# Testing the performance of TDT and Linkage methods using simulated virtual populations with polygenic diseases

Bo Peng[1,*] and Marek Kimmel[1]

[1]Department of Statistics, Rice University, 6100 Main St. MS138, Houston, TX, 77005

**Running Title:** Testing gene mapping methods using simulated populations

* To whom correspondence should be addressed

Department of Statistics
Rice University
6100 Main St. MS138
Houston, TX 77005
Phone: (713) 348-6045
Fax: (713) 348-5476
Email: bpeng@rice.edu

## Abstract

The evolution of complex polygenic diseases is modeled using forward-time individual-based simulations. In a typical simulation, individuals have 20 chromosomes with 400 microsatellite markers and 5 binary disease susceptibility loci (DSL). The simulation starts with a small population and proceeds through four stages of evolution processes including burn-in, disease introduction, split and growth, and mixing. Genetic drift, mutation, selection, migration and demographic changes shape the genotype of the resulting large multi-generation populations in which samples can be drawn using different ascertainment methods. Transmission Disequilibrium Test (TDT) and Linkage (LOD) method are applied to these samples and their performance are compared. For our simulated data, the TDT method turns out to be more sensitive than the LOD method. However, both methods seem either to miss DSL or to indicate wide chromosomal regions. The novelty of our study is that we provide a flexible way to generate datasets of polygenic diseases with different evolutionary backgrounds, and we are able to directly test not only different gene mapping methods, but also different experimental designs and ascertainment methods.

# 1 Introduction

In this study, we combine evolutionary simulation of genetic disease with statistical analyses aimed at discovering the genes causing this disease. Our purpose is to evaluate the effectiveness of these analyses.

Simulated datasets with known disease susceptibility genes and affectedness or trait values provide tools to test the performance of gene mapping methods. For example, genetic epidemiologists use computer-simulated datasets distributed by the Genetic Analysis Workshop (GAW, `http://www.gaworkshop.org`) to evaluate and compare statistical genetic methods.

Two main approaches exist to simulate such datasets, namely forward-time and backward-time (coalescent) simulations. (Kingman 1982) Backward-time simulations are sample based. The basic idea is that given a sample (of unknown genotype), we find the common ancestors of individuals and coalesce them according to a stochastic process characterized by evolutionary properties like mutation, recombination and migration. After the most recent common ancestor of all individuals is found, the process is run forward in time and it assigns genetic information to individuals in the coalescent tree. This method is fast because it only concerns individuals related to the final sample. It is also very flexible in that it can model many migration and mutation models. It becomes more complicated when arbitrary demographic models and recombination are involved. Even more, due to its theoretical basis as a neutral process, coalescent process can not handle selection well, despite some recent advances (e.g., Fearnhead 2003). Therefore, it is very difficult, if not impossible, to simulate the evolution of polygenic diseases with complex selection and penetrance models using this approach.

Forward-time simulations are simpler as an idea, and in implementation. Since evolution proceeds forward in time, all that is needed is to mimic this process as close as possible. A forward-time simulation usually starts from an initial population, and evolves it generation by generation, subject to arbitrary number of genetic or demographic changes. Samples are drawn from the last several generations. There is no limit on the type of disease, selection and penetrance models this approach can handle. The problem with this approach is its inefficiency. For example, to simulate the evolution of a rare simple disease, we have to keep track of all genotype in a very large population. From a sample point of view, most of the computing time is wasted on unaffected individuals. Because of their limited use in mimicking real populations, forward-time simulations have been used primarily for teaching purposes. However, due to the exponential growth of the power of personal computers and the availability of highly flexible forward-time simulation programs (Balloux 2001, Peng and Kimmel 2005), it is now feasible to simulate large populations with complex polygenic diseases.

In this paper, we will simulate the evolutions of polygenic diseases using a forward-time population genetic simulation environment simuPOP (Peng and Kimmel 2005). Although all parameters are customizable, a typical individual has 20 chromosomes with 400 microsatellite markers and 5 binary disease susceptibility loci (DSL). The disease is defined by the layout of the DSL and the single and multi-locus penetrance functions. During evolution, a small population runs through a long burn-in process and then expands exponentially to its current size. The disease is introduced at the beginning of the expansion and is nurtured until it achieves a common status with disease allele frequency greater than 5%. Mutation, selection, migration and demographic changes are applied to the population at appropriate generations. The results of the simulations are large multi-generation populations in which samples can be drawn using different ascertainment methods. We draw affected sib-pair samples from the population and apply Transmission Disequilibrium Test (TDT) and Linkage method to detect the DSL. The performance of these two methods is compared.

Our approach allows us to compare experimental designs and ascertainment methods. For example, we may draw both family-based sibpair samples and population-based case control samples and try to answer the question like 'what would be required sample size of a population based association study to achieve the power of family based linkage studies'? These will be the subject of further study.

In addition to pure simulation, we derive theoretical expression for characteristics such as allele frequencies in case and control groups and sibling recurrence risk, some of which being the multi-locus extension of the two-locus expressions derived by Risch (1990). This provides a desirable validation of simulations.

## 2 Simulation scenario

Using a forward-time population genetics simulation environment simuPOP (Peng and Kimmel 2005), we write a Python script that simulates the evolution of a polygenic disease using a flexible four-stage scenario. We note that the script can handle even more complicated scenarios.

1. Create an initial population of $N_0$ individuals. Each individual has 20 chromosomes each with 20 equal-spaced microsatellite markers. Five disease susceptibility loci (DSL) are placed on five different chromosomes, half-way between their adjacent markers.

2. Initialize each individual with two out of five initial haplotypes. Initialize DSL with wild type alleles. This leads to an initial population with complete linkage disequilibrium between markers.

3. Burn-in the population for $G_0$ generations, subject to symmetric stepwise mutation and recombination. Mutation and recombination will act on the population throughout the simulations.

4. In the next $G_1$ generations, disease alleles are introduced to the population by point-mutating disease loci of different individuals. An allele is re-introduced if it is lost because of genetic drift. Given a destined allele frequency, a disease allele is placed under strong (positive or negative) selection pressure until it stabilises around its destined frequency.

5. Split the population into $k$ subpopulations and expand it exponentially during the next

$G_2$ generations. Migration is not allowed so subpopulations evolve independently of each other. This allows population structure to build up.

6. Migration is allowed in the next $G_2$ generations. Depending on the migration rate and length of $G_2$, population structure is attenuated as a result of mixing.

7. In the last three generations, each mating event produces two offspring (instead of one as in previous generations). Pedigree information is recorded. The last three generations are saved as the final virtual population.

8. Draw affected sibpair samples from the large population and apply TDT and Linkage methods. DSL are removed from the samples.

The following subsections describe various aspect of this process in detail.

## Markers and disease susceptibility loci

Each individual has 400 microsatellite markers and 5 binary disease susceptibility loci. The microsatellite markers are spread evenly on 20 chromosomes. The microsatellite markers are initialized with alleles labeled 50 and then mutate following a symmetric stepwise mutation model. Although this mutation process has two absorbing boundaries 1 and 99, none of them is reached by any allele during our simulations.

Although the number of DSL can be arbitrary, we will use five DSL for all our simulations. These DSL are unlinked because they are put on different chromosomes. These DSL are placed half-way between their adjacent markers.

There are one wild ($N$) and one disease susceptibility ($S$) allele at each DSL. We do not model mutations between these alleles. Instead, we assume that all disease alleles are derived from one common ancestral allele introduced at the disease introduction stage.

Marker locations are not explicitly specified and are roughly determined by recombination rates between adjacent markers. For example, if recombination rate between adjacent markers is 0.0005, the map distance between these two markers is 0.05 centiMorgan and the length of chromosome is roughly one centiMorgan (using Haldane's mapping function $-\frac{1}{2}\ln(1-2\theta)$, where $\theta = 19 \times 0.0005$ ).

## Burn-in stage

Chromosomes of each individual are randomly assigned two out of five haplotypes ($nnnn\cdots$), $n = 50, 51, 52, 53, 54$ so that linkage disequilibrium between markers is complete ($D' = 1$). DSL are all initialized with the wild type allele. For the next $G_0$ generations, microsatellite markers are mutated under a symmetric stepwise mutation model. Recombinations between adjacent loci (including DSL) happen at a constant recombination rate.

The goal of this stage is to make LD between adjacent markers closer to the prevailing in the human population. For example, when $r = 0.0001$, the map distance between these two markers is around 0.01cM, roughly 10k base pair. After burning in for $G_0 = 400$ generations, $D'$ between adjacent markers will decline from 1 to a level comparable to that of human population, which is roughly 0.7 according to Dunning et al. (2000). This is the Linkage Disequilibrium scenario of Abdallah et al. 2003. Although it is possible to start with totally random alleles and let LD build up with time (the Linkage Equilibrium scenario). We have not investigated this possibility yet.

## Introduction of disease

Five mutants are introduced to the population to five different individuals at the beginning of the disease introduction stage. A mutant is re-introduced if it is lost because of genetic drift. To let disease alleles reach designated range of allele frequency (e.g. $5\% \sim 10\%$), strong advantageous or purifying selection is used to control the disease allele frequency at each DSL. This is an extension to the scenario adopted by Abdallah et al. (2003). Although other methods such as a bottleneck may be used to achieve the same high allele frequency, our method seems more convenient.

## Fitness models

We assume that the fitness of an individual is solely determined by his/her genotype, regardless of affectedness status or trait value. This allows us to disregard the penetrance information during evolution and only assign penetrance at the last several generations.

An additive fitness model is used at each DSL. That is, fitness at a DSL with genotype $NN$, $NS$ or $SS$ is 1, $1-s/2$ or $1-s$ respectively, where $s$ is the selection coefficient at this DSL, and it may vary from locus to locus. The overall fitness value is obtained using a multiplicative model (Pritchard 2001, Risch 1990). For example, if all DSL have the same selection coefficient $s$, an individual with genotype at five disease loci $NS, NN, NN, SS, NN$ will have fitness value $(1-s/2)\times 1\times 1\times(1-s)\times 1 \approx 1-3s/2$, compared to 1 for non-carriers.

## Population expansion and splitting

Human population has a complicated expansion and migration history. Most notably, human population grew from a quarter billion to 6.4 billion in a thousand years, roughly following an exponential increase model. Our increases the initial population from $N_0 = 10^4$ to $N_1 = 0.2\times10^6$ in 400 generations (8000 years if we assume 20 years per generation.) Although $N_1$ is nowhere close to our current census population size of $6\times10^9$, it seems enough to roughly mimic isolated regional populations such as that of Finland. Another reason to justify this population size is that human population does not follow random mating so the effective population size is much smaller than the census size.

In an attempt to mimic human migration, we split the population into ten subpopulations at the beginning of population expansion. The subpopulations first evolve separately and then start to mix with migrations following an island model. The length of the no-migration stage, mixing stage and intensity of migration determines the level of population structure at the final population. Note that disease prevalence varies from subpopulation to subpopulation and the disease may become extinct in some subpopulations. Since there is no mutation at DSL, the only way to re-introduce this allele into the subpopulation is through migration.

## Sample from the final population

True random mating is used almost all the time to ensure maximum effective population size. In this mating scenario, parents are chosen randomly and will produce one offspring at each mating event. Although a parent may be chosen multiple times and produce more than one offspring, it is very unlikely to observe full siblings since the probability of two offspring having the same parents is $1/N^2$ where $N$ is the population (or subpopulation) size. Because we use a family based ascertainment method (affected sibpairs), the number of offspring per mating is changed to 2 at the last two generations. The resulting population has $0.2 \times 10^6$ individuals in 10 equal-sized subpopulations. It has genotype and pedigree information of the last two generations so affected sibpairs and their parents can be sampled.

Affectedness of each individual is assigned according to a heterogeneity model (Risch 1990) superimposed on an additive model at each DSL. That is to say, the penetrance at a DSL with genotype $NN$, $NS$ or $SS$ is 0, $\delta$ or $2\delta$ respectively and the overall penetrance is determined using formula $1 - \prod_{i=1}^{5} (1 - d_i)$ where $d_i$ is the penetrance value at locus $i$. For example, if $\delta_i = 0.25$ for $i = 1, ..., 5$, the probability of being affected for an individual with genotype $NS$,$NN$,$NN$,$SS$,$NN$ is $1 - .75 \times 1 \times 1 \times .5 \times 1 = 0.625$. Note that if only one DSL has disease allele(s), the overall penetrance equals the penetrance at this DSL.

## TDT and Linkage (LOD score) method

Affected sibpairs and their parents are sampled. Genotypes at all DSL are removed from the samples so we are left with datasets of 400 microsatellite markers. Samples are drawn from the whole population regardless of the population structure. Since disease prevalence varies among subpopulations, the sibpairs from less diseased subpopulations may be underrepresented. This can be corrected by sampling equal number of families from each subpopulation but we will not discuss this possibility here.

These samples are saved in the Linkage format chromosome by chromosome so that they can be analyzed by TDT and Linkage methods. We use GeneHunter to perform these analyses. $p-$values at all markers are recorded.

## 3 Theoretical analysis

Because we know all the details about the evolutionary process and the penetrance models, it is possible to estimate some population properties theoretically. These estimates will be compared to simulated populations as a way to validate our simulations. It is also possible to use these properties to estimate the power of a particular gene mapping method.

We assume that

- There are five DSL. Although all the results will be obtained for the five DSL case, most of them can be easily extended to an arbitrary number of DSL.
- The allele frequency of the disease allele at locus $i$ is $f_i$, $i = 1, 2, 3, 4, 5$, which also will be called the size of a DSL. Since the DSL are binary, the frequencies of wild-type alleles are $1 - f_i$, $i = 1, 2, 3, 4, 5$.
- The genotype at a DSL is $g_i = g_i^1 g_i^2$, $i = 1, 2, 3, 4, 5$. $g_i^j$ can be set equal to $N$ or $S$. For numerical simplicity, we assume $N = 1$, $S = 2$.
- The frequency of a particular genotype is denoted by $F(g) = F(g_1g_2g_3g_4g_5)$.
- The penetrance of an individual is

$$P(g) = P(g_1g_2g_3g_4g_5) = 1 - \prod_{i=1}^{5}[1 - p_i(g_i)]$$

where

$$p_i(g_i) = \begin{cases} 0 & g_i = NN \\ \delta_i/2 & g_i = SN \text{ or } NS \\ \delta_i & g_i = SS \end{cases}$$

is the penetrance at DSL $i$.

In the following sections, we will estimate some population statistics such as disease prevalence, or sibling recurrence risk ratio in the final population. Note that some of the results are motivated by and are extensions of similar results for two DSL in Risch (1990).

## LD between adjacent markers

If we do not consider the impact of mutation and genetic drift, linkage disequilibrium ($D$ ) between two adjacent binary markers is equal to $D_t = D_0 (1-r)^t$. If we start from complete LD ($D_0 = \frac{1}{4}$), $D_{800} = 0.168$ when $r = 0.0005$.

Our case is more complicated since microsatellite markers are involved. In this case, LD between two adjacent markers is equal to the average of all $D$ values estimated using one allele at each locus, weighted by allele frequencies. More specifically,

$$D = \sum_i \sum_j f_i f_j |D_{ij}| = \sum_i \sum_j f_i f_j |F_{ij} - f_i f_j|$$

where $f_i$, $f_j$ are allele frequency of allele $i$ and $j$, $F_{ij}$ is the genotype frequency of genotype $ij$.

## Genotype frequency

At DSL $i$, the probability of having one disease allele is $2f_i(1-f_i)$, having two disease alleles is $f_i^2$. The genotype frequency $F$, assuming no interaction between DSL and homologous genotypes, is equal to

$$\begin{aligned} F(g) &= \prod_{i=1}^{5} F_i\left(g_i\right) \\ &= \prod_{i=1}^{5} f_i^{g_i^1-1}\left(1-f_i\right)^{2-g_i^1} f_i^{g_i 2-1}\left(1-f_i\right)^{2-g_1^2} \end{aligned}$$

where $g_i^1 g_i^2$ is the genotype at DSL $i$. $F_i\left(g_i\right)$ is the genotype frequency at DSL $i$, $g_i^1, g_i^2$ =1 (wild type) or 2 (disease allele).

## Overall disease prevalence

Let $G$ be all possible genotypes at the DSL ($4^5 = 1024$ distinct types). The population prevalence $K$ is

$$\begin{aligned} K &= \sum_{g \in G} F\left(g\right) P\left(g\right) \\ &= \sum_{g} \left(\prod_{i=1}^{5} f_i\left(g_i\right)\right)\left(1-\prod_{i=1}^{5}\left(1-p_i\left(g_i\right)\right)\right) \end{aligned} \qquad (1)$$

As a special case, if $f_i = f$ and $\delta_i = \delta$ for all $i = 1, ..., 5$,

$$K = 10f\delta - 40f^2\delta^2 + 80f^3\delta^3 - 80f^4\delta^4 + 32f^5\delta^5$$

When $f = 0.05$, $\delta = 0.25$,

$$K = 0.119$$

That is to say, if disease allele frequencies are 0.05 and penetrance is 0, 0.25 and 0.5 for genotype $NN$, $NS$ and $SS$ respectively, the population incidence rate should be 0.119. Table **??** displays $K$ under different $f_i$ and $\delta_i$ values.

## Prevalence summands

Equation 1 uses all five DSL. However, it is useful to have a look at prevalence caused by one DSL. Let

$$K_i \;=\; \sum_{g\in G} F\left(g\right) p_i\left(g\right)$$

and denote $G_i = \{g_i^1 g_i^2 \mid g_i^1, g_i^2 = 0 \text{ or } 1\}$. Without loss of generality, let $i = 1$. Because $\sum_{g_i\in G_i} F_i\left(g\right) = 1$, we have

$$\begin{aligned}
K_1 &= \sum_{g_1\in G_1}\sum_{g_2\in G_2}\sum_{g_3\in G_3}\sum_{g_4\in G_4}\sum_{g_5\in G_5} F\left(g\right) p_1\left(g\right) \\
&= \sum_{g_2\in G_2}\sum_{g_3\in G_3}\sum_{g_4\in G_4}\sum_{g_5\in G_5}\left[\sum_{g_1\in G_1} p_1\left(g\right) f_1\left(g\right)\right]\prod_{j=2}^{4} f_j\left(g\right) \\
&= \left[\sum_{g_1\in G_1} p_1\left(g\right) f_1\left(g\right)\right]\left(\sum_{g_2\in G_2} f_2\left(g\right)\sum_{g_3\in G_3} f_3\left(g\right)\sum_{g_4\in G_4} f_4\left(g\right)\sum_{g_5\in G_5} f_5\left(g\right)\right) \\
&= \sum_{g_1\in G_i} p_1\left(g\right) f_1\left(g\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
K &= \sum_{g\in G} F\left(g\right) P\left(g\right) = \sum_{g\in G} F\left(g\right)\left(1 - \prod_{j=1}^{5}\left(1 - p_j\left(g\right)\right)\right) \\
&= \sum_{g\in G} F\left(g\right) - \sum_{g\in G}\prod_{i=1}^{5} f_i\left(g\right)\left(1 - p_i\left(g\right)\right) \\
&= 1 - \sum_{g_1\in G_1}\sum_{g_2\in G_2}\sum_{g_3\in G_3}\sum_{g_4\in G_4}\sum_{g_5\in G_5}\prod_{i=1}^{5}\left(f_i\left(g\right)\left(1 - p_i\left(g\right)\right)\right) \\
&= 1 - \prod_{i=1}^{5}\left(\sum_{g_i\in G_i} f_i\left(g\right)\left(1 - p_i\left(g\right)\right)\right) = 1 - \prod_{i=1}^{5}\left(1 - K_i\right)
\end{aligned}$$

Under the additivity assumption, $K_i$ can be simply expressed as

$$K_i \;=\; 2\delta_i f_i$$

and (1) can be written as

$$K = 1 - \prod_{i=1}^{5} (1 - K_i) = 1 - \prod_{i=1}^{5} (1 - 2\delta_i f_i)$$

This formula is an extension of equation (15) in Risch (1990).

## Genotype frequency in case and control groups

If we sample cases and controls from the population, we should expect higher frequency of disease alleles in the case group and lower frequency in the control group. The magnitude of this difference is important in mapping the DSL using association based studies. For any genotype $g$,

$$\begin{aligned} Pr\left(g \mid \text{affected}\right) &= \frac{Pr\left(\text{affected}, g\right)}{Pr\left(\text{affected}\right)} = \frac{Pr\left(\text{affected} \mid g\right) Pr\left(g\right)}{Pr\left(\text{affected}\right)} \\ &= \frac{P\left(g\right) F\left(g\right)}{K} \\ Pr\left(g \mid \text{unaffected}\right) &= \frac{\left(1 - P\left(g\right)\right) F\left(g\right)}{1 - K} \end{aligned}$$

For example, when $f_i = f = 0.05$ , $\delta_i = \delta = 0.25$, $i = 1, ..., 5$

$$\begin{aligned} Pr\left([NN, NN, NN, NN, NS] \mid \text{affected}\right) &= 0.0663 \\ Pr\left([NN, NN, NN, NN, NS] \mid \text{unaffected}\right) &= 0.0268 \end{aligned}$$

By symmetry and independence, we know that the proportion of affected individuals having one disease allele is $6.63 \times 10 = 66\%$. Similarly, the proportion of affected individuals having two disease alleles is 28%, having three disease alleles is 5% and the rest of cases 1%. Among those individuals having two disease alleles, homozygotes account for 12% of the cases.

## Genotypic distribution at a single DSL

Given $Pr\left(g \mid \text{affected}\right)$, we can easily calculate the genotype frequencies at one DSL. For example,

$$\begin{aligned}
Pr\left(g_1 \mid \text{affected}\right) &= \sum_{g_2g_3g_4g_5} Pr\left(g_1g_2g_3g_4g_5 \mid \text{affected}\right) \\
&= \sum_{g_2g_3g_4g_5} \frac{P\left(g\right)F\left(g\right)}{K}
\end{aligned}$$

This formula can be expanded to

$$\begin{aligned}
Pr\left(\left(1,1\right) \mid \text{affected}\right) &= \left(1-f_1\right)^2 \frac{\sum_{i=2}^{5}\left(K_i\prod_{j=i+1}^{5}\left(1-K_j\right)\right)}{1-\prod_{i=1}^{5}\left(1-K_i\right)} \\
Pr\left(\left(1,2\right) \text{ or } \left(2,1\right) \mid \text{affected}\right) &= 2\left(1-f_1\right)f_1 \frac{\left[\delta_1+\sum_{1i=2}^{5}\left(\left(1-\delta_1\right)K_i\prod_{j=i+1}^{5}\left(1-K_j\right)\right)\right]}{1-\prod_{i=1}^{5}\left(1-K_i\right)} \\
Pr\left(\left(2,2\right) \mid \text{affected}\right) &= f_1^2\left(1-2\delta_1\right) \frac{\left[\sum_{i=2}^{5}\left(K_i\prod_{j=i+1}^{5}\left(1-K_j\right)\right)\right]}{1-\prod_{i=1}^{5}\left(1-K_i\right)}
\end{aligned}$$

Using the same example when $\delta_i = \delta = 0.25$ and $f_i = f = 0.05$, we have

$$\begin{aligned}
Pr\left(NN \mid \text{affected}\right) &= 0.731 \\
Pr\left(SN \text{ or } NS \mid \text{affected}\right) &= 0.257 \\
Pr\left(SS \mid \text{affected}\right) &= 0.012
\end{aligned}$$

Furthermore, with even more algebra, we can obtain that

$$\begin{aligned}
Pr\left(NN \mid \text{unaffected}\right) &= 0.926 \\
Pr\left(SN \text{ or } NS \mid \text{unaffected}\right) &= 0.073 \\
Pr\left(SS \mid \text{unaffected}\right) &= 0.001
\end{aligned}$$

When looking at a particular DSL, unlike in the case of rare Mendelian disease where most affected individual have affected alleles, only around a quarter of all affected individuals have at least one disease allele. The genotype of other three quarter of the affected individuals will contribute noise at this DSL and make mapping of this DSL difficult.

Further analysis shows that $Pr\left(\text{at least one diease allele} \mid \text{affected}\right)$ increases with increasing disease allele frequency, decreases with increasing penetrance coefficient. The impact of the size of DSL is much larger than that of penetrance coefficient though.

## Population versus Sample Allele Frequencies at DSL

In this simulation study, we have both population allele frequency (from the overall population) and sample allele frequency (from affected sibpair samples) available. It is of interest to see how large is the difference between these frequencies. In other words, what is$f_i' = P_r\left(\text{disease allele at DSL } i \mid \text{affected}\right)$?

Since we already know $P_r\left(g_1 \mid \text{affected}\right)$, we can estimate $f_i'$ by

$$f_i' \quad = \quad \frac{1}{2} Pr\left(SN \text{ or } NS \mid \text{affected}\right) + Pr\left(SS \mid \text{affected}\right)$$

The formula is too lengthy to list here. When $f_i = f = 0.05$, $\delta_i = \delta = 0.25$, $f_i' = 0.26/2 + 0.01 = 0.14$, larger than $f_i = 0.05$. More examples are listed in table **??** .

## Sibling recurrence risk

Let $X_P$,$X_S$ be the affectedness status of a proband and his/her sibling respectively (1 for affected and 0 for unaffected). Let $K_S = E\left(X_S \mid X_P = 1\right) = P_r\left(X_S = 1 \mid X_P = 1\right)$ be the recurrence risk for a sibling of an affected proband and $\lambda_S = K_S/K$ be the risk ratio for a sibling of an affected individual compared with population prevalence. Therefore,

$$K \times K_S = E\left(X_P\right) E\left(X_S \mid X_P = 1\right) = E\left(X_S X_P\right)$$

where $K = E(X_P)$ is the population prevalence of the disease. Following the same type of argument as that in Risch (1990), we can prove that (see the Appendix)

$$K \times K_S = 1 - 2\prod_{i=1}^{5}(1 - K_i) + \prod_{i=1}^{5}(1 - 2K_i + K_i K_{Si}) \tag{2}$$

where $K_i = \sum_{g \in G} F(g) P_i(g)$ and $K_{Si} = E(X_{Si} \mid X_{Pi} = 1)$ is $K_S$ restricted to DSL $i$ (using penetrance at DSL $i$ as individual penetrance).

Numerical values of sibling recurrence risk can be found in table **??**. $K_S$ for some $p_i$, $f_i$ configuration is listed in table **??**.

# 4 Results

A number of simulations were run using different combinations of parameters. Table 1 summarizes the expected and observed population statistics of some of the simulations. Note that disease allele frequencies can not be controlled exactly: They are adjusted to be close to anticipated values during disease introduction stage and then drift randomly as a result of genetic drift.

## 4.1 Population

Let us focus on two populations. Both populations are simulated using the following parameters: $N_0 = 10^4$, $N_1 = 2 \times 10^5$, $G_0 = 400$ (burn-in), $G_1 = 50$ (disease introduction), $G_2 = 300$ (without migration), $G_4 = 50$ (mixing). Mutation and migration rates are $10^{-4}$ and $10^{-3}$ respectively. The only difference between these two populations is the recombination rates, which are $5 \times 10^{-4}$ and $10^{-4}$ respectively. There are five disease susceptibility loci located after the 19th, 6th, 6th, 8th, 10th marker on chromosomes 1, 5, 7, 11, 16 respectively. The disease allele frequencies at these DSL before population expansion are controlled between 4.5% and 5.5%. The disease alleles are neutral in the sense that they are not subject to purifying selection after the disease introduction stage. The penetrance function at each DSL is additive with parameter 0.5, and the overall penetrance follows a heterogeneity model.

The resulting first population has allele frequencies 6.3%, 7.5%, 4.0%, 3.3%, 4.6% at the five DSL, which scatter around our expected allele frequency 5% as the result of modest level of genetic drift. There are 24359 affected individuals in the final population which form 2736 sibpairs. The sibling recurrence risk ( the probability that a sibling of an affected proband is also affected) is equal to $K_S = 18.2\%$, compared to population prevalence $K = 12.2\%$. This results in a sibling recurrence risk ratio $\lambda = \frac{K_S}{K} = 1.49$. Among all disease individuals, at any given DSL, about 73% have no disease allele and 25% have one disease allele. As depicted in Table 1, these observed population statistics match the theoretical expectations well.

Allele frequencies vary from subpopulation to subpopulation. Figure 2 plots the genotypes of all affected individuals in the final population. As we see from the figure, there are around 500 affected individuals in subpopulation 1 but around 3500 in subpopulation 8. Disease allele frequencies also vary: Most disease individuals have disease susceptibility allele 2 or 4 in subpopulation 1, but more than 50% of diseased individuals have disease susceptibility allele 1 in subpopulation 7.

With recombination rate $5 \times 10^{-4}$ or $1 \times 10^{-4}$, the linkage disequilibria between adjacent markers are still strong. From Figure 1, we can see that $D'$ between a DSL and its closest marker is close to 0.9 in the first population. LD between adjacent markers on a chromosome without DSL is around 0.6. These numbers increase to 0.95 and 0.84 for the second population because of a lower recombination rate.

## 4.2 TDT method

A sample of size 1000 (250 affected sibpairs with their parents) are drawn from each of the final populations, regardless of subpopulation structure. Single-locus TDT test is applied to both samples.

We use the basic transmission disequilibrium test provided by GeneHunter. This single-locus TDT method scans through all markers and looks for transmission distortion of parental alleles. A $p$-value is given at each marker, which is plotted in Figure 3 in the form of $-\log p$-value. Decimal logarithms are used.

We use Bonferroni method to adjust these $p$-values. At the significance level 0.05, we look for markers with $p$-values less than 0.05/400 (or equivalently with $-\log p$-value greater than $-\log \frac{0.05}{400} = 3.90$). This is the horizontal line in Figure 3 and all the following figures. Among five DSL, only DSL 1 and 2 are statistically significant. Note that disease allele frequency at DSL 2 is the highest among all DSL.

With a smaller recombination rate, LD between DSL and their adjacent markers, and between markers on the same chromosome are expected to be stronger in the second population. This results in lower $p$-values at markers around each DSL. Because the recombination rate between adjacent markers reflects map distances, population 2 reflects the result of a denser mapping than population 1: A chromosome of population 2 is roughly one fifth of that in the first population. The TDT method picks out three out of five DSL (2,4,5). Usually, several markers are significantly linked to a DSL at each picked-out DSL due to the strong linkage between these markers. Almost all markers are significant on chromosome 5.

From each of the two populations, we draw 10 samples and apply the TDT method. In population 1, disease locus one through five are picked out 8, 7, 1, 0, 0 times respectively. In population 2, disease loci are picked out 2, 10, 2, 3, 2 times. These numbers are closely related to disease allele frequencies which are 6.3%, 7.5%, 4.0%, 3.3%, 4.6% in the first population and are 3.3%, 5.4%, 3.1%, 5.0%, 2.8% in the second population.

## 4.3 LINKAGE (LOD Score) Method

The Linkage/LOD method is also applied to the samples, using GeneHunter. The underlying disease model behind LOD method is obviously not compatible with ours, but we would like to see if it can detect some of the DSL by treating it as a black-box method.

We try to feed the LOD method with correct information whenever possible. For example, we provide the true recombination rate and use allele frequency estimated from the whole virtual population as the allele frequencies for markers. Although our disease follows a heterogeneity multi-locus penetrance model, we provide the program with the true single-locus penetrance values (0, 0.25, 0.5). These penetrance values are mostly true when a DSL has one or two disease susceptibility alleles because few people have disease susceptibility alleles at more than one DSL. However, the zero penetrance when there is no disease susceptibility allele at a DSL is untrue for the simulated datasets, because the disease is caused by more than one DSL and disease susceptibility alleles at other DSL can also cause the disease.

$p$-values at each marker are collected and plotted in the format of $-\log p$-value (see Figure 4). Bonferroni multiple-testing correction is applied and the cutoff value is marked. The results show that at the 0.05/400 significance level, no DSL is detected in the ten samples we draw from each population.

# 5 Discussion

This paper employs a flexible way (Peng and Kimmel 2005) to generate large virtual populations from which various types of samples can be drawn and analyzed. We use evolutionary modeling in forward time to recreate non-observable characteristics of complex genetic disease in human population. This approach has been rarely used. We then explore two statistical genetics methods of disease gene mapping, using the simulated populations. Although many kinds of ascertainment and gene mapping methods can be tested, we only explore affected sibpair samples and the basic TDT and LOD methods.

This paper are not a comprehensive or exhaustive study of the methodology used for gene detection in complex diseases. We introduce the methodology and generate two examples of a complex disease, involving interaction of a number of genetic and demographic forces. The comparison between TDT and LOD is tentative, much more study is needed to elucidate the relative effectiveness of these methods in the context of complex diseases.

Some of the reasons for complex diseases being so difficult to handle are elucidated by numerical examples provided using the theoretical formulae derived in the paper: Naturally, when several DSL are involved, each of them is presented in the affected individual with a rather low probability.

It is considered that the TDT method is less sensitive to population structure. This might be the reason for its relatively better performance in our study.

In conclusion, we attempted to create a comprehensive and realistic model of evolution of complex genetic disease. Some of the details of our approach, such as the burn-in stage, may not be directly comparable to what is really happening in human population. We will further study this question.

# A Appendix: Proof of expression 2e

Let $g_S$ be the genotype of sibling of the proband and $g_p$ be the genotype of the proband, define

$$\tau\left(g_s \mid g_p\right) = Pr\left(g_s \mid g_p\right)$$

Conditioning on $g_p$, we have

$$\begin{aligned}
K \times K_S &= \sum_{g_p \in G} Pr\left(X_S = 1, X_P = 1 \mid g_p\right) F\left(g_p\right) \\
&= \sum_{g_P \in G} Pr\left(X_S = 1 \mid X_P = 1, g_p\right) P_r\left(X_P = 1 \mid g_p\right) F\left(g_p\right) \\
&= \sum_{g_p \in G} F\left(g_p\right) P\left(g_p\right) \sum_{g_S \in G} Pr\left(g_S \mid X_P = 1, g_p\right) Pr\left(X_S = 1 \mid g_s, g_p, X_p = 1\right) \quad (3) \\
&= \sum_{g_p \in G} F\left(g_p\right) P\left(g_p\right) \sum_{g_S \in G} \tau\left(g_s \mid g_p\right) P\left(g_S\right)
\end{aligned}$$

where $F$ and $P$ are genotype frequency and penetrance functions as defined before. To break (3) into single locus measures, we need to define several single DSL measures,

- Define

  $$\tau_i\left(g_{si} \mid g_{pi}\right) = Pr\left(g_{si} \mid g_{pi}\right) = Pr\left(g_{si}, g_{pi}\right)/f_i\left(g_{pi}\right)$$

  By conditioning on parental genotype $g_F$ (paternal) and $g_M$ (maternal), $\tau_i\left(g_{si} \mid g_{pi}\right)$ can be calculated as

  $$\tau_i\left(g_{si} \mid g_{pi}\right) = f_i\left(g_{pi}\right) \sum_{g_{Fi} \in G_i} \sum_{g_{Mi} \in G_i} Pr\left(g_{si}, g_{pi} \mid g_F, g_M\right) f_i\left(g_{Fi}\right) f_i\left(g_{Mi}\right) \qquad (4)$$

  where $Pr\left(g_{si}, g_{pi} \mid g_F, g_M\right) = Pr\left(g_{si} \mid g_F, g_M\right) \times Pr\left(g_{pi} \mid g_F, g_M\right)$ are obtained using a table like $P_r\left(g = (1,1) \mid g_F = (0,1), g_M = (0,1)\right) = \frac{1}{4}$.

- Define $K_{Si}$ analog to $K_S$ but with $g$ confined to DSL $i$ , (the meaning of $K_i$ and $K_{si}$ does not hold any more, they are used for notational and computational convenience.)

  $$\begin{aligned} K_{Si} &= \frac{1}{K_i} \sum_{g_{pi} \in G_i} \left[ f_i\left(g_{pi}\right) p_i\left(g_{pi}\right) \sum_{g_{Si} \in G_i} \tau_i\left(g_{si} \mid g_{pi}\right) p_i\left(g_{si}\right) \right] \\ &= \frac{1}{2}\left(1 + 3f_i\right) p_i \end{aligned}$$

  where $K_i = \sum_{g_{pi} \in G_i} f_i\left(g_{pi}\right) p_i\left(g_{pi}\right) = 2f_i p_i$.

Since $\tau_i$ are independent to each other, we have

$$\begin{aligned} \sum_{g_S} \tau_i\left(g_{si} \mid g_{pi}\right) &= \sum_{g_S} Pr\left(g_s \mid g_p\right) = 1 \\ \tau\left(g_s \mid g_p\right) &= \prod_{i=1}^{5} \tau_i\left(g_{si} \mid g_{pi}\right) \\ \sum_{g_{Si}} \tau_i\left(g_{si} \mid g_{pi}\right) &= \sum_{g_{Si}} Pr\left(g_{si} \mid g_{pi}\right) = 1 \end{aligned}$$

Therefore,

$$\begin{aligned}
K\times K_s &= \sum_{g_p\in G}\left[F\left(g_p\right)P\left(g_p\right)\sum_{g_S\in G}\tau\left(g_s\mid g_p\right)P\left(g_S\right)\right] \\
&= \sum_{g_p\in G}\sum_{g_s\in G}\left[\prod_{i=1}^{5}\left(f_i\left(g_{pi}\right)\tau_i\left(g_{si}\mid g_{pi}\right)\right)\left(1-\prod_{i=1}^{5}\left(1-p_i\left(g_{Pi}\right)\right)\right)\left(1-\prod_{i=1}^{5}\left(1-p_i\left(g_{Si}\right)\right)\right)\right] \\
&= \sum_{g_p\in G}\sum_{g_s\in G}\prod_{i=1}^{5}\left(f_i\left(g_{pi}\right)\tau_i\left(g_{si}\mid g_{pi}\right)\right)-\sum_{g_p\in G}\sum_{g_s\in G}\prod_{i=1}^{5}f_i\left(g_{pi}\right)\tau_i\left(g_{si}\mid g_{pi}\right)\left(1-p_i\left(g_{Pi}\right)\right) \\
&\quad -\sum_{g_p\in G}\sum_{g_s\in G}\prod_{i=1}^{5}f_i\left(g_{pi}\right)\tau_i\left(g_{si}\mid g_{pi}\right)\left(1-p_i\left(g_{si}\right)\right) \\
&\quad +\sum_{g_p\in G}\sum_{g_s\in G}\prod_{i=1}^{5}f_i\left(g_{pi}\right)\tau_i\left(g_{si}\mid g_{pi}\right)\left[\left(1-p_i\left(g_{Pi}\right)\right)\left(1-p_i\left(g_{Si}\right)\right)\right] \\
&= \text{item1-item2-item3+item4}
\end{aligned}$$

Notice that

$$\begin{aligned}
\text{item1} &= \sum_{g_p\in G}\sum_{g_s\in G}\prod_{i=1}^{5}\left(f_i\left(g_{pi}\right)\tau_i\left(g_{si}\mid g_{pi}\right)\right) \\
&= \sum_{g_p\in G}\left[f\left(g_p\right)\sum_{g_s\in G}\tau\left(g_s\mid g_p\right)\right]=\sum_{g_p\in G}F\left(g_p\right)=1,
\end{aligned}$$

$$\begin{aligned}
\text{item2} &= \sum_{g_p\in G}\sum_{g_s\in G}\prod_{i=1}^{5}f_i\left(g_{pi}\right)\tau_i\left(g_{si}\mid g_{pi}\right)\left(1-p_i\left(g_{Pi}\right)\right) \\
&= \prod_{i=1}^{5}\sum_{g_{p_i}\in G_i}\sum_{g_{si}\in G_i}f_i\left(g_{pi}\right)\tau_i\left(g_{si}\mid g_{pi}\right)\left(1-p_i\left(g_{Pi}\right)\right) \\
&= \prod_{i=1}^{5}\left(1-\sum_{g_{p_i}\in G_i}f_i\left(g_{pi}\right)p_i\left(g_{Pi}\right)\right)=\prod_{i=1}^{5}\left(1-K_i\right)
\end{aligned}$$

$$\begin{aligned}
\text{item3} &= \sum_{g_p \in G} \sum_{g_s \in G} \prod_{i=1}^{5} f_i\left(g_{pi}\right) \tau_i\left(g_{si} \mid g_{pi}\right)\left(1-p_i\left(g_{si}\right)\right) \\
&= \prod_{i=1}^{5} \sum_{g_{p_i} \in G_i} \sum_{g_{si} \in G_i} f_i\left(g_{pi}\right) \tau_i\left(g_{si} \mid g_{pi}\right)\left(1-p_i\left(g_{si}\right)\right) \\
&= \prod_{i=1}^{5} \left(1-\sum_{g_{p_i} \in G_i} \sum_{g_{si} \in G_i} f_i\left(g_{pi}\right) \tau_i\left(g_{si} \mid g_{pi}\right) p_i\left(g_{si}\right)\right) \\
&= \prod_{i=1}^{5} \left(1-\sum_{g_{p_i} \in G_i} \sum_{g_{si} \in G_i} Pr\left(g_{si}, g_{pi}\right) p_i\left(g_{si}\right)\right) \\
&= \prod_{i=1}^{5} \left(1-\sum_{g_{si} \in G_i} \left(\sum_{g_{p_i} \in G_i} Pr\left(g_{pi} \mid g_{s_i}\right)\right) f_i\left(g_{si}\right) p_i\left(g_{si}\right)\right) \\
&= \prod_{i=1}^{5} \left(1-K_i\right)
\end{aligned}$$

and remember that $K_i \times K_{Si} = \sum_{g_{pi} \in G_i} \sum_{g_{Si} \in G_i} \left[f_i\left(g_{pi}\right) p_i\left(g_{pi}\right) \tau_i\left(g_{si} \mid g_{pi}\right) p_i\left(g_{si}\right)\right]$, we have

$$\begin{aligned}
\text{item4} &= \sum_{g_p \in G} \sum_{g_s \in G} \prod_{i=1}^{5} f_i\left(g_{pi}\right) \tau_i\left(g_{si} \mid g_{pi}\right)\left(1-p_i\left(g_{Pi}\right)\right)\left(1-p_i\left(g_{Si}\right)\right) \\
&= \prod_{i=1}^{5} \sum_{g_{pi} \in G_i} \sum_{g_{si} \in G_i} \left(f_i\left(g_{pi}\right) \tau_i\left(g_{si} \mid g_{pi}\right) - f_i\left(g_{pi}\right) \tau_i\left(g_{si} \mid g_{pi}\right) p_i\left(g_{Pi}\right)\right. \\
&\quad - f_i\left(g_{pi}\right) \tau_i\left(g_{si} \mid g_{pi}\right) p_i\left(g_{Si}\right) + f_i\left(g_{pi}\right) \tau_i\left(g_{si} \mid g_{pi}\right) p_i\left(g_{Si}\right) p_i\left(g_{pi}\right) \\
&= \prod_{i=1}^{5} \left(1-2K_i+K_i K_{Si}\right)
\end{aligned}$$

Finally, we have

$$K \times K_S = 1 - 2\prod_{i=1}^{5}\left(1-K_i\right) + \prod_{i=1}^{5}\left(1-2K_i+K_i K_{Si}\right)$$

This is an extension to expression (16) of Risch (1990).

# References

J M Abdallah, Bruno Goffinet, C Cierco-Ayrolles, and Miguel Pierez-Enciso. Linkage disequilibrium fine mapping of quantitative trait loci: A simulation study. *Genet. Sel. Evol.*, 35:513–532, 2003.

F Balloux. Easypop, a computer program for the simulation of population genetics. *J. Heredity*, 92:301–302, 2001.

A Dunning, F Durocher, and C healey et al. The extent of linkage disequilibrium in four populations with distinct demographic histories. *American Journal of Human Genetics*, 67:1544–1554, 2000.

Paul Fearnhead. Ancestral processes for non-neutral models of complex diseases. *Theoret. Popul. Bio.*, 63:115–130, 2003.

JFC Kingman. The coalescent. *Stochastic Processes Appl.*, 13:235–248, 1982.

Bo Peng and Marek Kimmel. simupop: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18):3686–3687, 2005.

JK Pritchard. Are rare variants responsible for susceptibility to complex diseases. *Am. J. Hum. Genet.*, 69:124–137, 2001.

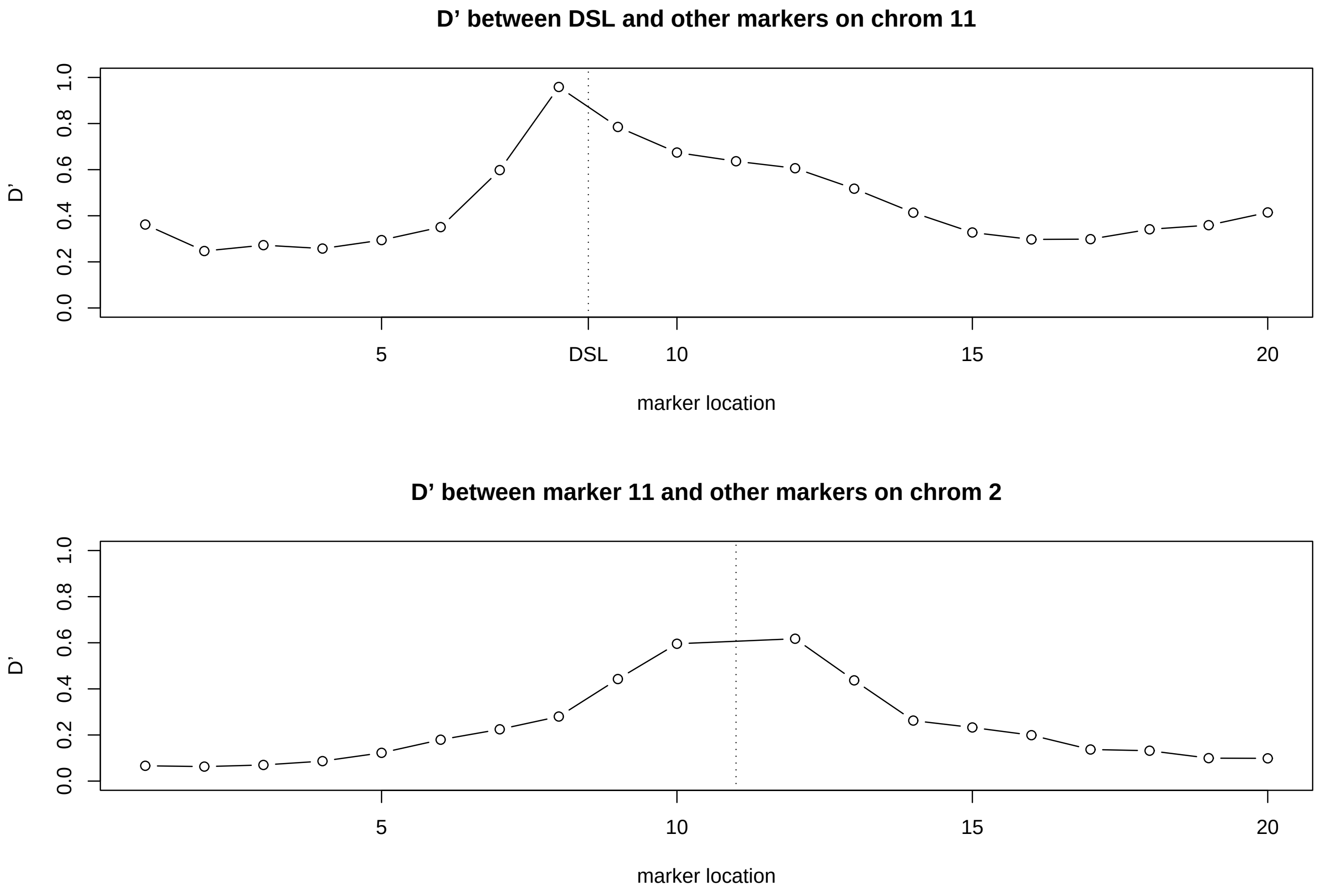
Neil Risch. Linkage strategies for genetically complex traits. i. multilocus models. *Am. J. Hum. Genet.*, 46:222–228, 1990.

Table 1: Theoretical versus simulated population statistics

| | $\delta = \delta_i$ | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $P_{11}$ | $P_{12}$ | $P_{22}$ | $f_1'$ | $K$ | $K_S$ | $\lambda_S$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| theor. | 0.5 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.731 | 0.257 | 0.012 | 0.140 | 0.119 | 0.174 | 1.463 |
| simul. | 0.5 | 0.053 | 0.067 | 0.053 | 0.053 | 0.046 | 0.730 | 0.248 | 0.022 | 0.155 | 0.122 | 0.182 | 1.49 |
| theor. | 0.7 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.735 | 0.254 | 0.011 | 0.138 | 0.163 | 0.236 | 1.445 |
| simul. | 0.7 | 0.053 | 0.067 | 0.053 | 0.053 | 0.046 | 0.763 | 0.224 | 0.014 | 0.125 | 0.173 | 0.264 | 1.52 |
| theor. | 1.0 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.740 | 0.249 | 0.011 | 0.136 | 0.226 | 0.321 | 1.419 |
| simul. | 1.0 | 0.053 | 0.067 | 0.053 | 0.053 | 0.046 | 0.762 | 0.224 | 0.014 | 0.126 | 0.239 | 0.356 | 1.490 |
| theor. | 0.5 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.786 | 0.219 | 0.002 | 0.108 | 0.025 | 0.095 | 3.854 |
| simul. | 0.5 | 0.006 | 0.023 | 0.010 | 0.009 | 0.011 | 0.882 | 0.115 | 0.003 | 0.062 | 0.029 | 0.098 | 3.390 |
| theor. | 0.5 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.917 | 0.082 | 0.001 | 0.042 | 0.073 | 0.108 | 1.481 |
| simul. | 0.5 | 0.016 | 0.024 | 0.035 | 0.031 | 0.072 | 0.882 | 0.114 | 0.004 | 0.061 | 0.085 | 0.147 | 1.72 |
| theor. | 0.5 | 0.10 | 0.05 | 0.05 | 0.05 | 0.05 | 0.551 | 0.410 | 0.039 | 0.244 | 0.141 | 0.202 | 1.426 |
| simul. | 0.5 | 0.115 | 0.056 | 0.064 | 0.046 | 0.086 | 0.550 | 0.38 | 0.065 | 0.258 | 0.171 | 0.234 | 1.370 |

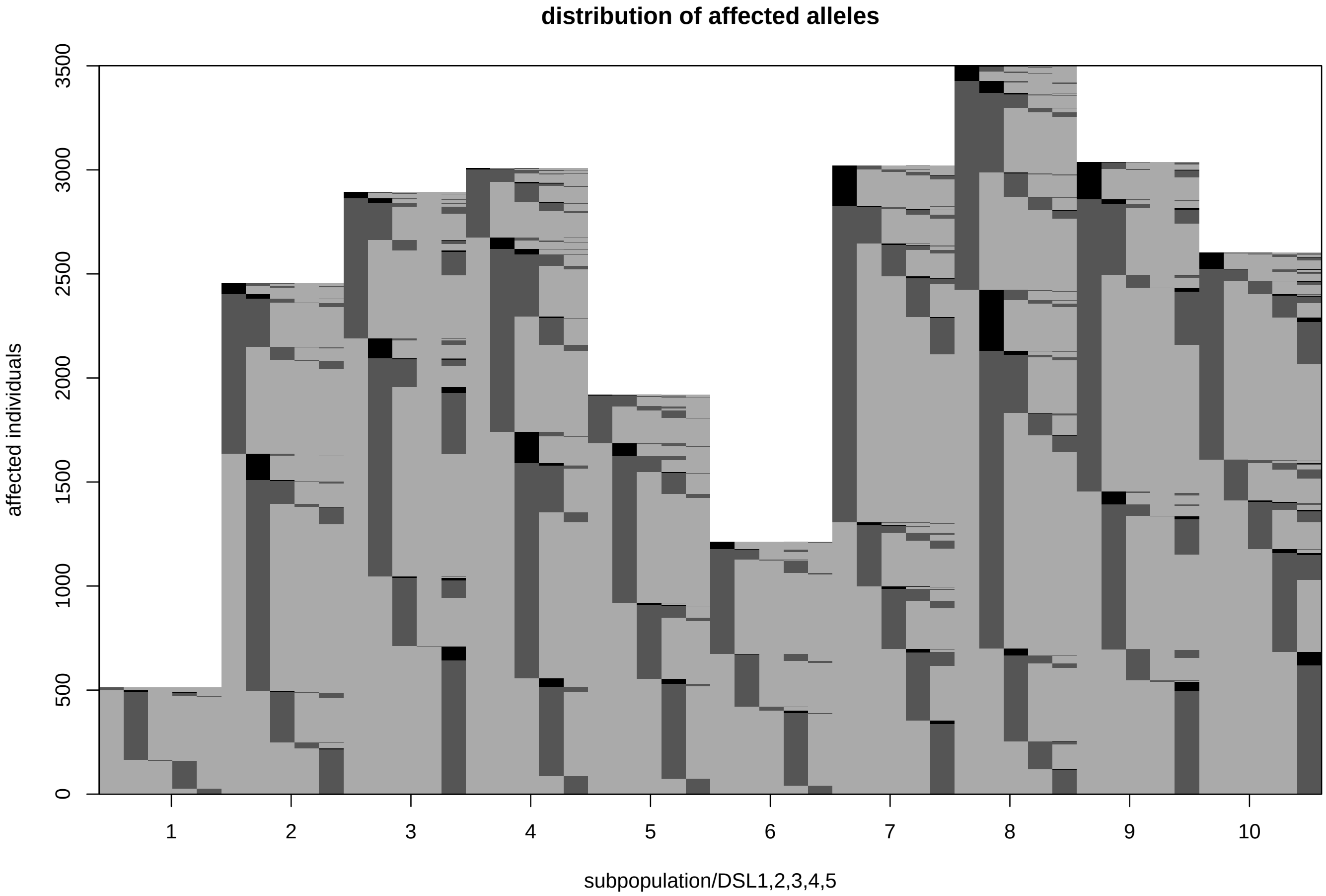
Expected and observed population statistics for various settings of $p$ (penetrance, the same for all DSL) and $f_i$ (allele frequency at DSL $i$), $i = 1, 2, ..., 5$. The statistics are: $K$: disease prevalence; $P_{11} = Pr\,(N, N \text{ at DSL } 1 \mid \text{affected})$; $P_{12} = Pr\,(N, S \text{ or } S, N \text{ at DSL } 1 \mid \text{affected})$; $P_{22} = P_r\,(S, S \text{ at DSL } 1 \mid \text{affected})$; $f_1'$ sample disease allele frequency at DSL 1. $K_s = E\,(X_s \mid X_p = 1)$ the probability of a sibling of an affected proband is affected, $\lambda_S = \frac{K_S}{K}$ risk ratio for a sibling of an affected proband to be affected compared with population prevalence. For each simulation, disease allele frequency are controled within $f_i \pm 0.002$ during disease introduction stage but then evolve freely afterwards.

Figure 1: Linkage disequilibrium on two chromosomes



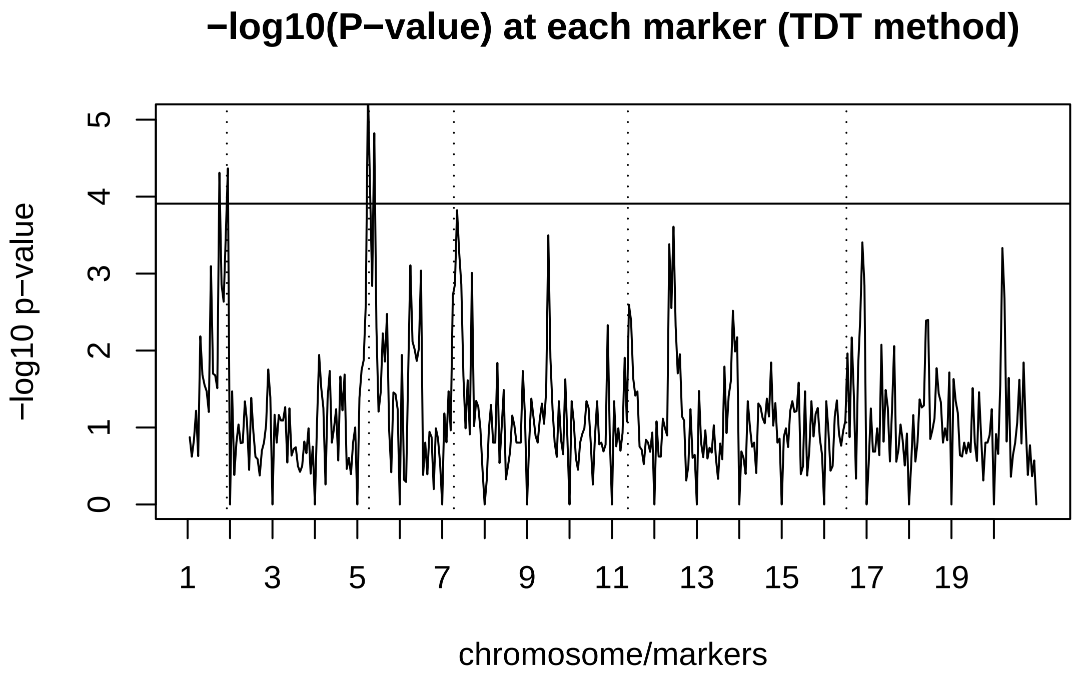
$D'$ value between a DSL and its surrounding markers on chromosome 11 (top), and $D'$ between marker 11 and other markers on chromosome two. Recombination rate is 0.0005.

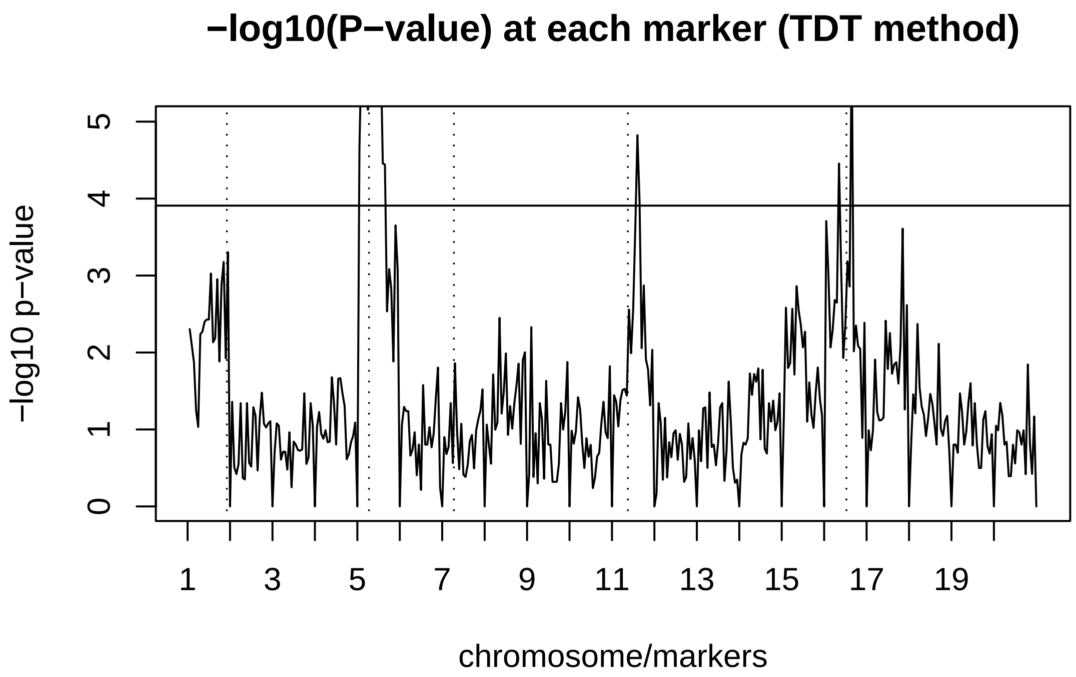
Figure 2: Alleles at disease susceptibility loci of affected individuals



Genotype at the five DSL of all affected individuals, arranged by subpopulation (1-10, from left to right). Genotype at each DSL of each affected individual is displayed from left to right, marked by light gray (NN), dark gray (NS) and black (SS). The single-locus penetrance model is an additive model with parameter 0.5.

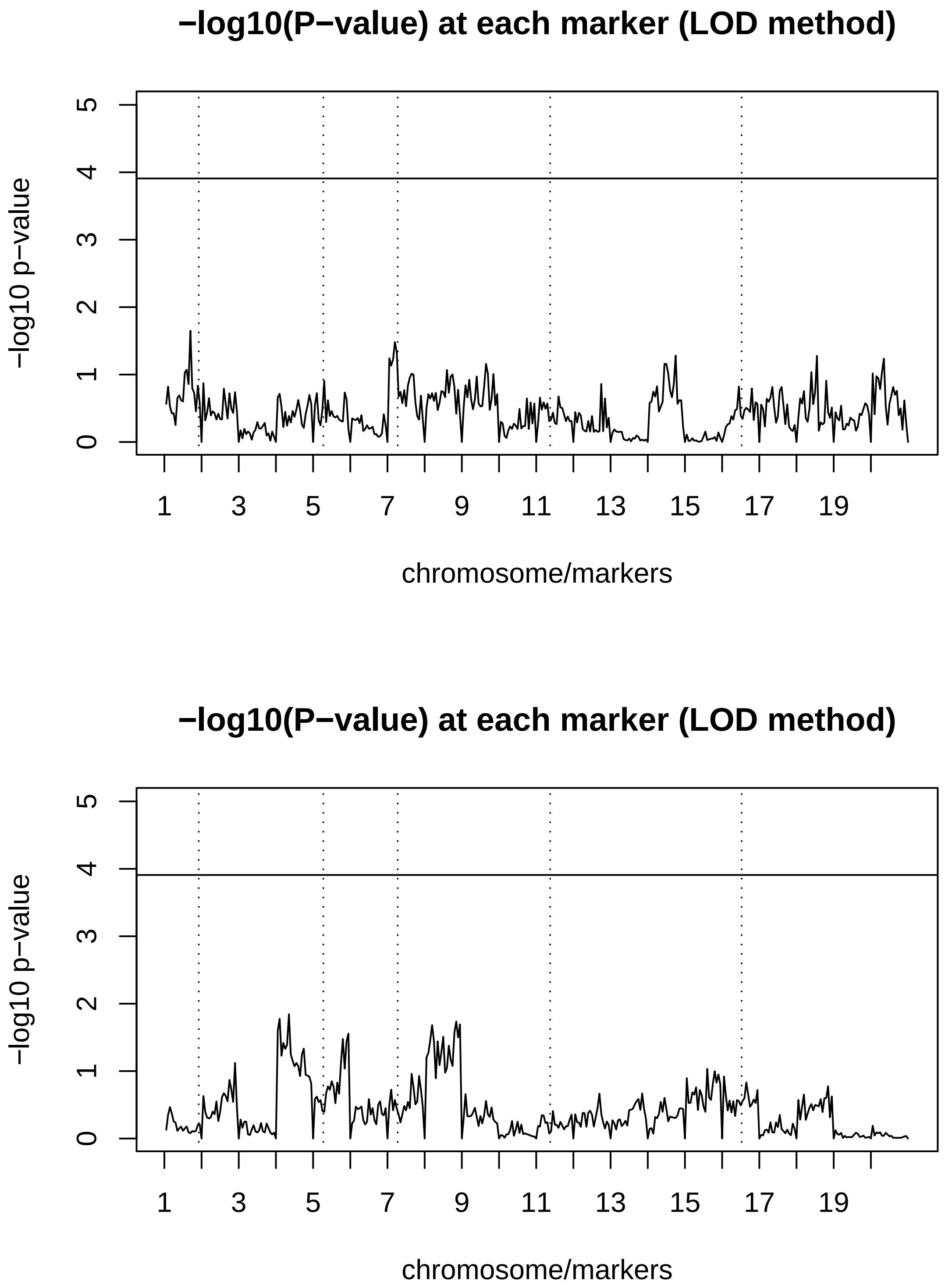
Figure 3: $p$-values of the TDT test





-log $p$-value obtained by applying TDT method to two samples that differ by recombination rate. Vertical dashed lines are location of the five DSL. Horizontal solid line is $-\log 0.05/400$ which is the Bonferroni-adjusted $p$-value. Top, a simulation with recombination rate $5\times10^{-4}$, bottom: recombination rate $1\times10^{-4}$.

Figure 4: $p$-values of the LOD test



−log10(P−value) at each marker (LOD method)

−log10(P−value) at each marker (LOD method)

-log $p$-value obtained by applying LOD method to two samples that differ by recombination rate. Vertical dashed lines are location of the five DSL. Horizontal solid line is $-\log 0.05/400$ which is the Bonferroni-adjusted $p$-value. Top, a simulation with recombination rate $5 \times 10^{-4}$, bottom: recombination rate $1 \times 10^{-4}$.