Multimarker Genetic Analysis Methods for High Throughput Array Data

by

Iuliana Ionita

A dissertation submitted in partial fulfillment of the requirements for the degree

of Doctor of Philosophy

Department of Computer Science

Courant Institute of Mathematical Sciences

New York University

May 2006

Bud Mishra

To Radu, for his love

Abstract

We describe three multi-marker/-locus statistical methods for analyzing high-throughput array data used for the detection of genes implicated in complex disorders.

1. Detection of Tumor Suppressor Genes and Oncogenes using Multipoint Statistics from Copy Number Variation Data

ArrayCGH is a microarray-based comparative genomic hybridization technique that has been used to compare a tumor genome against a normal genome, thus providing rapid genomic assays of tumor genomes in terms of copy number variations of those chromosomal segments, which have been gained or lost. When properly interpreted, these assays are likely to shed important light on genes and mechanisms involved in initiation and progression of cancer. Specifically, chromosomal segments, amplified or deleted in a group of cancer patients, point to locations of cancer genes, oncogenes and tumor suppressor genes, implicated in the cancer. In this chapter, we focus on automatic methods for reliable detection of such genes and their locations, and devise an efficient statistical algorithm to map oncogenes and tumor suppressor genes using a novel multi-point statistical score function. The proposed algorithm estimates the location of cancer genes by analyzing segmental duplications and deletions in the genomes from cancer patients and the spatial relation of these changes to any specific genomic interval. The algorithm assigns to an interval of consecutive probes a multipoint score

that parsimoniously captures the underlying biology. It also computes a p-value for every putative oncogene and tumor suppressor gene by using concepts from the theory of scan statistics. Furthermore, it can identify smaller sets of predictive probes that can be used as biomarkers for diagnosis, and therapeutics. We have validated our method using different simulated artificial datasets and one real dataset on lung cancer, and report encouraging results.

2. Multilocus Linkage Analysis of Affected Sib Pairs

The conventional Affected-Sib-Pair methods evaluate the linkage information at a locus by considering only marginal information. We describe a multilocus linkage method that uses both the marginal information and information derived from the possible interactions among several disease loci, thereby increasing the significance of loci with modest effects. Our method is based on a statistic that quantifies the linkage information contained in a set of markers. By a marker selection-reduction process, we screen a set of polymorphisms and select a few that seem linked to disease. We test our approach on a genome-scan data for inflammatory bowel disease and on simulated data. We show that our method is expected to be more powerful than single-locus methods in detecting disease loci responsible for complex traits.

3. We consider the problem of efficient inference algorithms to determine the haplotypes and their distribution from a dataset of unrelated genotypes.

With the currently available catalogue of single-nucleotide polymorphisms (SNPs) and given their abundance throughout the genome (one in about 500 bps) and low mutation rates, scientists hope to significantly improve their ability to discover

genetic variants associated with a particular complex trait. We present a solution to a key intermediate step by devising a *practical* algorithm that has the ability to infer the haplotype variants for a particular individual from its own genotype SNP data in relation to population data. The algorithm we present is simple to describe and implement; it makes no assumption such as perfect phylogeny or the availability of parental genomes (as in trio-studies); it exploits locality in linkages and low diversity in haplotype blocks to achieve a linear time complexity in the number of markers; it combines many of the advantageous properties and concepts of other existing statistical algorithms for this problem; and finally, it outperforms competing algorithms in computational complexity and accuracy, as demonstrated by the studies performed on real data and synthetic data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

High-throughput genetic data offer the promise of a better understanding of the genes and mechanisms implicated in complex diseases. Traditional methods dealing with one marker at a time have been successful for Mendelian (single-gene) diseases. The marker-by-marker approaches have been appropriate in the past, when the DNA markers available (RFLPs, microsatellites etc.) were relatively sparse, and hence there was little correlation among nearby markers. However, recently, with the discovery of millions of SNPs (single-nucleotide polymorphisms), and also with the interest moving from the study of single-gene diseases to more complex traits (multifactorial diseases), multi-marker/-locus methods are more appropriate and are expected to be more successful in detecting genes implicated in complex diseases. In this thesis we consider three problems in this area:

1. Mapping Tumor Suppressor Genes and Oncogenes using Multipoint Statistics

from Copy-Number Variation Data

2. Multilocus Linkage Analysis of Affected Sib-Pairs

3. A Practical Haplotype Inference Algorithm

The first problem (Chapter 2) concerns the detection of cancer genes from copy-number variation data. Copy-number variation data can be generated using a technology called arrayCGH (array comparative genomic hybridization). The normal copy number is two, one copy being inherited from the mother and the other from the father. Sometimes, there are copy number changes in the genome of an individual. There are deletions that can reduce the copy number from 2 to 1 (hemizygous deletions) or 0 (homozygous deletions). Another type of change are amplifications; in this case the copy number goes up, it can be 3, or 7 or 150, etc. These copy-number changes can in turn affect the expressions of the genes affected by these copy number changes. When these changes occur in certain genes (cancer genes), they can lead to cancer.

Two types of genes are involved in cancer: oncogenes and tumor suppressor genes. Oncogenes are usually due to gain-of-function mutations that lead to malignancy by mutations (e.g. amplifications) is genes whose functions may be in stimulating cell proliferation, increasing blood supply to the tumor, or inhibiting apoptosis. Tumor suppressor genes block tumor development by regulating cell growth. Loss-of-function (e.g. due to deletions) of proteins encoded by these tumor suppressor genes leads to uncontrolled cell division.

Our contribution is the introduction of a method for the detection of cancer genes from the data on deletions and amplifications in a group of patients, all suffering from the same cancer. Genomic regions enriched in deletions across many patients point to the location of tumor suppressor genes; and similarly regions with many amplifications point to the location of oncogenes.

The second problem (Chapter 3) concerns a multi-locus linkage method for affected sib-pairs. The affected-sib-pair (ASP) design is simple and popular, often used in the detection of genes involved in complex diseases. The basic idea is to collect sibs affected with a certain disease and their parents and genotype them at markers throughout the genome. Then statistical methods exploit the oversharing of genetic information at the disease loci in order to find the locations of these disease genes. The traditional statistical methods analyze the data marker-by-marker, thereby making use of only marginal information. However, the complex illnesses (like heart disease, diabetes, cancer) are, more often than not, caused by the complex interaction of multiple genetic and environmental risk factors. Therefore we propose a multilocus linkage method that is able to use both marginal, as well as interacting information to evaluate and find the disease genes.

The third problem (Chapter 4) concerns the inference of haplotypes from unphased genotype information in a sample of unrelated individuals. This problem is very important in genetics and has implications in the efficient design of genetic studies(HAPMAP project). Ideally we would like to be able to separate the genetic information (at multiple locations) coming from the mother from that coming from the father, the so-called haplotypes. They are important since they give in-

formation about recombination patterns and about correlations among markers. However, the genotyping experiments do not give us this information; they only tell us at each marker locus a set of two alleles, without differentiating between the mother and the father allele (the so-called unphased genotype). Therefore computational methods have been proposed to find the haplotypes from the unphased genotypes in a sample. We present a simple and efficient algorithm for this problem.

1.2 Thesis Outline

The rest of the thesis is organized as follows. In Chapter 2 we present the algorithm for the detection of cancer genes from copy-number variation data, with an application to lung cancer. In Chapter 3 we present a multilocus linkage method for the detection of disease genes using an affected-sib-pair design, with an application to inflammatory bowel disease. In Chapter 4 we address an important issue involved in the design of these disease-gene-detection studies: the haplotype inference problem. We conclude in Chapter 5 with a discussion of our results and future directions.

Chapter 2

Mapping Tumor Suppressor Genes and Oncogenes using Multipoint Statistics from Copy-Number Variation Data

SUMMARY: ArrayCGH is a microarray-based comparative genomic hybridization technique that has been used to compare a tumor genome against a normal genome, thus providing rapid genomic assays of tumor genomes in terms of copy number variations of those chromosomal segments, which have been gained or lost. When properly interpreted, these assays are likely to shed important light on genes and mechanisms involved in initiation and progression of cancer. Specifically, chromosomal segments, amplified or deleted in a group of cancer patients, point to locations of cancer genes, oncogenes and tumor suppressor genes, implicated in the

cancer. In this chapter, we focus on automatic methods for reliable detection of such genes and their locations, and devise an efficient statistical algorithm to map oncogenes and tumor suppressor genes using a novel multi-point statistical score function. The proposed algorithm estimates the location of cancer genes by analyzing segmental duplications and deletions in the genomes from cancer patients and the spatial relation of these changes to any specific genomic interval. The algorithm assigns to an interval of consecutive probes a multipoint score that parsimoniously captures the underlying biology. It also computes a p-value for every putative oncogene and tumor suppressor gene by using concepts from the theory of scan statistics. Furthermore, it can identify smaller sets of predictive probes that can be used as biomarkers for diagnosis, and therapeutics. We have validated our method using different simulated artificial datasets and one real dataset on lung cancer, and report encouraging results.

2.1 Introduction

The process of carcinogenesis imparts many genetic changes to a cancer genome at many different scales: point mutations, translocations, segmental duplications, and deletions. While most of these changes have no direct impact on the cellular functions, and may not contribute to the carcinogenesis in any obvious manner, few of these chromosomal aberrations have disproportionately significant impact on the cell's ability to initiate and maintain processes involved in tumor growth: namely, through its ability to proliferate, escape senescence, achieve immortality, and signal to neighboring cells. Two classes of genes are critically involved in

cancer development and are discernible in terms of their copy number variations: oncogenes that are activated or altered in function and tumor suppressor genes that are deactivated in cancer cells.

The effect of oncogenes is via gain-of-function mutations that lead to malignancy. For instance, a segmental amplification can increase the genomic copy number of a region containing an oncogene, thus leading to over-expression of the oncogene product. The mutation is dominant, i.e. only a mutated allele is necessary for the cell to become malignant, and needs to encompass the entire gene.

Tumor-suppressor genes, on the other hand, affect the cells via mutations (often involving segmental deletions) that contribute to malignancy by loss-of-function of both alleles of the gene. The "two-hit" hypothesis of Knudson (Knudson 1971 [1]) for tumorigenesis has been widely recognized as an important model of such losses of function involved in many cancers. Only portions of a tumor suppressor gene need to be deleted in order for the cell to become cancerous.

In the current whole-genome analysis setup, microarray techniques are being used successfully to measure fluctuations in copy number for a large number of genomic regions in one genome relative to a different but related genome sample. For example, array CGH can map copy number changes at a large number of chromosomal locations in one genome with respect to a reference genome, and from them extrapolate to infer segments of genome that have undergone same degree of amplifications or deletions. For some references to and discussions of algorithms that estimate these copy number variations (CNVs), see Daruwala et al. 2004 [2].

The approach here exploits spatial relations of these changes to any specific genomic interval in a group of cancer patients, and works by enumerating all short intervals in the genome and then evaluating them with a score function that measures the likelihood of an interval being the oncogene or tumor suppressor gene.

The rest of the chapter is organized as follows. We first propose simple probabilistic generative models assumed to have generated the copy number data, and then formulate our score functions based on these models. We then show how this function is used in evaluating whether a region represents an oncogene or a tumor suppressor gene. Next, we illustrate our method, using this score function and several sets of simulated data, computed under a wide variety of scenarios (Results section); we also assess the power of the method by examining how accurately it discovers the true location (which is known to the simulator) of the oncogene or tumor suppressor gene. Finally, we analyze and report the results from an arrayCGH dataset (using 100K Affy-chips), obtained from several lung cancer patients. We conclude with a discussion of the strength and weakness of the proposed method (Discussion section).

2.2 Methods

2.2.1 Statistical Model

We first describe a simple generative model assumed to model how the copy-number data arised. In this simplest parsimonious model, we model the breakpoints, defined as the starting points of amplification and deletion events, and the lengths of

these changes as follows. We assume that at any genomic location, a breakpoint may occur as a Poisson process at a rate $\mu \geq 0$; hence μ is the mean number of breakpoints per unit length. Starting at any of these breakpoints, a segmental change (amplification or deletion) occurs with length distributed as an Exponential random variable with parameter $\lambda \geq 0$; hence $\frac{1}{\lambda}$ is the average length of a change. We formulate two such models: one for amplifications (parameters μ_a, λ_a) and one for deletions (parameters μ_d, λ_d).

Having formulated this model, we can now compute for a genomic interval $I = [a, b]$ the background probabilities of I being deleted or amplified. Note that by I deleted we mean that any part of I is deleted, whereas I is amplified means that the entire interval I is amplified. This reflects our understanding as to how oncogenes become activated and tumor suppressor genes become inactivated.

Lemma 2.2.1 *Assuming the generative process described above:*

1. *The probability that an interval $I = [a, b]$ is amplified is as follows:*

$$\begin{aligned} P([a,b] \text{ amplified}) &= 1 - e^{-\mu_a a \frac{e^{-\lambda_a (b-a)} - e^{-\lambda_a b}}{2\lambda_a a}} \cdot \\ &\quad \cdot e^{-\mu_a (G-b) \frac{e^{-\lambda_a (b-a)} - e^{-\lambda_a (G-a)}}{2\lambda_a (G-b)}}, \end{aligned} \tag{2.1}$$

   where $[0, G]$ represents the region of interest (e.g., a chromosome).

2. *The probability that an interval $I = [a, b]$ is deleted is as follows:*

$$P([a,b] \text{ deleted}) = 1 - e^{-\mu_d (b-a)} e^{-\mu_d a \frac{1 - e^{-\lambda_d a}}{2\lambda_d a}} e^{-\mu_d (G-b) \frac{1 - e^{-\lambda_d (G-b)}}{2\lambda_d (G-b)}}, \tag{2.2}$$

where $[0, G]$ *represents the region of interest (e.g., a chromosome).*

Proof:

1. Given an interval $[a, b]$, it is easier to compute the following probability:

$$P([a, b] \text{ NOT amplified}) = P([a, b] \cap \text{ Ampl} = \phi)$$

We know that $[a, b] \cap$ Ampl $= \phi$ happens if and only if: each amplified interval starting from a breakpoint in $[0, a]$ does not contain $[a, b]$ AND each amplified interval starting from a breakpoint in $[b, G]$ does not contain $[a, b]$. Let P_1 be the probability of the first event and P_2 be the probability of the second event.

P_1 can be written as:

$$\begin{aligned} & P(\text{ no breakpoint in } [0, a]) \\ + \; & P(1 \text{ breakpoint in } [0, a]) \times P(\text{ amplified interval } \cap [a, b] = \phi) \\ + \; & P(2 \text{ breakpoints in } [0, a]) \times P(\text{ amplified intervals } \cap [a, b] = \phi) \\ + \; & \cdots \end{aligned}$$

We can now compute the terms in this sum.

The first term is:

$$P(\text{ no breakpoint in } [0, a]) = e^{-\mu a}$$

For the second term ($P(1$ breakpoint in $[0,a]) \times P($ amplified interval $\cap [a,b] = \phi)$) we have first:

$$P(1 \text{ breakpoint in } [0,a]) = \mu a e^{-\mu a}$$

and $P(\text{amplified interval } \cap [a,b] = \phi)$ is more complicated.

Suppose we divide the interval $[0,a]$ into many intervals $[x_i, x_{i+1}]$ for $i \in \{0,\ldots,n\}$. Then

$$\begin{aligned}
& P(\text{ amplified interval } \cap [a,b] = \phi) = \\
= & \sum_{i=0}^{n} P(\text{ breakpoint } x \text{ is in } [x_i, x_{i+1}] | 1 \text{ breakpoint in } [0,a]) \\
& \times P(\text{no overlap caused by} x) \\
= & \sum_{i=0}^{n} \frac{x_{i+1} - x_i}{a} \times P(\text{no overlap caused by } x)
\end{aligned}$$

Let us now compute the probability of no overlap caused by the amplification starting at $x \in [x_i, x_{i+1}]$.

$$\begin{aligned}
& P(\text{no overlap caused by } x \in [x_i, x_{i+1}]) \\
= & \frac{1}{2} + \frac{1}{2} \frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} \int_{0}^{b-x} \lambda e^{-\lambda t} dt dx \\
= & 1 - \frac{1}{2} \frac{1}{x_{i+1} - x_i} \frac{e^{-\lambda(b - x_{i+1})} - e^{-\lambda(b-x_i)}}{\lambda} \\
= & 1 - \frac{1}{2} \frac{G(x_{i+1}) - G(x_i)}{x_{i+1} - x_i}
\end{aligned}$$

where $G(x) = \frac{e^{-\lambda(b-x)}}{\lambda}$. Therefore

$$\begin{aligned} & P(\text{ amplified interval } \cap [a,b] = \phi) = \\ = & \sum_{i=0}^{n} \frac{x_{i+1} - x_i}{a} \times \left(1 - \frac{1}{2}\frac{G(x_{i+1}) - G(x_i)}{x_{i+1} - x_i}\right) \\ = & \int_0^a \frac{1}{a}\left(1 - \frac{1}{2}G'(x)\right) dx \\ = & 1 - \frac{e^{-\lambda(b-a)} - e^{-\lambda b}}{2\lambda a} \end{aligned}$$

So the second term is

$$\mu a e^{-\mu a} \cdot \left(1 - \frac{e^{-\lambda(b-a)} - e^{-\lambda b}}{2\lambda a}\right)$$

For the third term ($P(2$ breakpoints in $[0,a]) \times P($ amplified intervals $\cap [a,b] = \phi)$) we have first:

$$P(2 \text{ breakpoints in } [0,a]) = \frac{(\mu a)^2}{2!} e^{-\mu a}$$

and

$$\begin{aligned} & P(\text{ the two amplified intervals } \cap [a,b] = \phi) \\ = & \sum_{i<j} P(\text{ 1st break is in } [x_i, x_{i+1}] \text{ and 2nd break is in } [x_j, x_{j+1}] | \text{ two breaks in} [0,a]) \\ & \times P(\text{no overlap caused by the first break}) \\ & \times P(\text{no overlap caused by the second break}) \end{aligned}$$

Now

$$\begin{aligned} & P(\text{ 1st break is in } [x_i, x_{i+1}] \text{ and 2nd break is in } [x_j, x_{j+1}] | \text{ two breaks in} [0, a]) \\ = \quad & \frac{2(x_{i+1} - x_i)(x_{j+1} - x_j)}{a^2} \end{aligned}$$

by simple probability computations.

We have already computed

$$\begin{aligned} & P(\text{no overlap caused by } x \in [x_i, x_{i+1}]) \\ = \quad & 1 - \frac{1}{2} \frac{G(x_{i+1}) - G(x_i)}{x_{i+1} - x_i} \end{aligned}$$

and

$$\begin{aligned} & P(\text{no overlap caused by } x \in [x_j, x_{j+1}]) \\ = \quad & 1 - \frac{1}{2} \frac{G(x_{j+1}) - G(x_j)}{x_{j+1} - x_j} \end{aligned}$$

where

$$G(x) = \frac{e^{-\lambda(b-x)}}{\lambda}$$

Therefore:

$$
\begin{aligned}
& P(\text{ the two amplified intervals } \cap [a,b] = \phi) \\
= & \sum_{i<j} \frac{2(x_{i+1}-x_i)(x_{j+1}-x_j)}{a^2} \times \left(1 - \frac{1}{2}\frac{G(x_{i+1})-G(x_i)}{x_{i+1}-x_i}\right) \\
& \times \left(1 - \frac{1}{2}\frac{G(x_{j+1})-G(x_j)}{x_{j+1}-x_j}\right) \\
= & \frac{2}{a^2}\int_0^a \left(\int_0^y \left(1-\frac{1}{2}G'(x)\right)dx\right)\left(1-\frac{1}{2}G'(y)\right)dy \\
= & \frac{2}{a^2}\int_0^a (y - \frac{1}{2}G(y) + \frac{1}{2}G(0))(1-\frac{1}{2}G'(y))dy \\
= & 1 - \frac{1}{a^2}\int_0^a G(y)dy + \frac{e^{-\lambda b}}{\lambda a} - \frac{2}{a^2}\int_0^a (y - \frac{1}{2}G(y) + \frac{1}{2}G(0))\frac{1}{2}G'(y)dy
\end{aligned}
$$

Integrating by parts we obtain:

$$
\begin{aligned}
& \int_0^a G(y)dy + \int_0^a (y - \frac{1}{2}G(y) + \frac{1}{2}G(0))G'(y)dy \\
= & \int_0^a G(y)dy + (a - \frac{1}{2}G(a) + \frac{1}{2}G(0))G(a) - \int_0^a (1-\frac{1}{2}G'(y))G(y)dy \\
= & \int_0^a G(y)dy + (a - \frac{1}{2}G(a) + \frac{1}{2}G(0))G(a) - \int_0^a G(y)dy + \frac{1}{2}\int_0^a G'(y)G(y)dy \\
= & (a - \frac{1}{2}G(a) + \frac{1}{2}G(0))G(a) + \frac{G(a)^2}{4} - \frac{G(0)^2}{4} \\
= & aG(a) + \frac{1}{2}G(0)G(a) - \frac{G(a)^2}{4} - \frac{G(0)^2}{4}
\end{aligned}
$$

Hence

$$
\begin{aligned}
& P(\text{ the two amplified intervals } \cap [a,b] = \phi) \\
= & 1 + \frac{e^{-\lambda b}}{\lambda a} - \frac{1}{a^2}\left(aG(a) + \frac{1}{2}G(0)G(a) - \frac{G(a)^2}{4} - \frac{G(0)^2}{4}\right)
\end{aligned}
$$

We have $G(a) = \frac{e^{-\lambda(b-a)}}{\lambda}$ and $G(0) = \frac{e^{-\lambda b}}{\lambda}$ So:

$$\begin{aligned} & P(\text{ the two amplified intervals } \cap [a,b] = \phi) \\ = \; & 1 + \frac{e^{-\lambda b}}{\lambda a} - \frac{G(a)}{a} + \frac{1}{4a^2}(G(0) - G(a))^2 \\ = \; & \left(1 - \frac{G(a) - G(0)}{2a}\right)^2 \\ = \; & \left(1 - \frac{e^{-\lambda(b-a)} - e^{-\lambda b}}{2\lambda a}\right)^2 \end{aligned}$$

We can now say that the third term is:

$$\frac{(\mu a)^2}{2!} e^{-\mu a} \cdot \left(1 - \frac{e^{-\lambda(b-a)} - e^{-\lambda b}}{2\lambda a}\right)^2$$

Finally we obtain:

$$P_1 = e^{-\mu a}\left(1 + \mu a\left(1 - \frac{e^{-\lambda(b-a)} - e^{-\lambda b}}{\lambda a}\right) + \frac{1}{2!}\left(\mu a\left(1 - \frac{e^{-\lambda(b-a)} - e^{-\lambda b}}{\lambda a}\right)\right)^2 + \ldots\right)$$

Hence

$$P_1 = e^{-\mu a \frac{e^{-\lambda(b-a)} - e^{-\lambda b}}{2\lambda a}}$$

We need to compute P_2 also. P_2 is the probability that each amplified interval starting from a breakpoint in [b,G] does not contain [a, b]. This is very similar to P_1. We can show that:

$$P_2 = e^{-\mu(G-b) \frac{e^{-\lambda(b-a)} - e^{-\lambda(G-a)}}{2\lambda(G-b)}}$$

The final formula is:

$$P([a,b] \cap \text{ Ampl} = \phi) = e^{-\mu a \frac{e^{-\lambda(b-a)} - e^{-\lambda b}}{2\lambda a}} e^{-\mu(G-b)\frac{e^{-\lambda(b-a)} - e^{-\lambda(G-a)}}{2\lambda(G-b)}}$$

□

2. Given an interval $[a,b]$, it is easier to compute the following probability:

$$P([a,b] \text{ NOT deleted}) = P([a,b] \cap \text{ Del} = \phi)$$

We know that $[a,b] \cap \text{ Del} = \phi$ happens if and only if: there is no breakpoint in $[a,b]$ AND each deleted interval starting from a breakpoint in $[0,a]$ does not overlap with $[a,b]$ AND each deleted interval starting from a breakpoint in $[b,G]$ does not overlap with $[a,b]$.

Let P_1 be the probability of the first event, P_2 that of the second, and P_3 that of the last event. P_1, the probability of no breakpoint in $[a,b]$, is just

$$P_1 = e^{-\mu(b-a)}$$

The second probability, P_2, can be written as:

$$\begin{aligned} P_2 = & P(\text{ no breakpoint in } [0,a]) \\ & + \quad P(1 \text{ breakpoint in } [0,a]) \times P(\text{ deleted interval } \cap [a,b] = \phi) \\ & + \quad P(2 \text{ breakpoints in } [0,a]) \times P(\text{ deleted intervals } \cap [a,b] = \phi) \\ & + \quad \cdots \end{aligned}$$

We can now compute the terms in this sum.

The first term is:

$$P(\text{ no breakpoint in } [0,a]) = e^{-\mu a}$$

For the second term ($P(1 \text{ breakpoint in } [0,a]) \times P(\text{ deleted interval } \cap [a,b] = \phi)$) we have first:

$$P(1 \text{ breakpoint in } [0,a]) = \mu a e^{-\mu a}$$

and $P(\text{ deleted interval } \cap [a,b] = \phi)$ is more complicated.

Suppose we divide the interval $[0,a]$ into many small intervals $[x_i, x_{i+1}]$ for

$i \in \{0, \ldots, n\}$. Then

$$\begin{aligned}
& P(\text{ deleted interval } \cap [a,b] = \phi) \\
= & \sum_{i=0}^{n} P(\text{ breakpoint } x \text{ is in } [x_i, x_{i+1}] | 1 \text{ breakpoint in } [0,a]) \\
& \times P(\text{no overlap cause by } x) \\
= & \sum_{i=0}^{n} \frac{x_{i+1} - x_i}{a} \cdot P(\text{no overlap caused by } x)
\end{aligned}$$

Let us now compute the probability of no overlap caused by the deletion at $x \in [x_i, x_{i+1}]$.

$$\begin{aligned}
& P(\text{no overlap caused by } x \in [x_i, x_{i+1}]) \\
= & \frac{1}{2} + \frac{1}{2} \frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} \int_{0}^{a-x} \lambda e^{-\lambda t} dt dx \\
= & 1 - \frac{1}{2} \cdot \frac{1}{x_{i+1} - x_i} \frac{e^{-\lambda(a - x_{i+1})} - e^{-\lambda(a-x_i)}}{\lambda} \\
= & 1 - \frac{1}{2} \cdot \frac{G(x_{i+1}) - G(x_i)}{x_{i+1} - x_i}
\end{aligned}$$

where $G(x) = \frac{e^{-\lambda(a-x)}}{\lambda}$. Therefore

$$\begin{aligned}
& P(\text{ deleted interval } \cap [a,b] = \phi) = \\
= & \sum_{i=0}^{n} \frac{x_{i+1} - x_i}{a} \cdot \left(1 - \frac{1}{2} \cdot \frac{G(x_{i+1}) - G(x_i)}{x_{i+1} - x_i}\right) \\
= & \int_0^a \frac{1}{a} \left(1 - \frac{1}{2} G'(x)\right) dx \\
= & 1 - \frac{1 - e^{-\lambda a}}{2\lambda a}
\end{aligned}$$

So the second term in P_2 is

$$\mu a e^{-\mu a} \cdot \left(1 - \frac{1 - e^{-\lambda a}}{2\lambda a}\right)$$

For the third term ($P($2 breakpoints in $[0, a]) \times P($ deleted intervals $\cap [a, b] = \phi)$) we have first:

$$P(\text{2 breakpoints in } [0, a]) = \frac{(\mu a)^2}{2!} e^{-\mu a}$$

and

$$\begin{aligned} & P(\text{ the two deleted intervals } \cap [a, b] = \phi) \\ = & \sum_{i<j} P(\text{ 1st break is in } [x_i, x_{i+1}] \text{ and 2nd break is in } [x_j, x_{j+1}] | \text{ two breaks in } [0, a]) \\ & \times P(\text{no overlap caused by the first break}) \\ & \times P(\text{no overlap caused by the second break}) \end{aligned}$$

Now

$$\begin{aligned} & P(\text{ 1st break is in } [x_i, x_{i+1}] \text{ and 2nd break is in } [x_j, x_{j+1}] | \text{ two breaks in} [0, a]) \\ = & \frac{2(x_{i+1} - x_i)(x_{j+1} - x_j)}{a^2} \end{aligned}$$

by simple probability computations.

We have already computed

$$\begin{aligned} & P(\text{no overlap caused by } x \in [x_i, x_{i+1}]) \\ = \;& 1 - \frac{1}{2}\frac{G(x_{i+1}) - G(x_i)}{x_{i+1} - x_i} \end{aligned}$$

and

$$\begin{aligned} & P(\text{no overlap caused by } x \in [x_j, x_{j+1}]) \\ = \;& 1 - \frac{1}{2} \cdot \frac{G(x_{j+1}) - G(x_j)}{x_{j+1} - x_j} \end{aligned}$$

where $G(x) = \frac{e^{-\lambda(a-x)}}{\lambda}$. Therefore:

$$\begin{aligned} & P(\text{ the two deleted intervals } \cap [a,b] = \phi) \\ = \;& \sum_{i<j} \frac{2(x_{i+1} - x_i)(x_{j+1} - x_j)}{a^2} \times \left(1 - \frac{1}{2} \cdot \frac{G(x_{i+1}) - G(x_i)}{x_{i+1} - x_i}\right) \\ & \times \left(1 - \frac{1}{2}\frac{G(x_{j+1}) - G(x_j)}{x_{j+1} - x_j}\right) \\ = \;& \frac{2}{a^2}\int_0^a \left(\int_0^y \left(1 - \frac{1}{2}G'(x)\right) dx\right)\left(1 - \frac{1}{2}G'(y)\right) dy \\ = \;& \frac{2}{a^2}\int_0^a \left(y - \frac{1}{2}G(y) + \frac{1}{2}G(0)\right)\left(1 - \frac{1}{2}G'(y)\right) dy \\ = \;& 1 - \frac{1}{a^2}\int_0^a G(y)dy + \frac{e^{-\lambda a}}{\lambda a} - \frac{2}{a^2}\int_0^a \left(y - \frac{1}{2}G(y) + \frac{1}{2}G(0)\right)\frac{1}{2}G'(y)dy \end{aligned}$$

Integrating by parts we obtain:

$$\begin{aligned}
&\int_0^a G(y)dy + \int_0^a \left(y - \frac{1}{2}G(y) + \frac{1}{2}G(0)\right) G'(y)dy \\
&= \int_0^a G(y)dy + \left(a - \frac{1}{2}G(a) + \frac{1}{2}G(0)\right) G(a) - \int_0^a \left(1 - \frac{1}{2}G'(y)\right) G(y)dy \\
&= \int_0^a G(y)dy + \left(a - \frac{1}{2}G(a) + \frac{1}{2}G(0)\right) G(a) - \int_0^a G(y)dy + \frac{1}{2}\int_0^a G'(y)G(y)dy \\
&= \left(a - \frac{1}{2}G(a) + \frac{1}{2}G(0)\right) G(a) + \frac{G(a)^2}{4} - \frac{G(0)^2}{4} \\
&= aG(a) + \frac{1}{2}G(0)G(a) - \frac{G(a)^2}{4} - \frac{G(0)^2}{2}
\end{aligned}$$

Hence

$$\begin{aligned}
&P(\text{ the two deleted intervals } \cap [a,b] = \phi) \\
&= 1 + \frac{e^{-\lambda a}}{\lambda a} - \frac{1}{a^2}\left(aG(a) + \frac{1}{2}G(0)G(a) - \frac{G(a)^2}{4} - \frac{G(0)^2}{4}\right)
\end{aligned}$$

Since $G(a) = \frac{1}{\lambda}$ and $G(0) = \frac{e^{-\lambda a}}{\lambda}$, we obtain

$$\begin{aligned}
&P(\text{ the two deleted intervals } \cap [a,b] = \phi) \\
&= 1 + \frac{e^{-\lambda a}}{\lambda a} - \frac{G(a)}{a} + \frac{1}{4a^2}(G(0) - G(a))^2 \\
&= \left(1 - \frac{G(a) - G(0)}{2a}\right)^2 \\
&= \left(1 - \frac{1 - e^{-\lambda a}}{2\lambda a}\right)^2
\end{aligned}$$

We can now say that the third term in P_2 is:

$$\frac{(\mu a)^2}{2!}e^{-\mu a} \cdot \left(1 - \frac{1 - e^{-\lambda a}}{2\lambda a}\right)^2$$

Finally we obtain:

$$P_2 \;=\; e^{-\mu a}\left[1 + \mu a\left(1 - \frac{1 - e^{-\lambda a}}{2\lambda a}\right) + \frac{1}{2!}\left(\mu a\left(1 - \frac{1 - e^{-\lambda a}}{2\lambda a}\right)\right)^2 + \dots\right]$$

Hence:

$$P_2 = e^{-\mu a \frac{1-e^{-\lambda a}}{2\lambda a}}$$

The last probability we need to compute is P_3, the probability that each deleted interval starting from a breakpoint in $[b, G]$ does not overlap with $[a, b]$. Through computations similar to those for P_2, we can show that:

$$P_3 = e^{-\mu(G-b)\frac{1-e^{-\lambda(G-b)}}{2\lambda(G-b)}}$$

The final formula is:

$$P([a,b] \cap \text{ Del} = \phi) = e^{-\mu(b-a)}e^{-\mu a\frac{1-e^{-\lambda a}}{2\lambda a}}e^{-\mu(G-b)\frac{1-e^{-\lambda(G-b)}}{2\lambda(G-b)}}$$

and therefore

$$P([a,b] \text{ deleted}) = 1 - e^{-\mu(b-a)}e^{-\mu a\frac{1-e^{-\lambda a}}{2\lambda a}}e^{-\mu(G-b)\frac{1-e^{-\lambda(G-b)}}{2\lambda(G-b)}}$$

□

The way we have defined an interval I as amplified or as deleted, implies that short intervals are easier to get amplified than longer intervals; however short intervals are harder to get deleted than larger ones.

2.2.2 Genomic Data

In the preceding subsection, we have described a model for how the background copy numbers are distributed in a general population. Our aim is to devise a method that is able to infer useful information from cancer patients with regard to where the cancer genes are located.

Our assumption is that we have available a sample of patients, all affected by the same type of cancer. For each patient, copy number values are available at many positions throughout the genome. By comparing the distribution of copy number values in these patients at a particular genomic location, with that expected under the null hypothesis of no cancer gene at that position, we can infer the locations of cancer genes.

2.2.3 Multipoint Scores

Our method for the identification of oncogenes and tumor suppressor genes relies on a multipoint score function, computed over whole-genome analysis data for a sufficiently large group of patients suffering from the same form of cancer.

For any interval I (represented as a set of consecutive probes), we wish to quantify the strength of the association between copy number changes (amplifi-

cations and deletions) in I and the disease, by analyzing the genomic data for many diseased individuals. For this purpose, we select a metric, the relative risk (RR), as it compares and assigns a numerical value to the risks of disease in two populations with respect to each other: the first population comprises subjects whose genomes contain an amplification, respectively a deletion in the interval I (we call this event A) and the second comprises subjects whose genomes have no such segmental change in I (we call this event B).

$$\begin{aligned} \mathrm{RR}_I &= \ln \frac{P(\text{disease}|A)}{P(\text{disease}|B)} \\ &= \ln\left(\frac{P(A|\text{disease})}{P(B|\text{disease})} \times \frac{P(B)}{P(A)}\right) \\ &= \ln \frac{P(A|\text{disease})}{P(B|\text{disease})} - \ln \frac{P(A)}{P(B)} \end{aligned} \tag{2.3}$$

We have two scores: one for oncogenes and one for tumor suppressor genes. In the score for oncogenes, we replace A with "I amplified", and similarly for tumor suppressor genes, we replace A with "I deleted". The first term in (2.3) can be estimated from the tumor samples available:

$$\frac{P(A|\text{disease})}{P(B|\text{disease})} = \frac{n_A}{n_B}, \tag{2.4}$$

where n_A is simply the number of tumor samples in which A holds. Hence for oncogenes, n_A is the number of cancer samples with I amplified, and for tumor suppressor genes n_A is the number of cancer samples with I deleted.

The second part $-\frac{P(A)}{P(B)}$ incorporates prior information inherent in the statistical distribution of amplifications or deletions. We have computed $P(A)$ in the section

Statistical Model.

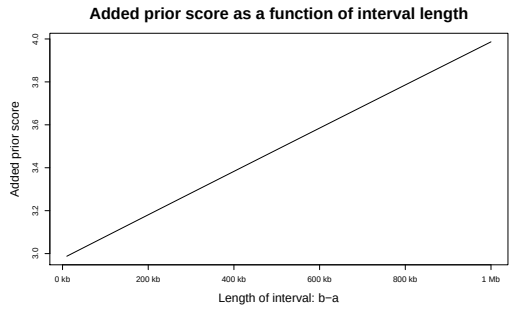
A difference arises between oncogenes and tumor suppressor genes with respect to how the prior information, included in the distribution of random unrelated deletions and amplifications in the genome, affects the evidence from the data. For oncogenes, the prior information is reflected through an advantage accrued to large intervals; in other words, assuming that the same strength of evidence exists in the data for different sizes, preference is given to the larger intervals. However, for tumor suppressor genes, the prior information gives advantage to the smaller intervals. The intuition is that is is harder to amplify large intervals and harder to delete small intervals. In Figure 2.1, we show how the two priors $-\ln \frac{P(A)}{P(B)}$ vary as a function of the length of the interval. All the parameters (μ_d, λ_d, G) are same as those in the simulation examples in the Results section.

For a genomic interval I, we can compute the score in 2.3. Clearly, we expect the high-scoring intervals determined by this method to be treated as candidates for tumor suppressor genes. We will still need to define precisely how and how many of these intervals are selected and then evaluated for their statistical significance.

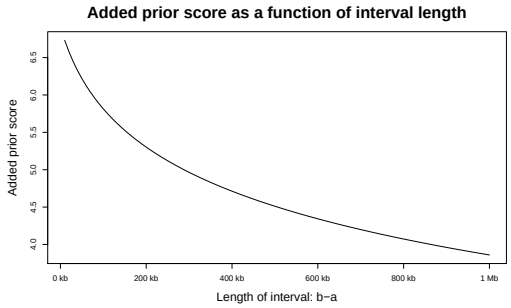
Note: If I_1 and I_2 are two intervals of the same length, with I_1 near one of the telomeres and I_2 near the center, then

$$\delta = \ln \frac{P(B_1)}{P(A_1)} - \ln \frac{P(B_2)}{P(A_2)} > 0.$$

This reveals an interesting fact. Intervals near the ends of the chromosome have a bigger prior advantage than intervals near the center. The reason for that is that



(a)



(b)

Figure 2.1: Prior score as a function of the length of the interval. (a) Oncogene (b) Tumor Suppressor Gene

intervals near the ends are harder to be mutated randomly than intervals near the center.

2.2.4 Estimating Parameters

In the preceding section, we have defined two scores for an interval I ($RR_{I \text{ is an oncogene}}$ and $RR_{I \text{ is a TSG}}$), which depend on extraneous parameters describing a background genome reorganization process. These parameters, namely λ_a, μ_a, λ_d and μ_d, must be estimated from array CGH data. We may recall that λ_a is the parameter of the exponential distribution for generating amplifications, i.e. $\frac{1}{\lambda_a}$ is the average length of an amplification, and μ_a is the parameter of the Poisson process used for

generating the breakpoints, i.e. μ_a is the mean number of breakpoints per unit length. The parameters λ_d and μ_d for the deletions are defined similarly.

Recently, several statistically powerful algorithms have been devised to analyze the array CGH data, and to render the underlying genome in terms of segments of regions of similar copy numbers. These algorithms readily yield an output that can be interpreted as alternating segments of normal and abnormal segments, with the abnormal segments falling into two groups: segmental losses and segmental gains. If these segments satisfy the assumptions regarding the breakpoint and length distributions, the desired parameters μ and λ can be estimated empirically from the segmentation of the data. Certain Bayesian algorithms, such as the one due to Daruwala et al. 2004 [2] and its variants (see Anantharaman, Sobel, and Mishra, unpublished data), include these assumptions in their prior and are thus able to estimate these parameters directly. The present algorithm builds on the latter class of segmentation algorithms, but is not limited by this requirement.

2.2.5 Estimating the Location of the Cancer Genes

We describe the estimation for the location of oncogenes and tumor suppressor genes.

The estimation procedure proceeds in a sequence of steps. In the first step, the algorithm computes the scores (RR_I) for all the intervals I with lengths taking values in a range determined by a lower and an upper bound, starting with small intervals containing a few markers and ending with very long intervals. We have evaluated two different methods, designed to estimate the location of cancer genes.

The first and the simplest method operates by simply choosing the maximum-scoring interval as the candidate cancer gene: namely, it selects the interval I with maximum RR_I in a small genomic region as the most plausible location of a cancer gene. We shall refer to this method as the "Max method."

The other method functions by estimating the locations of the left and the right boundaries of the gene, using two scoring functions, as described below. Two scores, SL_x and SR_x, are computed for every marker position $x \in [0, G]$. The first value, SL_x, is to be interpreted as the confidence that the point x is the left boundary of a cancer gene and symmetrically, the latter, SR_x, is the confidence that the point x is the right boundary of a cancer gene. These scores are defined more formally as follows:

$$SL_x = \sum_{I \in \mathcal{IL}_x} RR_I,$$

where $\mathcal{IL}_x$ is the set of intervals that are bounded by the marker x from the left. Similarly,

$$SR_x = \sum_{I \in \mathcal{IR}_x} RR_I,$$

where $\mathcal{IR}_x$ is the set of intervals with the right boundary exactly at x.

Using these two scores we can obtain an estimation of the true position of the cancer gene as the interval $[x_L^*, x_R^*]$, where x_L^* and x_R^* are chosen as

$$(x_L^*, x_R^*) = \underset{y>x, [x,y] \text{ small}}{\arg\max} \; SL_x + SR_y$$

. We refer to this method as the "LR method."

2.2.6 Significance Testing

Thus far, we have seen how to estimate the putative location of a cancer gene (oncogene or tumor suppressor gene) either by maximizing the relative risk scores over many intervals, or by estimating other related scores that characterize the boundaries of the gene. Irrespective of which method is chosen, the result is always an interval, consisting of some number of markers; in the following, the computed interval is referred to as $I_{\max}$. The final step of our algorithm determines whether this finding is statistically significant, i.e., it assigns a p-value to $I_{\max}$. The method for oncogenes differs slightly from the method for the tumor suppressor gene, so we first present the tumor suppressor gene case, and then the oncogene case.

Tumor Suppressor Genes

The algorithm computes the p-value from the observed distribution of breakpoints (points where deletions start) along the chromosome (as given by the segmentation algorithm). It uses a null hypothesis, which assumes that no tumor suppressor gene resides on the chromosome, and consequently, the breakpoints can be expected to be uniformly distributed. Note that if a detailed and complete understanding of a genome-wide distribution of breakpoints were available, then it would pose little difficulty in changing the following discussions and derivations *mutatis mutandis*. However, in order to avoid any unnecessary biases in our estimators, we have chosen, for the time being, to focus only on an uninformative prior, as reflected in

our assumptions. We may now note that if indeed $I_{\max}$ is a tumor suppressor gene, then its neighborhood could be expected to contain an unusually large number of breakpoints, thus signifying presence of a deviant region, which cannot be explained simply as random fluctuations in the null distribution of breakpoints. Therefore, after counting the number of breakpoints on the chromosome (counted as, say, N) and the number of breakpoints in the interval $I_{\max}$ (counted as, say, k) across all samples, we need to address the following question: how unusual is it to find k breakpoints in a region of length $w = |I_{\max}|$, given that there are N breakpoints uniformly distributed across the chromosome? We answer this question using results from the theory of scan statistics (Glaz et al. 2001 [3]), as follows.

Let S_w be the largest number of breakpoints in any interval of fixed length w (the interval contains a fixed number of markers). This statistic is commonly referred to as the scan statistic, and provides the necessary tool for our computation. Using this new notation, we answer the question we had posed: namely, how likely it is that we have k (out of N) breakpoints in any interval of length $w = |I_{\max}|$. The probability of this event is exactly $P(S_w \geq k)$.

Wallenstein and Neff (1987 [4]) have derived an approximation for $P(S_w \geq k)$, using the following notations. Let

$$\begin{aligned} b(k; N, w) &= \binom{N}{k} w^k (1-w)^{N-k}, \\ G_b(k; N, w) &= \sum_{i=k}^{N} b(i; N, w). \end{aligned}$$

Then

$$P(S_w \geq k) \approx (kw^{-1} - N - 1)b(k; N, w) + 2G_b(k; N, w), \tag{2.5}$$

which is accurate when $P(S_w \geq k) < 0.10$ and remains so, even for larger values.

Note that, for the above formula to be applicable, w must take values in $[0, 1]$. Therefore, in our derivation below, we use a normalized w, computed as the number of markers in the interval $I_{\max}$ divided by the total number of markers on the chromosome.

To illustrate how this approximation of the p-value performs, in Figure 2.2, we plot the calculated p-values against different numbers of breakpoints k, while examining the effect of different window-sizes, w. We have used the following assumptions: the total number of breakpoints is $N = 50$, $k \in \{1 \dots 20\}$ and $w \in \{\frac{1}{300}, \frac{1}{200}, \frac{1}{100}, \frac{1}{50}, \frac{1}{20}, \frac{1}{10}\}$. (Thus, w is normalized as the number of markers in the interval divided by the total number of markers on the chromosome).

Since the computation of p-values in (2.5) depends on the size of the interval w and since the size $w = |I_{\max}|$ of the interval $I_{\max}$ (found either by the Max or LR method) might not be the optimal length (e.g. because of underestimation of the length of the tumor suppressor gene), we also examine intervals overlapping $I_{\max}$, but of slightly different lengths, and then compute a p-value as before. From the resulting p-values, we choose the smallest (most significant) value to measure the statistical significance. To account for the fact that multiple window sizes have been tested, we apply a conservative Bonferroni adjustment for the p-values (we multiply the p-values by the number of window sizes; we use windows of lengths

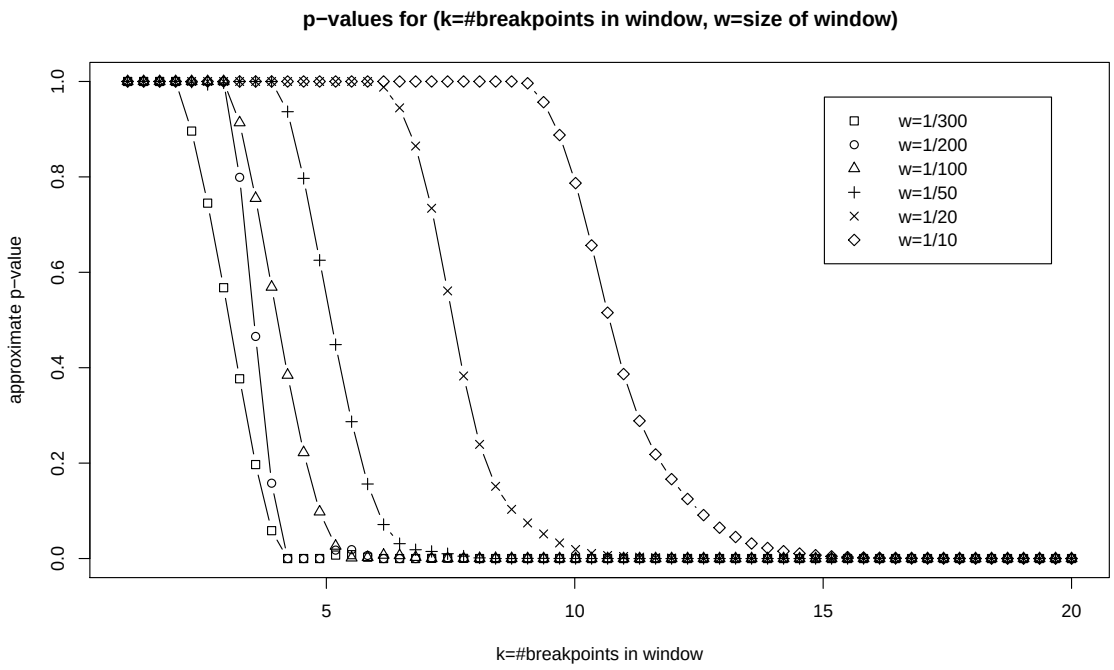


Figure 2.2: The tail probability $P(S_w \geq k)$ for different numbers of breakpoints k ($0 \leq k \leq 20$) and different window sizes w. S_w is the maximum number of breakpoints in a window of length w. The total number of breakpoints in the region is $N = 50$.

up to 10 markers in the analysis of both simulated and real data).

Oncogenes

The computation of p-values for the oncogene case is similar. However, since the amplifications encompass the entire oncogene, breakpoints tend to be far apart; in particular we do not expect breakpoints within the oncogene. Rather than looking at the distribution of breakpoints, we now consider the amplified interval as the event whose clustering in a small window may be unusual. If $I_{\max}$ is the interval with the highest score, then there are a number of samples (say k) that have this interval amplified. Let $w_{\max}$ be the length of this interval. Then if N is the total

number of amplifications of length at least $w_{\max}$ and w is some prespecified window size ($> w_{\max}$), then we can ask the question: how likely is it to have k amplifications out of N in a window of size w, if these amplifications were distributed uniformly in the genome? This way we have formulated our question in terms of a classic scan statistic problem.

The prespecified window size w is chosen such that it is expected to be larger than $w_{\max}$.

2.3 Results

We have applied our method to both simulated data as well as to a real dataset on lung cancer. Below, we describe the data sources, data qualities and computed results.

2.3.1 Simulated Data

We first describe a general simulation process used to generate the data. We use the tumor suppressor gene case to explain the process, and then shortly present the process for the oncogene case.

Tumor Suppressor Gene

We simulated data according to the generative process that was described earlier. The simulation works on a growing population of cells, starting with an individual normal cell, whose genome contains a single tumor suppressor gene at a known fixed position. As simulation proceeds, it introduces breakpoints at different positions in the genome, each occurring as a Poisson process with rate parameter, μ. At each of these breakpoints, it also postulates a deletion with length distributed as an Exponential random variable with parameter λ. Once, in some cell in the population, both copies of the tumor suppressor gene (TSG) become non-functional (either by homozygous deletion or hemizygous deletion in the presence of other mutations), the resulting pre-cancerous cell in the simulation starts to multiply indefinitely. Over time, the new progenitor cells also incur other independent "collateral damages" (i.e., deletions). Finally, the simulator randomly samples

the population for tumor cells, mimicking the micro-dissection process used by a physician, and thus, assuming that collected sample exhibits a composition made up of different tumor cells and some normal cells as well. In our simulations, we have assumed that even the normal cells have some random deletions, whereas the different tumor cells all come from the same ancestral pre-cancerous cell (Figure 2.3).

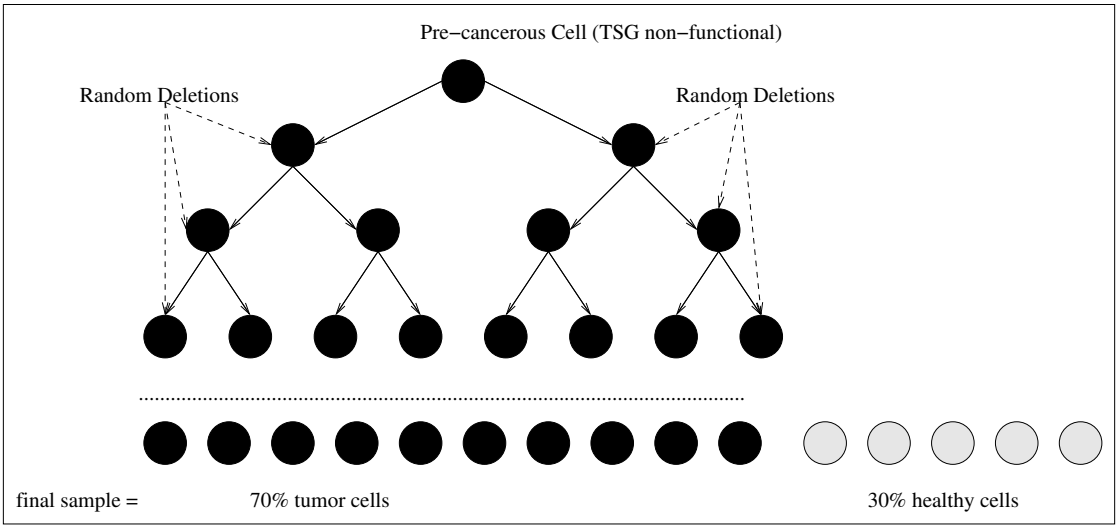


Figure 2.3: Depiction of the simulation process described in text. A single pre-cancerous cell (both copies of the TSG are non-functional) starts multiplying indefinitely. Over time, the new progenitor cells also incur other independent damage (i.e. deletions). The tumor sample that we collect is a composition of different tumor cells and some normal cells as well.

In all our simulations we have fixed the parameters, as listed below.

- $N = 50$ = Number of diseased individuals.
- $G = 100$ Mb = Length of the chromosome.
- $P = 10,000$ or $P = 5,000$ = Total number of probes (implying an average resolution of 10 Kb, and 20 Kb, respectively).

- $C = 100$ = Total number of cells per tumor sample with 70% tumor cells and 30% normal cells.
- $\mu \cdot G = 2$ = Mean number of breakpoints per cell. (This value corresponds to the background deletions occurring after the TSG becomes non-functional).
- $\frac{1}{\lambda} = 50$ Kb = Mean length of a deletion.
- TSG= $[10.0 \text{ Mb}, 10.1 \text{ Mb}]$. (TSG is represented by an interval starting at 10.0 Mb and has a length of 100 Kb.)

To the resulting copy numbers, we have added an independent Gaussian noise ($\sim N(0, 0.1^2)$). The simulated data were segmented using the publicly available software described in Daruwala et al. 2004 [2]. A segment was called deleted if $\log_2$ of the segmental-mean-ratio (test to normal) for that segment was less than a threshold value of $\log_2(1/2) = -1.0$.

Table 2.1 describes the different simulated scenarios, we have used. They all share the same set of parameters as described earlier, with an additional complexity to reflect differences in the composition of the starting population: some samples are assumed to be diseased because of mutations in the TSG ($p_{\text{homozygous}} + p_{\text{hemizygous}}$) and some samples are sporadic (p_{sporadic}). Among the samples with mutations in the TSG, some have only homozygous deletions ($p_{\text{homozygous}}$) and some have only hemizygous deletion of the TSG ($p_{\text{hemizygous}}$). Furthermore, the sporadic samples are assumed not to have deletions in the TSG under investigation; that is, they only have background deletions.

Model	$p_{\text{homozygous}}$	$p_{\text{hemizygous}}$	p_{sporadic}
1	100%	0%	0%
2	50%	50%	0%
3	0%	100%	0%
4	50%	0%	50%
5	25%	25%	50%
6	0%	50%	50%

Table 2.1: Six simulated models. $p_{\text{homozygous}}$ represents the percentage of samples in the dataset with homozygous deletions, $p_{\text{hemizygous}}$ is the percentage of samples with hemizygous deletions and p_{sporadic} is the proportion of samples with no deletion in the TSG under investigation (randomly diseased)

Oncogene

For oncogenes we use a similar model to simulate the data, with only a few differences: amplifications are generated instead of deletions; the precancerous cell has the entire oncogene amplified; and now we only simulate two models: model 1 where all samples have amplifications in the oncogene and model 2 where only 50% of the samples have amplifications in the oncogene (Table 2.2). The other parameters are the same as in the tumor suppressor gene case.

Model	p_{sporadic}
1	0%
2	50%

Table 2.2: Two simulated models. p_{sporadic} is the proportion of samples that only have random amplifications (randomly diseased)

Performance Measure

The performance of our method was evaluated by the Jaccard measure of overlap between the estimated position of the oncogene or tumor suppressor gene and the real position used in the simulation. Note that if E is the estimated interval and

T is the true one, then the Jaccard measure is defined simply as:

$$J(E,T) = \frac{|E \cap T|}{|E \cup T|},$$

where $|E \cap T|$ is the length of the interval common to both, i.e., the interval $E \cap T$.

We also tested the capacity of the inferred cancer gene as a possible biomarker for cancer detection or classification. More precisely, we measured, for a postulated oncogene or tumor suppressor gene, its sensitivity, which is defined as the percentage of diseased samples that have the estimated oncogene amplified or TSG deleted. For models that also contain sporadic samples, we considered, in our calculation of sensitivity, only the more meaningful situations, consisting only of samples that are diseased because of mutations in the cancer gene under investigation.

Results for Tumor Suppressor Gene

Tables 2.3 and 2.4 present our results, summarizing overlap and sensitivity measures for each of the 6 models outlined above and for the two marker resolutions simulated: 10 Kb and 20 Kb. The numbers, appearing in the table, are after averaging over 50 datasets simulated under the corresponding models. In all cases, the estimated p-value is very small (less than .001). These tables use the following abbreviations to denote the two competing methods for estimating the position of the TSG (for more details, see Methods section): "LR" refers to the LR method, which scores the boundaries (left and right) of intervals, and "Max" refers to the Max method, which scores only the intervals.

Model	Jaccard M. LR	Jaccard M. Max	Sensitivity LR	Sensitivity Max
1	0.82 ± 0.11	0.72 ± 0.23	0.80 ± 0.08	0.79 ± 0.10
2	0.84 ± 0.12	0.67 ± 0.24	0.69 ± 0.10	0.67 ± 0.13
3	0.84 ± 0.10	0.62 ± 0.30	0.56 ± 0.11	0.54 ± 0.13
4	0.74 ± 0.15	0.23 ± 0.19	0.80 ± 0.14	0.69 ± 0.12
5	0.73 ± 0.16	0.33 ± 0.25	0.69 ± 0.12	0.59 ± 0.16
6	0.74 ± 0.17	0.26 ± 0.25	0.54 ± 0.12	0.46 ± 0.12

Table 2.3: Overlap between true location and estimated location of the TSG and the resulting sensitivity for each of the six simulated models (Table 2.1). LR and Max refer to the two methods used to estimate the location of the TSG. Average inter-marker distance is 10 Kb.

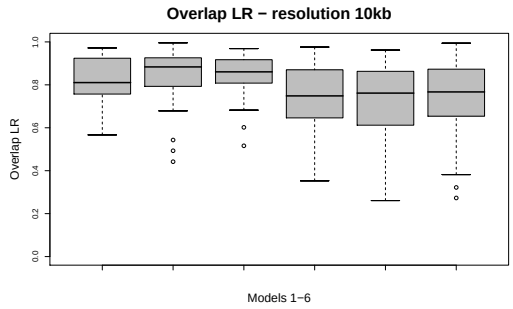
Model	Jaccard M. LR	Jaccard M. Max	Sensitivity LR	Sensitivity Max
1	0.70 ± 0.15	0.44 ± 0.27	0.59 ± 0.16	0.56 ± 0.16
2	0.70 ± 0.19	0.38 ± 0.30	0.46 ± 0.14	0.43 ± 0.15
3	0.68 ± 0.20	0.43 ± 0.30	0.38 ± 0.14	0.34 ± 0.16
4	0.60 ± 0.21	0.25 ± 0.21	0.60 ± 0.18	0.55 ± 0.15
5	0.65 ± 0.20	0.24 ± 0.22	0.46 ± 0.15	0.40 ± 0.14
6	0.58 ± 0.28	0.27 ± 0.28	0.37 ± 0.15	0.33 ± 0.14

Table 2.4: Overlap between true location and estimated location of the TSG and the resulting sensitivity for each of the six simulated models (Table 2.1). LR and Max refer to the two methods used to estimate the location of the TSG. Average inter-marker distance is 20 Kb.

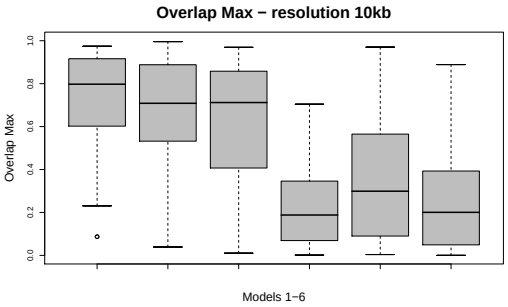
In order to present a better understanding of the entire distribution of scores, we have also plotted boxplots for the Jaccard measure and also for the sensitivity measure for all the simulated scenarios (see Figures 2.4, 2.5, 2.6, and 2.7).

Results for Oncogene

We also present the boxplots for the Jaccard measure of overlap for the two simulated models and the two resolutions simulated (Figures 2.8 and 2.9).



(a)

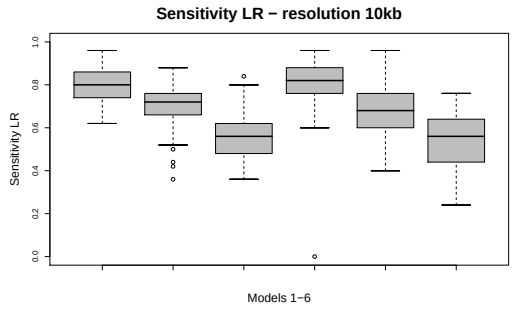


(b)

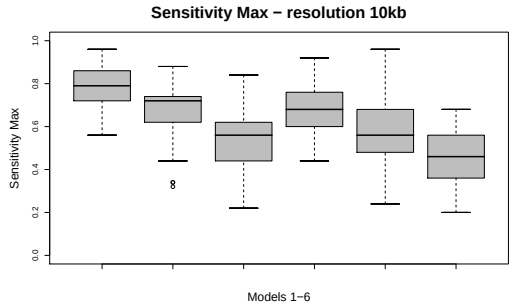
Figure 2.4: Boxplots of the Jaccard measure of overlap for each of the six models (Table 2.1). 50 datasets are simulated according to each model and the distribution of the resulting 50 overlap measures is depicted in each boxplot. Average inter-marker distance is 10 Kb. (a) LR, (b) Max.

2.3.2 Real Data (Lung Cancer)

Real data from cancer patients or cell lines, when examined with an available array technology, may contain other sources of error that may be correlated or may be nonstationary in a complicated manner that can never be modeled in the simulation; effects difficult to model include: degradation of genomic DNA, base-composition dependent PCR amplification in complexity reduction, presence of hyper-mutational regions, incorrect probes resulting from errors in reference genome assembly, contamination, cross-hybridization, and myriad others. Consequently, we cannot obtain full confidence in our methodologies, even though the



(a)

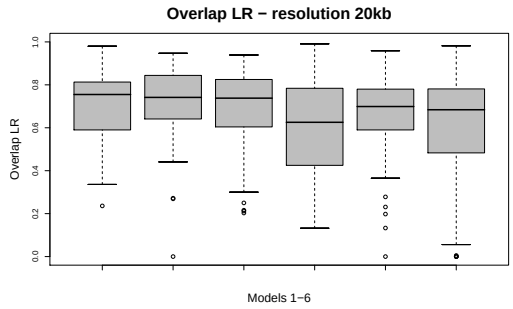


(b)

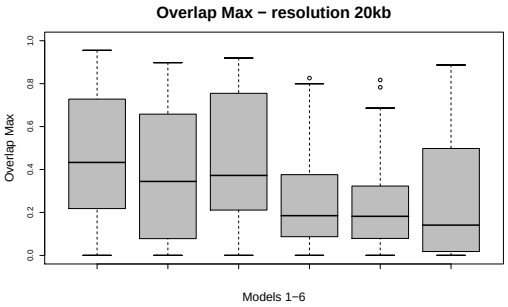
Figure 2.5: Boxplots of the sensitivity measure for each of the six models (Table 2.1). 50 datasets are simulated according to each model and the distribution of the resulting 50 sensitivity measures is depicted in each boxplot. Average inter-marker distance is 10 Kb. (a) LR, (b) Max.

results of the analysis of the simulated data were found very encouraging and even though the analysis showed that, in those ideal conditions underlying the simulation, our algorithm was able to detect with high accuracy and confidence the locations of the simulated disease genes.

In this section, we inspect the results of our method, when applied to a real dataset on lung cancer, which was originally published by Zhao et al. (2005 [5]). Seventy primary human lung carcinoma specimens were used in our analysis. For each sample, copy number changes at $\sim 115,000$ SNP loci throughout the genome were measured and recorded. We used an unpublished Affy normalization and



(a)

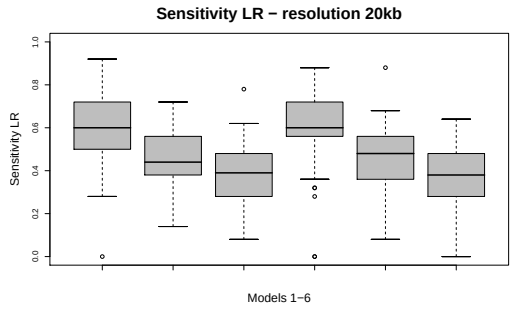


(b)

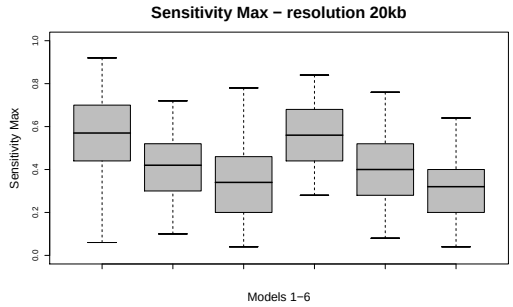
Figure 2.6: Boxplots of the Jaccard measure of overlap for each of the six models (Table 2.1). 50 datasets are simulated according to each model and the distribution of the resulting 50 overlap measures is depicted in each boxplot. Average inter-marker distance is 20 Kb. (a) LR, (b) Max.

summarization software (Mishra et al. 2006, unpublished data) to convert the raw data into genotypic copy number values. Next, as with the simulated data, we applied the segmentation algorithm (Daruwala et al. 2004 [2]) to the raw $\log_2$ signal ratio (test to normal) data and obtained a partition of the data into segments of probes with the same estimated mean. Since the previous steps were found to average out the random noises across groups of probe sets and neighboring probes, variance parameters were quite low and were discarded from further analysis.

For this dataset, we next determined that a chromosomal segment could be treated as deleted if the segment had an inferred $\log_2$ ratio less than a thresh-



(a)

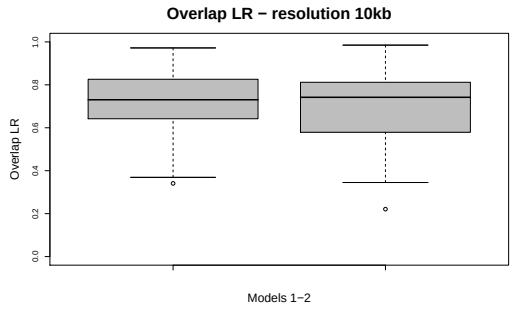


(b)

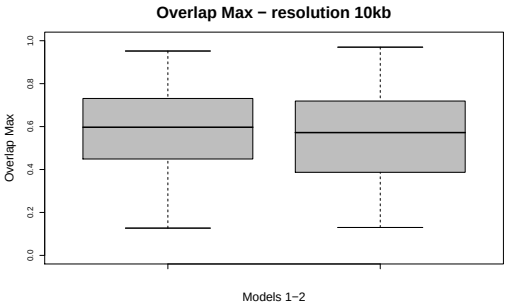
Figure 2.7: Boxplots of the sensitivity measure for each of the six models (Table 2.1). 50 datasets are simulated according to each model and the distribution of the resulting 50 sensitivity measures is depicted in each boxplot. Average inter-marker distance is 20 Kb. (a) LR, (b) Max.

old value of -1.0. Figure 2.10 depicts the histogram for the $\log_2$ ratio values for all SNPs in all 70 tumors, together with an empirical null density fitted to the histogram: $\mathrm{N}(\hat{\mu_0}, \hat{\sigma_0}^2)$. The overall threshold is defined as $\hat{\mu_0} - 2\hat{\sigma_0} = -1.0$. (Supplemental Information provides further details on the computation of this cutoff-threshold.)

For amplifications, the cutoff used was 0.5. There are few segments with $\log_2$ ratio larger than 1.0 (as the threshold resulting from $\hat{\mu_0} + 2\hat{\sigma_0} = 1.0$), and therefore we used $0.5 > \log_2\left(\frac{3}{2}\right)$ as a cutoff.



(a)

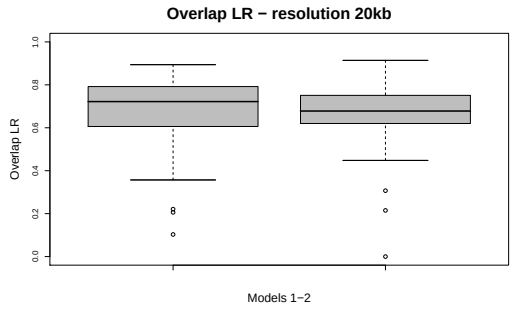


(b)

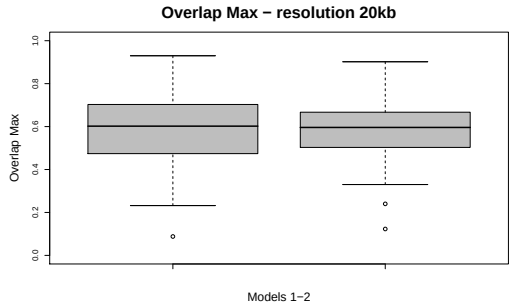
Figure 2.8: Boxplots of the Jaccard measure of overlap for each of the two models (Table 2.2). 50 datasets are simulated according to each model and the distribution of the resulting 50 overlap measures is depicted in each boxplot. Average inter-marker distance is 10 Kb. (a) LR, (b) Max.

Tumor Suppressor Gene

The significant regions (genome-wide significance level $< .01$) are presented in Tables 2.5 and 2.6. The intervals reported were computed using the Max method. Most of the regions detected have been previously reported as deleted in lung cancer (e.g. 5q21, 14q11). Most significantly, some of the intervals found overlap some good candidate genes, that may play a role in lung cancer (e.g. *MAGI3*, *HDAC11*, *PTPRD*, *PLCB1*). Also, the regions 3q25 and 9p23 have been found for the first time to be homozygously deleted by Zhao et al. (2005 [5]).



(a)



(b)

Figure 2.9: Boxplots of the Jaccard measure of overlap for each of the two models (Table 2.2). 50 datasets are simulated according to each model and the distribution of the resulting 50 overlap measures is depicted in each boxplot. Average inter-marker distance is 20 Kb. (a) LR, (b) Max.

Oncogene

The significant results for oncogenes are presented in Table 2.7. An interesting feature of this lung cancer dataset is that on average the amplifications are longer than the deletions and therefore the segments detected for oncogenes tend to contain more genes than the segments detected when hunting for tumor suppressor genes.

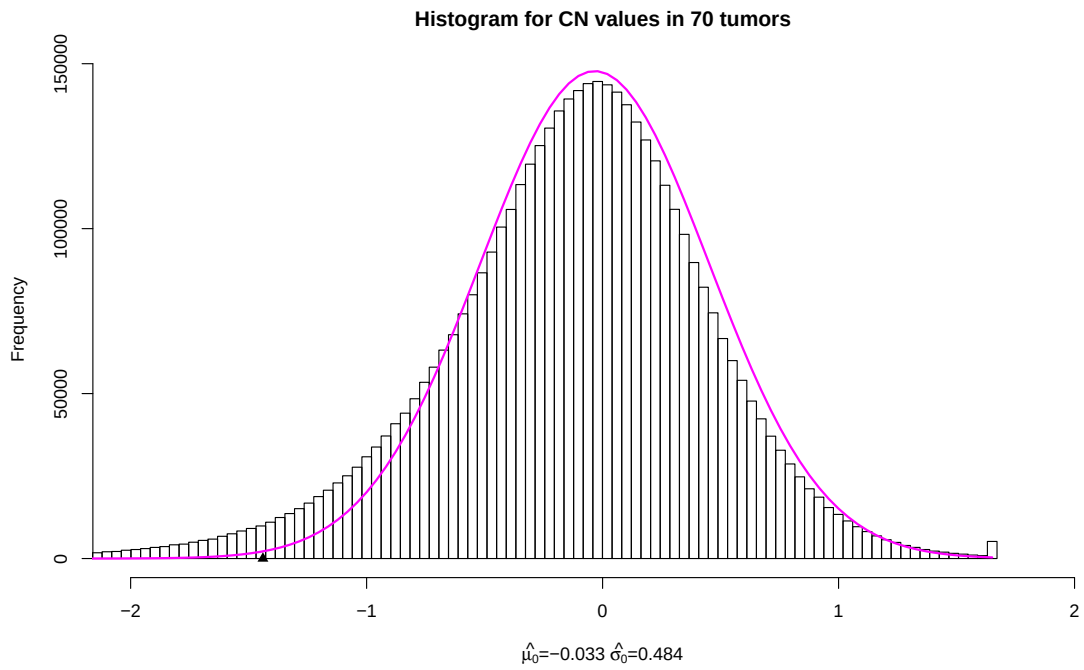


Figure 2.10: The histogram for the $\log_2$ ratio values for all SNPs in all 70 tumors, together with an empirical null density fitted to the histogram: $N(\hat{\mu_0}, \hat{\sigma_0}^2)$.

2.4 Comparison of Results for Oncogenes and Tumor Suppressor Genes

Chromosome	Exact interval	Comments
1p13.2	113.76 − 113.77Mb	*MAGI3* maps to this region[a]
3p25.1	13.51 − 13.56Mb	*HDAC11* maps to this region[b]
3q25.1	151.16 − 151.16Mb	[c]
4q34.1	173.46 − 173.46Mb	[d]
5q14.1	79.16 − 79.18Mb	
5q21.3	106.95 − 107.0Mb	[e]
6q14.1	78.50 − 79.02Mb	
7p15.3	20.48 − 20.49Mb	[f]
9p23	10.02 − 10.05Mb	[g]
9p21	32.85 − 32.85Mb	[h]

[a] *PTEN/MMAC* and *MAGI3* cooperate to modulate the kinase activity of *AKT/PKB* involved in the inhibition of apoptosis (Wu et al. 2000 [6])

[b] Frequent allelic losses have been reported in this region in lung and other solid tumors. Also in vitro studies suggest that this region is able to suppress growth of tumor cells (Rimessi et al. 1994 [7])

[c] Homozygous deletions in this region have been found using this dataset by Zhao et al. (2005 [5]) (not previously known)

[d] Deletions in this region have been reported in lung cancer (Shivapurkar et al. 1999 [8], Tseng et al. 2005 [9]). *LOC442117* (similar to *GalNAc transferase 10*) maps to this interval

[e] This region is known to be frequently deleted in lung cancer (Hosoe et al. 1994 [10])

[f] LOH has been found in this region in cancer (Inoue et al. 2003 [11])

[g] Homozygous deletions in this region have been found using this dataset by Zhao et al. (2005 [5]) (not previously known); this region is upstream of PTPRD (protein tyrosine phosphatase, receptor type, D), gene currently being investigated for its potential implication in lung cancer by Zhao et al.

[h] Deletions in this region have been reported in lung cancer

Table 2.5: Significant Deleted Regions in the Lung Cancer Dataset: Chromosomes 1-9

2.5 Discussion

The focus of this chapter has been a novel statistical method and its application to the problem of estimating the location of cancer genes from array-CGH data characterizing segmental changes in cancer genomes. The underlying algorithm computes a multipoint score for all intervals of consecutive probes. The computed score measures how likely it is for a particular genomic interval to be an oncogene or a tumor suppressor gene implicated in the disease. We propose two ways to estimate the location: the LR method and the Max method. In our experience,

Chromosome	Exact interval	Comments
10p13	$17.17 - 17.20$Mb	
10q24.1	$97.83 - 97.94$Mb	*BLNK* maps to this interval [a]
11p15.4	$4.9 - 5.0$Mb	[b]
12q14	$66.28 - 66.29$Mb	
14q11.2	$20 - 20.1$Mb	[c]
16q24	$82.8 - 82.8$Mb	*CDH13*, known TSG, is deleted in lung cancer
17q21	$39.5 - 39.6$Mb	*BRCA1, HDAC5* map to this region. [d]
19p13.3	$0.34 - 2$Mb	*LKB1* is deleted in lung cancer (Sanchez-Cespedes et al. 2002 [14])
20p12	$8.7 - 8.8$Mb	*PLCB1* maps to this region [e]
21q21.2	$23.27 - 23.38$Mb	[f]

[a] *BLNK* is a putative tumor suppressor gene (Flemming et al. 2003 [12])

[b] Deletions in this region have been found in several cancers (Karnik et al. 1998)

[c] Loss of heterozygosity in this region has been reported in lung cancer (Abujiang et al. 1998 [13]). *APEX1* maps to this region; this gene is implicated in the DNA repair mechanism, and in control of cell growth

[d] *HDAC5* play a critical role in transcriptional regulation, cell cycle progression, and developmental events

[e] This gene is important in the control of cell growth, therefore may be of interest in diseases like cancer that involve alterations of the control of cell growth (Peruzzi et al. 2002 [15])

[f] This region has been found deleted in lung cancer (Lee et al. 2003 [16])

Table 2.6: Significant Deleted Regions in the Lung Cancer Dataset: Chromosomes 10-22

both methods perform well, with the LR method being more accurate than the Max method in the simulation experiments, especially when the marker density is relatively high (i.e., $100,000$ or more probes spanning the human genome). However on the real data, due to the increased noise, we found that the Max method gives better (shorter) intervals.

We have evaluated the efficacy of our method by applying it to both simulated data and real data and concluded that the results are significant. In the ideal conditions, as in our simulations, our estimation method seems to perform exceedingly well. In particular, with an average inter-marker distance of 10–20 Kb, the overlap between the estimated position and the true position of the cancer gene is over

Chromosome	Exact interval	Comments
3q28	193.06 − 193.33Mb	[a]
5p15.3	1.81 − 2.06Mb	*LOC389267* maps to this region[b]
6p22.3	17.5 − 18.38Mb	
8q24	129.17 − 129.20Mb	*PVT1/MYC* map to this region[c]
11p15	4.93 − 4.93Mb	*OR51A2* maps to this interval [d]
12p11	32.63 − 33.10Mb	[e]
20q11.23	34.23 − 35.16Mb	[f]

[a]Overexpression of this region has been found in lung cancer (Tonon et al. 2005 [17]); *LOC647315*, similar to microtubule-associated protein 6 isoform 1, maps to this region

[b]*LOC389267* is similar to MUC-4; MUC-4 is known to be overexpressed in lung cancer (Nguyen et al. 1996 [18])

[c]*PVT1* oncogene homolog, MYC activator

[d]*OR51A2* is member of G-protein-coupled receptors (GPCRs) and among other functions, GPCRs are involved in cell growth stimulation and cell proliferation. Another member of this family, *OR51E2*, is overexpressed in prostate cancer

[e]Same interval found amplified by Zhao et al. 2005 [5]

[f]Amplification in this region has been found in lung cancer by Zhao et al. 2004 [19]

Table 2.7: Significant Amplified Regions in the Lung Cancer Dataset: Chromosomes 1-22

50%. While the simulations are only an attempt to approximate the real data, the results obtained show that our method is reliable in pinpointing the location of putative cancer genes. In addition, we also applied our method to a real dataset on lung cancer. We have obtained many regions that were previously reported as amplified or deleted in lung cancer. Most significantly, the intervals within the regions 3p25, 5p15, 8q24, 11p15, 16q24, 19p13, and 20p12 overlap some good candidate genes (*HDAC11*, *LOC389267*, *PVT1/MYC*, *OR51A2*, *CDH13*, *LKB1* and *PLCB1* respectively), that could play an important role in lung cancer. Several other regions have also been known to harbor amplifications or deletions in lung cancer patients. In addition we have detected a few regions, previously unreported, that warrant more detailed examination to understand their relation to lung cancer, for example 6q14, 6p22, 7p21.

We note that in comparative experimental settings such as those used by array CGH, one needs to keep track of the meaning of "normal genomes," since there are at least three kinds of "normal" genomes involved in this analysis: namely, the normal genome (or genomes) used in designing the ArrayCGH (or SNP) chips, the genomes from a population with similar distribution of polymorphisms (both SNPs and CNPs) as the patient under study, and finally, the genome from a normal cell in the same patient. The simplest situation in terms of statistical analysis is when the normal genome is the one from a normal cell from the same patient, and this is at the basis of the analysis we presented here. The other information can be augmented in pre-processing or post-processing steps, when the situation differs from this simplest one. Also, our scoring functions and the algorithm can be suitably modified, if it is deemed necessary that the polymorphisms in the probes and the population must be tracked. Other similar, but not insurmountable, complications arise, if one were to also model the "field effects" in the normal genomes from the patient.

In summary, we have formulated a general approach that is likely to apply to other problems in genetics, if a suitable generative model and an accompanying score function can be accurately formulated; the rest of the method works out *mutatis mutandis*. Unlike the classical approach, normally employed in most genetics studies, the proposed approach does not employ a locus-by-locus analysis and thus does not depend on linkages between a marker and genes, harboring causative mutations. The present algorithm exploits the fact that, when genome-wide high-density markers are studied, as with whole-genome arrays, one could look for the

interesting genes directly by examining every plausible genomic interval delineated by a group of consecutive markers. Such an interval-based analysis is more informative and allows it to assign significance-values to estimated intervals using scan statistics. We note that there have been other usages of scan statistics to genetics in different contexts, e.g., the work of Hoh and Ott (2000 [20]).

We also note that many variants of our method can be further enriched by augmenting other auxiliary information to the interval: underlying base-compositions (e.g., GC content, Gibbs free energy, codon bias) in the genomic interval, known polymorphisms (e.g., SNPs and CNPs), genes and regulatory elements, structures of haplotype blocks, recombination hot spots, etc. Note, however, that at present, in the absence of reliable and complete statistical understanding of these variables, it is safe to work only with uninformative and simple priors of the kind we have already incorporated in our algorithm.

Nonetheless, the utility of our algorithm will most likely be first validated with the simplest forms of array-CGH data and in the context of cancer: an area currently under intense study. We will gain more confidence as these methods are used for bigger datasets, larger number of patients and for many different cancers. There are few competing methods that bear some minor resemblance to our algorithm: For instance, the STAC method (Diskin et al., personal communication) also finds gene-intervals from array-CGH data, but it does not employ any generative model to compute a score to be optimized, nor does it compute a statistical significance based on such a model. (It uses a permutation approach to create a null-model). A detailed comparison will indicate how much statistical power is

gained when a more faithful but parsimonious generative model is used.

We recognize that a lot more remains to be done to completely realize all the potential of the proposed analysis. There may be more subtle correlations between intervals we detect, and such correlations (or anti-correlations) may hint at subtle mechanisms in play in cancer progression. If various regions of a polyclonal tumor can be analyzed separately, the distribution of important intervals may reveal much more details of the disease. There may be a critical need to stratify the patients into subgroups and analyze them separately in order to detect more subtle patterns. Once an important interval is detected (e.g., corresponding to a putative oncogene or tumor suppressor gene), one may wish to understand how the amplified or deleted intervals affecting the genes are spatially distributed. Such higher order patterns and motifs may paint a better picture about many varied genomic mechanisms responsible for the initiation and development of a cancer.

2.6 Web Resources

NCBI Human Genome Resources:

http://www.ncbi.nlm.nih.gov/genome/guide/human/

NYU Versatile MAP Segmenter:

http://bioinformatics.nyu.edu/Projects/segmenter/

2.7 Supplemental Material

2.7.1 Copy Number Distribution for Individual Samples

Our rule for calling a segment deleted is that the $\log_2$ of the segmental-mean-ratio (test to normal) for that segment be less than a certain threshold. To determine this threshold, we proceed as follows. For each sample in the dataset, we fit an empirical null density to the histogram of copy numbers; thus we obtain the null density $\mathrm{N}(\hat{\mu_0}, \hat{\sigma_0}^2)$. We then define the threshold for each individual sample as $\hat{\mu_0} - 2\hat{\sigma_0}$ and the average of these values is the overall cutoff we use. By this method, we obtain $c = -1.0$.

In Figures 2.11-2.15 we show for each sample the histogram of copynumbers together with the empirical null density fitted to the data. Below each plot are the estimated $\hat{\mu_0}$ (mean) and $\hat{\sigma_0}$ (standard deviation) for the empirical null density.

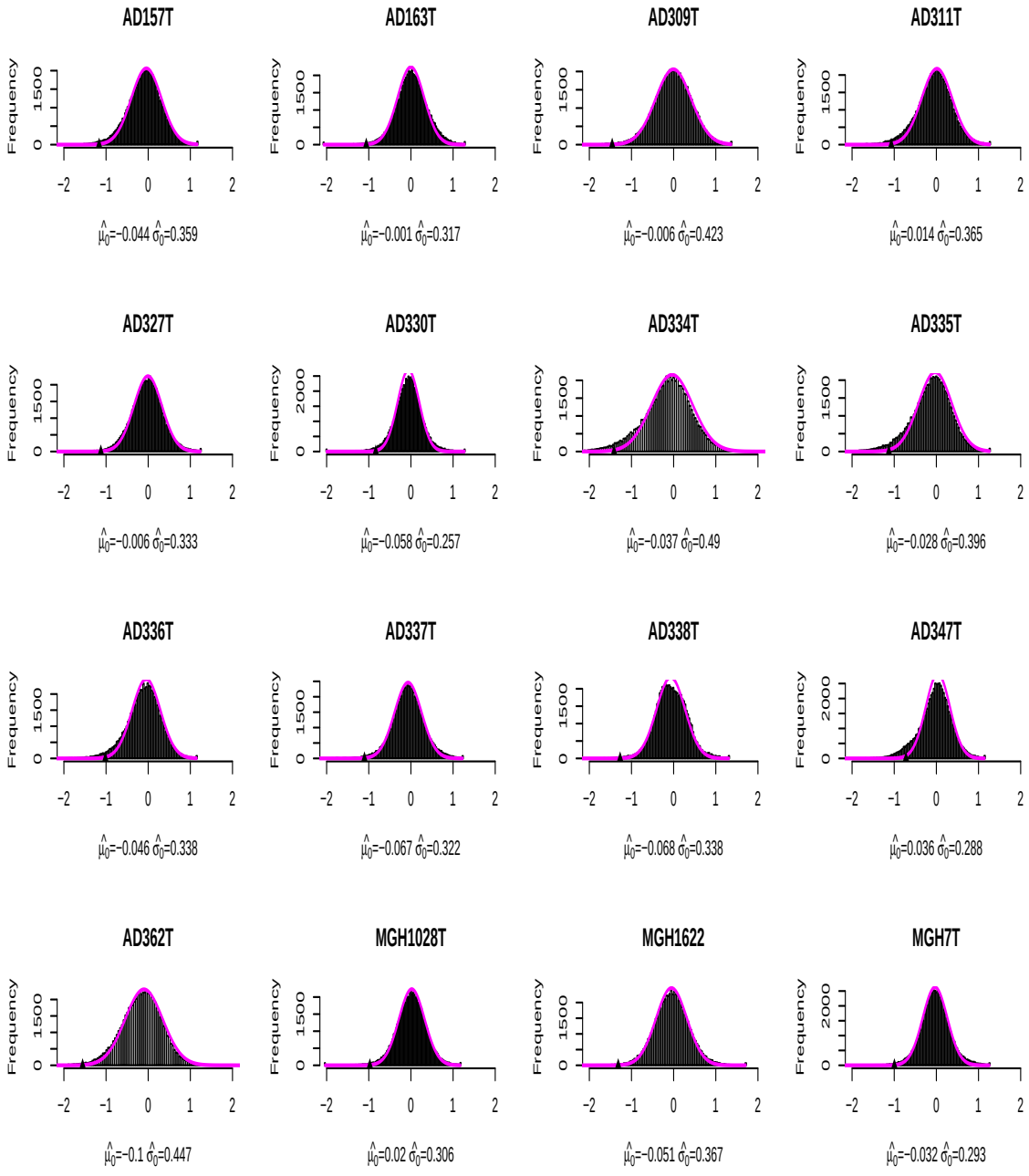


Figure 2.11: Copy number histogram; overlaid is the fitted empirical null density $\mathrm{N}(\hat{\mu_0}, \hat{\sigma_0}^2)$. Samples $1 - 16$

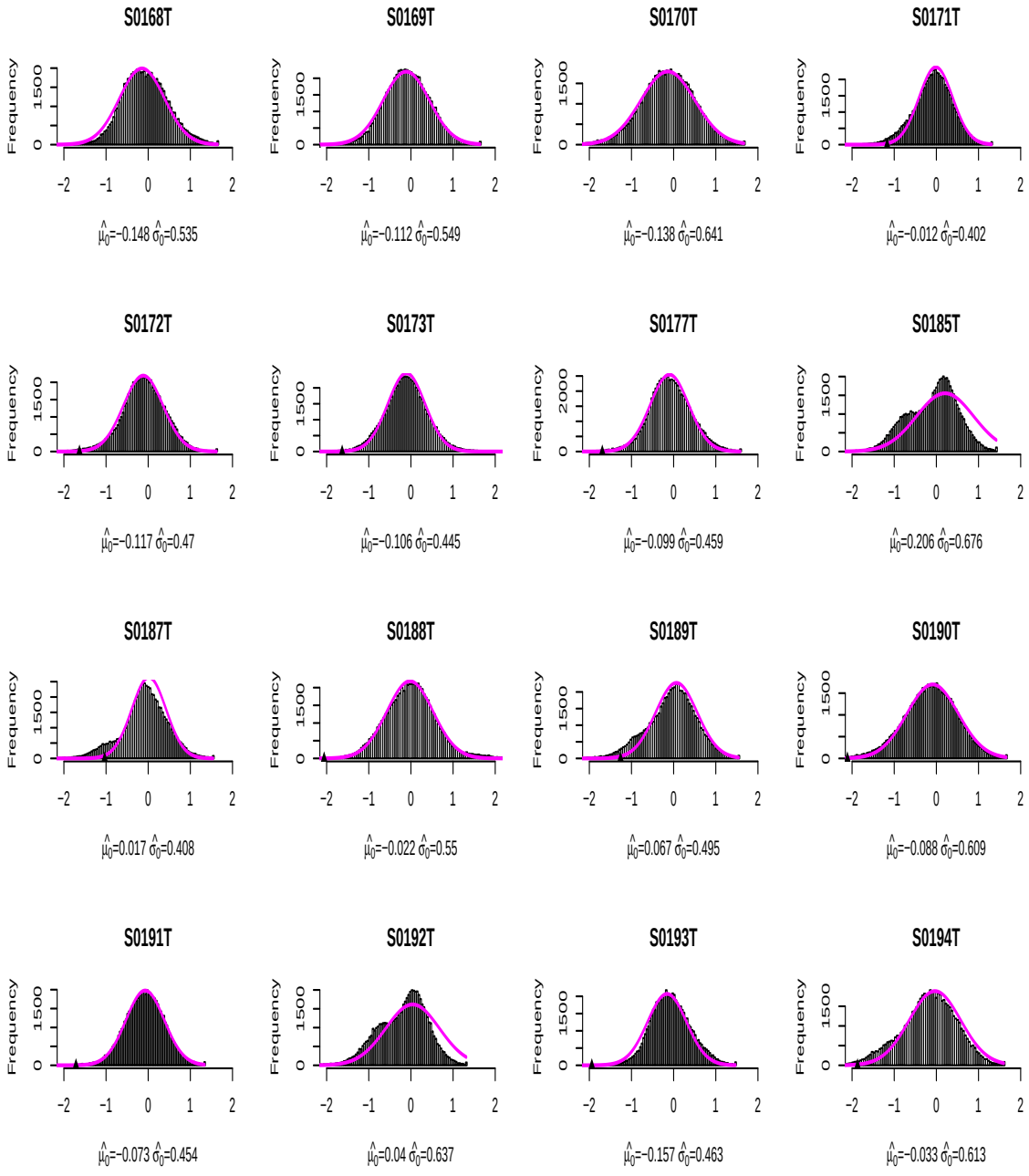


Figure 2.12: Copy number histogram; overlaid is the fitted empirical null density $\mathrm{N}(\hat{\mu_0}, \hat{\sigma_0}^2)$. Samples $17 - 32$

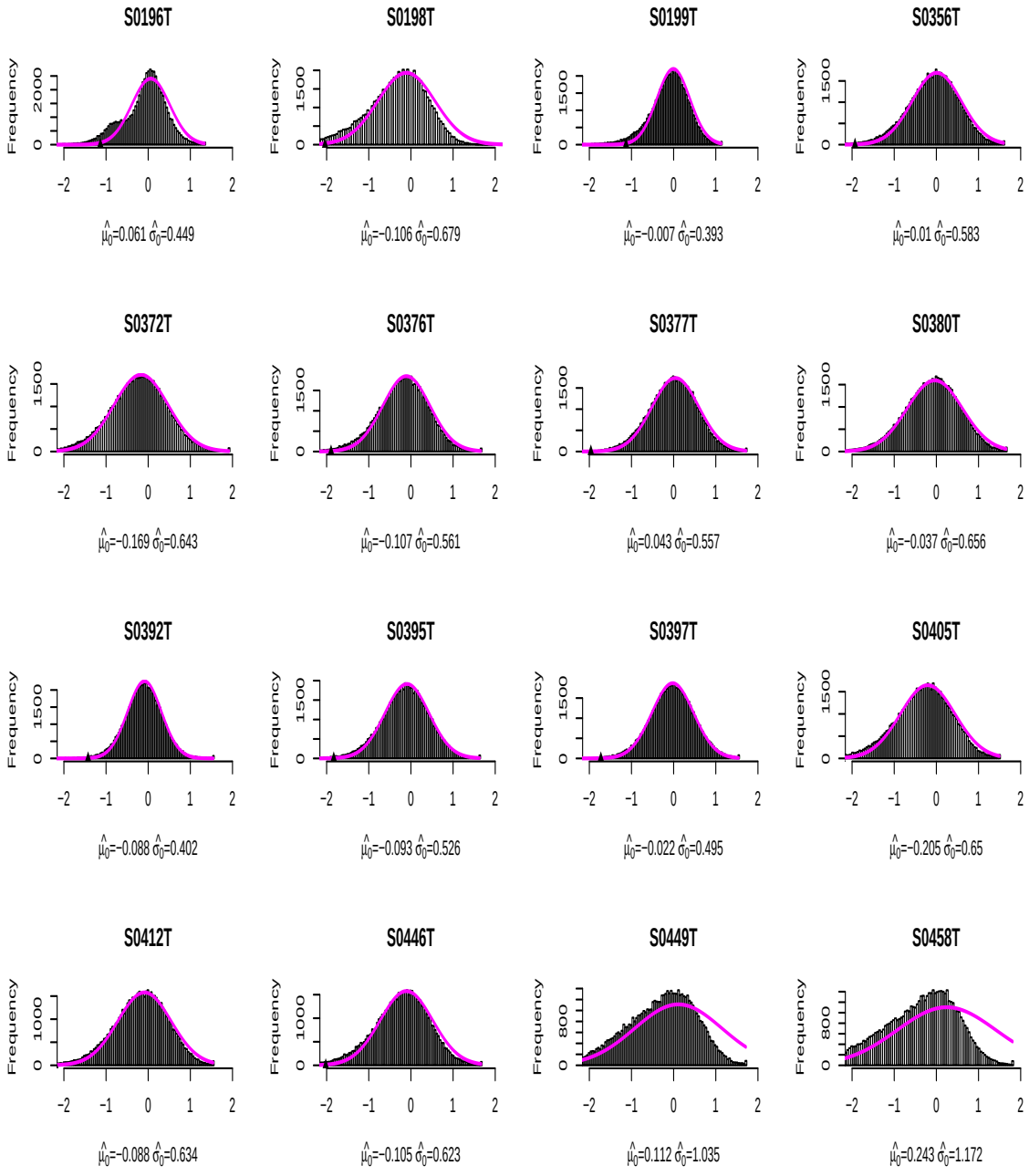


Figure 2.13: Copy number histogram; overlaid is the fitted empirical null density $\mathrm{N}(\hat{\mu_0}, \hat{\sigma_0}^2)$. Samples $33 - 48$

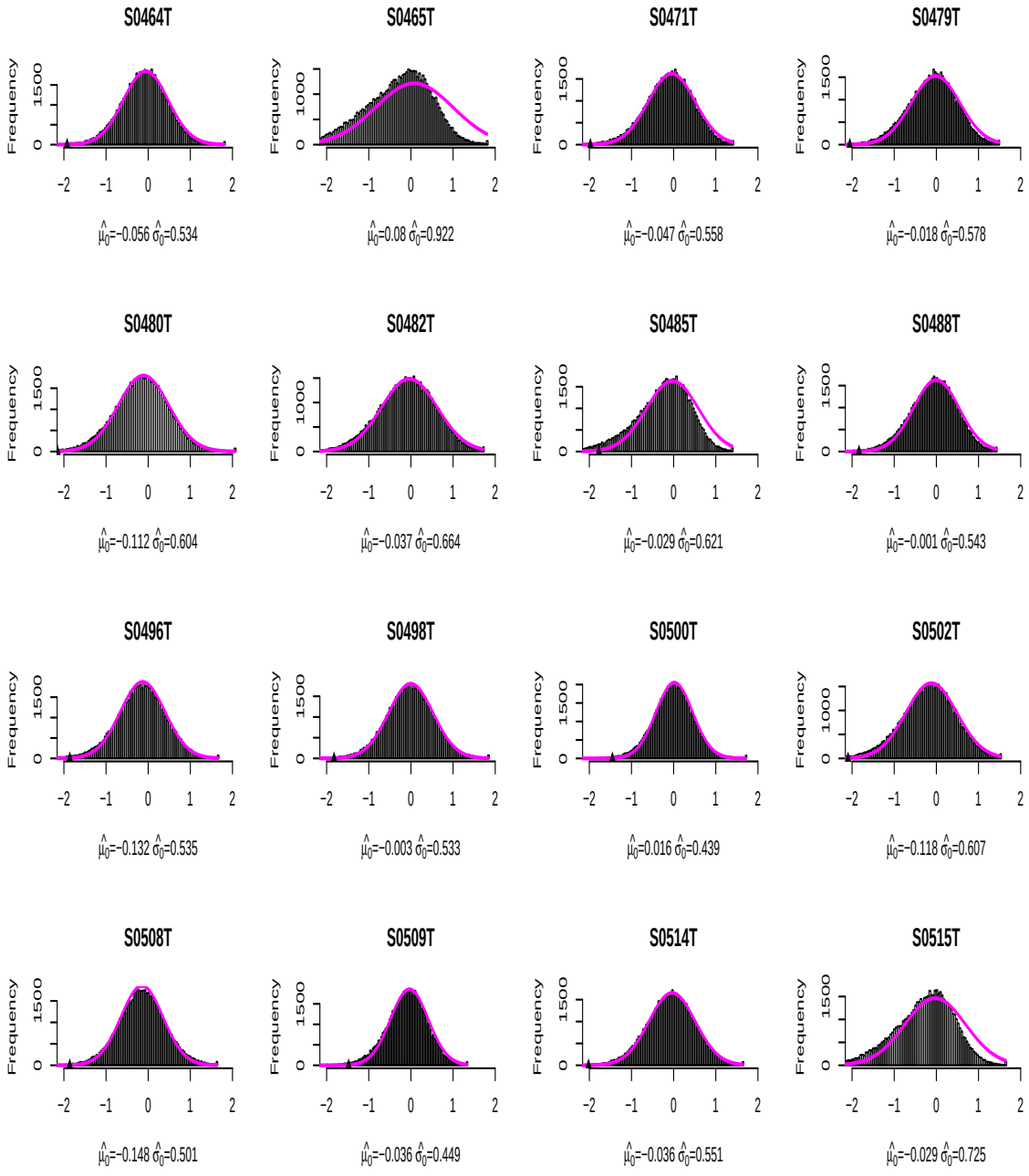


Figure 2.14: Copy number histogram; overlaid is the fitted empirical null density $\mathrm{N}(\hat{\mu_0}, \hat{\sigma_0}^2)$. Samples $49 - 64$

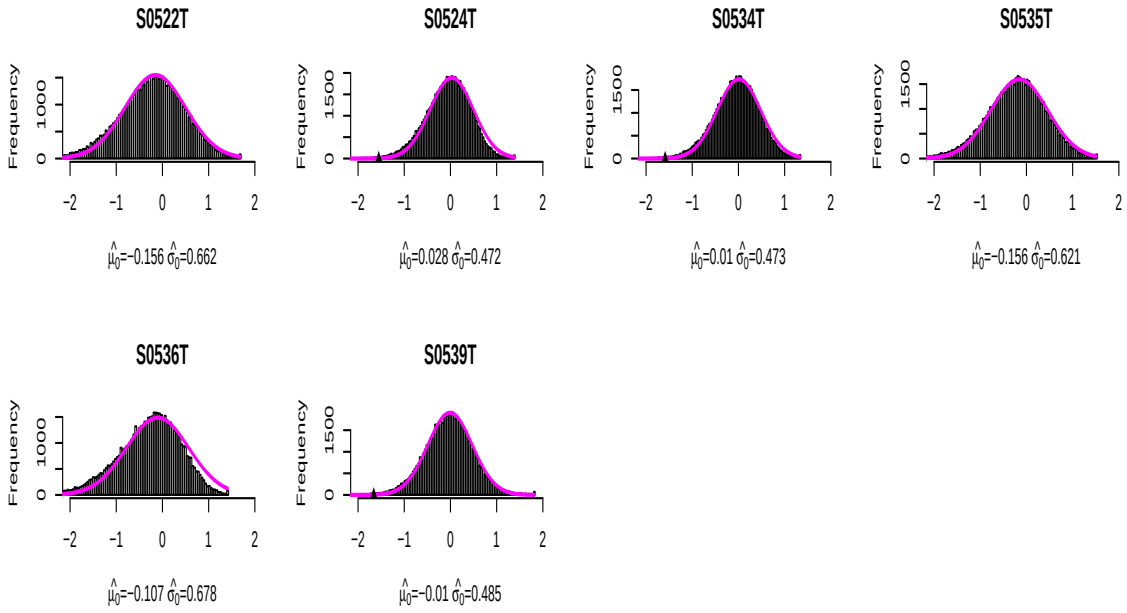


Figure 2.15: Copy number histogram; overlaid is the fitted empirical null density $\mathrm{N}(\hat{\mu_0}, \hat{\sigma_0}^2)$. Samples $65 - 70$

Chapter 3

Multilocus Linkage Analysis of Affected Sib-Pairs

SUMMARY: The conventional Affected Sib Pair methods evaluate the linkage information at a locus by considering only marginal information. We describe a multilocus linkage method that uses both the marginal information and information derived from the possible interactions among several disease loci, thereby increasing the significance of loci with modest effects. Our method is based on a statistic that quantifies the linkage information contained in a set of markers. By a marker selection-reduction process, we screen a set of polymorphisms and select a few that seem linked to disease. We test our approach on a genome-scan data for inflammatory bowel disease (InfBD) and on simulated data. On real data we detect 6 of the 8 known InfBD loci; on simulated data we obtain improvements in power of up to 40% compared to a conventional single-locus method. Our extensive simulations and the results on real data show that our method is in general

more powerful than single-locus methods in detecting disease loci responsible for complex traits. A further advantage of our approach is that it can be extended to make use of both the linkage and the linkage disequilibrium between disease loci and nearby markers.

3.1 Introduction

Traditional approaches to linkage analysis assign a score to each marker position by considering the linkage information given by that marker or a few nearby markers. These approaches have been very successfully applied to Mendelian diseases; however they have been less fruitful in the context of complex diseases. Because complex genetic diseases are caused by the action of several genes that can interact in a complicated manner, methods that can exploit interactions among multiple disease loci are expected to be more powerful. Here we report a novel linkage method for affected sib pairs (ASPs). Our approach screens a large number of polymorphisms and selects a few that appear to be linked to disease genes. The selection is based on an importance score assigned to each marker based on both marginal information as well as information coming from possible interactions among several disease loci.

Several multilocus linkage methods have been reported in the literature. These include model-based methods and model-free methods. The model-based methods calculate the full likelihood of disease and marker data under the assumed mode of inheritance (usually two-locus models, Schork et al. 1993 [21]). The model-free methods are based on comparing the observed allele-sharing among relatives

that are phenotypically alike with that expected under no linkage. Their main characteristic is that they do not assume a specific mode of inheritance. Examples include Cordell et al. (1995 [22]) and Farrall (1997 [23]) for ASPs. Their methods are based on computing a maximum likelihood statistic (MLS) and are restricted to two-locus models. More recently, Cordell et al. (2000 [24]) have presented a generalization of their earlier MLS method in Cordell et al. (1995 [22]) to several disease loci and affected relative pairs. Given linkage evidence at $m-1$ loci, the evidence at the mth locus is measured by the difference in MLS between the best fitting $m-1$ locus model and the best fitting m locus model. However, due to the sparseness of the data when m increases and the large number of parameters that need to be estimated in their model fitting procedure, the method is useful in practice only for the simultaneous analysis of at most ~ 3 disease loci. Also these methods are applicable only after a primary genome-screen has already been performed, when the number of loci under investigation is small.

Here we report a new screening method. Our method works on datasets with large number of markers and makes no assumption on the disease model, including the number of disease loci and their position in the genome. This new approach uses the interactions among several disease loci to help increase the importance of moderate effect disease loci relative to other noisy loci. The method is based on the repetition of a two-phase selection-reduction process. In the first step ("selection") we select a small set of markers at random from the available list of polymorphisms. In the next step ("reduction"), we remove the unimportant markers from the current set one by one in a stepwise fashion until all the remaining markers are

important or a single marker remains (we call these markers "returned"). At the end of this process we count how many times each marker was returned. Based on these counts we decide which markers are returned at significantly high frequency. The key technical aspect of this procedure is the definition of a statistic to measure the relative importance of a marker in the current set.

We apply this new approach to real data (Inflammatory Bowel Disease) as well as data simulated under several complex models. The results are very good. On the real data we confirm most of the known loci. On simulated data we show that our method is consistently more powerful than the single-locus methods currently in use.

The rest of the chapter is organized as follows. In Section 3.2 (Methods) we illustrate the theoretical aspects of our approach. In Section 3.3 (Results) we present our findings on a real dataset for Inflammatory Bowel Disease and on simulated data. We conclude in Section 3.4 (Discussion) with a discussion of our findings.

3.2 Methods

3.2.1 Linkage Measure

The core of our approach is the definition of a linkage measure for a set of markers. In this section we describe this measure.

Notation Most model-free methods for ASPs work with the genotypic identical-

by-descent (IBD) sharing at a locus, which can be 0, 1 or 2. In our approach we work with the allelic IBD status; in this case the IBD sharing can be 0 or 1, meaning the number of alleles a sibpair shares IBD transmitted from one of the parents. If the marker is not linked to disease, then the IBD sharing is 0 or 1 allele with equal probability 0.5. For several loci we define an IBD sharing vector, such that the ith component represents the sharing at the ith locus. For example, the IBD sharing vector 111 for three loci signifies that the sibpair shares 1 allele IBD at each of the three loci. Let n_{111}^{ijk} be the number of such sharing vectors in the dataset at three loci i, j, k.

Let $S = \{M_1, M_2, \ldots, M_k\}$ be a set of markers under evaluation. Then the measure is defined as:

$$\begin{aligned} H_{1\ldots k} &= w_k \Bigg[\frac{\sum_{i=1}^{k} \left(n_1^i - n_0^i\right)^2}{\binom{k}{1}} + \frac{\sum_{i<j} \left(n_{11}^{ij} - n_{00}^{ij}\right)^2}{\binom{k}{2}} + \ldots \\ &+ \frac{\sum_{i_1<\cdots<i_{k-1}} \left(n_{1\ldots1}^{i_1\ldots i_{k-1}} - n_{0\ldots0}^{i_1\ldots i_{k-1}}\right)^2}{\binom{k}{k-1}} + \left(n_{1\ldots1}^{1\ldots k} - n_{0\ldots0}^{1\ldots k}\right)^2 \Bigg] \end{aligned} \tag{3.1}$$

where

$$w_k = \frac{2^k}{2^k - 1} \tag{3.2}$$

The weight w_k is chosen such that when none of the markers in the set S is linked to disease, we have:

$$\mathrm{E}\left[H_{1\ldots k}\right] = \mathrm{E}\left[H_{1\ldots k-1}\right]$$

The rationale for this is that when S contains only unlinked markers, the linkage measure should remain constant when any marker is removed from the set (no drop or increase in the linkage measure). We assume further that the k markers in S are not linked among themselves. Under these assumptions: $p_{1\ldots1}^{1\ldots j} = p_{0\ldots0}^{1\ldots j} = \frac{1}{2^j}$.

Then we can write:

$$\frac{w_{k-1}}{w_k} = \frac{\mathrm{E}\left[\left(n_1^1 - n_0^1\right)^2\right] + \mathrm{E}\left[\left(n_{11}^{12} - n_{00}^{12}\right)^2\right] + \cdots + \mathrm{E}\left[\left(n_{1\ldots1}^{1\ldots k} - n_{0\ldots0}^{1\ldots k}\right)^2\right]}{\mathrm{E}\left[\left(n_1^1 - n_0^1\right)^2\right] + \mathrm{E}\left[\left(n_{11}^{12} - n_{00}^{12}\right)^2\right] + \cdots + \mathrm{E}\left[\left(n_{1\ldots1}^{1\ldots k-1} - n_{0\ldots0}^{1\ldots k-1}\right)^2\right]}$$

Since $(n_{0\ldots0}^{1\ldots j}, \ldots, n_{1\ldots1}^{1\ldots j})$ has a multinomial distribution with parameters N (twice the number of ASPs) and $(p_{0\ldots0}^{1\ldots j}, \ldots, p_{1\ldots1}^{1\ldots j})$, and $\mathrm{E}\left[n_{1\ldots1}^{1\ldots j}\right] = \mathrm{E}\left[n_{0\ldots0}^{1\ldots j}\right]$ we have:

$$\begin{aligned}
\mathrm{E}\left[\left(n_{1\ldots1}^{1\ldots j} - n_{0\ldots0}^{1\ldots j}\right)^2\right] &= \mathrm{Var}\left(n_{1\ldots1}^{1\ldots j} - n_{0\ldots0}^{1\ldots j}\right) = \mathrm{Var}\left(n_{1\ldots1}^{1\ldots j}\right) + \mathrm{Var}\left(n_{0\ldots0}^{1\ldots j}\right) \\
&\quad -2\,\mathrm{Cov}\left(n_{1\ldots1}^{1\ldots j}, n_{0\ldots0}^{1\ldots j}\right) \\
&= Np_{1\ldots1}^{1\ldots j}(1 - p_{1\ldots1}^{1\ldots j}) + Np_{0\ldots0}^{1\ldots j}(1 - p_{0\ldots0}^{1\ldots j}) + 2Np_{1\ldots1}^{1\ldots j}p_{0\ldots0}^{1\ldots j} \\
&= N(p_{1\ldots1}^{1\ldots j} + p_{0\ldots0}^{1\ldots j}) - N(p_{1\ldots1}^{1\ldots j} - p_{0\ldots0}^{1\ldots j})^2 = N(p_{1\ldots1}^{1\ldots j} + p_{0\ldots0}^{1\ldots j}) = \\
&= \frac{N}{2^{j-1}}
\end{aligned}$$

Considering this it is easy to see that:

$$\frac{w_{k-1}}{w_k} = \frac{\frac{2^k-1}{2^k}}{\frac{2^{k-1}-1}{2^{k-1}}}$$

hence the resulting weight w_k.

It is revealing to rewrite the linkage measure as follows:

$$H_{1\ldots k} = \frac{w_k}{w_{k-1}} \cdot \frac{H_{2\ldots k} + H_{13\ldots k} + \cdots + H_{1\ldots k-1}}{k} + w_k \left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2 \quad (3.3)$$

Then for $k \geq 4$ we have $w_k \approx 1$ and $w_k \approx w_{k-1}$. Hence we can write:

$$H_{1\ldots k} \approx \frac{H_{2\ldots k} + H_{13\ldots k} + \cdots + H_{1\ldots k-1}}{k} + \left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2 \quad (3.4)$$

Essentially our measure is defined recursively as follows. We start with the natural NPL-like measure for one marker $H_1 = 2(n_1 - n_0)^2$. The measure for k markers ($H_{1\ldots k}$) is obtained as the average of the measures for all possible k combinations of $k-1$ markers: $H_{2\ldots k}$, $H_{13\ldots k}$, $\ldots$, $H_{1\ldots k-1}$ plus an additional term $\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2$ that measures the interaction of all k markers together.

Notice that when none of the k markers is linked to disease we have: $\mathrm{E}\left[(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k})^2\right] = \frac{N}{2^{k-1}}$. Thus the interaction term tends to be small in this case ($O(N)$). However when all k markers are linked to disease, this term will become large (due to $\mathrm{E}^2\left[n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right] = O(N^2)$).

Remark Our experiments show that under the assumption of no specific interaction model (e.g. epistasis or heterogeneity), the other possible pieces of information that we could use in the definition of the measure (e.g. n_{10} , n_{01}, etc.) may introduce noise (e.g. in the case of disease loci that interact epistatically). Certainly n_{10} and n_{01} contain information in a two-locus heterogeneity model, but the choice of a consistent statistic that would work for different scenarios forces us to disregard these terms and instead focus on n_{11} and n_{00}. Notice that under both the epistatic

and the heterogeneity interaction model for two disease loci $\mathrm{E}(n_{11} - n_{00}) > 0$, whereas when none of the loci is linked to disease $\mathrm{E}(n_{11} - n_{00}) = 0$

3.2.2 Screening Algorithm

The screening procedure consists of a marker selection-reduction process described below. Suppose we have a list of many markers (hundreds in a whole-genome study). We proceed as follows:

- Step 0 Repeat steps $1 - 4$ B times ($B \geq 3000$ is a fairly large number).
- Step 1 Start by choosing a set of $k \approx 10$ markers at random from the available list of markers.
- Step 2 At each step compute for each marker in the current set the resulting change in the linkage measure when that marker is removed. For marker i:

$$\Delta_i = H_{1\ldots i-1\ i+1\ldots k} - H_{1\ldots k}$$

  If $\Delta_i < 0$, then the linkage measure decreases when removing marker i and therefore marker i is important relative to the other markers present in the current marker set. If $\Delta_i > 0$, then the linkage measure increases when removing marker i and therefore marker i is not important relative to the other markers present.

- Step 3 Remove the marker i (if any) with the largest positive Δ_i from the current set.

- Step 4 Do Steps $2-3$ until either all the markers in the current set are important (all Δ_i are negative) or only one marker remains. The returned markers are recorded.

- Step 5 We compute for each marker a final return count denoting the total number of times it was returned in Step 4. Based on these counts we separate the markers into two classes: the important/linked to disease markers and the unimportant/unlinked ones. The details of this statistical procedure are given in Section 3.2.4.

3.2.3 Why It Works

The behavior of the screening algorithm in Section 3.2.2 depends heavily on the properties of the statistic $H_{1\ldots k}$. We formulate these properties in the lemma below. The main idea is that in expectation only markers that are linked to disease are returned in Step 4 and markers that are not linked tend to be removed in Step 3. Let $S = \{1, \ldots, k\}$ be the current set. For the lemma below we make the simplifying assumption that the **k** markers are not linked among themselves.

Lemma 3.2.1 *The following properties are true:*

1. *If none of the markers is linked to disease, then for any marker i in S we have*

$$\mathrm{E}\left[H_{1\ldots i-1\ i+1\ldots k}\right] = \mathrm{E}\left[H_{1\ldots k}\right]$$

2. *If S contains one marker linked to disease (without loss of generality, assume this is the first marker) and the rest are unlinked, then for any unlinked*

marker u in S we have:

$$\mathrm{E}\left[H_{2\ldots k}\right] < \mathrm{E}\left[H_{1\ldots k}\right] < \mathrm{E}\left[H_{1\ldots u-1\ u+1\ldots k}\right]$$

3. *If the set S has some interacting markers, linked to disease, of similar relative importance and some unlinked markers, then for any linked marker l and any unlinked one u we have:*

$$\mathrm{E}\left[H_{1\ldots l-1\ l+1\ldots k}\right] < \mathrm{E}\left[H_{1\ldots k}\right] < \mathrm{E}\left[H_{1\ldots u-1\ u+1\ldots k}\right]$$

4. *If the current set S contains only markers linked to disease that are of similar relative importance and also have non-negligible interaction, then for any marker l in S*

$$\mathrm{E}\left[H_{1\ldots l-1\ l+1\ldots k}\right] < \mathrm{E}\left[H_{1\ldots k}\right]$$

Proof:

1. The first part is easy; we have chosen the weights in Section 2.1 such that when no marker is linked to disease, we have:

$$\mathrm{E}\left[H_{1\ldots i-1\ i+1\ldots k}\right] = \mathrm{E}\left[H_{1\ldots k}\right]$$

2. From (3.3) in Section 3.2.1 we can write:

$$\begin{aligned} & \mathrm{E}\left[H_{1\ldots k}\right] = \frac{w_k}{w_{k-1}} \cdot \frac{\mathrm{E}\left[H_{2\ldots k}\right] + \mathrm{E}\left[H_{13\ldots k}\right] + \cdots + \mathrm{E}\left[H_{1\ldots k-1}\right]}{k} + \\ + \quad & w_k \,\mathrm{E}\left[\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2\right] \\ = \quad & \frac{\left(\frac{w_k}{w_{k-1}} \,\mathrm{E}\left[H_{2\ldots k}\right] + k w_k \,\mathrm{E}\left[\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2\right]\right) + (k-1)\frac{w_k}{w_{k-1}} \,\mathrm{E}\left[H_{1\ldots u-1\ u+1\ldots k}\right]}{k} \end{aligned}$$

It suffices to show that

$$\mathrm{E}\left[H_{2\ldots k}\right] < \frac{w_k}{w_{k-1}} \cdot \mathrm{E}\left[H_{2\ldots k}\right] + k w_k \,\mathrm{E}\left[\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2\right] < \frac{w_k}{w_{k-1}} \,\mathrm{E}\left[H_{1\ldots u-1\ u+1\ldots k}\right] \quad (*)$$

(we also use $w_k < w_{k-1}$). We prove the first inequality, namely $\mathrm{E}\left[H_{2\ldots k}\right] < \frac{w_k}{w_{k-1}} \cdot \mathrm{E}\left[H_{2\ldots k}\right] + k w_k \,\mathrm{E}\left[\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2\right]$. Since markers $2, \ldots, k$ are not linked to disease and among themselves, one can easily show that $\mathrm{E}\left[H_{2\ldots k}\right] = 2N$ where N is twice the number of ASPs. Therefore we need to show that

$$2N < \frac{w_k}{w_{k-1}} \cdot 2N + k w_k \,\mathrm{E}\left[\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2\right] \Leftrightarrow \frac{N}{2^{k-1}} < k \,\mathrm{E}\left[\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2\right]$$

Now we have:

$$\begin{aligned} & \mathrm{E}\left[\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2\right] = \mathrm{E}^2\left[n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right] + \mathrm{Var}\left[n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right] \\ \approx \quad & \mathrm{E}^2\left[n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right] + N(p_{1\ldots 1}^{1\ldots k} + p_{0\ldots 0}^{1\ldots k}) \approx \mathrm{E}^2\left[n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right] + \frac{N}{2^{k-1}} \end{aligned} \tag{3.5}$$

where $p^{1\ldots k}_{1\ldots 1}$ is the probability of the IBD sharing vector $1\ldots 1$ at loci $1\ldots k$. Therefore we showed the first inequality.

For the second inequality in (*), $\frac{w_k}{w_{k-1}} \cdot \mathrm{E}\left[H_{2\ldots k}\right] + k w_k \,\mathrm{E}\left[\left(n^{1\ldots k}_{1\ldots 1} - n^{1\ldots k}_{0\ldots 0}\right)^2\right] < \frac{w_k}{w_{k-1}} \mathrm{E}\left[H_{1\ldots u-1\ u+1\ldots k}\right]$, we proceed as follows. By definition in Section 2.1 we have:

$$\begin{aligned} H_{2\ldots k} &= w_{k-1} \left[\frac{\sum_{i=2}^{k}\left(n^i_1 - n^i_0\right)^2}{\binom{k-1}{1}} + \frac{\sum_{\mathbf{i<j;i,j\neq 1}}\left(n^{ij}_{11} - n^{ij}_{00}\right)^2}{\binom{k-1}{2}} + \cdots + \right. \\ &\quad \left. + \left(n^{2\ldots k}_{1\ldots 1} - n^{2\ldots k}_{0\ldots 0}\right)^2\right] \\ H_{1\ldots u-1\ u+1\ldots k} &= w_{k-1} \left[\frac{\sum_{i=1,i\neq u}^{k}\left(n^i_1 - n^i_0\right)^2}{\binom{k-1}{1}} + \frac{\sum_{i<j;i,j\neq u}\left(n^{ij}_{11} - n^{ij}_{00}\right)^2}{\binom{k-1}{2}} + \cdots + \right. \\ &\quad \left. + \left(n^{1\ldots\hat{u}\ldots k}_{1\ldots 1} - n^{1\ldots\hat{u}\ldots k}_{0\ldots 0}\right)^2\right] \end{aligned} \tag{3.6}$$

It suffices to show that

$$\mathrm{E}\left[\left(n^1_1 - n^1_0\right)^2\right] > \mathrm{E}\left[\left(n^u_1 - n^u_0\right)^2\right] + k(k-1)\,\mathrm{E}\left[\left(n^{1\ldots k}_{1\ldots 1} - n^{1\ldots k}_{0\ldots 0}\right)^2\right] \text{ for } k \geq 2 \tag{3.7}$$

Using (3.2) it is easy to prove that $\mathrm{E}\left[\left(n^u_1 - n^u_0\right)^2\right] = N$ for any unlinked marker and $\mathrm{E}\left[\left(n^1_1 - n^1_0\right)^2\right] \approx \mathrm{E}^2\left[n^1_1 - n^1_0\right] + N$ for a linked marker. Also

$$\mathrm{E}\left[\left(n^{1\ldots k}_{1\ldots 1} - n^{1\ldots k}_{0\ldots 0}\right)^2\right] \approx \frac{1}{4^{k-1}}\,\mathrm{E}^2\left[n^1_1 - n^1_0\right] + \frac{N}{2^{k-1}}$$

where N is twice the number of ASPs. Hence what we need to prove is that:

$$\mathrm{E}^2\left[n_1^1 - n_0^1\right]\left(1 - \frac{k(k-1)}{4^{k-1}}\right) \;>\; \frac{k(k-1)N}{2^{k-1}}$$

If we let $p_1 = rp_0$ with $r > 1$ and since $p_1 + p_0 = 1$ we obtain:

$$N^2\left[p_1^1 - p_0^1\right]^2\left(1 - \frac{k(k-1)}{4^{k-1}}\right) > \frac{k(k-1)N}{2^{k-1}} \Leftrightarrow N\frac{(r-1)^2}{(r+1)^2} > \frac{1}{\frac{2^{k-1}}{k(k-1)} - \frac{1}{2^{k-1}}} \quad (**)$$

The latter inequality is true for N (twice the number of ASPs) large enough. For example when $r = \frac{p_1}{p_0} = 1.3$ and $k = 2$, a sample of 60 ASPs is sufficient. Hence we have shown that (3.7) is true.

Now we can complete the proof of the second inequality in (*). If marker 1 is linked to disease and marker u is not linked, then we have:

$$\begin{aligned}
\mathrm{E}\left[\left(n_1^1 - n_0^1\right)^2\right] &> \mathrm{E}\left[(n_1^u - n_0^u)^2\right] + k(k-1)\,\mathrm{E}\left[\left(n_{1\ldots1}^{1\ldots k} - n_{0\ldots0}^{1\ldots k}\right)^2\right] \\
&\quad (\text{ from (3.7)}) \\
\mathrm{E}\left[\left(n_{11}^{1j} - n_{00}^{1j}\right)^2\right] &> \mathrm{E}\left[\left(n_{11}^{uj} - n_{00}^{uj}\right)^2\right] \text{ for any } j \notin \{1, u\} \\
&\ldots \\
\mathrm{E}\left[\left(n_{11\ldots1}^{1j_1\ldots j_l} - n_{00\ldots0}^{1j_1\ldots j_l}\right)^2\right] &> \mathrm{E}\left[\left(n_{11\ldots1}^{uj_1\ldots j_l} - n_{00\ldots0}^{uj_1\ldots j_l}\right)^2\right] \text{ for any } j_1\ldots j_l \notin \{1, u\}
\end{aligned}$$

Using the definitions of $H_{2\ldots k}$ and $H_{1\ldots u-1\ u+1\ldots k}$ in (3.6) and together with the inequalities above we obtain the second inequality in (*). This completes our proof.

3. We assume $k \geq 4$ ($k = 2$ and $k = 3$ can be proved using case-by-case computations). We assume the first t markers are linked to disease and the rest $(k-t)$ (> 0) are unlinked.

   Let $\mathrm{E}\left[H_{1\ldots i-1\ i+1\ldots k}\right] = A$ for any $i \leq t$ and $\mathrm{E}\left[H_{1\ldots i-1\ i+1\ldots k}\right] = B$ for any $i > t$. Clearly, $\mathbf{A} < \mathbf{B}$. We now use the approximation in Section 2.1:

$$\begin{aligned}
\mathrm{E}\left[H_{1\ldots k}\right] &\approx \frac{\mathrm{E}\left[H_{2\ldots k}\right] + \mathrm{E}\left[H_{13\ldots k}\right] + \cdots + \mathrm{E}\left[H_{1\ldots k-1}\right]}{k} + \mathrm{E}\left[\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2\right] \\
&= \frac{tA + (k-t)B}{k} + \mathrm{E}\left[\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2\right]
\end{aligned}$$

   From this we have:

$$\mathrm{E}\left[H_{1\ldots k}\right] \approx A + \frac{k-t}{k}(B-A) + \mathrm{E}\left[\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2\right]$$

   and since $\mathbf{B} > \mathbf{A}$ we obtain:

$$\mathrm{E}\left[H_{1\ldots k}\right] > A$$

   Similarly:

$$\mathrm{E}\left[H_{1\ldots k}\right] \approx B + \frac{t}{k}(A-B) + \mathrm{E}\left[\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2\right]$$

We show that:

$$B - A \;>\; \frac{k}{t}\,\mathrm{E}\left[\left(n_{1\ldots1}^{1\ldots k} - n_{0\ldots0}^{1\ldots k}\right)^2\right] \tag{3.8}$$

and therefore:

$$\mathrm{E}\left[H_{1\ldots k}\right] < B$$

From (3.6) we can write:

$$\begin{aligned} B - A \;&>\; \frac{\mathrm{E}\left[\left(n_1^1 - n_0^1\right)^2\right] - \mathrm{E}\left[\left(n_1^u - n_0^u\right)^2\right]}{k-1} \qquad \text{where } u > t \\ &\approx\; \frac{\mathrm{E}^2\left[n_1^1 - n_0^1\right] + N - N}{k-1} = \frac{N^2\left(p_1^1 - p_0^1\right)^2}{k-1} \end{aligned} \tag{3.9}$$

With (3.8) and (3.9), we need to show:

$$\frac{N^2\left(p_1^1 - p_0^1\right)^2}{k-1} \;>\; \frac{k}{t}\,\mathrm{E}\left[\left(n_{1\ldots1}^{1\ldots k} - n_{0\ldots0}^{1\ldots k}\right)^2\right] \tag{3.10}$$

Now:

$$\begin{aligned} \mathrm{E}\left[\left(n_{1\ldots1}^{1\ldots k} - n_{0\ldots0}^{1\ldots k}\right)^2\right] \;&=\; \mathrm{E}^2\left[n_{1\ldots1}^{1\ldots k} - n_{0\ldots0}^{1\ldots k}\right] + \mathrm{Var}\left[n_{1\ldots1}^{1\ldots k} - n_{0\ldots0}^{1\ldots k}\right] \\ &\approx\; \mathrm{E}^2\left[n_{1\ldots1}^{1\ldots k} - n_{0\ldots0}^{1\ldots k}\right] + N\left(p_{1\ldots1}^{1\ldots k} + p_{0\ldots0}^{1\ldots k}\right) \\ &\approx\; N^2\left(p_{1\ldots1}^{1\ldots k} - p_{0\ldots0}^{1\ldots k}\right)^2 + N\left(p_{1\ldots1}^{1\ldots k} + p_{0\ldots0}^{1\ldots k}\right) \\ &\approx\; \frac{N^2}{4^{k-t}}\left(p_{1\ldots1}^{1\ldots t} - p_{0\ldots0}^{1\ldots t}\right)^2 + \frac{N}{2^{k-t}}\left(p_{1\ldots1}^{1\ldots t} + p_{0\ldots0}^{1\ldots t}\right) \end{aligned}$$

where we used the fact that the last $k-t$ markers are not linked to disease and among themselves (hence $p^{1\ldots k}_{1\ldots 1} = \frac{1}{2^{k-t}} p^{1\ldots t}_{1\ldots 1}$). From this and (3.10) follows that we need to prove that:

$$\frac{N^2 \left(p_1^1 - p_0^1\right)^2}{k-1} \quad > \quad \frac{k}{t} \cdot \frac{N^2}{4^{k-t}} \left(p^{1\ldots t}_{1\ldots 1} - p^{1\ldots t}_{0\ldots 0}\right)^2 + \frac{k}{t} \cdot \frac{N}{2^{k-t}} \left(p^{1\ldots t}_{1\ldots 1} + p^{1\ldots t}_{0\ldots 0}\right)$$

1. If t is small compared with k (i.e. $k-t$ is large) then since $\left(p^{1\ldots t}_{1\ldots 1} - p^{1\ldots t}_{0\ldots 0}\right)^2 < \left(p_1^1 - p_0^1\right)^2$ and $\left(p^{1\ldots t}_{1\ldots 1} + p^{1\ldots t}_{0\ldots 0}\right) < 1$ it is sufficient to show that:

$$N \left(p_1^1 - p_0^1\right)^2 > \frac{\frac{k}{t} \cdot \frac{1}{2^{k-t}}}{\frac{1}{k-1} - \frac{k}{t} \cdot \frac{1}{4^{k-t}}} = \frac{1}{\frac{t 2^{k-t}}{k(k-1)} - \frac{1}{2^{k-t}}}$$

   This inequality is similar to inequality (**) shown at point 2. of the lemma for the case $t = 1$. For N large enough and when t is small compared with k it is true.

2. If t is comparable to k and since $k \geq 4$ (from our assumption), then $p^{1\ldots t}_{1\ldots 1} \pm p^{1\ldots t}_{0\ldots 0}$ tend to be much smaller than $p_1^1 - p_0^1$ (say conservatively , $p^{1\ldots t}_{1\ldots 1} \pm p^{1\ldots t}_{0\ldots 0} < \frac{1}{2}\left(p_1^1 - p_0^1\right)$) and also

$$\frac{N}{k-1} > \frac{k}{4t} \cdot \frac{N}{4^{k-t}} + \frac{k}{2t\left(p_1^1 - p_0^1\right)} \cdot \frac{1}{2^{k-t}}$$

   For example, if $t = k-1$ and $p_1^1 - p_0^1 = 0.1$ then the inequality above is

$$\frac{N}{k-1} > \frac{N}{16} + \frac{1}{0.4}$$

which is true ($k \leq 10$). For t smaller than $k-1$ the inequality is even sharper.

This concludes our proof for t (the number of markers linked to disease) between 2 and $k-1$. Next we show the case $t = k$.

4. We prove the case $k \geq 4$. The cases $k = 2$ and $k = 3$ can be verified easily through direct case-by-case computations. We use the approximation in Section 2.1 and since the interacting disease loci have similar importance we obtain:

$$\begin{aligned} \mathrm{E}\left[H_{1\ldots k}\right] &\approx \frac{\mathrm{E}\left[H_{2\ldots k}\right] + \mathrm{E}\left[H_{13\ldots k}\right] + \cdots + \mathrm{E}\left[H_{1\ldots k-1}\right]}{k} + \mathrm{E}\left[\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2\right] \\ &= \mathrm{E}\left[H_{1\ldots l-1\ l+1\ldots k}\right] + \mathrm{E}\left[\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2\right] \end{aligned}$$

Since all k markers in the current set are assumed to be linked to disease and to interact together in a non-negligible fashion, $\mathrm{E}\left[\left(n_{1\ldots 1}^{1\ldots k} - n_{0\ldots 0}^{1\ldots k}\right)^2\right]$ is large enough to guarantee the inequality:

$$\mathrm{E}\left[H_{1\ldots l-1\ l+1\ldots k}\right] < \mathrm{E}\left[H_{1\ldots k}\right] \text{ for } k \geq 4$$

□

Remark We made the assumption that the k selected markers in Step 1 of the Screening Algorithm are unlinked among themselves. Given that in the majority of cases the k (≈ 10) markers chosen at random from a large number of markers are unlinked among themselves and also because of ease of computation, that

assumption is reasonable. However, even when some of the markers in the current set are linked, the effect tends to be very small and the screening algorithm behaves as desired.

To better illustrate these properties, we simulated a small dataset with 7 markers. The first two of these are each closely linked ($\theta = 0.01$) to a different disease gene. The other five are unlinked to disease. The disease model is epistatic RR, i.e. two mutations at each of the two disease loci are necessary to have disease. As shown in Figure 3.1(a), the measure H_{12347} decreases significantly when removing either one of the linked markers (1 or 2) and increases significantly when removing either of the unlinked markers (3, 4 or 7). In Figure 3.1(b) we see that when none of the markers in the current set is linked to disease, the values of the measure are small and not as well separated as the ones in Figure 3.1(a). In fact a random (unlinked) marker is removed. In this example (Figure 3.1(b)), marker 3 is removed (i.e. $\Delta_3 = H_{4567} - H_{34567}$ is the largest positive Δ_i).

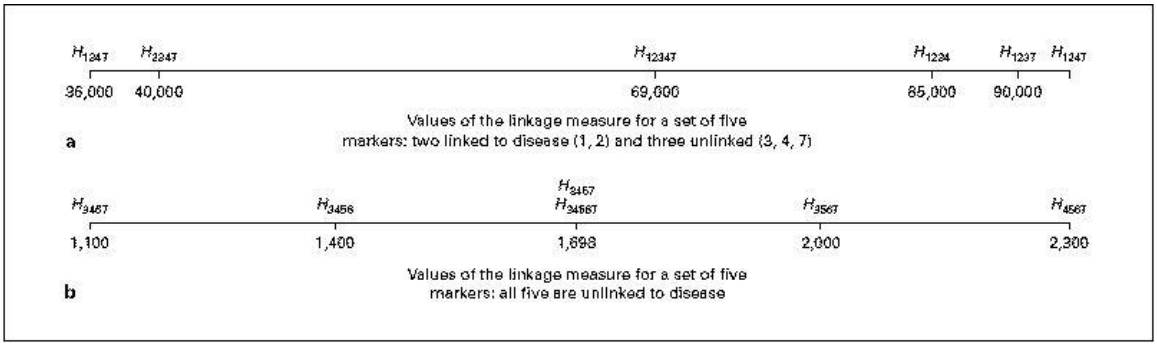


Figure 3.1: Linkage measure. Figure (a) and (b) illustrate the behavior of the measure for 5 markers when trying to remove each one of them in turn.

Note that we have:

$$H_{12347} \approx \frac{H_{1347} + H_{2347} + H_{1234} + H_{1237} + H_{1247}}{5} \quad \text{for (a)}$$
$$H_{34567} \approx \frac{H_{3467} + H_{3456} + H_{3457} + H_{3567} + H_{4567}}{5} \quad \text{for (b)}$$

Therefore according to (3.4) in Section 3.2.1 the interaction term $((n_{11111} - n_{00000})^2)$ is small in this case due to the presence of unlinked markers in the current set.

3.2.4 Important vs. Unimportant Markers

The goal of our method is to separate the important/linked to disease markers from the unimportant/unlinked markers. We present two different methods to achieve this goal. Both methods yield a good balance between false positive results and true positive results. In our experience the two methods behave similarly.

1. Normal-Mixture Method

   We first fit a two-component normal-mixture model to the histogram of return counts:

   $$p_1 N(\mu_1, \sigma_1^2) + p_2 N(\mu_2, \sigma_2^2)$$

   where $\mu_2 > \mu_1$ and $p_2 = 1 - p_1$; μ_2 and μ_1 are the means for the distribution of important and unimportant markers respectively. To control the false-positive rate (FPR), we select as threshold the $1 - \alpha$ percentile for the unimportant markers at a certain level α. The markers that have a return higher than this cutoff are claimed to be important (linked to disease genes).

2. Efron's Method

Another method to achieve this separation is based on an idea of Efron (2004 [25]). He proposes a method to divide the data values into two classes, interesting and uninteresting, when a large number of tests need to be evaluated as is the case in whole-genome scans. This is in contrast to the classical significant versus non-significant categorization used when the number of tests is small. The method first fits a natural spline to the histogram of return counts by Poisson regression. We call this curve: f (mixture density). Also an empirical null distribution is estimated, denoted by f_0 (empirical null density). Then for each marker M the *local false discovery rate* is defined as:

$$\text{locfdr}(M) = \frac{p_0 f_0(M)}{f(M)}$$

Controlling the false discovery rate suggests that the markers with locfdr$< \alpha$ be declared interesting (for a certain level α).

3.2.5 Choice of B and k

As explained in Section 3.2.2, our screening algorithm repeats B times the process of random selection of k markers and then evaluation of each of the markers in that set. We want to choose B and k large enough such that we get as clear a separation between the markers linked to disease and the unlinked ones as possible. We present a heuristic derivation of a formula for B below. The formula predicts conservatively that for 200 markers B should be about 8000 and for 500 markers $B \approx 20000$. The size of k influences the number of times certain markers are chosen

together in the random subset. It shouldn't be too small, since we want a good probability to select markers together. On the other hand, due to the sparseness of the data in large dimensions and also due to computational issues, k should not be too large. In our experience $k = 10$ works well.

A heuristic approach to estimating B was given in Lo and Zheng (2002 [26]). The derivation of B for the proposed method is very similar.

Suppose M is a marker linked to disease in the original set of polymorphisms. Let p_1 be the probability that a marker linked to disease is selected and returned in any single repetition (out of B) of the selection-reduction process; p_0 is the same probability but for markers not linked to disease. Assume $p_1 = rp_0$ with $r > 1$. Let X be the observed return count for marker M. Then $X \sim N(Bp_1, \sqrt{Bp_1(1-p_1)})$. In order to clearly separate the markers linked to disease from the ones not linked, we require:

$P(X \geq Bp_0 + 3.1\sqrt{Bp_0(1-p_0)}) \geq 99\%$. After some algebra, this can be written equivalently as:

$$B > \left(\frac{3.1 + 2.33\sqrt{r}}{r-1}\right)^2 \frac{1-p_0}{p_0}$$

We can estimate p_0, the probability for a marker not linked to disease to be selected and returned in a single repetition, as follows:

$$\begin{aligned} p_0 &= P(\text{returned}|\text{selected and unlinked}) \cdot P(\text{selected}|\text{unlinked}) \\ &\approx \frac{1}{k} \cdot \frac{\binom{n}{k-1}}{\binom{n}{k}} = \frac{1}{n-k+1} \approx \frac{1}{n} \end{aligned}$$

where n is the total number of markers and k is a small number of markers (say

10) selected to be evaluated; we assume conservatively that the probability that a selected marker, not linked to disease, is returned is $\frac{1}{k}$.

Therefore we obtain

$$B > \left(\frac{3.1 + 2.33\sqrt{r}}{r-1}\right)^2 (n-1)$$

r can be written as:

$$r = \frac{p_1}{p_0} = \frac{P(\text{returned}|\text{selected and linked})}{P(\text{returned}|\text{selected and unlinked})} \approx k(1-\epsilon)$$

where $1-\epsilon$ is an estimate for the probability that a linked marker, once selected, is returned. If conservatively we take $r = 2$ we obtain $B \approx 41(n-1)$.

3.3 Results

We evaluated our method on both simulated data and real data.

3.3.1 Simulated Data

We applied our method to two complex disease models.

First Simulated Disease Model

In the first disease model there are 9 unlinked disease loci. The disease is present when at least 5 of the 9 disease genes are mutated. The sample contains 200 ASPs genotyped at 50 markers, with 20% of the data sporadic (diseased because of nongenetic causes). Nine markers out of the total of 50 are linked to disease genes

($\theta = 0.05$), one marker for each disease gene. The rest are independent markers, not linked to disease and among themselves. The disease gene frequencies are all set to 0.05 and the marker frequencies are all 0.5. We assume we have complete data: the inference of the IBD sharing is without ambiguity. For each marker we compute two statistics:

- the single-locus statistic (the ASP mean test):

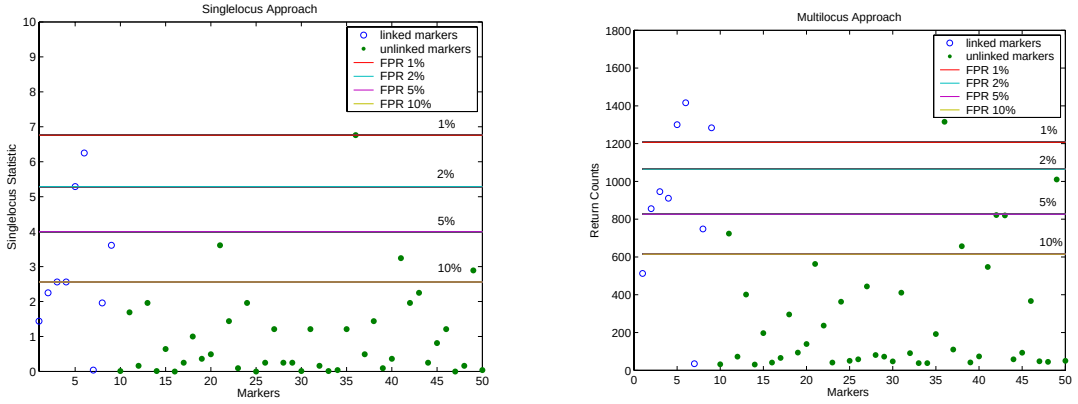
$$\frac{(n_1 - n_0)^2}{n_1 + n_0} \sim \chi_1^2$$

  where n_1 (n_0) is the number of 1 (0) IBD sharing at that particular marker.

- the return count computed by the proposed method ($B = 3000$ and $k = 10$ in our screening procedure).

For each of the two methods we report the number of loci above certain significance thresholds: $\{1\%, 2\%, \ldots, 10\%\}$ false positive rates. Since in the simulated data we know exactly which markers are linked to disease and which are unlinked, we can approximate the threshold corresponding to a specific false positive rate empirically by simulation.

Figure 3.2 shows an example of a simulated dataset according to the complex model outlined above. The horizontal lines represent the thresholds for the 1%, 2%, 5% and 10% FPR. It illustrates the advantage of our method; because the markers linked to disease are returned together in Step 4 of the screening algorithm in Section 3.2.2, they will separate better from the unlinked markers. Therefore the proposed method can be very powerful in increasing the importance of disease loci



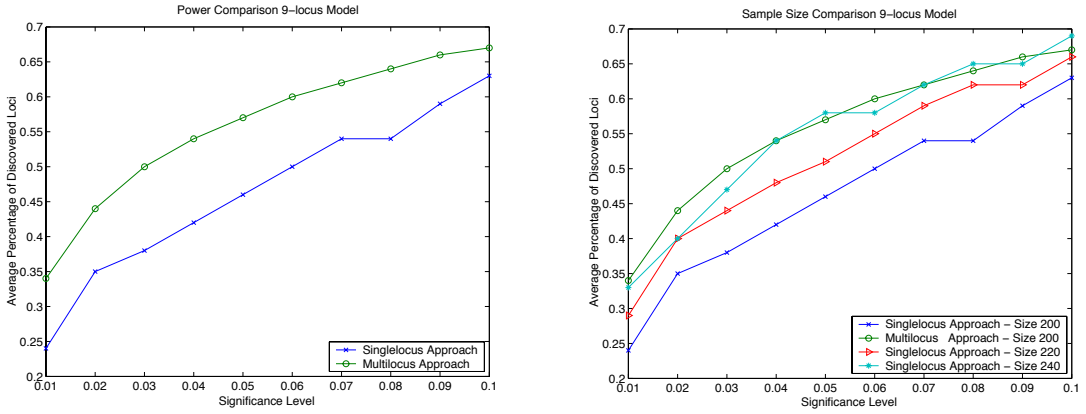
(a) (b)

Figure 3.2: Comparison between a simple single-locus method (a) and our new method (b) on a complex disease model with 9 disease loci. The figure illustrates how the multilocus approach can increase the significance of moderate effect loci.

of moderate effect by making use of interactions among disease loci.

To investigate the power, we generated 600 independent replicates. Figure 3.3(a) depicts the average percentage of disease loci selected by each of the two methods while keeping the false positive rate at the $\{1\%, 2\%, \ldots, 10\%\}$ level. Our method is more powerful than the single-locus method at all levels. At the 1% significance level, our method discovers on average 3.1 of the 9 disease loci, while the single-locus method finds only 2.2 loci. Similarly at the 3% level we detect on average 4.5 loci, while the single-locus method finds 3.4 loci.

Finally, we compared the increase in sample size necessary for the single-locus method to achieve similar power to that of the multilocus method. The results are depicted in Figure 3.3(b). For this particular model an increase in sample size of over 20% is necessary.



(a) (b)

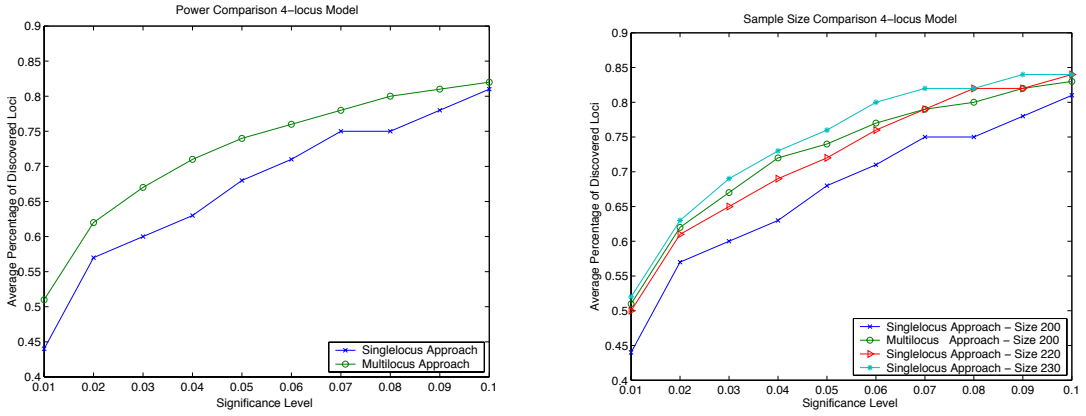
Figure 3.3: Average percentage of disease loci discovered with the single-locus method and the new multilocus method while controlling the FPR (a) and Sample Size Comparison (b) for the 9−locus disease model

Second Simulated Disease Model

We also simulated a similar disease model with 4 disease loci. Now the disease is present when at least 2 of the 4 disease genes are mutated.

Figure 3.4(a) depicts the average percentage of disease loci selected by each of the two methods while keeping the false positive rate at the $\{1\%, 2\%, \ldots, 10\%\}$ level. As we can see, our method is more powerful than the single-locus method at all significance levels. In Figure 3.4(b) we illustrate the increase in sample size necessary for the single-locus method to attain similar power to that of the multilocus method. For this simpler disease model, a $10\% - 15\%$ increase is necessary.

We then repeated the same simulations, but this time we introduced small linkage disequilibrium (LD) levels between some of the disease genes and the nearby linked markers. Namely, $\delta_1 = \delta_2 = 0.5$ and $\delta_3 = \delta_4 = 0$ where δ is the normalized LD measure. We compared the single locus approach to a modified version of



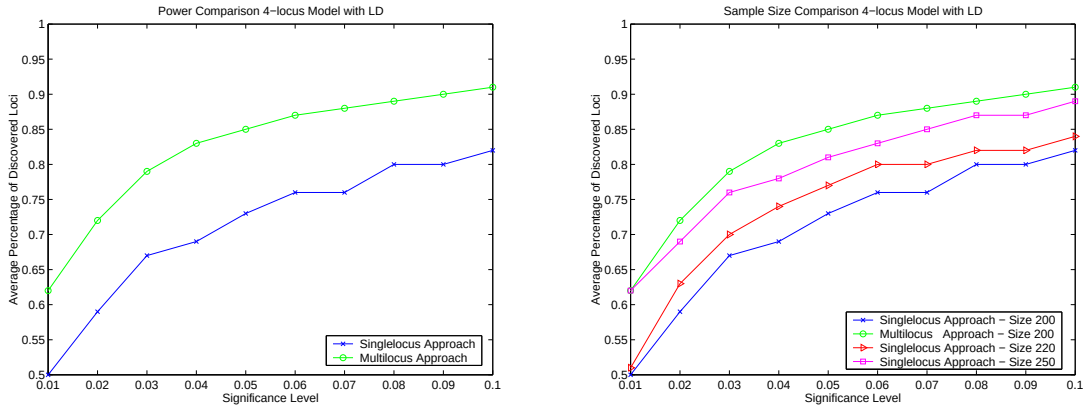
(a) (b)

Figure 3.4: Average percentage of disease loci discovered with the single-locus method and the new multilocus method while controlling the FPR (a) and Sample Size Comparison (b) for the 4−locus disease model

our multilocus linkage method (see Supplemental Material) that can also take advantage of mild linkage disequilibrium between disease loci and nearby markers. In this case the results (Figure 3.5) are even better compared to the ones obtained in Figure 3.4 where no linkage disequilibrium was present. The improvement at the 1% FPR is 23% and an increase in sample size of over 25% is necessary for the single locus linkage method to achieve similar performance as the modified multilocus linkage method.

3.3.2 Real Data (Inflammatory Bowel Disease)

We also analyzed a real dataset for Inflammatory Bowel Disease (InfBD) using our method. InfBD consists of two disorders: Crohn's Disease (CD) and Ulcerative colitis (UC). They are both inflammatory disorders of the gastrointestinal tract with a strong genetic contribution as revealed by epidemiological studies. Genome-wide searches for InfBD susceptibility loci have identified several regions of interest,



(a) (b)

Figure 3.5: Average percentage of disease loci discovered with the single-locus method and the new multilocus method while controlling the FPR (a) and Sample Size Comparison (b) for the 4−locus disease model with LD

showing that InfBD is a complex genetic disease caused by the action of several genes.

The present dataset is a genome screen of 106 ASPs (including parents) from Canada, affected with CD genotyped at 457 microsatellite markers; the average marker spacing is $\sim$ 10cM. These data have been previously analyzed in Rioux et al. (2000 [27]) and in Lo and Zheng (2004 [28]).

In order to apply our new approach to these data, we first inferred the IBD (identity-by-descent) sharing probabilities for each ASP using the program GENEHUNTER 2.0 (Daly et al. 1998 [29]). Since our method requires complete IBD sharing information, we probabilistically impute the IBD sharing values. More exactly, for each sib pair under study we generate the IBD value at each position in the sharing vector according to the corresponding sharing probabilities (as calculated by GENEHUNTER 2.0); for example if the sharing probabilities at a certain position are $(0.2, 0.5, 0.3)$ for sharing 0, 1 and 2 alleles respectively, then

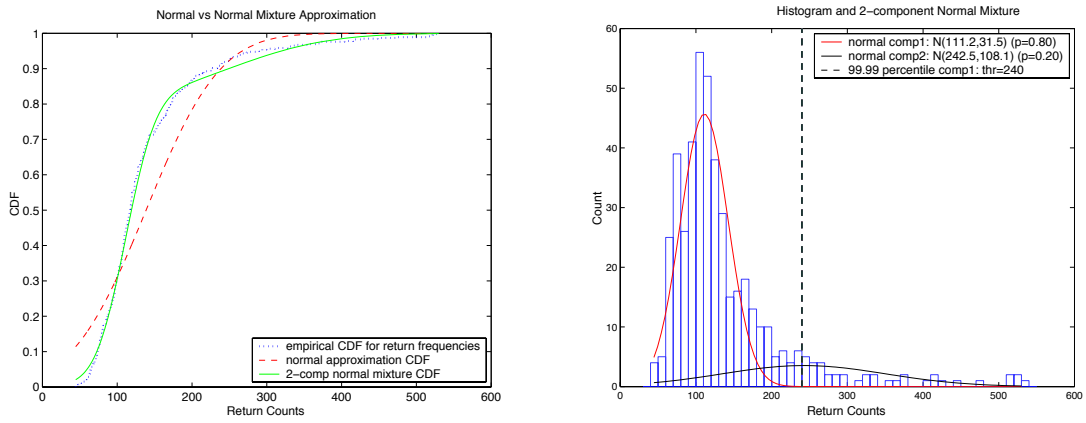
we generate the IBD value 0, 1 or 2 according to this distribution. In order to minimize the bias due to these probabilistic imputations, we do it 100 times, each time generating a new dataset.

We applied our algorithm on each of the 100 generated datasets. We use $B = 20000$ and $k = 10$ in our screening procedure. The return counts (averaged over the 100 datasets) for all markers together with the fitted two-component normal mixture are depicted in Figure 3.6(b). By controlling the false positive rate at a stringent level we obtain that markers with return count above 240 should be reported as important. In figure 3.6(a) we depict the cumulative distribution function (CDF) for a single normal fitted to the data versus the CDF for a mixture of two normals.

We also applied Efron's approach. In Figure 3.7 we give the results. On the left hand side, the histogram of the return counts together with the fitted empirical null density $f0(z)$ and the mixture density $f(z)$ are depicted. On the right hand side, the localfdr (local false discovery rate) plot is added to the histogram (scaled up by a factor of 50). A return count of 242 corresponds to a local fdr of 1%.

In Figure 3.8 we show the return counts plotted versus marker locations in the genome. The mean return count is 140 and is marked by a horizontal solid line. The threshold for declaring a marker important is 240 and is marked by a broken line.

The results we obtain are extremely significant. We validated 6 (IBD1, IBD3, IBD5, IBD6, IBD7, IBD8) of the 8 known InfBD loci. Additionally we found



(a) CDFs (b) Normal Mixture

Figure 3.6: Normal mixture approximation and histogram of the return counts

several other interesting regions.

1. The region 1q21 contains a cluster of genes influencing epidermal differentiation. This region is linked to other inflammatory diseases, e.g. psoriasis; psoriasis can occur in association with Inflammatory Bowel Disease (Crohn's disease), suggesting that they may share common genetic risk factors.

2. The locus on chromosome 2p11 (D2S1790) is located $\sim$ 10 cM from the gene IL1R1 (interleukin 1 receptor, type 1). There is evidence for the activation of the mucosal immune system and the production of inflammatory cytokines, i.e. interleukin (IL)-1ra and IL-1beta, in the Inflammatory Bowel Disease (Heresbach et al. 1997 [30]).

3. The region 2q32 harbors the STAT1 and STAT4 genes (signal transducers and activators of transcription), which are candidate genes for Inflammatory Bowel Disease (Barmada et al. 2004 [31]).

4. The locus on chromosome 3p: suggestive linkage in this region was found in

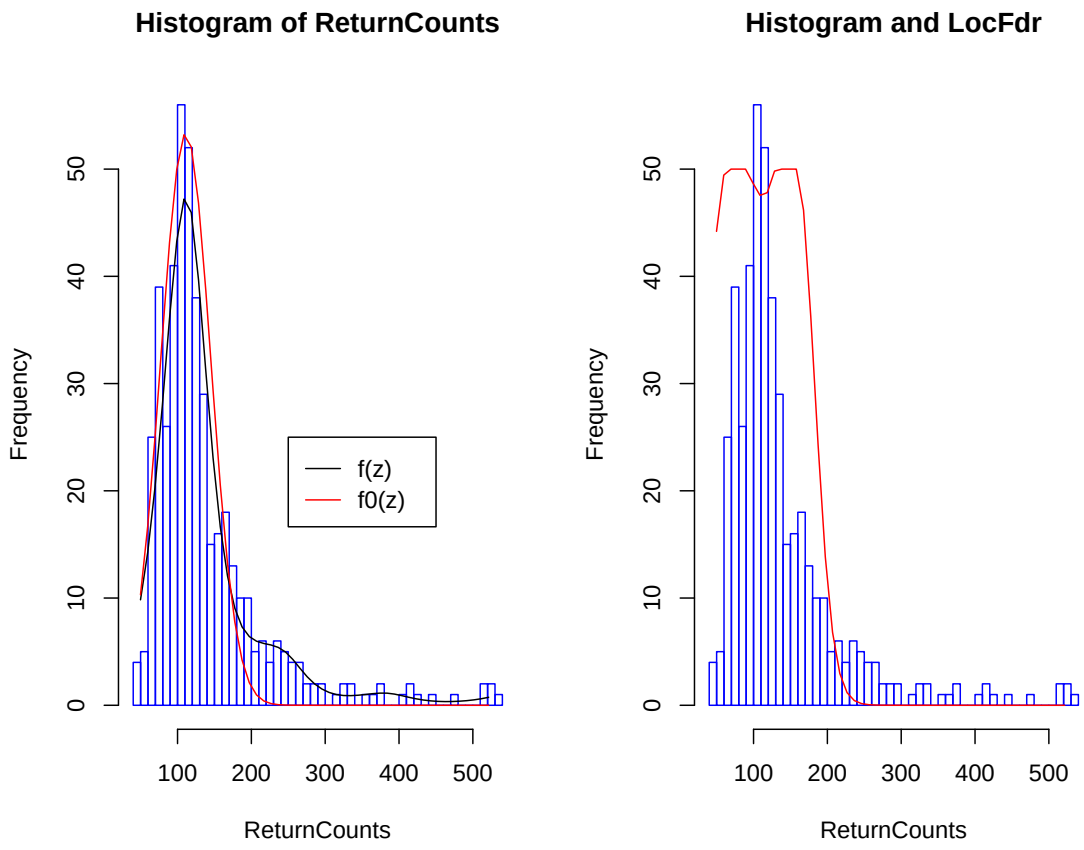


Figure 3.7: Efron's approach to separating the linked markers from the unlinked markers

Rioux et al. (2000 [27]).

5. The locus on chromosome 7p13: gene IGFBP3 (insulin-like growth factor binding protein 3) maps to this region. Katsanos et al. (2001 [32]) found that the serum IGFBP3 levels are reduced in patients with Inflammatory Bowel Disease.

6. The locus on chromosome 21q22.2 (D21S1809) is close to the TFF1 and TFF2 (trefoil factor 1 and 2) genes. These genes, located on 21q22.3, are expressed in the gastrointestinal mucosa. Increased levels of TFF1 and TFF2 have been

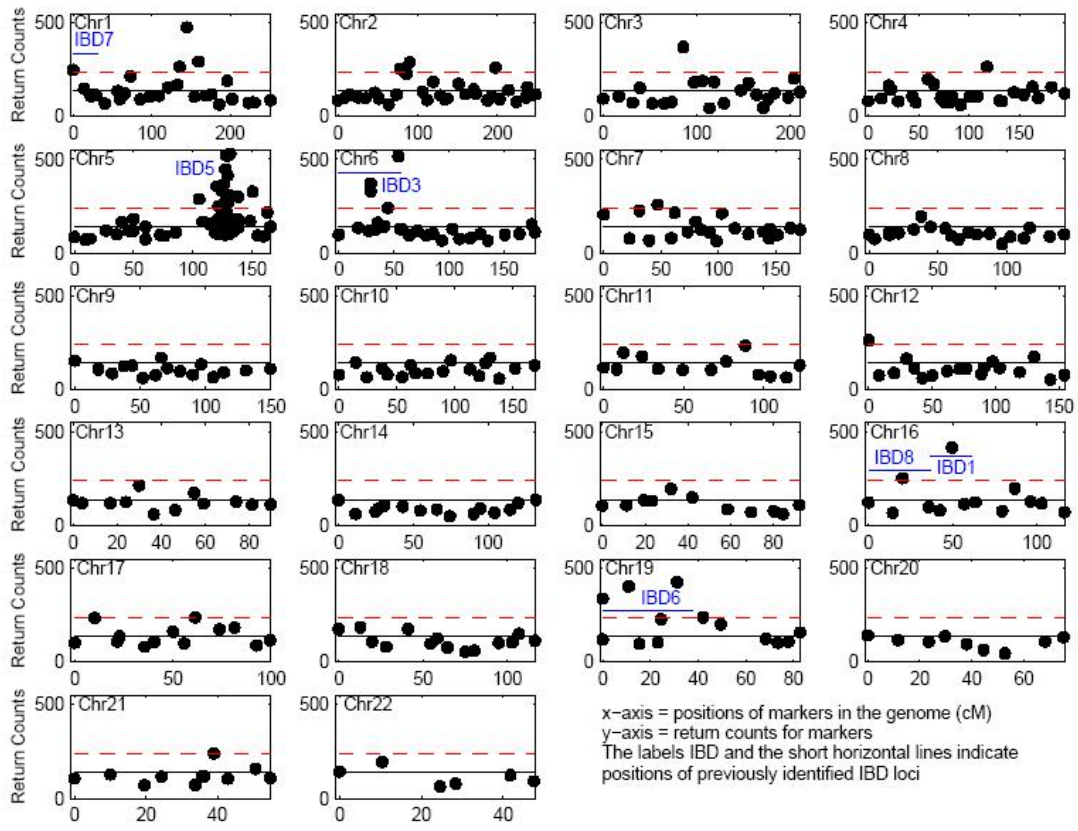


Figure 3.8: Results of the multilocus linkage method on the InfBD data

found in serum from Inflammatory Bowel Disease patients (Vestergaard et al. 2004 [33]).

Table 3.1 lists the markers we claim important together with their chromosomal position (InfBD dataset).

3.4 Discussion

We presented a new model-free multilocus linkage method for affected sib pairs. Our approach selects from a large number of polymorphisms a small number that appear to be linked to disease. No assumption is made on the disease model, in-

chr	selected marker	region
1	D1S1612	IBD7
	D1S534	1q21
	D1S1595	1q21
	D1S1677	1q23
2	D2S1394	2p11
	D2S1790	2p11
	D1S1649	2q32
3	D3S1285	3p
4	D4S2394	4q23-4q28
5	D5S500	IBD5
5	other 18 from the region	IBD5
6	DRB1	IBD3
6	DQB1	IBD3
6	D6S1017	IBD3
7	GATA31A10	7p13
12	D12S372	12p
16	D16S2619	IBD8
16	D16S2753	IBD1
19	D19S591	IBD6
19	GATA21G05	IBD6
19	D19S714	IBD6
21	D21S1809	21q22.3

Table 3.1: Selected important markers

cluding number of disease loci or their positions in the genome. It uses both the marginal linkage information as well as information coming from the possible interaction among several disease loci to boost the significance of modest single-locus effects. A further advantage of our method is that it can be naturally extended to take into consideration small linkage disequilibrium levels between disease loci and nearby markers, thereby gaining even greater increases in power over single locus linkage methods.

We evaluated our method on both simulated data and real data. The extensive simulations that we did show consistently that the proposed approach is more

powerful than the conventional single-locus linkage methods at all significance levels (up to 40% increase in power). The improvement in power increases when the number of interacting disease loci increases. In the absence of interactions our method performs similarly to the single-locus methods. Also the results on the real data are highly significant. We validated 6 of the 8 known InfBD loci and also found a few interesting loci, some of which have been already implicated in Inflammatory Bowel Disease pathogenesis.

Our method is also very general; the disease loci can be anywhere in the genome (possibly on different chromosomes) and they can interact in complex, unknown ways. We did make a simplifying assumption, namely we assumed that the selected markers in the current set are unlinked among themselves. It is clear however that the effect of linkage between two unimportant markers is superseded by the presence in the current set of a marker linked to disease. This point is best illustrated on real data, where we see that markers close together do not tend to have return counts higher than expected (e.g. chromosome 4 in Figure 3.7).

Given the complex nature of the common diseases and the many challenges in genomewide scans, we believe that our approach is very relevant; by using both the marginal and the interaction information, our method performs better than the traditional single-locus methods. Also due to its generality, the proposed method is applicable to a large number of situations.

3.5 Supplemental Material

3.5.1 Extension of the Multilocus Linkage Method

We give a natural extension of the multilocus linkage method so that mild linkage disequilibrium (LD) levels between disease loci and marker loci can be used in addition to linkage to obtain even greater increases in power over single-locus linkage methods. The extension is based on combining the multilocus linkage method with a similar association method (the BHTA algorithm, Lo and Zheng 2002 [26]). They are both based on the screening procedure in Section 2.2. In what follows we denote by ΔL_i (called Δ_i in the main text) and ΔLD_i (defined in Lo and Zheng 2002 [26]) the change in linkage and association information respectively when removing marker i from the current set.

- Step 0 Repeat steps $1-4$ B times.
- Step 1 Start by choosing a set of $k \approx 10$ markers at random from the available list of markers.
- Step 2 At each step compute for each marker in the current set the resulting change in both the linkage and association measure respectively when that marker is removed. For marker i:

$$\begin{aligned} \Delta L_i &= L_{1\ldots i-1\ i+1\ldots k} - L_{1\ldots k} \\ \Delta LD_i &= LD_{1\ldots i-1\ i+1\ldots k} - LD_{1\ldots k} \end{aligned}$$

- Step 3 Remove the marker i (if any) such that both $\Delta L_i > 0$ and $\Delta LD_i > 0$ (i.e. both the linkage measure and the association measure deem marker i unimportant) and that has the largest $\Delta L_i + \Delta LD_i$.

- Step 4 Do Steps $2 - 3$ until either all the markers in the current set are important (for each remaining marker i not both ΔL_i and ΔLD_i are positive) or only one marker remains. We return marker i R_i times, depending on the linkage and the association evidence as follows:

$$R_i = 1_{\Delta LD_i<0} + 1_{\Delta L_i<0}$$

  where $1_{\Delta LD_i<0}$ is the indicator random variable for the event $\Delta LD_i < 0$; $1_{\Delta L_i<0}$ is defined similarly.

- Step 5 We compute for each marker a final return count denoting the total number of times it was returned in Step 4. Based on these counts we separate the markers into two classes: the unimportant (unlinked) markers and the important (linked AND/OR associated) ones.

This simple procedure guarantees that when evaluating a marker we consider both the marginal information, as well as the interaction information contained in a dataset. Also it takes into account two pieces of information, usually treated separately: linkage information and linkage disequilibrium information.

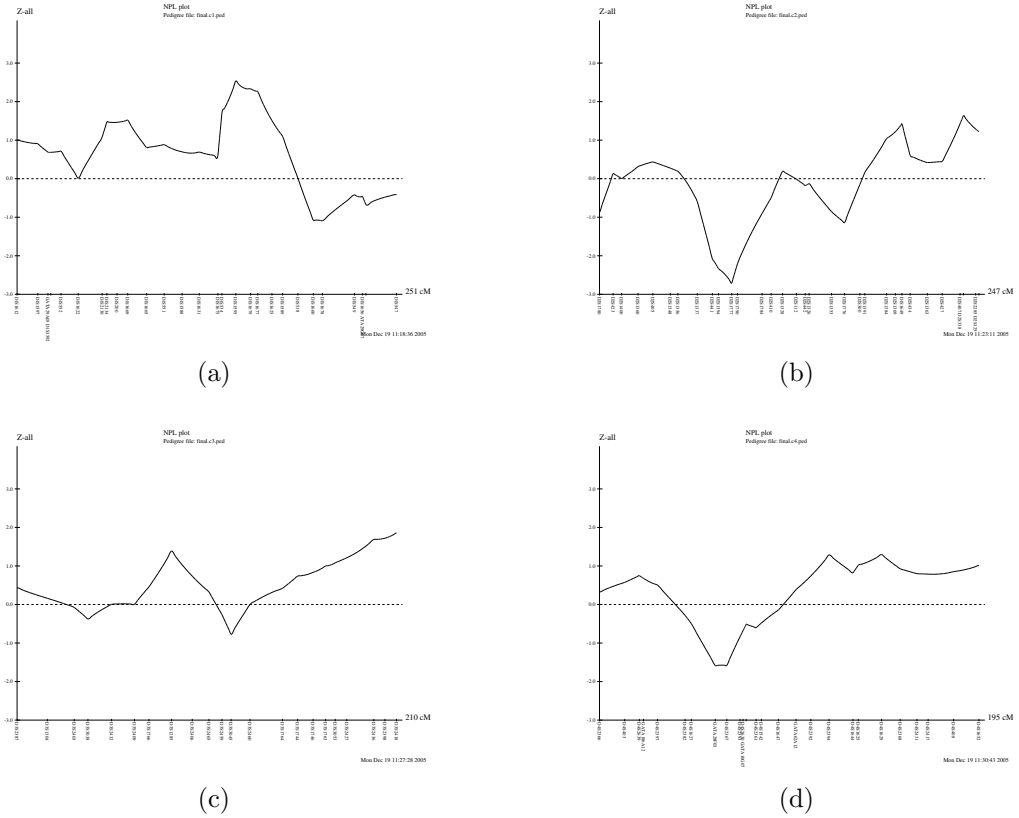


Figure 3.9: NPL results for chromosomes 1-4

3.5.2 Inflammatory Bowel Disease Data - NPL Results

We also applied the conventional NPL statistic (GENEHUNTER 2.0, Daly et al. 1998 [29]) to the same dataset and the results are illustrated below. Noteworthy is the fact that IBD1 on chromosome 16 (CARD15 gene) could not be detected using the conventional methods, whereas our proposed method did detect it.

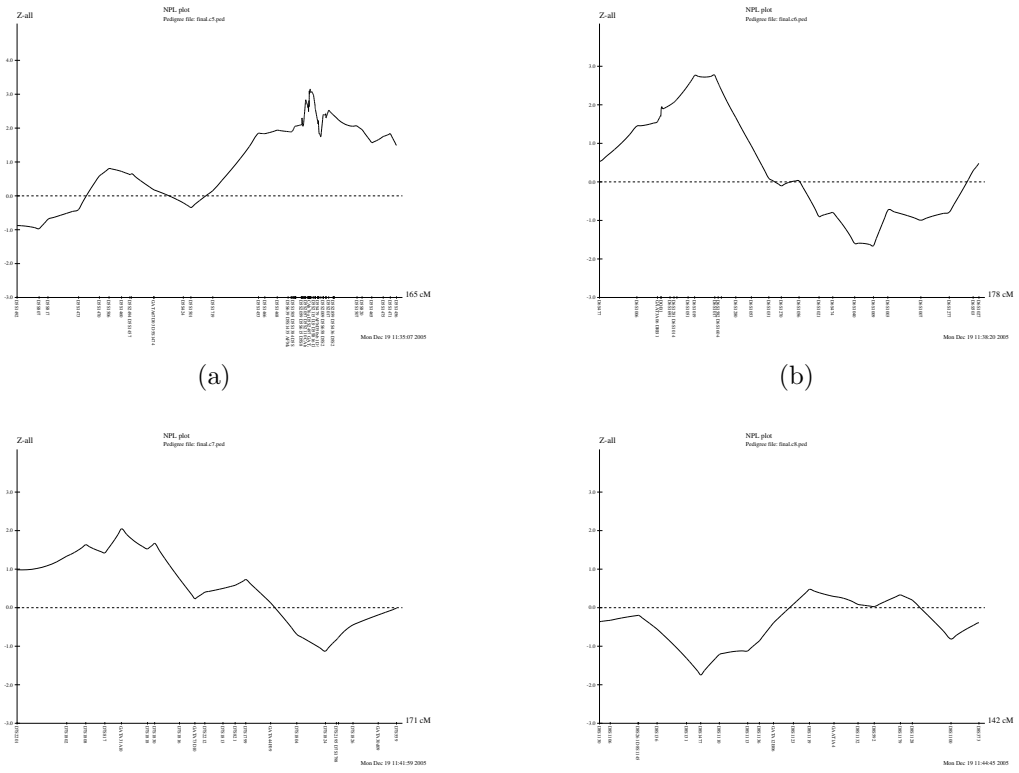


Figure 3.10: NPL results for chromosomes 5-8

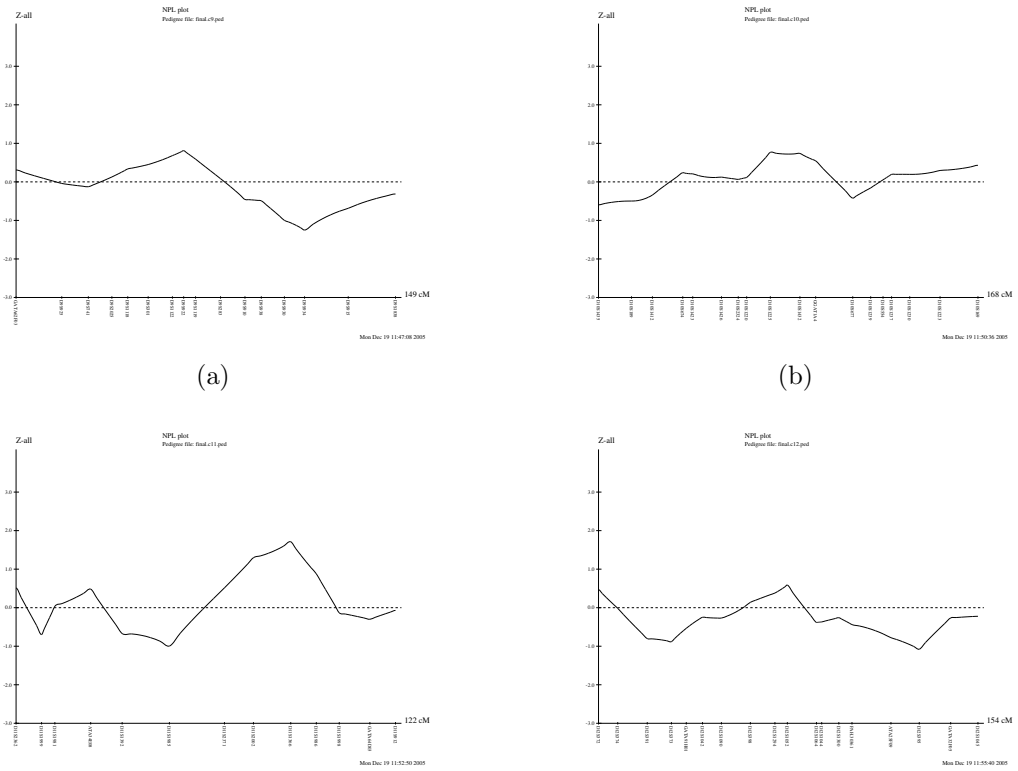
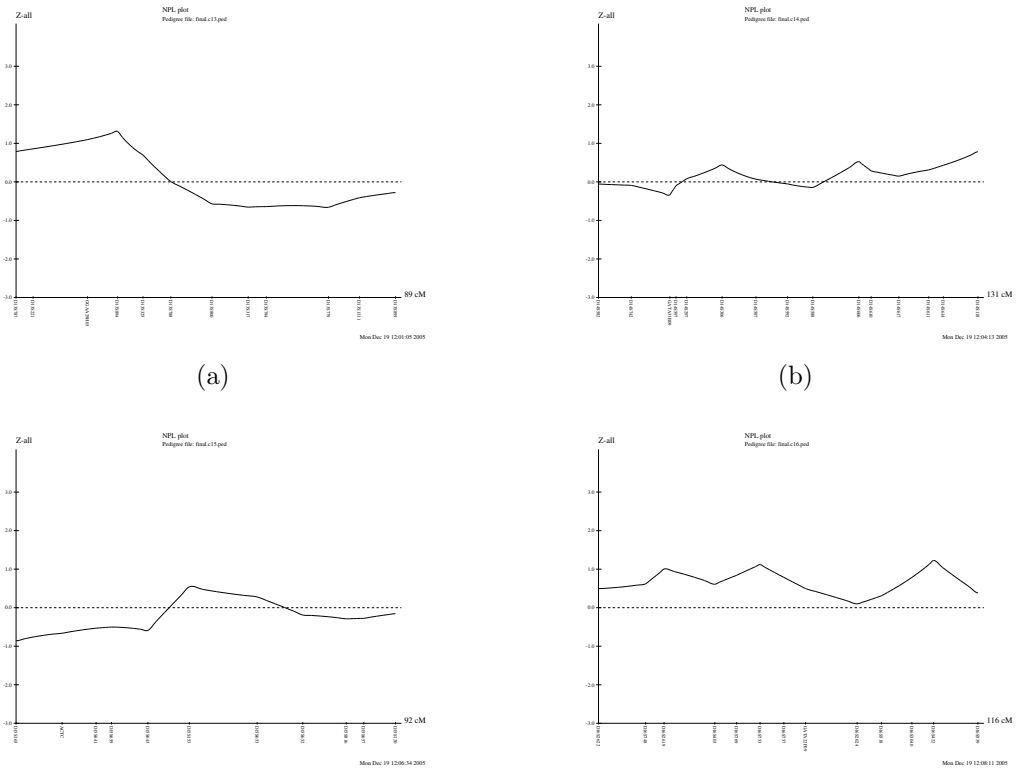


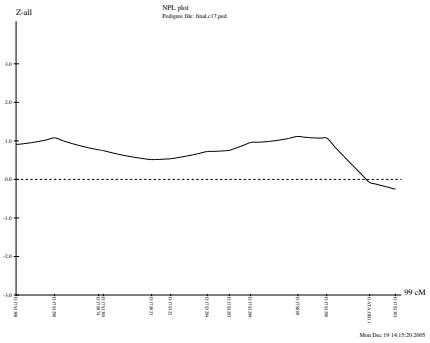
Figure 3.11: NPL results for chromosomes 9-12



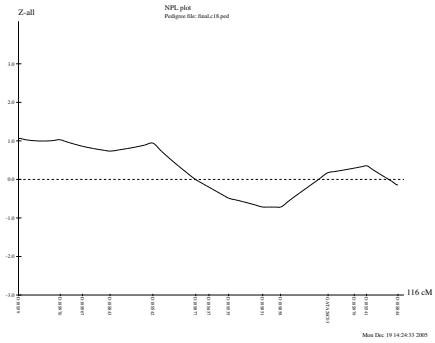
(a) (b)

(c) (d)

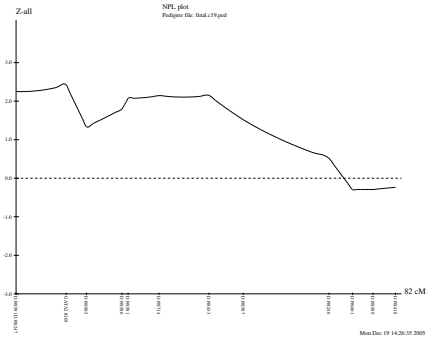
Figure 3.12: NPL results for chromosomes 13-16



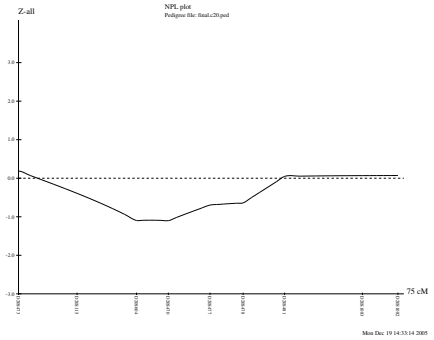
(a)



(b)

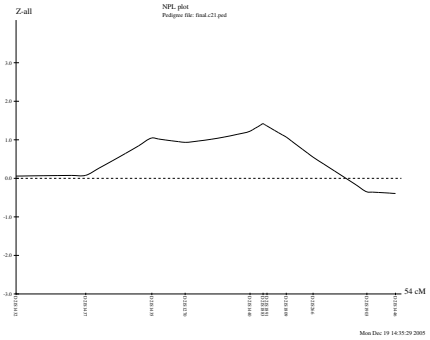


(c)

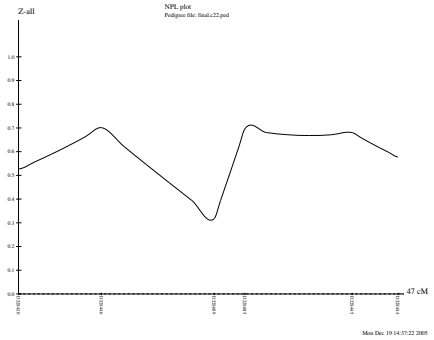


(d)

Figure 3.13: NPL results for chromosomes 17-20



(a)



(b)

Figure 3.14: NPL results for chromosomes 21-22

Chapter 4

A Practical Haplotype Inference Algorithm

SUMMARY In this chapter, we consider the problem of efficient inference algorithms to determine the haplotypes and their distribution from a dataset of unrelated genotypes.

With the currently available catalogue of single-nucleotide polymorphisms (SNPs) and given their abundance throughout the genome (one in about 500 bps) and low mutation rates, scientists hope to significantly improve their ability to discover genetic variants associated with a particular complex trait. We present a solution to a key intermediate step by devising a *practical* algorithm that has the ability to infer the haplotype variants for a particular individual from its own genotype SNP data in relation to population data. The algorithm we present is simple to describe and implement; it makes no assumption such as perfect phylogeny or the availability of parental genomes (as in trio-studies); it exploits locality in linkages and low diver-

sity in haplotype blocks to achieve a linear time complexity in the number of markers; it combines many of the advantageous properties and concepts of other existing statistical algorithms for this problem; and finally, it outperforms competing algorithms in computational complexity and accuracy, as demonstrated by the studies performed on real data and synthetic data. Furthermore, the basic ideas generalize to other related problems: HCNP (Haplotype-Copy-Number-Polymorphism) problem (work in progress) and IHRM (Individual-Haplotype-Restriction-Map) problem (Anantharaman et al. 2005 [34]).

4.1 Introduction

Each diploid individual has two copies for each chromosome: one inherited from the mother and the other from the father. The material on each copy is called a haplotype. Since haplotypes are the material that is transmitted from a parent to offspring and so are diseases, we are interested in finding the haplotypes and establish correlations between disease and specific haplotypes.

Current high-throughput genotyping methods only determine which two alleles are present at a locus, but lack information as to which of the two chromosomes (mother or father) each allele belongs to. For instance, if the two haplotypes for an individual are ACG and TCA, then the result of the genotyping experiment is: $\{A, T\}$ $\{C, C\}$ $\{G, A\}$. It is easy to see that this ambiguity causes problems if the individual is heterozygous for more than one locus, since for an individual with k ambiguous (heterozygous) loci, there are 2^{k-1} possible haplotype pairs that resolve that same genotype. In the example above, the two possible haplotype pairs are:

ACG/TCA (the true pair) and ACA/TCG. Note that for a given genotype of a single individual, there are potentially an exponential number of haplotype pairs, obscuring the true pair we wish to discover.

Thus we must turn to either novel experimental methods that map or sequence two homologous chromosome pairs separately or to computational approaches that exploit the evolutionary history of the chromosomes, which can be discerned from a population or familial relationship. Here we focus on algorithms for population data, rather than family data.

It is important to mention that the problem we face would be hopeless if the SNPs under consideration would be in linkage equilibrium. Linkage disequilibrium fades away with distance, so we can only hope to determine the correct haplotypes from genotypes for SNPs that are relatively close together. The prospect becomes less dismal when we consider several recent studies (Daly et al. 2001 [37], Gabriel et al. 2002 [42], Patil et al. 2001 [52]) that suggest that the linkage disequilibrium extent in several analyzed regions is larger than expected. These studies show that several regions can be partitioned into *blocks* of size up to 100 kb such that in each block there is very little variation across the population. More specifically, in each block only a few haplotypes (2-4) account for over 90% of the haplotypes in the sample. We exploit this fact in devising an asymptotically linear-time statistical algorithm for this problem.

Note that the traditional approach to genetic mapping has been to examine each SNP one at a time and associate a correlation or score (e.g., LOD score) to quantify its contribution to a trait. However the information gained from single

marker scores turns out to be very noisy. Thus with genome-wide study of the haplotypic patterns for several adjacent SNPs, one can now hope for a more robust method of mapping disease genes based on the algorithm described here.

4.2 Related Literature

The experimental solutions — *Single Molecule Dilution* (Ruano et al. 1990 [54]), *Asymmetric PCR Amplification* (Michalatos-Beloin et al. 1996 [49]), *Isolation of Single Sperm Cells* (Li et al. 1988 [48]) — are low-throughput, expensive, and difficult to automate. A new approach based on single-molecule based individual haplotype maps appears promising (Anatharaman et al. 2005 [34]), but not yet available. Faced with these disadvantages inherent to the laboratory methods, scientists have begun to explore the computational approaches as a viable alternative.

The competing computational methods fall into the following categories and are described briefly:

The first method due to Clark (Clark 1990 [36]) is based on a heuristic that starts with the list of haplotypes that can be unambiguously inferred from the genotype data, (i.e. the ones coming from homozygous or single-site heterozygous individuals), and then tries to solve the phase ambiguous individuals by using these already determined haplotypes, while adding new haplotypes to the list (when trying to resolve a genotype with a haplotype in the list and a new, yet-to-be discovered haplotype). While this algorithm is rather simple and remains popular, it suffers from a few problems —namely, it might have trouble getting started; it

might fail to resolve all individuals; and the solution depends on the order in which this algorithm examines the genotypes.

Another class of methods is based on the so-called *Perfect Phylogeny (PP) model* of haplotype evolution — namely a model assuming no recombination and the usual infinite-site mutation process as in population genetics. Given the putative existence of these blocks of extended linkage disequilibrium, the PP model works well on an individual block, but not applicable to whole genome study. (See Gusfield's algorithm (Gusfield 2002 [43]) based on a reduction of the haplotype inference problem to the graph realization problem, or other much simpler solutions, but with less efficient quadratic time complexity, e.g., algorithms devised independently by Bafna et al. 2002 [35] and by Eskin et al. 2002 [39].) More recently, algorithms that are able to allow for small deviations from the PP model have been developed (Halperin et al. 2004 [44] and Song et al. 2005 [55]).

A third class of methods consists of *statistical methods*, to which our current algorithm belongs. The *Maximum Likelihood Estimation* (Excoffier and Slatkin 1995 [40]) approach estimates *the haplotype frequencies* by maximizing the likelihood of the frequencies given the data, using the EM (expectation-maximization) algorithm. Starting with some initial frequencies, the following two steps are repeated until convergence to a limit distribution:

- The E-step computes for each genotype the probability of resolving it into each possible haplotype pair:

$$P(h_1, h_2 |\ g) \sim p_{h_1} \cdot p_{h_2},$$

using the haplotype frequencies estimated at the previous step ($(h_1,\ h_2)$ haplotype pair explains the genotype g).

- The M-step updates the haplotype frequencies using the estimates obtained in the E-step.

$$p_h = \frac{1}{2n} \sum_{j=1}^{m} n_j \sum_{i=1}^{c_j} \delta_{ih} P(h_{i1}, h_{i2} |\ g_j)$$

where $2n$ = the total number of haplotypes in the sample, n_j = the number of genotypes of type j, c_j = the number of possible haplotype pairs that explain genotype g_j (**exponential** in the number of heterozygous sites) and δ_{ih} = an indicator equal to the number of times haplotype h is present in the pair $(h_{i1},\ h_{i2})$.

This algorithm is accurate in estimating the haplotype frequencies, especially in large sample sizes (Fallin and Schork 2000 [41]), but, unfortunately, exponential in the number of heterozygous loci.

Several *Bayesian algorithms* devised for this problem use Gibbs samplers but differ with respect to priors: e.g., Dirichlet prior, or a coalescent-based prior. See Stephens et al. 2001 [56]. These are MCMC algorithms that construct a Markov chain whose stationary distribution is $P(H |\ G)$, where H is a set of haplotype pairs that can explain the set of known genotypes G. Starting with an initial guess of haplotypes H^0, an individual is repeatedly chosen at random from the ambiguous individuals and its haplotype pair is estimated given the estimated haplotypes for the other individuals: sample (h_{i1}, h_{i2}) from $P((h_1, h_2) |\ G, H_{-i})$ where H_{-i} are

the estimated haplotypes for the other individuals. This process is repeated until convergence.

More practical Bayesian algorithms exploit locality in linkages as we do: using a *divide and conquer* strategy developed by Niu et al. 2002 [50], one such algorithm uses two important *computational tricks*: partition-ligation and prior annealing that help reduce running times and also the mixing of the MC.

Partition Step: Partition the region into small (around 8 SNPs) continuous blocks. On each block, apply a Gibbs sampler as before [56].

Ligation Step: Next, combine estimates for adjacent blocks using the same Gibbs sampler to obtain estimates for the entire region. These tricks have been applied to other algorithms as well: the one in Stephens and Donnelly (2003 [57]) and to EM (Qin et al. 2002 [51]).

4.3 Methods

We propose an EM-based simple algorithm that shares similar accuracy as the algorithms mentioned above, while being able to handle large datasets involving hundreds of sites and genotypes. This algorithm is shown below:

Algorithm FastHI:

    Find-Distributions;

    Repeat the following steps several times

        Generate-Blocks;

        Solve-on-Blocks;

        Stitch-Blocks;

    End Repeat

4.3.1 Find-Distributions

The algorithm starts by approximating the distribution of haplotypes on each possible block (formed by consecutive SNPs) with length between 4 SNPs and some upper limit, e.g., 25 SNPs (the upper limit depends on the number of SNPs in the dataset). An efficient implementation proceeds as follows: start by applying the classic EM method on all small blocks of length 4 and determine the possible haplotypes and their distribution on these small blocks.

Figure 4.1: Dividing the genomes into blocks.

From every pair of consecutive small blocks of length 4, say consisting of SNPs $i \ldots i+3$ and respectively $i+1 \ldots i+4$, compute a list of possible haplotypes on the larger block of length 5, consisting of SNPs $i \ldots i+4$. In this list, include the haplotypes that result from the merging of two haplotypes which on both blocks of length 4 have a frequency greater than a threshold.

We call the haplotypes obtained in this way **extended** haplotypes. Having this small list of haplotypes, we repeatedly estimate their frequencies by EM. Since this list has far fewer haplotypes than the **exponential** number of haplotypes possible in a block, and since we only extend from blocks of length l to length $l+1$, FastHI

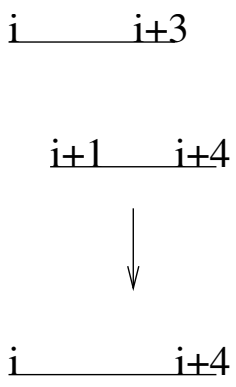


Figure 4.2: Merging a pair of consecutive blocks.

obtains a good approximation of the haplotype frequencies in each possible block of length between 4 and some upper bound. Note that this process does not lose any of the **common** haplotypes (i.e., with a real frequency larger than some threshold; according to our simulations –data not shown– 5% seems a safe cutoff), considering the fact that EM accurately estimates the frequencies of the common haplotypes on small regions.

Lemma 4.3.1 *Suppose a haplotype has (real) frequency $p \geq \theta$ in a block of length $l > 4$. Then by the process described above, this haplotype will be included in the list of estimated haplotypes for that block.*

Proof by induction: On blocks of length 4, EM can be shown to perform well in estimating the haplotype distribution. Assume we have a block of length $l > 4$. Let H be a haplotype on this block with real frequency $p = \text{realfreq}(H)$. Then if h_1 and h_2 are the two haplotypes of length $l-1$, which by merging give rise to H,

we have:

$$\theta \leq p = \text{realfreq}(H) \leq \text{realfreq}(h_i),$$

with $i \in \{1, 2\}$. By the induction hypothesis, h_1 and h_2 will be in the list of estimated haplotypes and hence H will be too.

Note that we may still lose rare haplotypes that occur only in combination with other rare haplotypes, but this is true for all general statistical methods.

4.3.2 Generate-Blocks

This step randomly generates break points along the region, creating a partition such that each block (a region between two consecutive breakpoints) of a sequence of SNPs has length between 4 and some upper bound (e.g. 25).

4.3.3 Solve-on-Blocks

On each block partial solutions are generated using the distribution obtained in the first step (Find-Distributions).

For each individual we can determine the solution on blocks by selecting the most plausible hypothetical solution. Formally, for a genotype g we choose the following resolution:

$$\arg \max_{(h_1,h_2)\in H_g} P(h_1, h_2|g)$$

where $P(h_1, h_2|g) \sim p_{h_1} \cdot p_{h_2}$, H_g is the set of all possible resolutions with haplo-

types in our list for genotype g, and p_{h_1} and p_{h_2} are the estimated frequencies of haplotypes h_1 and h_2, respectively.

4.3.4 Stitch-Blocks

Finally, the FastHI algorithm merges the partial solutions on blocks to get the solution for the entire genotype. For a given block partition, say consisting of b blocks, we have to make $b-1$ decisions about how to connect the adjacent haplotypes.

Formally, for an individual genotype having local resolutions h_i^1 and h_i^2 on block i, and h_{i+1}^1 and h_{i+1}^2 on block $i+1$, we have to choose between $((h_i^1,\ h_{i+1}^1),\ (h_i^2,\ h_{i+1}^2))$ and $((h_i^1,\ h_{i+1}^2),\ (h_i^2,\ h_{i+1}^1))$. It is implemented using a very efficient EM estimator to decide which pair has a higher likelihood. By repeatedly joining on each pair of consecutive blocks from left to right, we obtain for each individual genotype a solution.

We remark that the solution obtained after just one such single partition into blocks is not reliable, simply because the partition is random and so even the local solutions might not be accurate. In order to avoid this problem, we execute multiple independent random partitions (say 50), and in the end we choose for each individual the solution that occurs most often. For the real datasets that we used as well as for the simulated ones, the accuracy results are very similar to the ones obtained by **Haplotyper** (Niu et al. 2002 [50]) and **modified Phase** (Stephens et al. 2003 [57]), two state of the art statistical algorithms. However, FastHI is much faster and thus, handles large datasets.

4.3.5 Error Metrics

We use the following measures to characterize the error in our inference, which assumes that the correct haplotypes have been determined by some other independent means, e.g. experimental methods for the real datasets and known from simulations for the simulated data.

Mean Individual Error

This metric is defined as the proportion of individuals whose inferred haplotypes do not exactly match the correct haplotypes. This metric is rather stringent and the error rate tends to one rapidly as the number of SNP markers increases (due to the decay of the linkage disequilibrium), independent of the underlying algorithms.

Single-Site Error

This metric is defined as the proportion of phases (of all *possible* phase errors) that are wrongly inferred.

Note that the number of *possible* phase errors is half the number of heterozygous entries in the data set. This measure is better in comparing different methods, since it gives more insight about how the algorithms perform locally. For instance, a mistaken individual may have only one phase error, or at the opposite end all the possible phase errors.

Switch Error

This metric is defined as the proportion of heterozygous positions that are wrongly phased with respect to the previous heterozygous position.

This measure is better for datasets with a large number of markers and for which we do not have enough statistical information to correctly infer the phases at a global level. For instance, imagine that there is a hot spot of recombination somewhere in the middle. Then the left part and the right part are independent and even an optimal algorithm would produce on average 50% errors for the mean individual error. The new measure would detect that for each wrongly phased individual the left and the right part are correct. Moreover, globally, we need only a phase switch to produce the correct solution.

4.4 Results

4.4.1 Real Data

The real datasets examined are as follows:

β_2 AR (β_2-Adrenergic Receptor) (Drysdale et al. 2000 [38])

This dataset contains 13 SNPs from the human β_2-adrenergic receptor gene (believed to be related to asthma) genotyped in 121 Caucasian patients with asthma. There are only 10 different haplotypes present in this dataset of 121 genotypes. All three methods correctly phase all individuals. This is probably to be expected of such large datasets with little variability.

ACE (Angiotensin Converting Enzyme) (Rieder et al. 1999 [53])

This dataset contains 52 SNPs from the ACE gene (implicated in cardiovascular diseases) genotyped in 11 individuals. The errors for all three algorithms are similar.

CFTR (Cystic Fibrosis Transmembrane-Conductance Regulator) (Kerem et al. 1989 [46])

This dataset consists of 23 SNPs genotyped in a 1.8 Mb region on chromosome 7q31, believed to be implicated in Cystic Fibrosis, a common recessive disease that occurs about once per 2000 births. As in Niu et al. 2002 [50], we selected the 57 haplotypes without missing data from the 94 haplotypes in the diseased individuals. We generated 100 datasets of 28 genotypes by randomly permuting the 57 haplotypes. The large errors in the inferred haplotypes from all methods are because of the limited number of genotypes and the great diversity (there are 30 different haplotypes and 28 genotypes).

The results on all three datasets are summarized in Table 4.1 (FastHI is the proposed algorithm). For these real datasets of rather small size, all algorithms are fast.

4.4.2 Simulated Data

We also performed an extensive simulation study in order to be able to analyze larger datasets than the real datasets we have had access to.

We generated haplotypes according to the following model: the parameters in

	β_2 AR	ACE	CFTR[a]
Mean Indiv. Error			
Haplotyper [b]	0	.18	.42
modified Phase [c]	0	.18	.47
FastHI	0	.18	.43
Single-Site Error			
Haplotyper	0	.07	.32
modified Phase	0	.06	.35
FastHI	0	.06	.32
Avg Switch Error[d]			
Haplotyper	0	2.5	1.68
modified Phase	0	1.5	1.60
FastHI	0	3	1.68

[a]average error rates for 100 data sets generated by randomly pairing 56 haplotypes
[b]error rate for the best of 20 independent runs
[c]error rate for the best of 5 independent runs; for ACE data we did 100 runs
[d]defined as the average number of switch errors on a mistaken individual

Table 4.1: Comparison of Error Rates of Haplotyper, Phase and FastHI algorithms on the 3 real datasets

our simulations are the number of distinct haplotypes present in the dataset, the number of SNPs analyzed and the distribution of these haplotypes. The details are as follows: we first specify a number of distinct haplotypes present in the dataset. Each such haplotype is a random haplotype, independent of the other haplotypes. Each position (SNP) is a Bernoulli trial with $p = .5$. Then for each generated type we specify its frequency in the dataset. We then generated 20 independent datasets of genotypes. Each dataset is obtained by generating a set of haplotypes and randomly generating genotypes by pairing up haplotypes.

We simulated several types of datasets: SD1, SD2, SD3, SD4 and SD5, all described below. All have 150 genotypes, 100 SNPs, but they differ in the number of haplotypes and their distribution.

SD1 Simulation Data

There are 150 genotypes, 100 SNPs, 50 different haplotypes in the dataset. The distribution of the haplotypes is uniform (expected count for each haplotype is 6).

SD2 Simulation Data

There are 150 genotypes, 100 SNPs, 100 different haplotypes in the dataset. The distribution of the haplotypes is non-uniform, with 90 haplotypes having the same frequency .007 (expected count is 2), 5 haplotypes having frequency .02 (expected count is 6) and the remaining 5 haplotypes have the following frequencies: .1 (expected count is 30), .07 (expected count is 20), .07 (expected count is 20), .03 and .03 (expected count is 10).

SD3 Simulation Data

There are 150 genotypes, 100 SNPs, 100 different haplotypes in the dataset. The distribution of the haplotypes is uniform (expected count for each haplotype is 3).

SD4 Simulation Data

There are 150 genotypes, 100 SNPs, 150 different haplotypes in the dataset. The distribution of the haplotypes is non-uniform.

SD5 Simulation Data

There are 150 genotypes, 100 SNPs, 150 different haplotypes in the dataset. The distribution of the haplotypes is uniform.

	SD1[a]	SD2[b]	SD3[c]	SD4[d]	SD5[e]
Mean Indiv. Error					
Haplotyper	–	–	–	–	–
modified Phase	.018	.13	.12	.193	–
FastHI	.0016	.069	.058	.126	.50
Single-Site Error					
Haplotyper	–	–	–	–	–
modified Phase	.002	.0368	.02	.057	–
FastHI	.00045	.027	.02	.069	.22
Average Time on Dataset[f]					
Haplotyper	–	–	–	–	–
modified Phase	270	630	780	1080	–
FastHI	8	15	20	21	24

[a] 150 genotypes, 100 SNPs, 50 different haplotypes, uniform distribution
[b] 150 genotypes, 100 SNPs, 100 different haplotypes, non-uniform distribution
[c] 150 genotypes, 100 SNPs, 100 different haplotypes, uniform distribution
[d] 150 genotypes, 100 SNPs, 150 different haplotypes, non-uniform distribution
[e] 150 genotypes, 100 SNPs, 150 different haplotypes, uniform distribution
[f] in minutes

Table 4.2: Comparison of Error Rates of Haplotyper, Phase and FastHI algorithms on simulated datasets (averages over 20 simulated datasets)

The results of the simulations and the time performance (in minutes) are given in Table 4.2. As can be seen from the table, our algorithm is much faster (minutes versus hours) and has similar (even better) performance. Also on the large dataset of 150 genotypes, 100 SNPs and 150 different haplotypes, FastHI is the only algorithm that works. These results strengthen our belief that this simple and fast algorithm can handle very large datasets with reasonable accuracy.

4.5 Discussion

We have presented a simple and fast algorithm, FastHI, for inferring haplotypes from genotypes of diploid individuals. The main advantage FastHI enjoys over

its rivals is that it effortlessly handles very large datasets for inference without sacrificing accuracy.

More importantly, the underlying ideas are powerful and generalizes to combinations of other polymorphisms. In particular, application of these ideas to "Haplotype-Copy-Number-Polymorphims" problem in analyzing arrayCGH (Comparative Genomic Hybridization) data suggests an interesting opportunity (work in progress). In particular, successful analysis of these data will provide us with needed information to understand LOH (loss of heterozygousity), chromosomal aberrations and copy number fluctuations in cancer, auto-immune disorders and other related diseases.

Chapter 5

Future Work

In this thesis we described three important problems dealing with the analysis of large-scale genetic datasets. The important contribution is that they try to use the available information on multiple markers and multiple disease loci, so that as much information as possible is used from the datasets. New statistical methods are needed in the context of the massive datasets that are being generated and the problems addressed in this thesis are important in this context.

There are two main extensions that we are currently pursuing:

1 Combining the gene expression with copy number variation data for the detection of cancer genes.

The algorithm we proposed makes use only of copy-number-variation data. However, the deletions or amplification may not be very precise in pinpointing oncogenes and tumor suppressor genes. The gene expression data contribute important information as to which disease genes are involved in cancer. We are currently working on a method to incorporate gene expression informa-

tion, when available, in addition to aCGH data.

2 Combining linkage and association information

The linkage method in Chapter 3 only works for affected-sib-pairs. However Lo and Zheng have proposed similar methods for association for the ASP design and for a case-control design. Combining the linkage and association signal at marker loci should give a superior method for the detection of disease susceptibility loci.

Bibliography

[1] Knudson AG (1971) Mutation and Cancer: Statistical Study of Retinoblastoma. Proc Natl Acad Sci USA 68(4): 820–823

[2] Daruwala RS, Rudra A, Ostrer H, Lucito R, Wigler M, Mishra B (2004) A Versatile Statistical Analysis Algorithm to Detect Genome Copy Number Variation. Proc Natl Acad Sci USA 101(46): 16292–16297

[3] Glaz J, Naus J, Wallenstein S (2001) Scan Statistics. Springer

[4] Wallenstein S, Neff N (1987) An Approximation for the distribution of the scan statistic. Statistics in Medicine 6: 197–207

[5] Zhao X, Weir BA, LaFramboise T, Lin M, Beroukhim R, Garraway L, Beheshti J, Lee JC, Naoki K, Richards WG, Sugarbaker D, Chen F, Rubin MA, Janne PA, Girard L, Minna J, Christiani D, Li C, Sellers WR, Meyerson M (2005) Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. Cancer Res 65(13): 5561–70

[6] Wu Y, Dowbenko D, Spencer S, Laura R, Lee J, Gu Q, Lasky LA (2000)

Interaction of the tumor suppressor PTEN/MMAC with a PDZ domain of MAGI3, a novel membrane-associated guanylate kinase. J Biol Chem 275(28): 21477–21485

[7] Rimessi P, Gualandi F, Morelli C, Trabanelli C, Wu Q, Possati L, Montesi M, Barrett JC, Barbanti-Brodano G (1994) Transfer of human chromosome 3 to ovarian cancer cell lines identifies three region on 3p involved in ovarian cancer. Oncogene 9: 3467–3474

[8] Shivapurkar N, Virmani AK, Wistuba II, Milchgrub S, Mackay B, Minna JD, Gazdar AF (1999) Deletions of chromosome 4 at multiple sites are frequent in malignant mesothelioma and small cell lung carcinoma. Clin Cancer Res 5: 17–23

[9] Tseng RC, Chang JW, Hsien FJ, Chang YH, Hsiao CF, Chen JT, Chen CY, Jou YS, Wang YC (2005) Genomewide loss of heterozygosity and its clinical associations in non small cell lung cancer. Int J Cancer 117(2): 241–7

[10] Hosoe S, Ueno K, Shigedo Y, Tachibana I, Osaki T, Kumagai T, Tanio Y, Kawase I, Nakamura Y, Kishimoto T (1994) A frequent deletion of chromosome 5q21 in advanced small cell and non-small cell carcinoma of the lung. Cancer Res. 54(7): 1787–90

[11] Inoue M, Starostik P, Zettl A, Strobel P, Schwarz S, Scaravilli F, Henry K, Willcox N, Muller-Hermelink HK, Marx A (2003) Correlating genetic aberrations with World Health Organization-defined histology and stage across the spectrum of thymomas. Cancer Res. 63(13): 3708–15

[12] Flemming A, Brummer T, Reth M, Jumaa H (2003) The adaptor protein SLP-65 acts as a tumor suppressor that limits pre-B cell expansion. Nat Immunol. 4(1): 38–43

[13] Abujiang P, Mori TJ, Takahashi T, Tanaka F, Kasyu I, Hitomi S, Hiai H (1998) Loss of heterozygosity (LOH) at 17q and 14q in human lung cancers. Oncogene 17(23): 3029–33

[14] Sanchez-Cespedes M, Parrella P, Esteller M, Nomoto S, Trink B, Engles JM, Westra WH, Herman JG, Sidransky D (2002) Inactivation of LKB1/STK11 is a common event in adenocarcinomas of the lung. Cancer Res. 62(13): 3659–62

[15] Peruzzi D, Aluigi M, Manzoli L, Billi AM, Di Giorgio FP, Morleo M, Martelli AM, Cocco L (2002) Molecular characterization of the human PLC beta1 gene. Biochim Biophys Acta 1584(1): 46–54

[16] Lee EB, Park TI, Park SH, Park JY (2003) Loss of heterozygosity on the long arm of chromosome 21 in non-small cell lung cancer. Ann Thorac Surg. 75(5): 1597–600

[17] High-resolution genomic profiles of human lung cancer (2005) Tonon G, Wong KK, Maulik G, Brennan C, Feng B, Zhang Y, Khatry DB, Protopopov A, You MJ, Aguirre AJ, Martin ES, Yang Z, Ji H, Chin L, Depinho RA. Proc Natl Acad Sci U S A. 102(27): 9625–30

[18] Nguyen PL, Niehans GA, Cherwitz DL, Kim YS, Ho SB (1996) Membrane-

bound (MUC1) and secretory (MUC2, MUC3, and MUC4) mucin gene expression in human lung cancer. Tumour Biol. 17(3): 176–92

[19] Zhao X, Li C, Paez JG, Chin K, Janne PA, Chen TH, Girard L, Minna J, Christiani D, Leo C, Gray JW, Sellers WR, Meyerson M (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res. 64(9): 3060–71

[20] Hoh J, Ott J (2000) Scan statistics to scan markers for susceptibility genes. Proc Natl Acad Sci USA 97: 9615–9617

[21] Schork NJ, Boehnke M, Terwilliger JD, Ott J (1993) Two-Trait Locus Linkage Analysis: A Powerful Strategy for Mapping Complex Genetic Traits. Am J Hum Genet 53: 1127–1136

[22] Cordell HJ, Todd JA, Bennett ST, Kawaguchi Y, Farrall M (1995) Two-Locus Maximum LOD Score Analysis of a Multifactorial Trait: Joint Consideration of IDDM2 and IDDM4 with IDDM1 in Type 1 Diabetes. Am J Hum Genet 57: 920–934

[23] Farrall M (1997) Affected Sibpair Linkage Tests for Multiple Linked Susceptibility Genes. Genet Epidemiol 14: 103–115

[24] Cordell HJ, Wedig GC, Jacobs KB, Elston RC (2000) Multilocus Linkage Tests Based on Affected Relative Pairs. Am J Hum Genet 66: 1273–1286

[25] Efron B (2004) Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. J Am Stat Assoc 99: 96–104

[26] Lo SH and Zheng T (2002) Backward Haplotype Transmission Association (BHTA) Algorithm - A Fast Multiple Marker Screening Method. Hum Hered 53: 197–215

[27] Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, McLeod RS, Griffiths AM, Green T, Brettin TS, Stone V, Bull SB, Bitton A, Williams CN, Greenberg GR, Cohen Z, Lander ES, Hudson TJ, Siminovitch KA (2000) Genomewide Search in Canadian Families with Inflammatory Bowel Disease Reveals Two Novel Susceptibility Loci. Am J Hum Genet 66: 1863–1870

[28] Lo SH and Zheng T (2004) A Demonstration and Findings of a Statistical Approach through Reanalysis of Inflammatory Bowel Disease Data. Proc Natl Acad Sci USA 101: 10386–10391

[29] Daly MJ, Kruglyak L, Pratt S, Houstis N, Reeve Mp, Kirby A, Laner ES (1998) GENEHUNTER 2.0 - a complete linkage analysis system. Am J Hum Genet Suppl 63: A286

[30] Heresbach D, Alizadeh M, Dabadie A, Le Berre N, Colombel JF, Yaouanq J, Bretagne JF, Semana G (1997) Significance of interleukin-1beta and interleukin-1 receptor antagonist genetic polymorphism in inflammatory bowel diseases. Am J Gastroenterol. 92(7): 1164–9

[31] Barmada MM, Brant SR, Nicolae DL, Achkar JP, Panhuysen CI, Bayless TM, Cho JH, Duerr RH (2004) A Genome Scan in 260 Inflammatory Bowel Disease-Affected Relative Pairs. Inflammatory Bowel Disease 10(1): 15–22

[32] Katsanos KH, Tsatsoulis A, Christodoulou D, Challa A, Katsaraki A, Tsianos EV (2001) Reduced serum insulin-like growth factor-1 (IGF-1) and IGF-binding protein-3 levels in adults with inflammatory bowel disease. Growth Horm 11(6): 364–7

[33] Vestergaard EM, Brynskov J, Ejskjaer K, Clausen JT, Thim L, Nexo E, Poulsen SS (2004) Immunoassays of human trefoil factors 1 and 2: measured on serum from patients with inflammatory bowel disease. Scand J Clin Invest 64(2): 146–156

[34] Anantharaman TS, Mysore V, Mishra B (2005) Fast and Cheap Genome wide Haplotype Construction via Optical Mapping. Proceedings of PSB 385–96

[35] Bafna V, Gusfiled D, Lancia G, Yooseph S (2002) Haplotyping as Perfect Phylogeny: A Direct Approach. UC Davis, Computer Science Technical Report CSE-2002-21

[36] Clark AG (1990) Inference of Haplotypes from PCR-Amplified Samples in Diploid Populations. Mol Biol Evol 7: 111–122

[37] Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-Resolution Haplotype Structure in the Human Genome. Nat Genet 29 : 229–232

[38] Drysdale CM et al. (2000) Complex Promoter and Coding Region β_2-Adrenergic Receptor Haplotypes Alter Receptor Expression and Predict *in vivo* Responsiveness. Proc Natl Acad Sci USA 97 : 10483–10488

[39] Eskin E, Halperin E, Karp R (2002) Efficient Reconstruction of Haplotype Structure via Perfect Phylogeny. Technical Report, UC Berkeley Computer Science Dept

[40] Excoffier L, Slatkin M (1995) Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population. Mol Biol Evol 12 : 947–959

[41] Fallin D, Schork N (2000) Accuracy of Haplotype Frequency Estimation for Biallelic Loci, via the Expectation-Maximization Algorithm for Unphased Diploid Genotype Data. Am J Hum Genet 67 : 947–959

[42] Gabriel S et al. (2002) The Structure of Haplotype Blocks in the Human Genome. Science 296: 2225–2229

[43] Gusfield D (2002) Haplotyping as Perfect Phylogeny: Conceptual Framework and Efficient Solutions (extended abstract). Proceedings of the 6th International Conference on Computational Molecular biology (RECOMB)

[44] Halperin E, Eskin E (2004) Haplotype reconstruction from genotype data using Imperfect Phylogeny. Bioinformatics 20(12): 1842–9

[45] Hudson RR (2002) Generating Samples Under a Wright-Fisher Neutral Model of Genetic Variation. Bioinformatics 18 : 337–338

[46] Kerem B et al. (1989) Identification of the Cystic Fibrosis Gene: Genetic Analysis. Science 245 : 1073–1080

[47] Lander E and Schork N et al. (1994) Genetic Dissection of Complex Traits. Science 265: 2037–2048

[48] Li H et al. (1988) Amplification and Analysis of DNA Sequences in Single Human Sperm and Diploid Cells. Nature 335 : 414–417

[49] Michalatos-Beloin S et al. (1996) Molecular Haplotyping of Genetic Markers 10kb Apart by Allele-Specific Long Range PCR. Nucleic Acids Res 24 : 4841–4843

[50] Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms. Am J Hum Genet 70: 157–169

[51] Qin ZS, Niu T, Liu JS (2002) Partition-Ligation-Expectation-Maximization for Haplotype Inference with Single Nucleotide Polymorphisms. Am J Hum Genet 71: 1242–1247

[52] Patil N et al. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294(5547): 1719–23

[53] Rieder MJ et al. (1999) Sequence Variation in the Human Angiotensin Converting Enzyme. Nat Genet 22: 59–62

[54] Ruano G et al. (1990) Haplotype of Multiple Polymorphisms Resolved by Enzymatic Amplification of Single DNA Molecules. Proc Natl Acad Sci USA 87: 6296–6300

[55] Song YS, Wu Y, Gusfield D (2005) Algorithms for Imperfect Phylogeny Haplotyping (IPPH) with a Single Homoplasy or Recombination Event. Proceedings of WABI 2005

[56] Stephens M, Smith NJ , Donnelly P (2001) A New Statistical Method for Haplotype Reconstruction from Population Data. Am J Hum Genet 68: 978–989

[57] Stephens M, Donnelly P (2003) A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. Am J Hum Genet 73(6): 1162–9

[58] Zhang K et al. (2002) A Dynamic Programming Algorithm for Haplotype Block Partitioning. Proc Natl Acad Sci USA 95: 7335–7339