Thesis for the Degree of Licentiate of Philosophy

Modelling Hereditary Diseases Using Galton-Watson Processes

Ulrica Olofsson

Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
Göteborg, Sweden 2000

Modelling Hereditary Diseases Using Galton-Watson Processes
Ulrica Olofsson

ISSN 0347-2809/NO 2000:59
Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone +46-(0)31-772 1000

[Chalmers University of Technology]
Göteborg, Sweden 2000

Abstract

In this thesis we consider two applications of Galton-Watson branching processes. The first application concerns the estimation of the age of a disease-causing founder mutation of autosomal dominant inheritance. A Galton-Watson branching model is proposed, resulting in an estimate of the mutation age. Also, the problem of combining estimates based on several markers is discussed. Simulations illustrate the various methods and the mutation age estimate is applied to real data. The subject of homozygosity mapping, a method using differences in marker lengths to find the disease locus, is our second application.

Keywords: Age estimation - Galton-Watson process - composite likelihood - homozygosity mapping - BRCA1

Acknowledgements

First I would like to express my deepest gratitude to all those people who, during this work, have been supportive, encouraging and who never stopped believing in me.

Especially I wish to thank:

Ziad Taib	my supervisor, for his enthusiasm, excellent ideas and for guiding me through the work on this thesis
Bill Amos	for proposing some interesting ideas and for giving me inspiration to start
Annika Bergman and Tommy Martinsson	for giving me the opportunity to use my estimate on real data sets
family and friends and colleagues	for being there whenever I needed

Ulrica Olofsson,
Härryda, October 2000

Contents

1 Introduction **1**

2 Basic Genetics **5**
2.1 Introduction 5
2.2 Markers and Microsatellites 6
2.2.1 Microsatellites 7
2.3 Mutation Mechanisms 7
2.3.1 IAM - Infinite alleles model 7
2.3.2 SMM - Stepwise mutation model 8
2.4 Genetic Linkage 9

3 Estimating the Age of a Disease using Galton-Watson Processes **11**
3.1 Introduction 11
3.2 The Model 12
3.2.1 Galton-Watson 12
3.2.2 Finding $\boldsymbol{M}^t$ 14
3.3 The Estimate 15
3.3.1 Method of Moments Estimate 15
3.3.2 Maximum Likelihood Estimate 16
3.4 Preliminary Simulations 18
3.4.1 Simulation 1 19
3.4.2 Simulation 2 20

4 Haplotypes and Composite Likelihood **21**
4.1 Introduction 21
4.2 Haplotypes 21
4.3 Combining Estimates 26

4.3.1 Simulation example - combination of estimates 26
4.4 Composite likelihood 27

5 The example of a BRCA1 mutation 31
5.1 Background 31
5.2 Results from family data 32

6 Mutations and Homozygosity Mapping 37
6.1 Introduction 37
6.2 The Model 38
6.3 Homozygosity mapping 40

CHAPTER 1

Introduction

One of the matters discussed in this work is the problem of estimating the time since a disease-causing mutation arose in a population. The model proposed for solving this problem is a multitype Galton-Watson branching process, earlier described by (Pankratz, 1998).

We consider an autosomal dominant founder-mutation causing a rare disease, and assume that the disease-gene location is known. In Chapter 3, a Galton-Watson branching process model is constructed, describing the evolution of some genetic marker in the disease-carriers along the family tree. An estimate of the mutation age is derived using the method of moments and, further, a couple of simulation examples illustrate the method.

In the literature, other methods for estimating the age of a founder mutation are proposed. One of them can be found in (Risch *et al.*, 1995). This method is based on the probability that a disease-mutation-bearing chromosome does not carry a founder marker allele, the frequency of that allele among normal individuals (p_{ni}) and the recombination frequency (θ) between the marker and the disease locus. It is used in several earlier age estimations in rare diseases among others in Finnish populations (Moisio *et al.*, 1996) and among Ashkenazi Jews (Risch *et al.*, 1995).

A model using haplotypes is introduced in Chapter 4, resulting in formulas containing large matrices. As early as in the two-marker case, the formulas become cumbersome. There seems to be no simple formula describing the marker allele transitions among disease-carriers for t generations. This method seems to be a dead end, but can be used in various computational investigations.

The problem of combining several age estimates, one for each marker, is mentioned. The simplest way of combining estimates, and the one used in the simulation examples to follow, is to use the mean estimate. It is also possible, and probably better, to use weights, appropriately chosen. A simulation study is included to illustrate some results.

Another way of finding a combined estimate using several markers at the same time is the method of composite likelihood. The estimate is obtained by adding log likelihood functions and finding the maximum. An example comparing the method of moments estimate and the composite likelihood estimate is included. As is also the conclusion in (Pankratz, 1998), this method does not yield great results in estimating the age of a disease.

In chapter 5, data from a mammary cancer study, conducted in the western part of Sweden, is analysed using the herein described method.

Eighteen families with a specific cancer-causing mutation have been collected and genotyped for thirteen polymorphic markers. The haplotypes are found using the GENEHUNTER software. From family data, $18-20$ disease-allele carrying individuals are collected as a sample of the disease population, and $31-38$ healthy relatives are used as the normal population sample. The distances from the genetic linkage map used in the study are translated to recombination frequencies using the Haldane map function. Because of some constraints on our estimate, all markers do not yield an estimate. The mean value of our 10 estimates results in an age estimate of 128 generations.

One simulation of 20 and another of 50 generations have been performed, using marker distances and marker allele frequencies obtained in family data. The resulting estimates were 72.1858 and 90.0115 generations respectively. Thus, the estimate obtained above is most likely much larger than the true value.

In chapter 6, mutations are introduced in the earlier G-W model. The microsatellite markers are assumed to follow the Stepwise Mutation Model. Again a formula for the transitions of marker alleles in t generations is deduced.

The concept of homozygosity mapping is then introduced, being our second Galton-Watson application. Our goal here is to localise a disease gene. Shortly described, homozygosity mapping works by using differences in marker allele lengths to compare normal and disease populations to find the most probable location for the disease gene. Again, we consider a disease caused by a founder mutation, but with the distinction that we are dealing with a recessive disease. Thus, affected individuals are carrying two disease alleles, inherited one from each parent.

The concept used is that due to homozygosity by descent, individuals carrying the disease allele are homozygous in the area of the chromosome surrounding the disease locus in much larger extent than expected (cf. (Kruglyak *et al.*, 1995) and (Lander and Botstein, 1987)).

CHAPTER 2

Basic Genetics

2.1 Introduction

The genetic blueprint of every living organism is found in the *nucleotides*, stacked in *chromosomes* which are gathered in the nucleus of each cell as the entire *genome* of that organism. The nucleotides are paired together, two by two, forming a string of nucleotides attached to a singularly determined complementary string of nucleotides, building up the chromosome in the form of a ladder.

When an organism has one set of chromosomes, it is said to be a *haploid* organism, and if it has two sets of chromosomes, i.e. pairs of chromosomes, it is called *diploid*.

The process that produces *gametes* (a gamete is a haploid cell that fuses with a corresponding haploid cell from the opposite sex to form a new individual) is called *meiosis* (see Figure 2.1). During this process, certain *crossovers* may occur, where the sister chromosomes exchange parts. If an odd number of crossovers have occurred between two loci, we call that a *recombination* event.

The chromosomes carry genetic information in the protein-coding regions called the *genes*. Non-coding regions are not totally information free, since certain *polymorphic* sequences can be used as *genetic markers*. This matter is further explained in the next section. By the term polymorphic, we mean that in the population, several phenotypic forms (*alleles*) exist at a certain site (*locus*). As guides to the genome, *genetic maps* are constructed, containing markers and distances. The unit used is mostly cM, where the distance of 1 cM

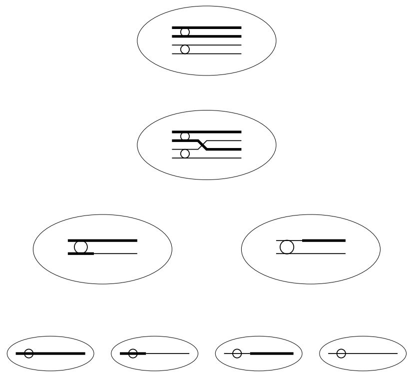


Figure 2.1: Simplified scheme over the meiosis

between two loci corresponds to the recombination probability between them of 0.01 (i.e. one product of meiosis out of 100 will in average be recombinant).

Genotype is what we call the specific allelic arrangement of one or several genes and the *phenotype* is the detectable outward manifestation of a specific genotype.

A *recessive* trait will only be visible in an individual carrying two copies of that trait, i.e. inherited from both parents. If only one copy of the disease allele is enough to express the trait it is said to be *dominant*. *Autosomal dominant* is when the trait is not located on the sex chromosomes, X and Y.

For more detailed information, we refer to existing literature. Two sources used in this introduction are (Griffiths *et al.*, 1996) and (Liu, 1998).

2.2 Markers and Microsatellites

Genetic markers are DNA subsequences that exist in several polymorphic forms in individuals within a species. Several kinds of markers have been detected and are used for various reasons. One of these are further explained in the next subsection.

A large number of certain molecular markers have been collected to be able to describe and find our way around in the genome. These are detected by probes: one short sequence finding the beginning of the marker and another to find the ending, which 'cut out' our marker of interest and show us the contents, simply explained. These markers, being evolutionary neutral, have a high level of heterozygosity for some type of neutral DNA variation (i.e. polymorphic).

2.2.1 Microsatellites

As mentioned earlier, DNA contains a vast amount of information in regions coding for proteins but there is also a huge amount of pieces that seem to carry no information at all. Among these are short strings of DNA called *microsatellites* or *simple sequence repeats* (SSR). As the name suggests, they consist of several repeats of short, simple sequences. As a rule, a microsatellite is a short sequence of 1 to 6 basepairs repeated after each other up to 100 times (Liu, 1998).

The most common kind of microsatellites is repetitive DNA based on dinucleotide repeats, like the CA-repeat. This microsatellite consists of the two nucleotides C and A repeated n times, and is sometimes denoted by $(CA)_n$.

Even though these DNA pieces seem to be nonsense, we can use them as markers when trying to make something out of the enormous amount of information contained in the genome. The reason why these pieces of DNA are useful is that the number of repeats differ from one chromosome to another. We inherit one set of chromosomes from each parent and thereby also one marker from each. It is feasible to visually see the difference in length of markers using appropriate methods, such as the *polymerase chain reaction*, PCR. The microsatellites are useful when, among other things, we want to reconstruct phylogenetic relationships, for forensic science, parentage testing and as it seems, in population genetics (cf. (Harding *et al.*, 1993)).

2.3 Mutation Mechanisms

2.3.1 IAM - Infinite alleles model

A motivation for using this model is the following. Consider a gene coding for a protein with 300 aminoacids - it has the length of 900 nucleotides. We thereby have $4^{900} = 10^{542}$ possible ways of placing one of four nucleotides at each position. That is, we have such large number of possible combinations that the probability that the same combination will occur more than once is negligible. Thus, when a new mutation occurs, a new allele is created that did not exist in the population earlier.

This model is called the *infinite alleles model*, since there are infinitely many states to which an allele can mutate and every mutation state is equally probable irrespective of the state of the allele. Since we have infinitely many possibilities and the probability of mutating to an already existing number of repeats is tiny, we have few states represented in the population and every mutation will be unique.

What makes this model less useful, though, is that the difference in the number of repeats between two markers gives us information about how far back in time they have a common ancestor allele. This information is not taken into consideration by this model (Goldstein *et al.*, 1995). When a mutation occurs, we get a new mutant that is related to the 'original' allele, with respect to the number of repeats. Mutation tends to change the size of the allele very little, by a few repeats more or less. These two facts are the reasons why we need a model better suited for microsatellites.

2.3.2 SMM - Stepwise mutation model

This model can be interpreted in two slightly different ways.

Strict stepwise mutation model

This model states that when a microsatellite loci mutates, it always changes size with one repeat more or less, i.e., an allele with i repeats mutates to $i-1$ or $i+1$ repeats with the same probability $\theta/2$ respectively.

The probabilities for adding and deleting one repeat does not necessarily need to be the same. This model can also be generalised to permit other variants of mutations.

Stepwise mutation model

The more general model allows other changes in the number of repeats than the strict stepwise mutation model. It is thus consistent with

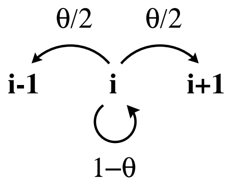


Figure 2.2: The transition probabilities in the SMM model

the distribution of alleles at microsatellite loci, (Valdes *et al.*, 1993), but at the same time also a bit more difficult to implement.

Mutations resulting in a state already represented in the population are the most common ones. In the SMM, there are fewer possible alleles than in the IAM for the same population size and mutation degree. Since our alleles consist of tandemly repeated sequences, this model is more appropriate than the IAM. Also, this model does not predict as high levels of heterozygosity as the IAM (cf. (Harding *et al.*, 1993)).

2.4 Genetic Linkage

Genetic linkage is what we call the association of genes located on the same chromosome. For genes on the same chromosome, the segregation ratios for the genotypes depart from the Mendelian independent assortment ratios. If the position of a marker is close to the disease locus, resulting in a low probability of crossover at meiosis, we say that we have linkage between them. Thus, linked to the disease locus are some markers. Being linked implies that they are inherited together with the disease allele most of the time, but recombinations and mutation events occasionally break this inheritance pattern.

CHAPTER 3

Estimating the Age of a Disease using Galton-Watson Processes

3.1 Introduction

We consider an autosomal dominant disease which arose in a normal, homogeneous population by a genetic mutation in the DNA of an individual, (from now on called the founder), some unknown time ago. This disease has seemingly no negative effect on the fertility (at least until the carrier has reached a certain age) so the disease-causing mutation has spread to the offspring due to normal segregation for generations. The propagation of the disease is conveniently modelled using a Galton-Watson branching process (further explained in 3.2.1). As of today, we know the disease did not die out, so the branching process considered is supercritical.

On the same chromosome as the disease locus we have several linked genetic markers.

In this chapter, we only have interest in one polymorphic microsatellite marker at a time, chosen at a suitable, known distance from the disease locus.

We assume that the disease-causing mutation is relatively recent, so the population of individuals carrying normal alleles is considerably larger than the disease allele carrying population.

Our interest herein lies in finding an estimate of the time since the disease allele appeared in the population. We will later in this chapter propose an estimate, found both using the method of moments and the maximum likelihood method, for the number of generations since the founder mutation arose in the population. This estimate is based on information gained from the marker allele frequencies.

3.2 The Model

We fix a suitable genetic marker linked to the disease locus. The marker considered is a polymorphic microsatellite whose position in relation to the disease locus is exactly known.

We need to know the true marker allele frequencies in the normal population, assumed being stable and homogeneous.

The marker allele will not always be inherited together with the disease allele due to recombination events. In the case of cross-over, we presume the new marker allele randomly chosen according to the allelic distribution in the normal population due to panmixia (i.e. random mating in a population).

Our disease should not be too old. The size of the disease population need to be considerably smaller than the normal population, i.e., the disease is not so widely spread in the population. We can then assume that the probability of two carrier chromosomes meeting is negligible. We also disregard mutation events since the disease appeared relatively recent in time and the normal allele distribution is thought to be stable.

Time is measured in number of generations. The fact that the generation span can differ a lot in different branches of the family tree is not taken into consideration.

We use the Galton-Watson branching process to describe the spread of the disease through time, tracing the disease chromosome labelled by the marker allele.

3.2.1 Galton-Watson

Let $M_1, M_2, \dots M_k$ denote the k marker alleles. Let further r be the recombination frequency between the loci of the disease gene and the marker and Z_t is the number of individuals in generation t. The

Z_t variables are based on earlier events as

$$Z_{t+1} = \sum_{j=1}^{Z_t} X_{jt},$$

where X_{jt} denotes the number of offspring of individual j in generation t. To emphasise the multitype nature of this process we take

$Z_t(i) = \#$ individuals carrying allele i in generation t, $i \in \{1, 2, ..., k\}$.

We assume that the reproduction law is such that $\mathsf{E}[X_{jt}] = 1 + \lambda$ for all i and t and that the process is supercritical, i.e. that $\lambda > 0$. Moreover we assume that the X_{jt}:s are independent.

Denote $\boldsymbol{M} = (m_{ij})_{i,j}$ as the matrix consisting of

$$m_{ij} = \mathsf{E}[\#\text{ offspring of type } j \text{ from a type } i \text{ individual}].$$

Now it is easily shown that

$$m_{ij} = \begin{cases} (1+\lambda) r p_{nj} & i \neq j \\ (1+\lambda)[(1-r) + r p_{nj}] & i = j \end{cases}$$

where p_{ni} is the proportion of individuals carrying marker allele i in the normal population. If the marker allele is changed, it is due to recombination. When instead the progeny inherits the same marker allele it can be for one out of two reasons: either no recombination has occurred or a recombination occurred giving back the same marker allele, randomly chosen according the normal allele distribution.

Let $\boldsymbol{Z}_t = [Z_t(1), Z_t(2), ..., Z_t(k)]$, where $Z_t(i)$ is as above. A well known result for Galton-Watson processes (cf. (Jagers, 1975)) allows us to write

$$\mathsf{E}[\boldsymbol{Z}_t | \boldsymbol{Z}_0] = \boldsymbol{Z}_0 \boldsymbol{M}^t$$

where $\boldsymbol{M}^t$ stands for the matrix $(m_{ij}^{(t)})_{i,j}$ where $m_{ij}^{(t)}$ is the expected number of progeny of type j born to an individual of type i t generations later.
Define the founder generation as

$$\boldsymbol{Z}_0 = [1, 0, \ldots, 0]$$

We are interested in the first row of the matrix $\boldsymbol{M}^t$, that is $\boldsymbol{Z}_0 \boldsymbol{M}^t$. The reason being that we assume that the probable founder allele is placed in the first position. First, we need to find a formula describing $\boldsymbol{M}^t$. This is done in the next section.

3.2.2 Finding $\boldsymbol{M}^t$

It is not difficult to see that in this case

$$\boldsymbol{M} = (1+\lambda)[(1-r)\boldsymbol{I} + r\boldsymbol{P}]$$

(recall the formula for m_{ij}). As earlier, $1+\lambda$ is the expectation of the offspring distribution. The $\boldsymbol{P}$-matrix gives the transition probabilities between the different alleles

$$\boldsymbol{P} = \begin{pmatrix} p_{n1} & p_{n2} & \cdots & p_{nk} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nk} \end{pmatrix}$$

where p_{nj} is the proportion of individuals with allele j in the normal population. Since it is assumed that the population in which the disease appeared is stable and homogeneous, p_{nj} is a fixed quantity from generation to generation. The corresponding quantities in the disease population, p_{dj} are defined as the proportion of individuals with allele j in the disease population in the present time, but the true frequencies are unknown. As an estimate we use the observed allele proportions.

If we expand the model, adding the possibility of mutations, we get another kind of matrix describing transitions from one allele to another. This case will be discussed in a chapter 6.

We now have the following marker allele distribution for the offspring:

Recall that

$$\boldsymbol{M} = \mathsf{E}[\#\text{ offspring with marker allele } j \text{ when the founder had allele } i]$$

where $i, j \in \{1, 2, ..., k\}$ and notice that

$$\begin{aligned} \frac{\boldsymbol{M}^t}{(1+\lambda)^t} &= [(1-r)\boldsymbol{I} + r\boldsymbol{P}]^t \\ &= r^t\boldsymbol{P}^t + \sum_{i=1}^{t-1} \binom{t}{i} [(1-r)\boldsymbol{I}^i (r\boldsymbol{P})^{t-i}] + (1-r)^t \boldsymbol{I}^t \\ &= r^t\boldsymbol{P}^t + \boldsymbol{P}\sum_{i=1}^{t-1} \binom{t}{i} (1-r)^i r^{t-i} + (1-r)^t \boldsymbol{I}^t \\ &= r^t\boldsymbol{P} + (1 - r^t - (1-r)^t)\boldsymbol{P} + (1-r)^t\boldsymbol{I} = \\ &= (1-r)^t\boldsymbol{I} + (1-(1-r)^t)\boldsymbol{P} \end{aligned}$$

where we use that $\sum_{i=0}^{t}\binom{t}{i}(1-r)^i r^{t-i} = 1$ and that $\boldsymbol{P}^t = \boldsymbol{P}$.

We now make a simple but crucial approximation:

$$\frac{\boldsymbol{M}}{1+\lambda} \simeq \Big(\mathsf{P}(\text{offspring has marker } j, \text{ founder had marker } i)\Big)_{i,j}.$$

From this we get

$$\frac{\boldsymbol{M}^t}{(1+\lambda)^t} \simeq \Big(\mathsf{P}(\text{offspring has marker } j \text{ after } t \text{ generations})\Big)_{i,j}$$

which can be obtained by replacing the actual empirical ratios by the ratios of expectations.

3.3 The Estimate

First assume that we have a sample of n_d individuals from generation t carrying the disease allele. Let n_{dj} be the number of individuals carrying marker allele j in the sample. We can then use that information to estimate p_{dj}, namely as $\hat{p}_{dj} = \frac{n_{dj}}{n_d}$.

We will now show how to find the same estimate

$$\hat{t} = \frac{\ln(\frac{p_{d1}-p_{n1}}{1-p_{n1}})}{\ln(1-r)}$$

using two different methods: The method of moments and the maximum likelihood method.

3.3.1 Method of Moments Estimate

Using the formula for t generations $(1-r)^t\boldsymbol{I} + (1-(1-r)^t)\boldsymbol{P}$ obtained earlier, we have two cases:

$i = 1$: Using the first row of the above matrix, corresponding to the founder allele, we get $(1-r)^t + (1-(1-r)^t)p_{n1} = p_{n1} + (1-p_{n1})(1-r)^t$ and thus

$$p_{d1} = p_{n1} + (1-p_{n1})(1-r)^t$$

where p_{d1} is replaced by its estimate $\hat{p}_{d1}$. The equation is solved for the unknown variable t, as

$$\hat{t} = \frac{\ln(\frac{\hat{p}_{d1} - p_{n1}}{1 - p_{n1}})}{\ln(1-r)}$$

Note that we require that $\hat{p}_{d1} > p_{n1}$.

$i \neq 1$: Here we use the i^{th} row in the matrix above:

$$p_{di} = p_{ni} - p_{ni}(1-r)^t$$

which also gives us an estimate of t as:

$$\hat{t} = \frac{\ln(\frac{p_{ni} - \hat{p}_{id}}{p_{ni}})}{\ln(1-r)}$$

In this case we have the requirement of $p_{ni} > \hat{p}_{di}$.

As we see here, we get an estimate of t also in the case where the individual does not carry the original marker allele. This case will not be further studied in this work.

3.3.2 Maximum Likelihood Estimate

The same estimate as above can also be derived from the Maximum Likelihood method. One condition for using this method is that the individuals are independent of each other. In reality, however, we have individuals who are related, e.g. siblings, cousins, second cousins etc, and hence not independent. When sampling from the last generation, we will get individuals who are likely to be less related, and we get approximately independence, which should suffice.

As earlier, let

$$\pi_j = \begin{cases} p_{ni} + (1-p_{ni})(1-r)^t & i = 1 \\ p_{ni} - p_{ni}(1-r)^t & i \neq 1 \end{cases}$$

The likelihood function is

$$L(n_{d1}, \ldots, n_{dk}; t) = \prod_{j=1}^{k} \pi_j^{n_{dj}} = (p_{n1} + (1-p_{n1})(1-r)^t)^{n_{d1}} \cdot \prod_{j=2}^{k} (p_{nj} - p_{nj}(1-r)^t)^{n_{dj}}$$

and the log likelihood is

$$l = \ln L = n_{d1} \cdot \ln(p_{n1} + (1 - p_{n1})(1-r)^t) + \sum_{j=2}^{k} n_{dj} \cdot \ln(p_{nj} - p_{nj}(1-r)^t) =$$

$$n_{d1} \cdot \ln(p_{n1} + (1 - p_{n1})(1-r)^t) + \sum_{j=2}^{k} n_{dj} \cdot \ln(p_{nj}(1 - (1-r)^t))$$

Maximum is obtained by solving the equation

$$0 = \frac{\delta \ln L}{\delta t} =$$

$$\frac{n_{d1}(1-p_{n1})(1-r)^t \ln(1-r)}{p_{n1} + (1-p_{n1})(1-r)^t} - \sum_{j=2}^{k} \frac{n_{dj}(1-r)^t \ln(1-r)}{1-(1-r)^t} =$$

$$\frac{n_{d1}(1-p_{n1})}{p_{n1} + (1-p_{n1})(1-r)^t} - \frac{\sum_{j=2}^{k} n_{dj}}{1-(1-r)^t}$$

where $\sum_{j=2}^{k} n_{dj} = n_d - n_{d1}$. Thus, we have that

$$\frac{n_{d1}(1-p_{n1})}{p_{n1}(1-p_{n1})(1-r)^t} = \frac{n - n_{d1}}{1-(1-r)^t}$$

Solving this equation for t (in several steps) yields

$$n_{d1}(1-p_{n1})(1-(1-r)^t) = (n_d - n_{d1})(p_{n1} + (1-p_{n1})(1-r)^t)$$

and then

$$n_{d1} - n_{d1}(1-r)^t + n_{d1}p_{n1}(1-r)^t =$$

$$np_{n1} + n(1-p_{n1})(1-r)^t - n_{d1}(1-p_{n1})(1-r)^t.$$

Now, divide by n_d on both sides and use that $\hat{p}_{d1} = n_{d1}/n_d$,

$$\hat{p}_{d1} - p_{n1}(1-r)^t + \hat{p}_{d1}p_{n1}(1-r)^t =$$

$$p_{n1} + (1 - p_{n1}(1-r)^t - \hat{p}_{d1}(1-p_{n1})(1-r)^t$$

Notice that $\hat{p}_{d1}(1-p_{n1})(1-r)^t = (\hat{p}_{d1} - \hat{p}_{d1}p_{n1})(1-r)^t$, which is a term on the left side of the equation. Dividing on both sides yields

$$\hat{p}_{d1} = p_{n1} + (1-p_{n1})(1-r)^t$$

which can be solved for t, resulting in the ML-estimate

$$\hat{t} = \frac{\ln(\frac{\hat{p}_{d1} - p_{n1}}{1 - p_{n1}})}{\ln(1 - r)}$$

Notice that this estimate uses the distance between p_{d1} and p_{n1} in an intuitively natural manner. It also has a certain resemblance to the so called p_{excess} as is described in (Weir, 1996). To investigate the behaviour of the estimate we have performed a simulation study.

3.4 Preliminary Simulations

The simulation source code was written in MATLAB. Some details can be found in Appendix A.

We start with the first generation with one founder individual carrying the marker allele named 1. The subsequent generation will consist of a certain number of disease-allele carrying individuals (obtained by a formula described below), some with other marker alleles but most with the same as the founder. If a recombination changes a marker allele from the founder allele, it is possible to retrieve it by recombination since this allele also exists in the normal population. The family tree is then built up, generation by generation, consisting only of disease allele carriers, where individuals are represented by their marker allele.

As described in the previous section and using the same notation, we use the Galton-Watson branching process resulting in the following model:

Take $Z_t(i) = \#$ individuals in generation t carrying marker allele i. As before, the marker allele carried by the founder is named allele 1 and the other alleles are numbered $2, \ldots, k$. The founder generation can be written as

$$Z_0(1) = 1,\ Z_0(2) = 0\,,\ Z_0(3) = 0\,,\ \ldots,\ Z_0(k) = 0$$

or in vector form

$$\boldsymbol{Z}_0 = [1, 0, \ldots, 0]\,.$$

The $(t+1)^{th}$ generation is generated according to the formula (cf. (Kaplan and Weir, 1997))

$$Z_{t+1}(i) \sim \mathsf{Poi}\left((1+\lambda)\left[(1-r)\, Z_t(i) + r p_{ni}\, Z_{tot}(t)\right]\right)$$

where $i = 1, \ldots, k$, $\lambda > 0$ is the population growth factor in a supercritical Galton-Watson branching process, r is the recombination frequency between the marker and disease loci, p_{ni} is, as described earlier, the frequency of marker allele i in the normal population and

$$Z_{tot}(t) = \sum_{j=1}^{k} Z_t(j) = \# \text{ of individuals in generation } t.$$

The expected number of children of each individual is $1 + \lambda$ as earlier.

3.4.1 Simulation 1

The first simulation is of a population of 20 generations and the population growth rate was set as 1.50. That is, each individual will get the expected number of 1.50 disease carrying children. The recombination probability between the marker and disease loci is 0.03. We fix the number of marker alleles in the population to be 5. The simulation was iterated 2000 times, but all simulations where the disease population died out before the 20^{th} generation were removed, here resulting in 1077 full iterations.

Using the previously described estimate for each iteration, we get the age distribution in Figure 3.1 below, with a mean value of 18.9809 and standard deviation 15.3634.

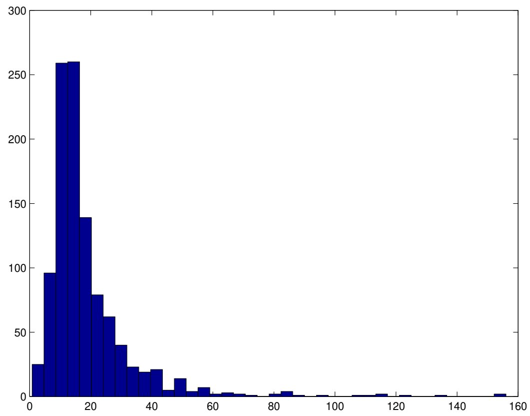


Figure 3.1: Distribution of sample estimates

3.4.2 Simulation 2

The second simulation is also of a population of 75 generations, but to illustrate the importance on the choice of marker locus, we here illustrate the case when the distance between the marker and disease loci is smaller than in the previous simulation. The recombination probability is set to 0.01. The other parameters are the same as in the first simulation example. Here we got the estimated number of generations 62.9196.

To get a feeling for the robustness of the estimate especially when based on a small number of individuals, 1000 samples of size 20 was collected. Individuals, here represented by their marker alleles, are chosen with a probability obtained from the observed allele frequencies in the last generation of the simulation. 1000 samples of size 20 from the last generation resulted in the mean estimate 65.5629 with the standard deviation 28.1057.

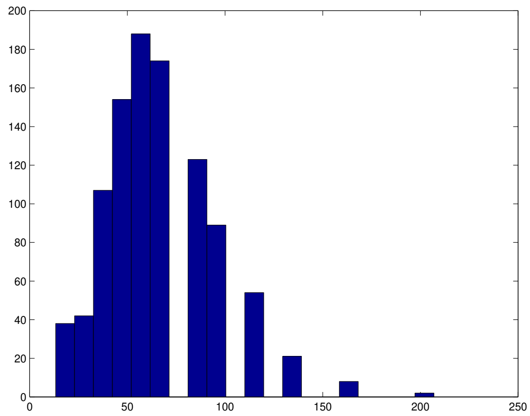


Figure 3.2: Estimates

As is evident in the Figure 3.2, we here got an underestimation of the number of generations. The probable cause could be that the disease is somewhat older and the marker locus is chosen a bit too closely to the disease locus. Many recombinations have had time to occur, and the allele distribution has already begun to resemble the one from the normal population.

CHAPTER 4

Haplotypes and Composite Likelihood

4.1 Introduction

In this chapter, we consider the same situation as in the previous chapter except for the fact that we allow for more than one marker.

Our starting point is that we know how to estimate the age of a disease using every marker separately. The question now is how such estimates could be combined.

In what follows we will consider three possible ways to do this:

1. Studying the haplotypes, using multi-type Galton Watson processes

2. Using weighted averages

3. Using composite likelihood

4.2 Haplotypes

Rather than using both sister chromosomes in the diploid individuals, we concentrate on the sister chromosome carrying the dominant disease. Thereby we get a useful simplification of the model: Shortly explained, we trace the disease carrying chromosome throughout the generations.

Let us first examine the case with two markers and a disease locus, as illustrated in the Figure 4.1 below.

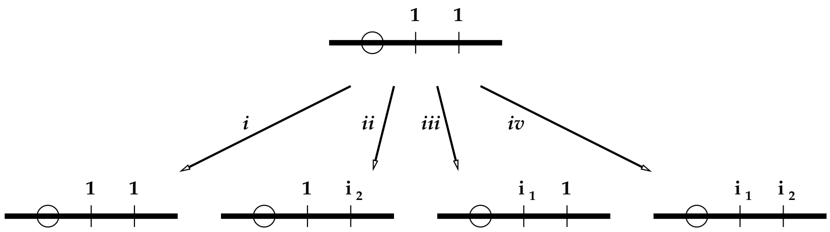


Figure 4.1: Two markers

As can be seen, there are five possible haplotypes resulting from a disease-allele carrying individual. The fifth case, not shown in Figure 4.1, is the case where the recombination results in an individual who did not inherit the disease allele and our interest lies only in individuals carrying the disease.

i) The first case seems trivial since the haplotype is identical to the original. Considering that recombinations can have given back the same marker/markers as before, there are four different possibilities to acquire this result.

ii) One of the markers seem to be unchanged. Thus, case *ii* implies a recombination between the two markers, disregarding the possibility that the recombination gave in return the same allele on one of the two markers. Thus, there might instead have been a recombination between disease locus and the first marker, leaving the first marker unchanged by chance.

iii) The third case can occur either by a recombination between the disease locus and the first marker where the new piece is carrying the same allele at marker 2 as the founder, or by two recombinations, exchanging the piece between the disease locus and the second marker.

iv) To obtain case *iv* with two entirely different markers but still the disease locus, the recombination has to have occurred between the disease locus and the first marker.

For simplicity, let us use the notation described in Figure 4.2 below, when describing the possible transitions. The aim of the following is to show how complicated things become in this relatively simple case with only two markers. The results arrived at can however be used as the basis of various computational investigations.

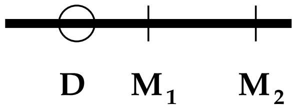


Figure 4.2: The notation

Transitions from the marker allele pair (i, j) to the four different haplotypes are shown below with the corresponding probabilities.

$$ij \rightarrow \left\{ \begin{array}{lll} ij: & (1-r_1)(1-r_2) & \text{no recombination} \\ & (1-r_1)r_2 p_{\cdot j} & \text{rec. between } M_1 \text{ \& } M_2 \\ & r_1 r_2 p_{i\cdot} & \text{rec. between } D \text{ \& } M_1 \text{ and } M_1 \text{ \& } M_2 \\ & r_1(1-r_2)p_{ij} & \text{rec. between } D \text{ \& } M_1 \\ ib: & (1-r_1)r_2 p_{\cdot b} & \text{rec. between } M_1 \text{ \& } M_2 \\ & r_1(1-r_2)p_{ib} & \text{rec. between } D \text{ \& } M_1 \\ aj: & r_1 r_2 p_{a\cdot} & \text{rec. between } D \text{ \& } M_1 \text{ and } M_1 \text{ \& } M_2 \\ & r_1(1-r_2)p_{aj} & \text{rec. between } D \text{ \& } M_1 \\ ab: & r_1(1-r_2)p_{ab} & \text{rec. between } D \text{ \& } M_1 \end{array} \right.$$

Thus, we have the following four terms: $(1-r_1)(1-r_2)$, $(1-r_1)r_2$, $r_1(1-r_2)$ and $r_1 r_2$. Also note that $p_{i\cdot} = \sum_j p_{ij}$ and $p_{\cdot j} = \sum_i p_{ij}$.

The expectation matrix $\boldsymbol{M}$ will get the following appearance:

$$
\boldsymbol{M} = \begin{pmatrix}
m_{11,11} & \cdots & m_{11,1n_2} & \cdots & m_{11,n_11} & \cdots & m_{11,n_1n_2} \\
\vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\
m_{1n_2,11} & \cdots & m_{1n_2,1n_2} & \cdots & m_{1n_2,n_11} & \cdots & m_{1n_2,n_1n_2} \\
m_{21,11} & \cdots & m_{21,1n_2} & \cdots & m_{21,n_11} & \cdots & m_{21,n_1n_2} \\
\vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\
m_{2n_2,11} & \cdots & m_{2n_2,1n_2} & \cdots & m_{2n_2,n_11} & \cdots & m_{2n_2,n_1n_2} \\
\vdots & & \vdots & & \vdots & & \vdots \\
m_{n_11,11} & \cdots & m_{n_11,1n_2} & \cdots & m_{n_11,n_11} & \cdots & m_{n_11,n_1n_2} \\
\vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\
m_{n_1n_2,11} & \cdots & m_{n_1n_2,1n_2} & \cdots & m_{n_1n_2,n_11} & \cdots & m_{n_1n_2,n_1n_2}
\end{pmatrix}
$$

Recall that the elements $m_{ij,ab}$ of $\boldsymbol{M}$ can be seen as the transitions from allele i to a in the first marker locus and from allele j to b in the second. Here, $i, a = 1, 2, \ldots, n_1$ and $j, b = 1, 2, \ldots, n_2$. These transitions are further described in

$$
m_{ij,ab} = \begin{cases}
(1-r_1)(1-r_2) & (a,b) = (i,j) \\
(1-r_1)r_2p_{\cdot b} & a \neq i \\
r_1(1-r_2)p_{ab} & (a,b) \neq (i,j) \\
r_1r_2p_{a\cdot} & b \neq j
\end{cases}
$$

The allele frequencies in the normal population are assembled in the $\boldsymbol{P}$ matrix:

$$
\boldsymbol{P} = \begin{pmatrix}
p_{11} & \cdots & p_{1n_2} & p_{21} & \cdots & p_{2n_2} & \cdots & p_{n_11} & \cdots & p_{n_1n_2} \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\
p_{11} & \cdots & p_{1n_2} & p_{21} & \cdots & p_{2n_2} & \cdots & p_{n_11} & \cdots & p_{n_1n_2}
\end{pmatrix}
$$

The dimensions of the matrices $\boldsymbol{M}$ and $\boldsymbol{P}$ are $n_1n_2 \times n_1n_2$.

We need a formula describing the $\boldsymbol{M}$ matrix. This is done a bit easier by first rewriting the form of the matrices $\boldsymbol{M}$ and $\boldsymbol{P}$. Let

$$
\boldsymbol{M} = \begin{pmatrix}
\boldsymbol{M}_{11} & \boldsymbol{M}_{12} & \cdots & \boldsymbol{M}_{1n_1} \\
\boldsymbol{M}_{21} & \boldsymbol{M}_{22} & \cdots & \boldsymbol{M}_{2n_1} \\
\vdots & \vdots & \ddots & \vdots \\
\boldsymbol{M}_{n_11} & \boldsymbol{M}_{n_12} & \cdots & \boldsymbol{M}_{n_1n_1}
\end{pmatrix}
$$

where $\boldsymbol{M}_{ij}$ is an $n_2 \times n_2$ matrix. Thus, the mutation matrix $\boldsymbol{M}$ is decomposed into several matrices $\boldsymbol{M}_{ij}$, where

$$\boldsymbol{M}_{ij} = \begin{pmatrix} m_{i1,j1} & m_{i1,j2} & \cdots & m_{i1,jn_2} \\ m_{i2,j1} & m_{i2,j2} & \cdots & m_{i2,jn_2} \\ \vdots & \vdots & \ddots & \vdots \\ m_{in_2,j1} & m_{in_2,j2} & \cdots & m_{in_2,jn_2} \end{pmatrix}$$

In the same way, the $\boldsymbol{P}$ matrix is split up in elements $\boldsymbol{P}_i$ with the same dimension $n_2 \times n_2$

$$\boldsymbol{P}_i = \begin{pmatrix} p_{i1} & p_{i2} & \cdots & p_{in_2} \\ \vdots & \vdots & & \vdots \\ p_{i1} & p_{i2} & \cdots & p_{in_2} \end{pmatrix}$$

giving us the new form of $\boldsymbol{P}$ as

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{P}_2 & \cdots & \boldsymbol{P}_{n_1} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{P}_1 & \boldsymbol{P}_2 & \cdots & \boldsymbol{P}_{n_1} \end{pmatrix}$$

In the formula for $\boldsymbol{M}$ we also need the following matrix

$$\boldsymbol{P}. = \begin{pmatrix} p_{\cdot 1} & p_{\cdot 2} & \cdots & p_{\cdot n_2} \\ \vdots & \vdots & & \vdots \\ p_{\cdot 1} & p_{\cdot 2} & \cdots & p_{\cdot n_2} \end{pmatrix}$$

And so, the mutation matrix consists of smaller matrices of the following form

$$\boldsymbol{M}_{ij} = \begin{cases} (1-r_1)(1-r_2)\boldsymbol{I}_{n_2\times n_2} + (1-r_1)r_2\boldsymbol{P}.+ & \\ r_1 r_2\, p_{j}.\boldsymbol{I}_{n_2\times n_2} + r_1(1-r_2)\boldsymbol{P}_i & i = j \\ & \\ r_1 r_2\, p_{j}.\boldsymbol{I}_{n_2\times n_2} + r_1(1-r_2)\boldsymbol{P}_i & i \neq j \end{cases}$$

where as usual, $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix.
The problem starts when trying to see what happens in t generations. There seem to be no simple formula for describing $\boldsymbol{M}^t$.

4.3 Combining Estimates

With many markers we can use the formula

$$\hat{t} = \frac{\ln(\frac{p_{d1} - p_{n1}}{1 - p_{n1}})}{\ln(1 - r)}$$

to obtain an estimate of t for each marker. A simple way to combine these is to take the average. Note, however, that if a marker is situated too close to the disease locus, few recombinations will occur and thus the age estimate will become nonsense. If instead the distance is too large, several recombinations have had time to occur and the time estimate will be less accurate. That is why the different estimations should be weighted.

One way of obtaining a combined estimate of the time t using weights, is to set $\hat{t} = \sum_i w_i \hat{t}_i$. The w_i's should be chosen so that $\sum_i w_i = 1$. One criteria of choosing the weights is to require the variance $\mathsf{Var}(\hat{t})$ to be minimised subject to the condition $\sum_i w_i = 1$.

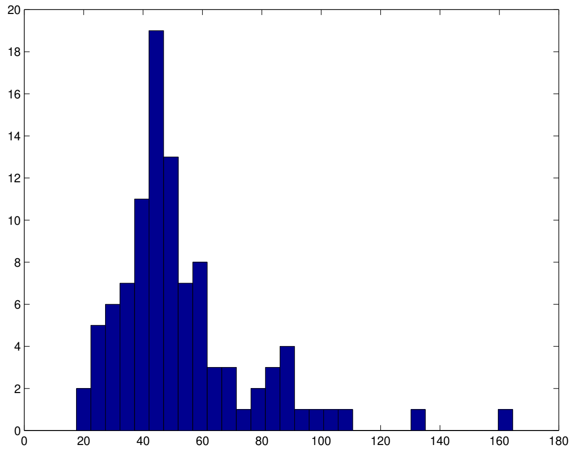
In the simulations to follow, the mean of the estimates is used as combined estimate. If the markers are chosen at nice symmetrical distances, we should get an acceptable estimate.

4.3.1 Simulation example - combination of estimates

The computer language used in this example to implement this method is JAVA. Further descriptions of this application can be found in the Appendix B. As long as either the number of generations or the probability of having two children is kept low, it runs painlessly.

The simulation is repeated 100 times. Having the pairwise map distance of 1 cM, we fix the number of alleles for each of the eight markers to 5. The probability of having one offspring is 0.85 and we continue the simulation for 50 generations.

We get the following distribution:
The simulation resulted in the mean estimate 52.9184 and the standard deviation 23.5796.



4.4 Composite likelihood

To be able to use information from several markers at the same time, the method of composite likelihood is convenient. An estimate is obtained by combining information from several (possibly dependent) sources by adding together conditional or marginal log likelihoods (cf. (Devlin *et al.*, 1996) and (Pankratz, 1998)).

Let $L_i(t)$ be the likelihood function for marker i. Then, $l_i(t)$ is the log likelihood function (marginal or conditional). These are combined as the composite likelihood function according to

$$CL(t) = \sum_{i=1}^{k} l_i(t).$$

The value of the parameter t that maximises the composite likelihood function solves the equation

$$\frac{\delta CL(t)}{\delta t} = \frac{\delta\left(\sum_{i=1}^{k} l_i(t)\right)}{\delta t} = \sum_{i=1}^{k} \frac{\delta\, l_i(t)}{\delta t} = 0.$$

The terms $S_i(t) = \frac{\delta\, l_i(t)}{\delta t}$ can also be called the (composite) score functions.

Another way to obtain the most probably value of t, is to find the maximum of the $CL(t)$ function,

$$CL(t) = \sum_{i=1}^{k} l_i(t),$$

where the log likelihood functions $l_i(t)$:s are found using calculations similar to those in Chapter 3.3.2.

We can use what we already know, and find a somewhat more simplified equation to maximise:

$$l_i(t) = n_i^{(i)} \ln(p_{ni}^{(i)} + (1-p_{ni}^{(i)})(1-r^{(i)})^t) + \sum_{i=2}^{k} n_j^{(i)} \ln(p_{nj}^{(i)}(1-(1-r^{(i)})^t)) =$$

$$n_i^{(i)} \ln(p_{ni}^{(i)} + (1-p_{ni}^{(i)})(1-r^{(i)})^t) +$$

$$\underbrace{\sum_{j=2}^{k} n_j^{(i)} \ln(p_{nj}^{(i)})}_{\text{independent of t}} + \sum_{j=2}^{k} n_j^{(i)} \ln(1-(1-r^{(i)})^t)$$

Thus, maximum of $CL(t)$ is obtained at the same parameter value of t as in the somewhat simpler

$$l_i(t) = n_i^{(i)} \ln(p_{ni}^{(i)} + (n-n_1^{(i)})(1-p_{ni}^{(i)})(1-r^{(i)})^t) + \ln(1-(1-r^{(i)})^t).$$

The Figure 4.3 illustrates the composite likelihood estimate of the age of a disease, where data is taken from a simulation of two markers where the first marker is chosen at the distance of 1 cM from the disease gene and the second marker is located 1 cM from the first marker. Both markers have 5 alleles each. 50 generations are simulated and the result of the method of moment estimate of 886 iterations are shown in the Figure 4.4. The data used in the above composite likelihood example is chosen from the simulation iteration for which the method of moment yielded an age estimate of 50.1053 generations. It is possible to get several composite likelihood estimates using the simulation data, but we use only this example to illustrate that the composite likelihood method seem to give rather low estimates of the mutation age. The same result is obtained in (Pankratz, 1998).

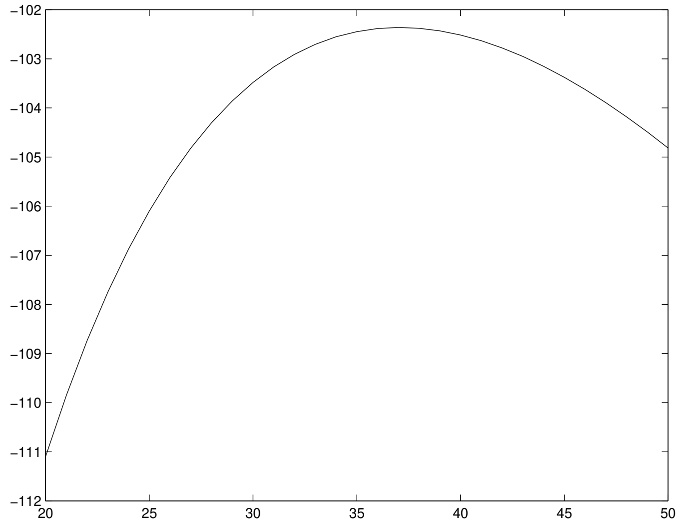


Figure 4.3: The maximum is the composite likelihood estimate of t

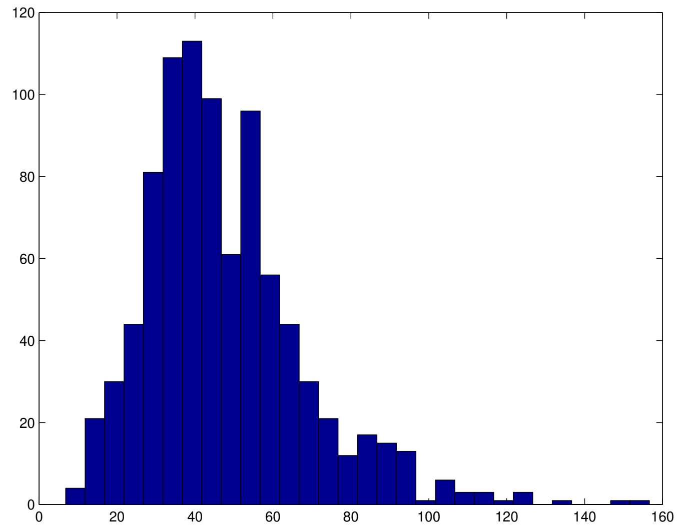


Figure 4.4: Estimates using method of moments

CHAPTER 5

The example of a BRCA1 mutation

5.1 Background

In this chapter, the model presented in the previous chapters is applied to real data to find an age estimate of a founder mutation found on the west coast of Sweden resulting in mammary cancer.

The study consists of 18 apparently unrelated families, where at least one member in each family carries a specific mutation in the BRCA1 gene (called 3171ins5). All families seem to have their geographical origin in the western part of Sweden.

Thirteen polymorphic markers have been genotyped for all available family members and by using the software GENEHUNTER ((Kruglyak *et al.*, 1996) and (Kruglyak and Lander, 1998)) the haplotypes were derived.

To estimate the time t since the appearance of the mutation, an estimate found using the method of moments, based on the theory of Galton-Watson branching processes is used:

$$\hat{t} = \frac{\ln(\frac{\hat{p}_{d1} - p_{n1}}{1 - p_{n1}})}{\ln(1 - r)}.$$

Time is measured in the number of generations since the appearance of the mutation in the population.

For every marker, the founder allele frequency in the sample can be estimated by the observed allele frequency, $\hat{p}_{d1} = \frac{n_{d1}}{n_d}$, where n_{d1} is the number of individuals carrying both the disease allele and the founder allele and n_d is the sample size.

We assume that the normal population is stable, i.e. that the allele frequencies p_{ni}, are unchanged from generation to generation. Here, $i \in \{1, \ldots, k\}$ and k is the number of alleles at that certain locus in the population. As estimates of the normal allele frequencies, the observed frequencies are used.

Here, the allele frequencies were based on the haplotypes of $18-20$ disease-allele carrying individuals and $31-38$ individuals assumed normal (healthy family members not carrying the disease mutation). The locations of the markers in relation to the disease locus is shown in the linkage map in Figure 5.1.

The recombination frequency r is found using the Haldane map function (see (Haldane, 1919)), translating map distances to recombination frequencies:

$$r_{AB} = \frac{1}{2}(1 - e^{-2d_{AB}/100})$$

where d_{AB} is the distance between two loci A and B, measured in cM and hence r_{AB} is the recombination frequency.

The criteria for choosing founder allele is the most common allele among the families with confirmed disease-associated allele.

5.2 Results from family data

Using the methods described in the previous chapters, the following table gives estimates of the age of the mutation, based on the frequencies of the different marker alleles available.

Using the mean of these estimates result in an estimate of 127.7 generations. Assuming that a generation is 25 years long, we conclude that the disease is about 3190 years old. One of the markers yield an unreasonable result, probably out of one of the reasons stated below. Removing that particular marker, gives us an estimate of 85.4 generations, implying an age of 2135 years. When removing all markers situated at a distance less than 1 cM, we get the estimate 60.3 generations, or 1508 years.

Marker	Age estimate	Founder allele
D17s1872	33.88	133
D17s946	61.55	126
D17s250	118.24	156
D17s800	99.47	173
D17s1299	62.83	192
D17s846	223.37	234
D17s1321	508.08	144
D17s855	-	144
D17s1325	122.92	195
D17s902	29.56	152
D17s588	16.74	154
D17s790	-	187
D17s787	-	137

Comments:

> Our estimate can only be used when $\hat{p}_{d1} > p_{n1}$, i.e. when the founder allele is more common in the disease population than among the normal population. Thus, the markers D17s855, D17s790 and D17s787 will not provide any information.
>
> The markers located at a distance less than 1 cM from the disease locus result in estimates much larger than expected. This effect can be out of several reasons, some of them mentioned in the sources of error section below.
>
> To get a feeling of the estimate variance, bootstrapping should be performed. This will be done in a coming article by (Bergman *et al.*, 2000).

There are several sources of error:

Founder allele: The choice of founder allele is naturally of importance, since the difference between allele frequencies in the two populations is the foundation of the estimate.

Genotyping: Genotyping errors caused in the laboratory will result in errors in the observed allele frequencies and hence give us a skew estimate.

Haplotype: Haplotyping errors caused by the software GENEHUNTER will result in errors in the observed allele frequencies and hence give us a skew estimate.

Recombination frequency: If a marker is set at a location closer to the disease locus than it is in reality, we will get a larger estimate, implying an older mutation. The estimate is very sensitive for changes in this parameter.

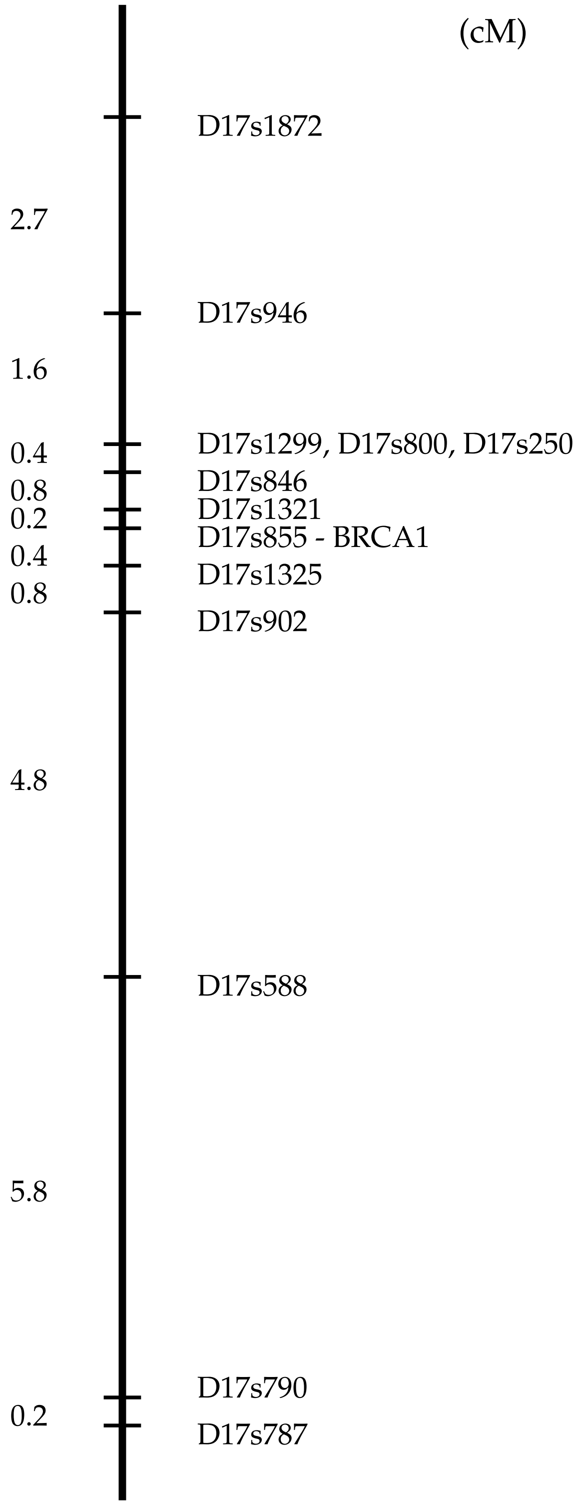


Figure 5.1: The linkage map: chromosome 17q21

CHAPTER 6

Mutations and Homozygosity Mapping

6.1 Introduction

Loosely speaking, homozygosity mapping is concerned with locating the locus of the disease gene for a recessive disease. Since an affected individual is necessarily homozygous for the disease, we expect such an individual to be homozygous also for markers surrounding the disease gene.

The method of homozygosity mapping is generally used for mapping rare recessive traits in children of consanguineous marriages. The disease is more probable to arise in inbred children due to the phenomenon homozygosity by descent in the adjacent region around the disease locus.

Homozygosity mapping is known to be convenient to use for mapping recessive diseases using data based on a very small number of families with few affected offspring in each, since data from unrelated individuals can be combined (cf. (Lander and Botstein, 1987)).

Recently (Amos *et al.*, 1999) have proposed the use of homozygocity mapping based on population (as opposed to family) data. In this chapter we propose a Galton-Watson model to describe such data. This model has two new characteristics: We consider the case of a recessive single-locus disease and assume that the disease is old. The second assumption implies that mutations in the marker can no longer be ignored. For the exact nature of the marker, see below.

The use of Galton-Watson processes as models of recessive traits has been criticised in the literature (Lange and Fan, 1997), but as shown in (Kaplan and Weir, 1997) and (Pankratz, 1998) the results based on such models are quite good.

6.2 The Model

Again, we have the transition matrix $\boldsymbol{P}$ as earlier. We assume a stepwise mutation model (SMM) as described in the Chapter 2 and the mutation matrix $\boldsymbol{U}$,

$$\boldsymbol{U} = \begin{pmatrix} 1-v & v & \cdots & 0 \\ u & 1-u-v & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1-u \end{pmatrix}$$

We now have two cases:

1. Mutation occur before the recombination
$\boldsymbol{M}_{mr} = (1+\lambda)\boldsymbol{U}[(1-r)\boldsymbol{I} + r\boldsymbol{P}] = (1+\lambda)[(1-r)\boldsymbol{U} + r\boldsymbol{U}\boldsymbol{P}] = (1+\lambda)[(1-r)\boldsymbol{U} + r\boldsymbol{P}]$

2. Recombination occur before the mutation
$\boldsymbol{M}_{rm} = (1+\lambda)[(1-r)\boldsymbol{I} + r\boldsymbol{U}]\boldsymbol{U} = (1+\lambda)[(1-r)\boldsymbol{U} + r\boldsymbol{P}\boldsymbol{U}]$

So, do mutations occur before or after the recombination event? Most likely, the mutations should occur in the process of DNA replication and since replication occur before recombination, then mutations should also occur before recombination (cf. (Pankratz, 1998)). I will thereby use $\boldsymbol{M}_{mr}$ as mutation matrix, simply calling it $\boldsymbol{M}$.

Assume that mutation events occur independently of recombination events. The formula $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{U}$ gives us the stationary distribution of $\boldsymbol{U}$. Here, $\boldsymbol{\pi}$ is the stationary distribution and $\boldsymbol{U}$ is the mutation matrix. Our transition matrix $\boldsymbol{P}$ will thereby get this form: $\boldsymbol{P} = \begin{pmatrix} \boldsymbol{\pi} \\ \boldsymbol{\pi} \\ \vdots \\ \boldsymbol{\pi} \end{pmatrix}$.

Since every row of $\boldsymbol{P}$ is the stationary distribution of $\boldsymbol{U}$, the following is valid: $\boldsymbol{PU} = \boldsymbol{P}$, $\boldsymbol{UP} = \boldsymbol{P}$ and $\boldsymbol{P}^2 = \boldsymbol{P}$.

Now, the interesting question was: *What does $\boldsymbol{M}^t$ look like?* Some long calculations show that

$$\frac{\boldsymbol{M}^t}{(1+\lambda)^t} = (1-r)^t \boldsymbol{U}^t + (1-(1-r)^t)\boldsymbol{P}$$

where $\boldsymbol{M} = (1-r)\boldsymbol{U} + r\boldsymbol{P}$. The i^{th} row is obtained as

$$\boldsymbol{e}_i \frac{\boldsymbol{M}^t}{(1+\lambda)^t} = (1-r)^t \boldsymbol{U}_{ij}^{(t)} + (1-(1-r)^t)p_{nj}$$

where $\boldsymbol{e}_i = (0, \dots, 0, 1, 0, \dots, 0)$, i.e. a 1 at the i^{th} position, 0s otherwise. The reason for our interest in the i^{th} row in this matrix is that we wish to find the distribution of the founder marker allele which we placed in the i^{th} position.

The transitions of the SMM model are described in Figure 6.1. As

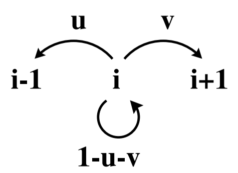


Figure 6.1: Transitions

an example of how $\boldsymbol{U}$, $\boldsymbol{\pi}$ etc can look like, we consider an example. Assume now that the evolution of the marker allele follows a SMM model (cf. Chapter 2). Since we assume that a long time has passed since the appearance of the disease, it is plausible to require that the values of the marker have reached stationarity. Using the formula $\boldsymbol{\pi} = \boldsymbol{\pi U}$ we obtain

$\pi_2 = \frac{v}{u}\pi_1$

$\pi_2 = v\pi_1 + (1-u-v)\pi_2 + u\pi_3$

$\Leftrightarrow u\pi_3 = u\pi_2 + v\pi_2 - v\pi_1 = v \cdot \frac{v}{u}\pi_1$

$\Leftrightarrow \pi_3 = (\frac{v}{u})^2\pi_1$

Thus, $\pi_j = \frac{v}{u}\pi_{j-1} = \frac{v}{u} \cdot (\frac{v}{u})^{j-2}\pi_1 = (\frac{v}{u})^{j-1}\pi_1$.

Let $\rho = \frac{v}{u}$ and hence $\pi_j = \rho^{j-1}\pi_1$. Since $\sum_{i=1}^{k} p_{ni} = 1$, where $p_{ni} = \pi_1\rho^{i-1}$, we get that

$$1 = \sum_{i=1}^{k} p_{ni} = \sum_{i=0}^{k-1} \pi_1\rho^i = \pi_1 \sum_{i=0}^{k-1} \rho^i = \pi_1 \frac{1-\rho^k}{1-\rho} \Rightarrow \pi_1 = \frac{1-\rho}{1-\rho^k} = c.$$

Finally, we get an expression for our transition matrix $\boldsymbol{P}$

$$\boldsymbol{P} = \begin{pmatrix} p_{n1} & p_{n2} & \cdots & p_{nk} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nk} \end{pmatrix} = \begin{pmatrix} c & c\rho & \cdots & c\rho^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ c & c\rho & \cdots & c\rho^{k-1} \end{pmatrix}.$$

Recall that r is the recombination frequency, that is, $(1-r)^t = \mathsf{P}$(no recombination in t generations) and thus $1-(1-r)^t = \mathsf{P}$(at least one recombination). The formula

$$\frac{\boldsymbol{M}^t}{(1+\lambda)^t} = (1-r)^t\boldsymbol{U}^t + (1-(1-r)^t)\boldsymbol{P}$$

admits the following interpretation. It can be viewed as a probability distribution f consisting of a mixture of two distributions, f_1 and f_2, according to

$$f(x) = \phi f_1(x) + (1-\phi) f_2(x).$$

Misusing the notation we can represent the expectation and the variance of $f(x)$ as

$$\mathsf{E}[f(x)] = \phi\mathsf{E}[f_1(x)] + (1-\phi)\mathsf{E}[f_2(x)]$$

and

$$\mathsf{Var}(f(x)) = \phi^2\mathsf{Var}(f_1(x)) + 2\phi(1-\phi)\mathsf{Cov}(f_1(x), f_2(x)) + (1-\phi)^2\mathsf{Var}(f_2(x)).$$

6.3 Homozygosity mapping

The reasoning behind this method is shown in Figure 6.2. Individuals with a recessive disease being homozygous for the disease gene should also be homozygous for markers in the vicinity of that gene. This provides a criteria for mapping the disease genes. Moreover comparing individuals homozygous for the disease with normal individuals allows us to find regions that differ considerably between the two populations and hence to find the disease locus. Due to homozygosity by descent, individuals carrying the disease, should also be homozygous in the adjacent area of the disease locus to much larger extent than expected (cf. (Kruglyak *et al.*, 1995) and (Lander and Botstein, 1987)).

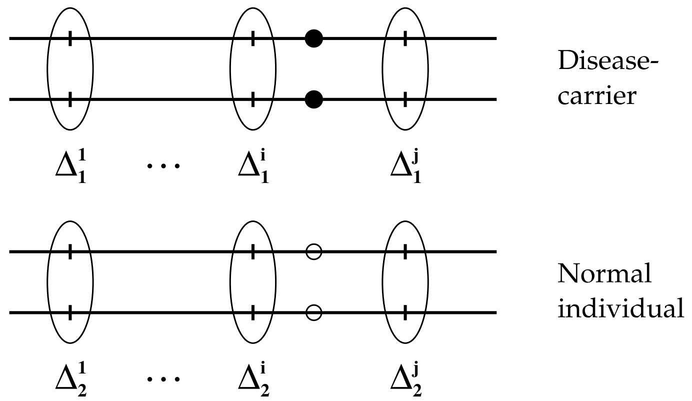


Figure 6.2: Difference in length between marker alleles

We have thus two ways of mapping a disease gene:

i) search for approximately homozygous markers bearing in mind that homozygocity can be lost due to mutations.

ii) to amplify the affect in i we focus on the difference in length of marker allele pairs (in Figure 6.2 called Δ_1^m for the m^{th} marker in the disease population and Δ_2^m for the same marker in the normal population), comparing this measure between individuals in the disease population and in the normal population. For markers tightly linked to the disease locus, the Δ_1^i distance will be very small due to homozygosity by descent but the Δ_2^i distance will be normal. Looking instead at markers situated at the other end of the chromosome, hence not linked to the disease locus, the distances Δ_1 and Δ_2 will both be normal.

To find a measure to help us search for the disease locus by comparing these two populations, we can form the distance measure $\Delta = |\Delta_1 - \Delta_2|$. This measure is calculated for a sample of individuals from each population by performing a gene-scan. When Δ is large in some sense for a marker, it will be close to the disease locus.

Consider an individual homozygous for the disease locus, and thus hopefully for the marker locus. Let $D_1 = I - J$, where I is the length of the marker allele on the first chromosome in the pair, and J is the length of the marker allele on the second. I and J are independent and have the distribution in the formula above $\left(\frac{\boldsymbol{M}^t}{(1+\lambda)^t}\right)$. Then D_1 can be considered as a distance between the two alleles. Since $\mathsf{E}[D_1] = 0$, D_1 cannot give us that much information so we need to find a better measure, for example

$$\Delta_1 = D_1^2 = (I - J)^2 = I^2 + J^2 - 2IJ.$$

Hereby, we get that

$$\mathsf{E}[\Delta_1] = 2\mathsf{E}[I^2] - 2(\mathsf{E}[I])^2 = 2(\mathsf{E}[I^2] - (\mathsf{E}[I])^2) = 2\mathsf{Var}(I).$$

Now we need the variance of I, that is the variance of the length of the marker allele. Recall the distribution of I, $f(t) = (1-r)^t\boldsymbol{U}^t + (1-(1-r)^t)\boldsymbol{P}$. That is, we have a mixture of $f_1(t) = \boldsymbol{U}^t$ and $f_2(t) = \boldsymbol{P}$. Let E_1 and Var_1 refer to the former and E_2 and Var_2 to the latter. We have that

$$\mathsf{E}_2[I] = \sum k\pi_k = \mu_2,$$

where π_k is the stationary distribution, and

$$\mathsf{Var}_2(I) = \sum (k - \mu_2)^2\pi_k = \sigma_2^2$$

is the variance.

Let $I = \sum_i X_i$, where X_i is the mutation in step i. We use the strict stepwise mutation model (SMM) with equal probabilities of adding and subtracting one, that is, $\mathsf{P}(X_i = 1) = \mathsf{P}(X_i = -1) = \theta/2$ and $\mathsf{P}(X_i = 0) = 1 - \theta$ (see Figure 2.2) and, of course, the X_i's are considered independent. We get

$$E_1[I] = 0$$

and

$$\mathsf{Var}_1(I) = \mathsf{E}_1[I^2] = \mathsf{Var}_1\Big(\sum_i X_i\Big) = \sum_{i=1}^{t} \mathsf{Var}_1(X_1) = t\theta.$$

Notice that this formula shows that, knowing $\mathsf{Var}_1(I)$ and θ, t can be estimated.

Now, the expectation of Δ_1 can be calculated:

$$\mathsf{E}[\Delta_1] = 2\,Var(I) = 2(1-r)^t t\theta + 2(1-(1-r)^t)\sigma_2^2 + 2(1-r)^t(1-(1-r)^t)\mu_2^2.$$

As mentioned earlier, the effect of homozygosity mapping can be amplified by looking for regions with large Δ-values between affected and normal individuals. For this reason, it is of interest to consider the case of a person showing no symptoms of the disease. We are interested in a rare disease, and thus very few people in the population should be carrying the disease allele in any of the two loci. We thereby assume that the people in this category carry some other allele than the disease allele.

Let $D_2 = I - J$ as before. I and J follow the distribution P, which is the distribution of alleles of the marker loci for 'normal' persons.

Again, $\mathsf{E}[D_2] = 0$, so we use $\Delta_2 = D_2^2$ instead. $\mathsf{E}[\Delta_2] = 2\mathsf{Var}(I) = 2\sigma_2^2$ is the expectation.

Under the assumption of no linkage,

$$\mathsf{E}[\Delta_1] - \mathsf{E}[\Delta_2] \approx 0.$$

No linkage: $r = 1/2 \Rightarrow \mathsf{E}[\Delta_1] - \mathsf{E}[\Delta_2]$

$$\begin{aligned}
&= 2(1/2)^t t\,\theta + 2(1-(1/2)^t)\,\sigma_2^2 + 2(1/2)^t(1-(1/2)^t)\,\mu_2^2 - 2\sigma_2^2 \\
&= 2(1/2)^t t\,\theta - 2(1/2)^t\,\sigma_2^2 + 2(1/2)^t(1-(1/2)^t)\,\mu_2^2 \\
&= 2(1/2)^t(t\,\theta - \sigma_2^2 + (1-(1/2)^t)\mu_2^2
\end{aligned}$$

What we hope to find, using a gene scan, is a marker linked to the disease allele.

In (Amos *et al.*, 1999), a simulation study is performed which confirms the usefulness of this method. The same reference contains an application to real data.

Bibliography

Amos, W. *et al.* (1999). Manuscript. (personal communication).

Bergman, A., Einbeigi, Z., Olofsson, U., Taib, Z., Wallgren, A., Karlsson, P., Wahlström, J., Nordling, M., and Martinsson, T. (2000). Manuscript.

Devlin, B., Risch, N., and Roeder, K. (1996). Disequilibrium mapping: Composite likelihood for pairwise disequilibrium. *Genomics*, **36**, 1–16.

Goldstein, D., Linares, A., Cavalli-Sforza, L., and Feldman, M. (1995). An evaluation of genetic distances for use with microsatellite loci. *Genetics*, **139**, 463–471.

Griffiths, A., Miller, J., Suzuki, D., Lewontin, R., and Gelbart, W. (1996). *An Introduction to Genetic Analysis, 6th Ed.* W.H. Freeman and Company.

Haldane, J. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet*, **8**, 299–309.

Harding, R., Boyce, A., Martinson, J., Fling, J., and Clegg, J. (1993). A computer simulation study of VNTR population genetics: Constrained recombination rules out the Infinite Alleles Model. *Genetics*, **135**, 911–922.

Jagers, P. (1975). *Branching Processes with Biological Applications.* Wiley.

Kaplan, N. and Weir, B. (1997). The use of linkage disequilibrium for estimating the recombination fraction between a marker and a disease gene. In P. Donnelly and S. Tavare', editors, *Progress in Population Genetics and Human Evolution. The IMA volumes in Mathematics and its Applications*, volume 87, pages 207–219. Springer Verlag.

Kruglyak, L. and Lander, E. (1998). Faster multipoint linkage analysis using fourier transforms. *J Comput Biology*, **5**, 1–7.

Kruglyak, L., Daly, M., and Lander, E. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *Am J Hum Genet*, **56**, 519–527.

Kruglyak, L., Daly, M., Reeve-Daly, M., and Lander, E. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *Am J Hum Genet*, **58**, 1347–1363.

Lander, E. and Botstein, D. (1987). Homozygosity mapping: A way to map human recessive traits with the DNA of inbred children. *Science*, **236**, 1567–1570.

Lange, K. and Fan, R. (1997). Branching process models for mutant genes in nonstationary populations. *Theor Pop Biol*, **51**, 118–133.

Liu, B. (1998). *Statistical Genomics Linkage, Mapping and QTL Analysis*. CRC Press LLC.

Moisio, A.-L., Sistonen, P., Weissenbach, J., de la Chapelle, A., and Peltomäki, P. (1996). Age and origin of two common MLH1 mutations predisposing to hereditary colon cancer. *Am J Hum Genet*, **59**, 1243–1251.

Pankratz, V. (1998). *Stochastic Models and Linkage Disequilibrium: Estimating the Recombination Coefficient*. PhD thesis, Rice University.

Risch, N., de Leon, D., Ozelius, L., Kramer, P., Almasy, L., Singer, B., Fahn, S., Breakefield, X., and Bressman, S. (1995). Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nature Genetics*, **9**, 152–159.

Valdes, A., Slatkin, M., and Freimer, N. (1993). Allele frequencies at microsatellite loci: The Stepwise Mutation Model revisited. *Genetics*, **133**, 737–749.

Weir, B. (1996). *Genetic Data Analysis II.* Sinauer Associates, Inc.

APPENDIX A

Simulation, one marker

Consider the case with one marker locus linked to the disease locus, chosen at an appropriate distance.

Parameters:

k $= \#$ alleles at marker locus

t $= \#$ generations

r is the recombination frequency

$1 + \lambda$ is the expected number of offspring of any given person

Assumptions:

- We keep track of the number of individuals in each generation carrying one of the k marker alleles.
- For simplicity, the marker allele quantities are assumed to be uniformly distributed in the normal population.
- The $(t+1)^{th}$ generation is generated according to the formula

$$Z_{t+1}(i) \sim \mathsf{Poi}\left((1+\lambda)\left[(1-r)\, Z_i(t) + r p_{ni}\, Z_{tot}(t)\right]\right),$$

where

$$Z_{tot}(t) = \sum_{j=1}^{k} Z_t(j) = \# \text{ of individuals in generation } t.$$

- The age of the disease is estimated using the formula

$$\hat{t} = \frac{\ln(\frac{\hat{p}_{d1} - p_{n1}}{1 - p_{n1}})}{\ln(1 - r)}$$

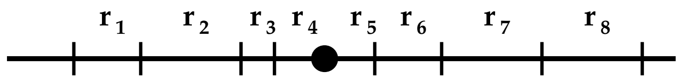
  where $\hat{p}_{d1}$ is the observed allele frequency in the disease population.

- We are sampling individuals from the last generation.

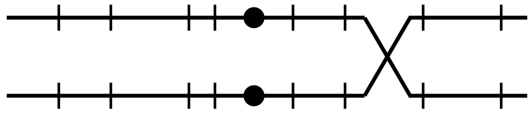
APPENDIX B

Simulation, eight markers

- Each individual is represented by haplotypes. We do not need information of the sister chromosome not carrying the disease. The rest of the genome is also irrelevant in our case.
- Since we are only interested in the carriers in the last generation, we only include carriers in the simulation. Each individual begets one or two offspring carrying the disease allele.
- We track eight markers surrounding the disease locus, four on each side of the disease locus, preferably chosen at somewhat symmetrical distances surrounding the disease locus.



- Recombination events are allowed.



The implementation is as follows: for each interval, we compare a random number with the recombination probability and if it results in a recombination event, the marker alleles in the area from the interval of interest and the end, seen in the opposite direction of the disease locus, is exchanged. The new marker alleles are chosen from the stable population of normal alleles, here assumed to be uniformly distributed.

- The map distances are translated to recombination frequencies using the Haldane map function (see (Haldane, 1919)):

$$r_{AB} = \frac{1}{2}(1 - e^{-2d_{AB}/100})$$

Note that d_{AB} is the distance between the two loci A and B, measured in cM.

- We calculate eight estimates of the number of generations, t, using the formula: $\hat{t} = \frac{\ln(\frac{\hat{p}_{d1} - p_{n1}}{1 - p_{n1}})}{\ln(1-r)}$. As a final estimate, we use the observed mean value of these.

- Only the last generation is saved to save computer memory, so we cannot backtrack our simulation.

The computer language used to implement this method is JAVA. As long as either the number of generations or the probability of having two children is kept low, it runs painlessly.

The simulation parameters are:

p = P(individual begets one offspring)

Thus, $1 - p$ = P(individual begets two offspring)

g = # of generations

c is a vector describing the number of alleles of each of the eight markers.

r is the vector containing the eight distances between pairs of markers.