distance and the recombination fraction does not stay linear any more. This degeneration is due to the fact that the probability of multiple crossovers also increases with increasing genetic distances and, since we can only detect odd number of crossovers between any two loci, we would end up missing any even number of crossovers that might have occurred between them. As a result, the observed number of crossovers will be less than the actual number of crossovers [11].

For genetic distances greater than 10 cM one has to use mathematical equations called map functions, in order to establish a relationship between the genetic distance and recombination fractions. A genetic map function ‘M’ is defined as one which establishes a relation $r = M(d)$, connecting the recombination fractions ‘r’ and genetic map distances ‘d’ between pairs of loci along a chromosome arm [10].

We shall now discuss a slightly generalized and rather simple mapping function, in order to illustrate the method by which mapping functions are created. We also briefly discuss a mapping function proposed in [27] and its two special cases, Kosambi [13] and Haldane [12]. The derivation of the mapping function we have used in our research, the Carter-Falconer mapping function [15], is rather complex and out of the scope of this document, however the following discussion should give a good picture of mapping functions in general. We shall first define a few necessary terms.

Since double recombinations cannot be detected between two loci, a three point cross experiment is considered. A three point cross is a cross involving three loci. Let *r(x)* be the recombination fraction over an interval '*x*', whose length is measured in Morgans. Let *dx* represent an infinitesimally small genetic length and *r(dx)* be the recombination fraction over the length *dx*. Since the recombination fraction and genetic distance feature a one-to-one correspondence for small genetic intervals, and *dx -> 0*, we can say that *r(dx) = dx*. Let *c(x,dx)* be the coefficient of confidence over the interval *(x,dx)*. The coefficient of confidence is defined as a ratio between the expected and observed number of double crossovers (DCO). Crossover interference hampers the occurrence of double crossovers according to the mathematical relation *i = 1 – c*, where, *i* represents interference and *c* represents the coefficient of confidence, the ratio of the observed number of DCOs with the theoretical number of DCOs. If $x < 10$ cM then $r(x+dx) = r(x) + r(dx)$, but, if $x > 10$cM the we have to calculate $r(x+dx)$ by considering the possibility of a double crossover in the interval.

The following derivation [27] is direct extension of the Kosambi mapping function [13]. It is a simple mapping function with just one parameter (confidence coefficient) to determine the effect of interference. This mapping function is derived by adding a very short interval *dx* to an interval *x* and finding the recombination fraction for the total interval.

This is done as follows;

$$r(x+dx) = r(x) + r(dx) - (observed\ number\ of\ DCO) \quad (1)$$

But $c(x, dx) = (\text{observed number of DCO}) / (\text{theoretical number of DCO})$ (2)

and, *theoretical recombination fraction* $= r(x)r(dx)$ and hence;

theoretical number of DCO $= 2r(x)r(dx)$ (3)

Plugging (3) in (2), we get;

observed number of DCO $= c(x, dx)\left(2r(x)r(dx)\right)$ (4)

Plugging (4) in (1), we get;

$$r(x + dx) = r(x) + r(dx) - c(x, dx)\left(2r(x)r(dx)\right) \quad (5)$$

Since, $dx << 10cM$; we have $r(dx) = dx$;

This mapping function makes the assumption of linearity in *r(x),* and assumes that when *x* is small, *C(x, dx)* = *K,* an arbitrary constant and when *x* is large, *C(x, dx)* must

be *1,* which then results in the following differential equation;

$$dr/dx = 1 - 2\,K\,r + 4\,(K - 1)\,r^2$$

Solving this differential equation results in the following mapping function;

$$x = \frac{1}{2(K-2)} \ln\left(\frac{1-2r}{1-2(K-1)r}\right)$$

and the inverse of this function is given by,

$$r = (1/2)\left(\frac{1 - e^{2(K-2)x}}{1 - (K-1)e^{2(K-2)x}}\right)$$

Figure 2.2 shows a graph of the mapping function for different values of '*K*'. It is worth noting that when K = 1 the equation (7) results in the Haldane mapping function and when K = 0 in equation (6), it results in the Kosambi mapping function.

Some commonly used mapping functions are listed below:

The Haldane map function: assumes no crossover interference [12]

$$r = \frac{1}{2}\left(1 - e^{-2m}\right)$$

The Kosambi mapping function: similar to the level of interference in humans [13]

$$m_k = \frac{1}{4}\left[\ln(1+2r) - \ln(1-2r)\right]$$

The Carter-Falconer mapping function: similar to the level of interference in mice [15]

$$m_{FC} = \frac{1}{4}\left\{\frac{1}{2}\left[\ln(1+2r) - \ln(1-2r)\right] + \tan^{-1}(2r)\right\}$$

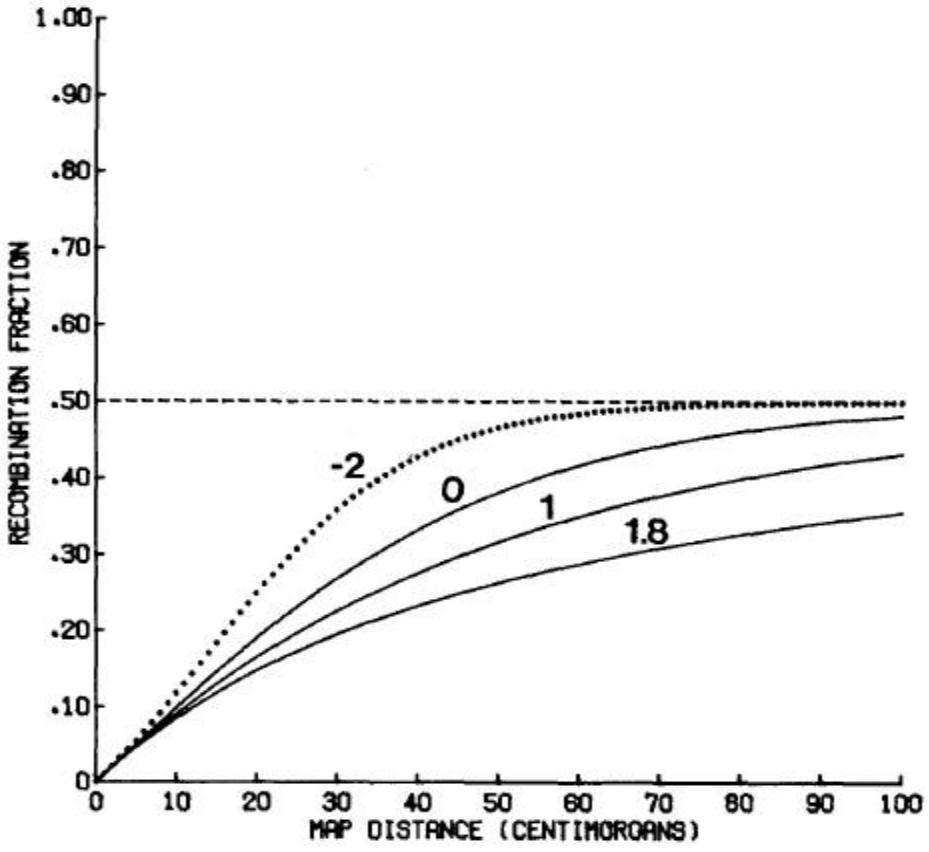


Figure 2.2 Recombination fraction as a function of map distance [27].

2.4 Recombination fractions using binomial distribution

We start with a simplistic assumption that there is at most one crossover between $(i-1)^{th}$ and i^{th} loci, where 'i' represents the index of a locus, when they are discretised and sequentially arranged. We then define the following:

D_i= 1 implies that a crossover occurs between locus i-1 and locus i

D_i= 0 implies that a crossover does not occur between i-1 and i

We now define the probability of a crossover between the first two loci as 'p_0' and hence the probability of no crossover between the same loci is '1 - p_0'. Thus:

$$P(D_i = 1) = p_0$$

$$P(D_i = 0) = 1 - p_0$$

With this setup we can use the binomial distribution to find the recombination fractions as shown below: [http://coe.math.keio.ac.jp/2005/cherrybud/pdffiles/sugaya.pdf]

$$\theta(p_0, n) = P\left(\sum_{i=1}^{n} X_i = odd\right)$$

$$= \sum_{k:odd} \binom{n}{k} p_0^k (1 - p_0)^{n-k}$$

$$= \frac{1}{2}\left\{1 - (1 - 2p_0)^n\right\}$$

2.5 Markov property

We will use the same notation and assumptions as above. Since crossovers are not independent events we can use the principles of conditional probability, and define the following four equations. [http://coe.math.keio.ac.jp/2005/cherrybud/pdffiles/sugaya.pdf].

These considerations have led to algorithms based on Hidden Markov Models, capable of estimating the recombination fractions when some genotypes are missing.

$$P(D_i = 1) = p_0$$
$$P(D_i = 0) = 1 - p_0$$
$$P(D_i = 1 \mid D_{i-1} = 0) = p_1$$
$$P(D_i = 0 \mid D_{i-1} = 0) = 1 - p_1$$
$$P(D_i = 1 \mid D_{i-1} = 1) = q_1$$
$$P(D_i = 0 \mid D_{i-1} = 1) = 1 - q_1$$
$$\Theta = P\left(\sum_{i=1}^{n} D_i = \text{odd}\right)$$

Where, P (x | y) represents the conditional probability of the event 'x' occurring given that 'y' has already occurred and, $P(x \mid y) = P(x \cap y) / P(y)$.

2.6 R/QTL

R/Qtl [30] is an extensible, interactive environment for mapping quantitative trait loci (QTLs) in experimental crosses. It is implemented as an add-on package for the statistical language/software R [http://www.r-project.org/]. R is a programming language and environment designed for statistical computing and graphics. It is an integrated suite of software facilities for data manipulation, calculation and graphical display. The R/Qtl library is built on the R platform to take advantage of the basic mathematical and statistical functions, and its powerful graphics capabilities. For our research purposes, we installed the R and R/Qtl packages on our Linux server and also made it available to our collaborators at the Biology Department.

We opted for the R/Qtl package to exploit its statistical and graphical functions provided by R in general, and the built in functions to estimate genetic maps provided in the R/Qtl library in particular. The most important component of this tool for estimating genetic maps is the Lander-Green algorithm [14] which uses the hidden Markov model [31] technology for dealing with missing genotype data.

The key features of the R/Qtl package pertaining to our project are its capability of dealing with missing genotype data and allowance for the presence of genotyping errors, for backcrosses.

CHAPTER 3

DATA COLLECTION AND ANALYSIS

To study the recombination rates we first gathered eight mapping panel datasets, all of backcross animals. Backcross greatly simplifies the interpretation of genetic data because it allows us to directly interpret the recombination fractions from the genotypes of offspring, as discussed above. We looked at the recombination rates in all chromosomes except Y, as our goal was to pursue the gene which has female-specific effect on recombination levels.

Due to the amount of information, it was important that we automate the process of data collection as much as possible. We have thus written several Perl scripts to crawl the web and fetch the reference genetic distance and the physical distance of each marker. We have assembled the panels in the format required by R/QTL, which we used to re-calculate the genetic distances based on the data from our panels only. We then estimated the recombination fractions between all pairs of markers. Since we only used backcrosses, we simply had to count just the number of recombination events.

Our goal was to build genetic maps specific to the genotypes of the animals in the selected panels, as opposed to the reference data which was already available, averaged over many animals and thus having only a rough fit to any given species of mice. For this purpose, we used the Lander-Green algorithm based on the Hidden Markov Model technology [14]. The details of the functions we have used for this purpose are given in the following section. In order to accurately estimate the genetic maps, we used the Carter-Falconer mapping function, as it fits the level of crossover interference in mice the best [15]. We have estimated the genetic maps in order to compare the corresponding genetic distances with the actual physical intervals between the markers, and hence arrive at recombination rates.

We compared the genetic distances with the corresponding physical distances by taking their ratio. We have calculated this ratio for all mouse autosomes (1 through 19), and the

X-chromosome. We gathered the recombination rates for all the chromosomes from all eight panels. Equipped with these data, our goal was to find the distributions of the rates of recombination for each chromosome and compare them across all the eight panels, in an effort to establish the variability of recombination rates among the eight different backcrosses.

Since genotype data rarely span the entire lengths of the chromosome arm, we calculated the physical distance between the first and the last marker for each chromosome and attempted to estimate the amount of data we were missing. When we compared the physical distance between the markers flanking each chromosome in each panel to the actual size of the corresponding chromosomes, we determined that we were missing a significant amount of data from the ends of some chromosomes. This presented a substantial problem, as it is known that the recombination rates are significantly higher in the telomere region of mouse chromosomes [16]. There was little point in estimating the distributions of recombination rates based on the gathered data as further analysis based on these panels would probably lead to flawed analysis. However, we could still use the gathered information for other types of analysis towards our general goal, that of modeling the recombination interference.

To model the interference we looked at the double recombinants in the same set of mapping panels. Since it is not possible to know the exact position of the crossover we can only isolate the interval between two markers in which it had occurred. Similarly, for double recombinations we can only isolate the flanking markers of both crossovers. To gather the intervals, we have obtained the physical distances of the markers flanking the double-recombination events. For each chromosome featuring two recombination events, we obtained four numbers corresponding to the positions of the flanking markers. For a given haplotype, if the first recombinant falls between the distances (u1 , v1) and the second between (u2 , v2), it implies that the inter-recombination's distance, 'd', must be bound within the interval (l , r) = (u2 - v1 , v2 - u1). Such data are said to be double-censored [21]. After we obtained the bounding intervals for all double recombinants, we used graph theory to simplify the problem of estimating the distributions of distances between the crossover events. Censored data can be

represented as intersection graphs, by first representing every double recombination interval as a vertex and then connecting the intersecting intervals by an edge. Once we obtained the intersection graphs for all panels, we identified the maximal cliques in these graphs. If V is the set of all vertices in a graph then a clique C is defined as a subset of V such that every vertex in C is connected to every other vertex in C. A clique is said to be maximal if it is not a subset of another clique. In case of univariate (i.e. our case, since we were only interested in the distribution of interval lengths) censored data every interval will belong to at least one maximal clique. For the maximum likelihood estimation, all that matters are the sets of intervals which participate in these maximal cliques, and not the actual representation of the cliques [21]. After we obtained the maximal cliques for all datasets, we represented them as clique matrices. A clique matrix is a matrix in which each vertex participating in the corresponding maximal clique is represented by '1' and the the others are represented by '0'. In the context of interval graphs, clique matrices are also known as Petrie matrices [28].

We have built a C# program to identify all double recombinants and to extract the corresponding markers. We have used our Perl scripts to automatically fetch the physical distances for the necessary markers.

3.1 Methods and Algorithms for Data Collection

In this section we describe the methods involved in data collection, preprocessing and estimating genetic maps and estimating the recombination rates.

In our study we have assembled all panels of reasonable quality which could be collected in the public domain, a total of eleven, and here we provide their list. Unfortunately, we could not use the panels form the Mouse Genome Informatics source (list number 2), as the physical distances for most of their markers were unavailable. The data sources for the panels are mentioned in the header. Any modifications to the original panels are described for the ones which needed changes.

1. Center for Genome Dynamics at the Jackson Laboratory (http://cgd.jax.org/):
    - BALB X (BALB X C57BL/6)

- (C57BL/6 x C3H) x C57BL/6 (Modifications: Removed unfinished genotypes Rows 41 – 203)
- (C3H X C57BL/6) X C3H (Modifications: Removed unfinished genotypes Rows 53 – 259)
- (SM X NZB) x NZB (Modifications: Removed unfinished genotypes Rows: 3, 4, 8-10,15-17, 21, 22, 29, 30, 37, 38, 40, 42, 44, 51, 56, 61, 62, 66-2, 74, 77, 78, 81-86)

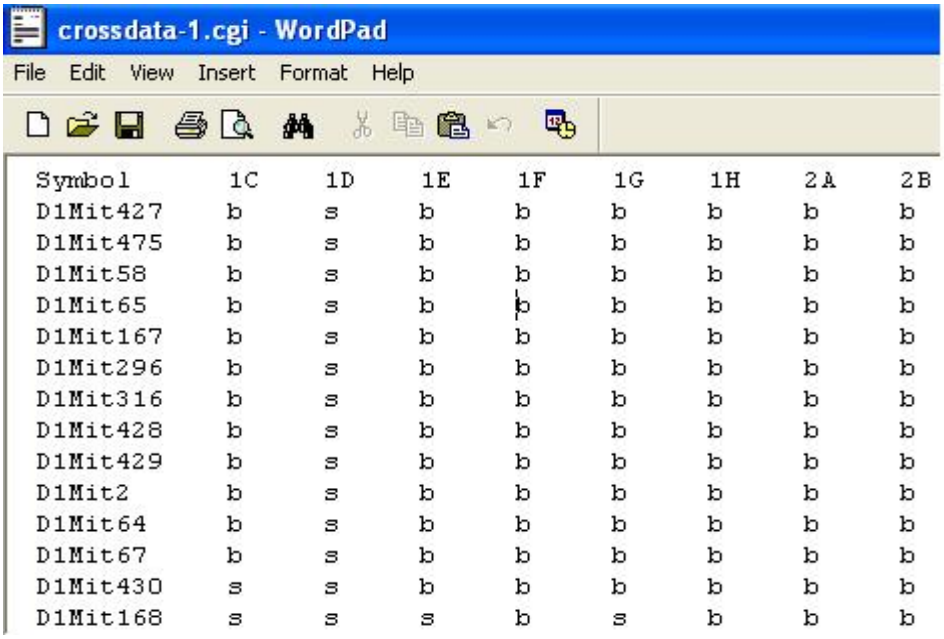
2. Mouse Genome Informatics (http://www.informatics.jax.org/)
    - JAX-BSB (C57BL/6J x M. spretus)F1 x C57BL/6J
    - JAX-BSS (C57BL/6JEi x SPRET/Ei)F1 x SPRET/EiJ
    - JAX-BCB Mouse Mutant Resource (C57BL/6J x CAST/Ei) F1 x C57BL/6J
3. Folami Datasets: [23] B2.xls: Genotypes of [(PERC x DDK) F1 x C57BL/6] N2 females, chr 1-X
    - B11-9.xls: Genotypes of [(PERA x DDK)F1 x C57BL/6]N2 females, chr 1-9
    - B110-X.xls: Genotypes of [(PERA x DDK)F1 x C57BL/6]N2 females, chr 10-X

Less and more skewed: [21]

- Cross: (C57BL/6.Cg-Pgk1a x DDK) x C57BL/6 (less XCI skewed)
- Cross: (C57BL/6.Cg-Pgk1a x DDK) x C57BL/6 (more XCI skewed)

Most of the mapping panels were in Comma Separated Values (CSV) format or tab delimited format with the first column specifying the list of markers and the first row providing animal IDs. The rest of the data are the corresponding genotypes, as shown in Figure 3.1.

For our purposes we needed to transpose the above format and record the chromosome number and genetic distances as the second and the third row, respectively. We have also populated the datasets with the genetic distances obtained for each marker, and we show a screenshot of the modified panel data in Figure 3.2.



Symbol	1C	1D	1E	1F	1G	1H	2A	2B
D1Mit427	b	s	b	b	b	b	b	b
D1Mit475	b	s	b	b	b	b	b	b
D1Mit58	b	s	b	b	b	b	b	b
D1Mit65	b	s	b	b	b	b	b	b
D1Mit167	b	s	b	b	b	b	b	b
D1Mit296	b	s	b	b	b	b	b	b
D1Mit316	b	s	b	b	b	b	b	b
D1Mit428	b	s	b	b	b	b	b	b
D1Mit429	b	s	b	b	b	b	b	b
D1Mit2	b	s	b	b	b	b	b	b
D1Mit64	b	s	b	b	b	b	b	b
D1Mit67	b	s	b	b	b	b	b	b
D1Mit430	s	s	b	b	b	b	b	b
D1Mit168	s	s	s	b	s	b	b	b

Figure 3.1 Screenshot Depicting the Raw Data.

	I	J	K	L	M	N	O	P	Q
1	D1Mit318	D1Mit445	D1Mit117	D2Mit237	D2Mit94	D2Mit307	D2Mit456	D3Mit21	D3Mit49
2	1	1	1	2	2	2	2	3	3
3	18.5	92.3	106.1	28	48.1	74.9	86.3	19.2	41
4	H	H	H	H	H	H	A	H	H
5	A	A	A	H	H	H	H	A	A
6	A	A	A	A	A	H	H	A	H
7	A	H	H	H	H	A	A	H	H
8	H	H	A	H	H	H	H	A	H
9	A	H	H	A	A	A	A	A	H
10	A	A	A	A	A	A	H	H	A
11	A	A	A	A	A	A	A	H	H
12	H	H	A	A	A	A	H	H	H
13	A	A	H	A	A	A	H	A	H
14	A	A	A	A	A	A	H	A	H
15	A	H	A	H	H	H	A	H	H
16	H	H	H	H	A	A	A	H	H
17	A	H	H	A	A	A	A	H	H
18	H	H	H	A	A	A	A	H	H
19	A	A	H	H	A	A	A	A	A
20	H	H	A	H	H	H	H	A	A

Figure 3.2 A Screenshot of the Modified Panel.

The genetic distances for each marker were obtained from Mouse Genome Informatics (MGI) web site [http://www.informatics.jax.org/]. We have used Perl scripts to crawl the MGI site and parse out the genetic distances. For this case, we generated a list of URLs by appropriately inserting the marker name to the URL of the Genes and Markers Form in the MGI website. The crawler script then serially goes to each of these web pages listed as URLs in the input file and parses out the corresponding genetic distances. Some of the marker names in our list represent syntenic genes (genes that occur on the same chromosome are called syntenic genes). For such markers the Genes and Markers Form does not provide any information about their physical distances. For a given webpage, if the script does not encounter any physical distance, it looks for the string "Syntenic". If it finds this string, it returns an appropriate message; otherwise, it returns a "Null" string. For our analysis, we simply eliminated the markers that represented syntenic genes.

The pseudo-code for the script implementing this strategy, cM_Grabber, is provided in Figure 3.3. In the pseudo code, italicized words are variables and strings in quotes are messages returned. The '+' is used to represent more than one returned string, a message and a value, for instance.

Input file: list of all URLs corresponding to each marker whose genetic distance is to be found.

Output file: list of marker names and corresponding genetic distance/error message.

Obtaining the physical distances for the flanking markers for each chromosome in each panel is a more complex task. For this, we have used the same strategy as in the cM_Grabber. We generated a URL file and crawled the MGI site and parsed the physical distance in base pairs, for all markers for which it was available. If the physical distance for a marker is not available, then the script will automatically look for the primers used to identify that particular genomic sequence. If the crawler failed to fetch the physical distance from the page represented by the URL, it looks for the string "PCR" and captures the hyperlink embedded in a number next to it. The crawler then goes to the web page represented by this URL and looks for a link embedded, this time, in the string "PCR". Now the script follows this new link to its

corresponding webpage and captures another link embedded in the marker name listed under the string “Probe”. Following this URL leads the crawler to the webpage containing the forward and reverse primers for that particular marker.

```
Start
For each URL in the list
do
        If URL not valid
                Return “URL not valid!” + marker, Continue..
                Go to the webpage specified by the URL
                Begin crawl (looking for cM position)
                If  centi Morgan position found
                        Return marker + cM position
                Elseif “Syntenic” found
                        Return  “Syntenic marker!” + marker
                Else
                        Return "Genetic distance not found!" + “Null”
End
```

Figure 3.3 Pseudo-code for the Script Implementing cM_Grabber.

The script captures these two primers and stores them in a different file with their corresponding marker name. A list of such primer pairs and their corresponding marker names are generated and stored serially.

Once the script captures all the physical distances using the first strategy, it will then automatically submit the forward and reverse primers from the primer list to the UCSC Genome Browser [29] for In-Silico PCR [http://genome.ucsc.edu/cgi-bin/hgPcr]. Here, the “Genome”, and “Assembly” fields are set, as defaults, to “Mouse” and “Feb. 2006” respectively. The “Max Product Size”, “Min Perfect Match” and “Min Good Match” were set to 4000, 15, and 15, respectively.

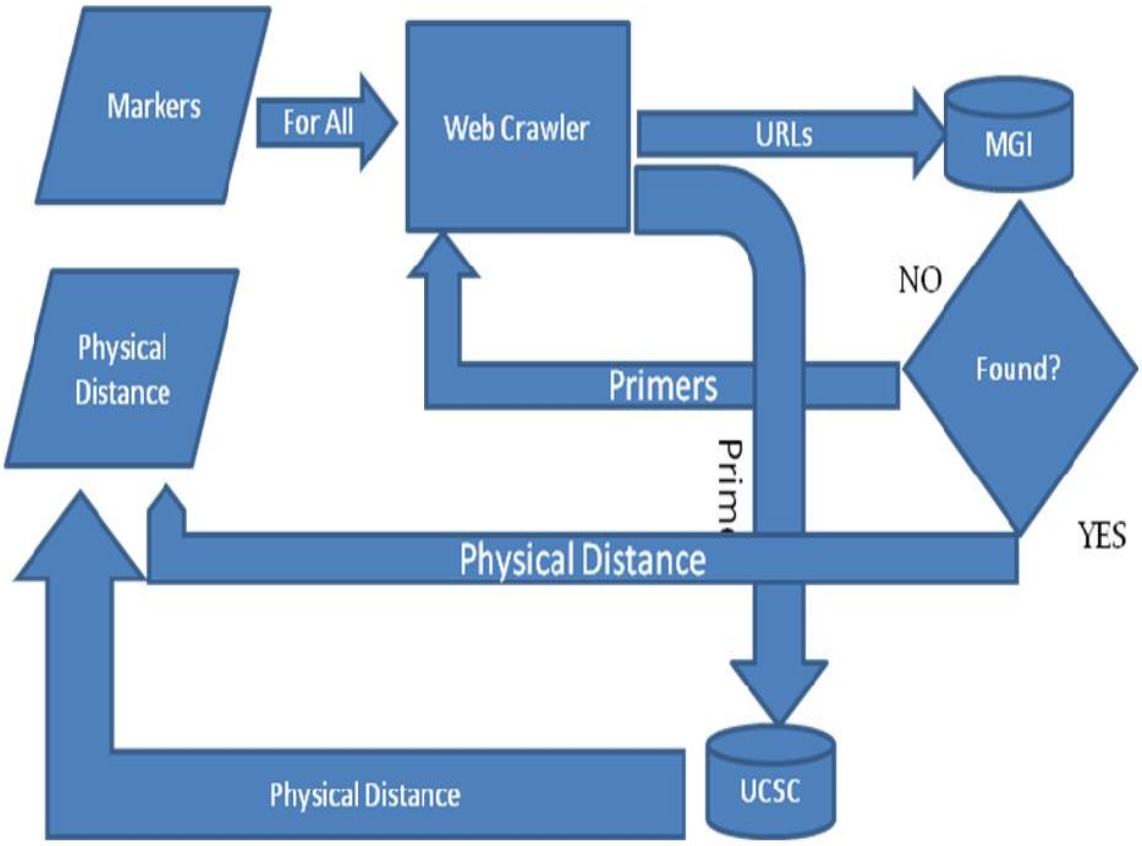


Figure 3.4 Flow Chart Representing the Collection of Physical Distances for all Markers.

If these parameters fail to generate a hit, then they are changed to 10000, 10, and 10, respectively. In case of a hit we flag it with an asterisk in the output file, so that we can manually check the output later. If the script fails to fetch the physical distance through ePCR, it will automatically submit the first primer to Blat [24]. If that generates a hit it will move to the next marker, else it submits the second primer. If both fail, the script returns the marker itself. The entire procedure is shown in Figure 3.4 as a flowchart.

3.2 Estimating Genetic Maps

To estimate the genetic maps, we have calculated the recombination fractions between all pairs of markers, using R/QTL functions. Since we used backcrosses only, it is sufficient to just count the number of recombination events. We then used the Lander-Green algorithm to re-estimate the genetic map based on the genotype data of the experimental cross. We set the genotyping error rate used in the calculation of the penetrance Pr (observed genotype | true genotype) to 0.001, for all panels. For the mapping function we used Carter-Falconer, as it best fits the level of interference in mice [15]. Figure 3.5 illustrates this process in a form of a flow chart.

We conclude this section with a detailed description of every step performed in estimating genetic maps based on the genotype data:

Step 1: We first include the QTL library in R. Then using the *read.cross* function we input a file formatted as mentioned earlier and convert it into an object of class *cross*. As parameters to this function we give the path to the file, file format, genotype characters, the character denoting the missing genotypes, and we also set another parameter, *convertXdata*, to TRUE. If this is set to TRUE, any X chromosome genotype data is converted to the internal standard, using columns ‘sex’ and ‘pgm’ in the phenotype data, if they are available, or by inference if they are not. If FALSE, the X chromosome data are read as is.

Sometimes, there can be more than one marker at the same position; we then use the function *jittermap* for adjustment. This function simply adds a small number to the positions of each of the overlapping markers, and hence makes them distinct.

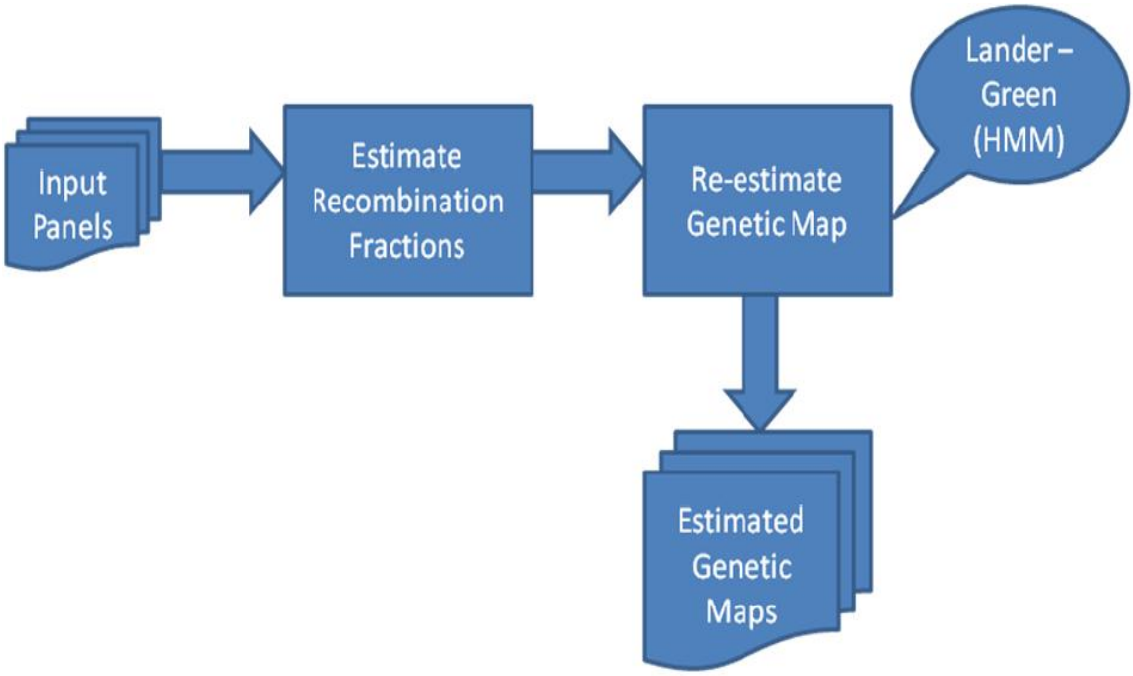


Figure 3.5 The Process of Estimating Genetic Maps.

Step 2: We use the est.rf function to estimate the recombination fraction between all pairs of markers. Since we are using backcross, this simply counts the number of recombination events.

Step 3: We then use the est.map function which implements the Lander-Green algorithm, in order to re-estimate the genetic map based on the genotype data of the experimental cross. For this function we use three parameters. First is the name of the map. Second, called error.prob, is the assumed genotyping error rate used in the calculation of the penetrance Pr (observed genotype given true genotype). The third, map.function, specifies the function used to map genetic distance to recombination fraction. We use the Carter-Falconer function as it best fits the level of interference in mice.

Step 4: We can use the plot.map feature to visualize the difference between the reference map based on the markers (averaged for many animals of different strains) and the

one estimated from the genotype data of a particular panel. This function shows the comparison chromosome by chromosome, as illustrated in Figure 3.6.

Step 5: The function summary (map name) outputs the final result as a table, as illustrated in Table 3.1.

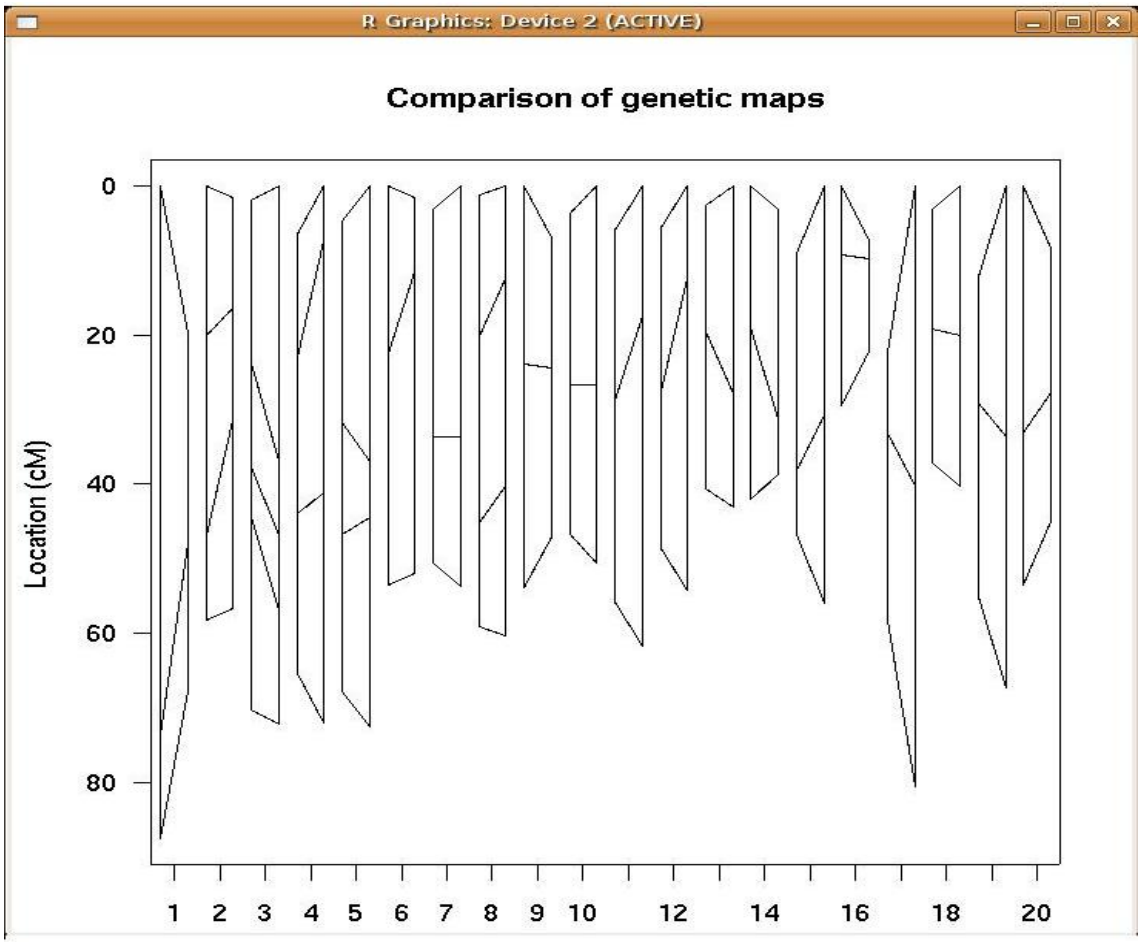


Figure 3.6 Comparing Genetic Maps Based on Reference Data and the Newly Estimated Values for a Particular Panel.

Table 3.1 The Estimated Genetic Map.

chro #	n.mar	length	ave.spacing
1	3	47.9	24
2	4	55.3	18.4
3	5	72.3	18.1
4	4	72	24
5	4	72.6	24.2
6	3	50.3	25.2
7	3	53.8	26.9
8	4	60.4	20.1
9	3	40.1	20.1
10	3	50.5	25.3
11	3	61.7	30.9
12	3	54.2	27.1
13	3	43.2	21.6
14	3	35.6	17.8
15	3	56	28
16	3	14.9	7.5
17	3	80.7	40.3
18	3	40.3	20.2
19	3	67.4	33.7
20	3	36.6	18.3
overall	66	1065.7	23.2

3.3 Estimating Recombination Rates

Once we have the genetic lengths for all the chromosomes, our next task is to obtain the recombination fractions. We calculate them for each chromosome and for each panel by taking the ratio of genetic distance in cM to the physical distance in MB (please see column cM/MB of Table 3.2). We also record the physical distances for the first and last marker of every chromosome, in columns MB_Start and MB_End, respectively. The column Length_MB gives the physical distance between the first and the last markers. This value should be compared with the last column, recording the chromosome size.

Table 3.2 An Example of Final Compiled Results.

Modifications: Removed unfinished genotypes (Rows 41 - 203) *Cross: (B6 X C3H) X B6*									
Cross BC_to_B6	Number of Markers	cM length	Average spacing	*Mb start*	*Mb end*	*Length Mb*	*cM/Mb*	*Chrom.Size m36*	*Data Covered %*
1	3	47.9	24	33.7	191.6	157.9	0.303	197.1	80.1
2	4	55.3	18.4	40.8	168.6	127.8	0.433	182.0	70.2
3	5	72.3	18.1	37.3	157.9	120.6	0.600	159.9	75.4
4	4	72	24	38.6	141.8	103.2	0.698	155.0	66.6
5	4	72.6	24.2	32.3	142.0	109.7	0.662	152.0	72.2
6	3	50.3	25.2	48.7	146.4	97.7	0.515	149.5	65.4
7	3	53.8	26.9	36.9	128.4	91.5	0.588	145.1	63.1
8	4	60.4	20.1	33.5	129.9	96.5	0.626	132.1	73
9	3	40.1	20.1	36.6	120.1	83.6	0.480	124.0	67.4
10	3	50.5	25.3	28.8	117.6	88.7	0.569	130.0	68.2
11	3	61.7	30.9	44.6	112.3	67.7	0.911	121.8	55.6
12	3	54.2	27.1	36.3	114.3	78.0	0.695	120.5	64.8
13	3	43.2	21.6	44.8	113.9	69.2	0.625	120.6	57.3
14	3	35.6	17.8	28.3	105.5	77.1	0.461	124.0	62.2
15	3	56	28	31.9	98.9	67.0	0.836	103.5	64.7
16	3	14.9	7.5	34.8	87.5	52.7	0.283	98.3	53.6
17	3	80.7	40.3	34.4	91.0	56.6	1.426	95.2	59.5
18	3	40.3	20.2	28.9	77.1	48.1	0.837	90.7	53.1
19	3	67.4	33.7	15.6	54.9	39.4	1.713	61.3	64.2
20	3	36.6	18.3	47.9	160.7	112.8	0.325	165.6	68.1
Overall	66	1065.7	23.2		Total=	1745.8	0.610	2628.2	66.4

3.4 Modeling Recombination Interference

To model the recombination interference we have looked for all double recombinants in all mapping panel datasets we have assembled. For each instance of double recombination we have located the physical distances of the flanking markers, i.e. for each chromosome with two recombination events, we recorded four numbers corresponding to the position of the flanking markers. For a given haplotype, if the first recombination happened between the positions u1 and v1, and the second between u2 and v2, it implies that the inter-recombination distance 'd',

falls in the interval (l, r) = (u2 - v1 , v2 - u1). In that case the distance d is said to be double-censored.

For doubly censored data we can use graph theoretic approach to simplify the problem of finding the nonparametric maximum likelihood estimation (NPMLE) of the underlying distribution. We represent the censored data as intersection graphs, then use a combinatorial algorithm to obtain the clique matrix.

We have used the same mapping panels as before, and captured all intervals of double recombination events, for every chromosome and for every panel. We have constructed the intersection graphs using all the double recombination intervals from each panel, then used the algorithm proposed by Gentleman and Vandal [22] to identify the maximal cliques in each graph. Based on these maximal cliques we have constructed Petrie matrices for the panels.

We conclude this section with a detailed description of every step performed in modeling the recombination interference based on the genotype data:

For each panel, do step 1 through 4.

Step 1. Record all intervals (L, R), where L is the minimal distance and R is the maximal distance between four flanking markers.

Step 2. Compare these intervals and form a graph with each interval as a vertex, and an edge connecting the vertices if and only if the corresponding intervals intersect.

Step 3. Identify all maximal cliques using the perfect elimination scheme, and then using the combinatorial algorithm described in [22].

Step 4. Form a Petrie/clique matrix based on the maximal cliques.

End

We now elaborate on the above steps based on example data.

Step 1: Let us consider the following intervals:

A= [1, 3]

B = [5, 7]

C = [2, 5]

D = [7, 9]

E = [6, 8]

F = [1, 8]

G = [9, 11]

H = [12, 14]

Step 2: We can manually compare the above intervals and determine the intersecting ones. For example, the intervals of A and B do not intersect, whereas the intervals B and C intersect. We compare each interval with every other interval to form a graph, such that every interval is a vertex and intersecting intervals are connected by an edge. Figure 3.7 shows the graph thus obtained.

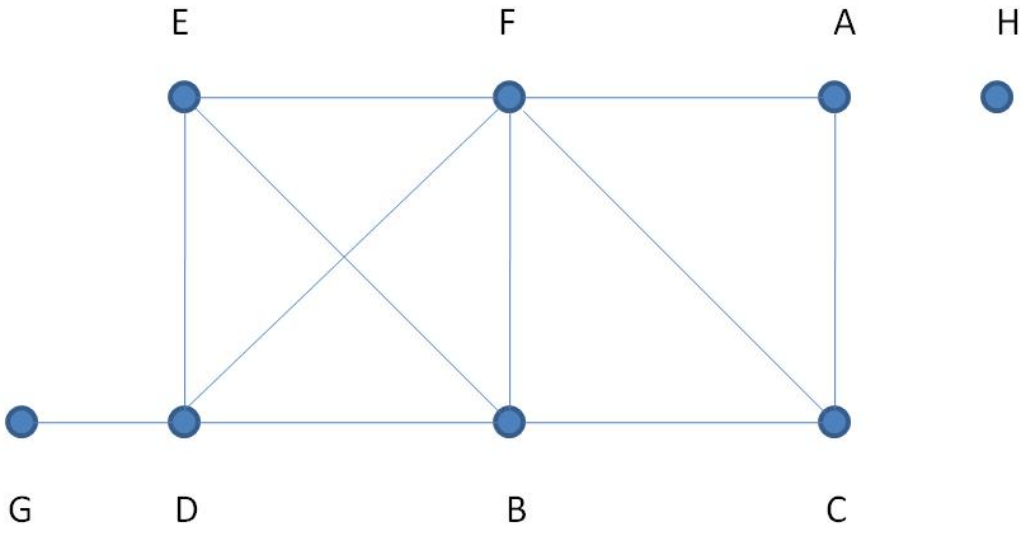


Figure 3.7 An Intersection Graph.

Step3: Having obtained the intersection graph, we incorporate a perfect elimination scheme to find the maximal cliques. For univariate interval-censored data a particularly useful perfect elimination scheme is obtained from the ordering of the right endpoints of the data in increasing order [22]. Therefore, we sort the interval data based on the right value, and input it to our implementation of the combinatorial algorithm of [22].

Step 4: After identifying all maximal cliques, we form the Petrie matrices for each panel as follows:

1. We build a matrix with vertices heading the columns and identified maximal cliques as rows.

2. We populate the matrix with '1' for each vertex participating in the corresponding maximal clique.

3. We populate the rest of the matrix with '0'.

A screenshot of the matrix obtained for one of our panels is shown in Figure 3.8.

More	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
Max_Clique 1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
Max_Clique 2	0	0	1	1	0	1	1	1	1	0	1	1	0	1	0	0
Max_Clique 3	0	0	1	1	0	1	1	1	1	0	1	1	0	1	0	0
Max_Clique 4	0	0	1	1	0	1	1	1	1	0	1	1	0	1	0	0

Figure 3.8 A Screenshot Depicting Petrie Matrix.

CHAPTER 4

DISCUSSION

During the initial phase of this project, we spent a substantial effort in collecting genetic data from various sources. For this purpose, we have implemented several scripts and programs, as described in the previous chapters. Furthermore, there was a major need for preprocessing of the collected data to make them useable for our study. Much of the computational work has been performed using a variety of pre-existing software commonly used for bioinformatics tasks, such as R, R/QTL, ePCR and BLAT, which needed to be installed and connected into a customized environment through glue scripts. Overall, it was a true exercise in bioinformatics and computational support of the work of investigators in life sciences.

Unfortunately, not everything in science goes smoothly, and as we gathered more data we realized that the irregularity of the genetic data currently available was a huge obstacle. In particular, the lack of flanking markers covering the chromosome end to end created a situation in which we could not use our data for reliable conclusions. For example, for the (C57BL/6 always x C3H) x B6 dataset, the flanking markers spanned an average of only 65.25% of the chromosome length. We have observed a maximum coverage of 80% for chromosome 1 and a minimum coverage of 53% for chromosome 18 (Table 3.2), and all the missing data were at the ends of the chromosomes. Since it has already been established that the recombination in mice occurs at significantly higher rates near the telomeres, we have found ourselves unable to derive reliable conclusions.

In order to keep pursuing our original goals, we had to think about a fresh approach, so we decided to take the recombination interference route to model the distribution of recombination rates. We have thus identified all intervals of double recombination in our original panels, and used the idea of censored data from the order theory. With the aid of graph

theoretical techniques we have generated the intersection graphs based on these intervals of double recombination. Eventually, we came up with fast and efficient ways of achieving data collection, preprocessing and analysis of the acquired datasets.

Equipped with the interval graphs and the Petrie matrices for all panels, the products of our work, the biology/statistics part of our team, Dr. Elena de la Casa-Esperon and Dr. J.C. Loredo-Osti, intend to find the Non Parametric Maximum Likelihood Estimate (NPMLE) of the underlying distribution. This will facilitate the comparison the empirical distributions of the double recombinations in each cross from the non-interference distribution (modeled as Poisson) and hence analyze the deviation between the two models. Comparing these deviations between crosses will provide a means to compare interference differences between them. In this way we can study the variation of the levels of meiotic recombination, and fulfill the goals of our original project.

REFERENCES

1. Alberts, Bruce et al. The Molecular Biology of the Cell, 4th ed., Garland Science, 2002, ISBN 0-8153-3218-1.

2. Brown TA Genomes 2, 2nd edition. The Methodology for DNA Sequencing, Chapter 6: 6.1, ISBN 1 85996 228 9. Section 2, 2002

3. Brown TA. Genomes 2, 2nd edition. Assembly of a Contiguous DNA Sequence, Chapter 6: 6.2 ISBN 1 85996 228 9 Section 2, 2002.

4. Service RF. The Race for the $1000 Genome. Science 311 (5767): 1544 - 1546. doi:10.1126/science.311.5767.1544, 2006.

5. Tushar Kumar Jayantilal, An Application of Parallel and Distributed Computing Methods to Approximate Pattern matching of Genetic Regulatory Motifs,Master's Thesis, UT Arlington, 2005.

6. SF Altschul, W Gish, W Miller, EW Myers, DJ Lipman, Basic local alignment search tool, Journal of Molecular Biology, 215: 403-410, 1990.

7. C. Burge and S. Karlin, Prediction of complete gene structures in human genomic DNA, Journal of Molecular Biology, 268: 78-94, 1997.

8. Kent, W.J., Sugnet, C. W., Furey, T. S., Roskin, K.M., Pringle, T. H., Zahler, A. M., and Haussler, D. The Human Genome Browser at UCSC. Genome Res. 12(6), 996-1006 (2002).

9. Karl W. Broman, Lucy B. Rowe, Gary A. Churchill, and Ken Paigen. Crossover Interference in the Mouse, 2002.

10. Reproduced form the Encyclopedia of Biostatistics, 2nd Edition. © John Wiley & Sons, Ltd. ISBN: 0-470-84829-4.

11. Lee M. Silver, Mouse Genetics - Concepts and Applications, Oxford University Press 1995

12. Haldane, J. B. S, The mapping function, J. Genet. 8: 299-309, 1919

13. Kosambi, D. D, The estimation of map distances from recombination values. Ann. Eugenics 12: 172-175., 1944

14. Lander, E. S., and P. Green, 1987 Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA 84: 2363–2367.

15. Carter, T. C., and Falconer, D. S, Stocks for detecting linkage in the mouse and the theory of their design, J. Genet. 50: 307-323, 1951.

16. M. W. Nachman and G. A. Churchill, Heterogeneity in Rates of Recombination Across the Mouse Genome, 1995

17. de Boer. P and Groen. A, Fertility and meiotic behavior of male T70H tertiary trisomics of the mouse (Mus musculus). A case of preferential telomeric meiotic pairing in a mammal. Cytogenet. Cell Genet. 13: 489-510, 1974.

18. Zimmerer, E. J., and Passmore, H. C, Structural and genetic properties of the Eb recombinational hotspot in the mouse. Immunogenetics 33: 132-40, 1991.

19. Bryda, E. C., DePari, J. A., SantAngelo, D. B., Murphy, D. B., and Passmore, H. C, Multiple sites of crossing over within the Eb recombinational hotspot in the mouse. Mamm Genome 2: 123-9, 1992.

20. Davisson, M. T., Roderick, T. H, and Doolittle, D. P, Recombination Percentages and Chromosomal Assignments In Genetic Variants and Strains of the Laboratory Mouse, Lyon, M. F. and Searle, A. G., eds. (Oxford University Press, Oxford), pp. 432-505, 1989.

21. de la Casa-Esperon, Elena, Loredo-Osti, J Concepcion, Pardo-Manuel de Villena, Fernando, Briscoe, Tammi L., Malette, Jan Michel, Vaughan, Joe E., Morgan, Kenneth, Sapienza, Carmen, X Chromosome Effect on Maternal Recombination and Meiotic Drive in the Mouse Genetics 161: 1651-1659, 2002

22. Robert Gentleman, Alain C. Vandal, Computational Algorithms for Censored-Data Problems Using Intersection Graphs, Journal of Computational and Graphical Statistics, Vol. 10, No. 3, pp. 403-421, Sep. 2001, The Canadian Journal of Statistics / La Revue Canadienne de Statistique, Vol. 30, No. 4, pp. 557-571, Dec. 2002

23. Folami Y. Ideraabdullah , Kuikwon Kim, Daniel Pomp, Jennifer L. Moran, David Beier, and Fernando Pardo-Manuel de Villena. Rescue of the Mouse DDK Syndrome by Parent-of-Origin-DependentModifiers. Biol Reprod.76(2):286-29, Feb. 2007.

24. W. James Kent, BLAT-The BLAST-Like Alignment Tool, Genome Res., Apr 2002; 12: 656 - 664.

25. Sturtevant, A.H., the Behavior of the Chromosomes as Studied through Linkage, Z. Induk. Abstammungs Vererbungsl. 13:234-287. (1915).

26. Luke E. Berchowitz and Gregory P. Copenhaver, Division of labor among meiotic genes, Nature Genetics 40, 266 - 267 (2008)

27. Felsenstein J., A mathematically tractable family of genetic mapping functions with different amounts of interference. Genetics. 1979;91:769–775

28. Vandal, A. C., Order theory and nonparametric analysis for interval censored data. PhD thesis, University of Auckland, 1998

29. Kent, W.J., Sugnet, C. W., Furey, T. S., Roskin, K.M., Pringle, T. H., Zahler, A. M., and Haussler, D. The Human Genome Browser at UCSC. Genome Res. 12(6), 996-1006 (2002).

30. Broman KW, Wu H, Sen Ś, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889-890

31. Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, 77 (2), p. 257–286, February 1989.

BIOGRAPHICAL INFORMATION

Vishnukumar Galigekere Nagabhushana joined the University of Texas at Arlington in the fall of 2005, for pursuing his M.S. degree in Computer Science and Engineering. He started his research career in the field of bioinformatics from fall of 2007. His areas of interest include bioinformatics, machine learning, algorithms, pattern recognition and computer vision. He received his Bachelor's degree in Electronics from Bangalore University, Bangalore, India. He then received his Masters degree in Computer Applications from Visvesvaraya Technological University, Belgaum, India. He briefly worked as a software engineer at Siemens, Bangalore, India before starting his Masters in Computer Science and Engineering. He received his Master of Science degree in Computer Science, in August, 2008. His future plan is to obtain a Doctoral degree in Computer Science and pursue a career in research and academia.