Statistical Methodology for Sequence Analysis

Thesis advisor: Prof. Christoph Lange Kaustubh Adhikari

Statistical Methodology for Sequence Analysis

Abstract

Rare disease variants are receiving increasing importance in the past few years as the potential cause for many complex diseases, after the common disease variants failed to explain a large part of the missing heritability. With the advancement in sequencing techniques as well as computational capabilities, statistical methodology for analyzing rare variants is now a hot topic, especially in case-control association studies.

In this thesis, we initially present two related statistical methodologies designed for case-control studies to predict the number of common and rare variants in a particular genomic region underlying the complex disease. Genome-wide association studies are nowadays routinely performed to identify a few putative marker loci or a candidate region for further analysis. These methods are designed to work with SNP data on such a genomic region highlighted by GWAS studies for potential disease variants. The fundamental idea is to use Bayesian methodology to obtain bivariate posterior distributions on counts of common and rare variants. While the first method uses randomly generated (minimal) ancestral recombination graphs, the second method uses ensemble clustering method to explore the space of genealogical trees that represent the inherent structure in the test subjects.

In contrast to the aforesaid methods which work with SNP data, the third chapter deals with next-generation sequencing data to detect the presence of rare variants in a genomic region. We present a non-parametric statistical methodology for rare variant association testing, using the well-known Kolmogorov-Smirnov framework adapted for genetic data. it is a fast, model-free robust statistic, designed for situations where both deleterious and protective variants are present. It is also unique in utilizing the variant locations in the test statistic.

Contents

List of Figures

Acknowledgments

First, I sincerely acknowledge the guidance of my advisor Christoph Lange, and my committee members Nan Laird and Gary King.

Chapter 1: Thanks to Taofik AlChawa, Kerstin Ludwig and Elisabeth Mangold for providing the cleft lip dataset. Yufeng Wu has provided the code of his TMARG program to generate minARGs randomly. Sebastian Zöllner, David Balding and Saurabh Ghosh have provided many helpful comments.

Chapter 2: Thanks to Justin Grimmer and Brandon Stewart for the clustering code.

Chapter 3: Thanks to Benjamin Raby and Michael Cho for their advice on real datasets. Rajarshi Mukherjee has also provided some helpful comments.

1

Is it rare or common?

Many Genome-Wide Association Studies (GWAS) have signals with unknown etiology. This paper addresses the question — is such an association signal caused by rare or common variants that lead to increased disease risk? For a genomic region implicated by a GWAS, we use Single Nucleotide Polymorphism (SNP) data in a case-control setting to predict how many common or rare variants there are, using a Bayesian analysis. Our objective is to compute posterior probabilities for configurations of rare and/or common variants. We use an extension of coalescent trees — the Ancestral Recombination Graphs (ARG) — to model the genealogical history of the samples based on marker data. As we expect SNPs to be in Linkage Disequilibrium (LD) with common disease variants, we can expect the trees to reflect on the type of variants. To demonstrate the application, we apply our method to candidate gene sequencing data from a German case-control study on nonsyndromic cleft lip with or without cleft palate (NSCL/P).

1.1 Introduction

The Common-Disease-Common-Variant (CDCV) hypothesis [Balding et al., 2007] extended the simple model of one-gene-one-disease applicable only to Mendelian disorders, the notion of common disease variants being that a few common variants underly a common disease by leading to increased disease susceptibility. Common variants were defined as variants with $> 5\%$ Minor Allele Frequency (MAF). But as common variants could not explain a large part of the heritability for many common diseases, the rare variants hypothesis [Bodmer and Bonilla, 2008] [Schork et al., 2009] was put forward as an explanation [Dickson et al., 2010].

As rare variants have very low LD with the SNP markers typically used for GWAS, such studies are generally under-powered to detect the presence of rare disease variants [Asimit and Zeggini, 2010]. So, while still using the SNP data, we aim to answer a fundamental question — does this genomic region contain rare variants for increasing disease risk?

If we answer this correctly, then we can either continue doing SNP association studies if only common disease variants are present, or go into sequencing studies to detect those rare disease variants. Thus, it seems that a proper answer would be to predict the number of common and rare disease variants in that region from the data. More generally, we will provide a posterior distribution of the number of common and rare variants in the region.

We suggest that the SNP data does in fact contain information about this problem. It is commonly known that SNPs contain useful information about the genealogical history of the samples [Balding et al., 2007], which is used to construct the genealogical tree on which the samples are arranged. The common variants will shape the coarser structure of the tree, while the rare mutations will come into play in the lower branches — how the affected and unaffected are clustered in the lower subtrees will tell us if there are some rare disease variants for those group of subjects. The diagram in the next section illustrate two such scenarios.

Our Bayesian approach is the following — we want to obtain posterior probabilities for having different configurations of rare and/or common disease variants. To do so, we use SNP data to generate min-ARG's [Wu, 2008] (an ARG [Griffiths and Marjoram, 1996] with minimum number of recombinations) to model the genealogical history of the sample, without regard to their case-control status. On these trees, to perform a Monte Carlo integration, we generate different configurations of disease mutations, and calculate the likelihood of the observed disease status. That is then coupled with priors to generate the posterior distribution. To illustrate, we apply our method to a real dataset, and observe that the posterior mode indicates the presence a few rare variants — it is discussed in detail in the real data analysis section. A flowchart showing the steps is presented in the next section, and a summarized algorithm is presented at the end of the methods section. We also explain the workings of the method by a toy example presented later.

This method can be thought as an extension to the analysis in [Zöllner and Pritchard, 2005] or [Morris et al., 2002], where a single disease variant (which is unobserved) within one of the SNP-intervals with the highest posterior probability was detected. (Similarly, we too assume that the Disease Susceptibility Loci (DSL) are not the SNP markers themselves.) Here, we allow for multiple disease variants, both common and rare. Moreover, we want to make an overall conclusion regarding presence or absence of rare variants, so we aggregate the common and rare variants by not trying to determine their location within the gene segment (which would get particularly difficult for rare variants).

1.2 Materials and Methods

The genealogical tree has been a common approach [Balding et al., 2007] to model the ancestral history of a set of individuals. It is generally accepted [Zöllner and Pritchard, 2005], [Gusfield et al., 2004] that the ARGs are good approximations of the true unknown genealogy for case-control data when we have sufficient number of SNPs. The purpose of using them is to distinguish excess sharing of disease allele from allele sharing due to relatedness. In this way, the genealogical tree presents the information in the marker SNPs to the case-control association study, thereby increasing efficiency.

In the following diagram (figure 1.1), we illustrate the hypothetical situation of two (complex) diseases via two different genealogical trees, one being driven by a common variant and the other by rare variants.

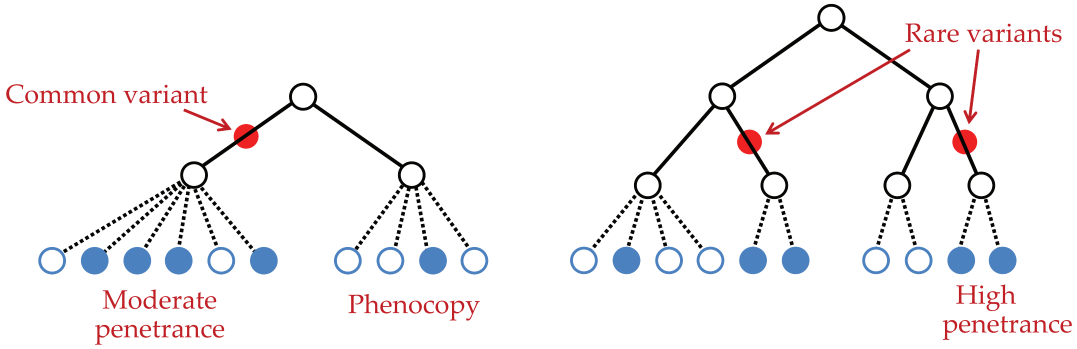


Figure 1.1: Two contrasting scenarios where a complex disease is caused by either a common variant or two rare variants.

This distinction between common and rare variants is driven largely by the disease model that we will specify, because the analysis certainly depends on how we define those variants and their effects. We specify all the components of the disease model while describing the likelihood, which has three main components. After the likelihood, we define the priors to be used in conjunction, and then show the steps to compute the posterior. It is important to remember that, as the final posterior probability, we are interested in the presence or absence of variants, rather than their locations. We explain our method through the flowchart on the next page.

1.2.1 Flowchart of the Method

A flowchart (figure 2.1) illustrating the steps of our method is presented here. It is the diagrammatic representation of the algorithm in section 2.2.4. Here, we show that, at first, genealogical trees (actually, AR graphs) are generated from the SNP genotype (phased) data. Then, disease information is added for the subjects, and the disease likelihood is modeled, for which we simulate potential disease mutations at different branches of the tree. Using all these, the likelihood is computed, and then using appropriate priors, the posterior is calculated. The posterior is aggregated over simulated trees and mutations to give the final posterior distribution of rare and common variant counts.

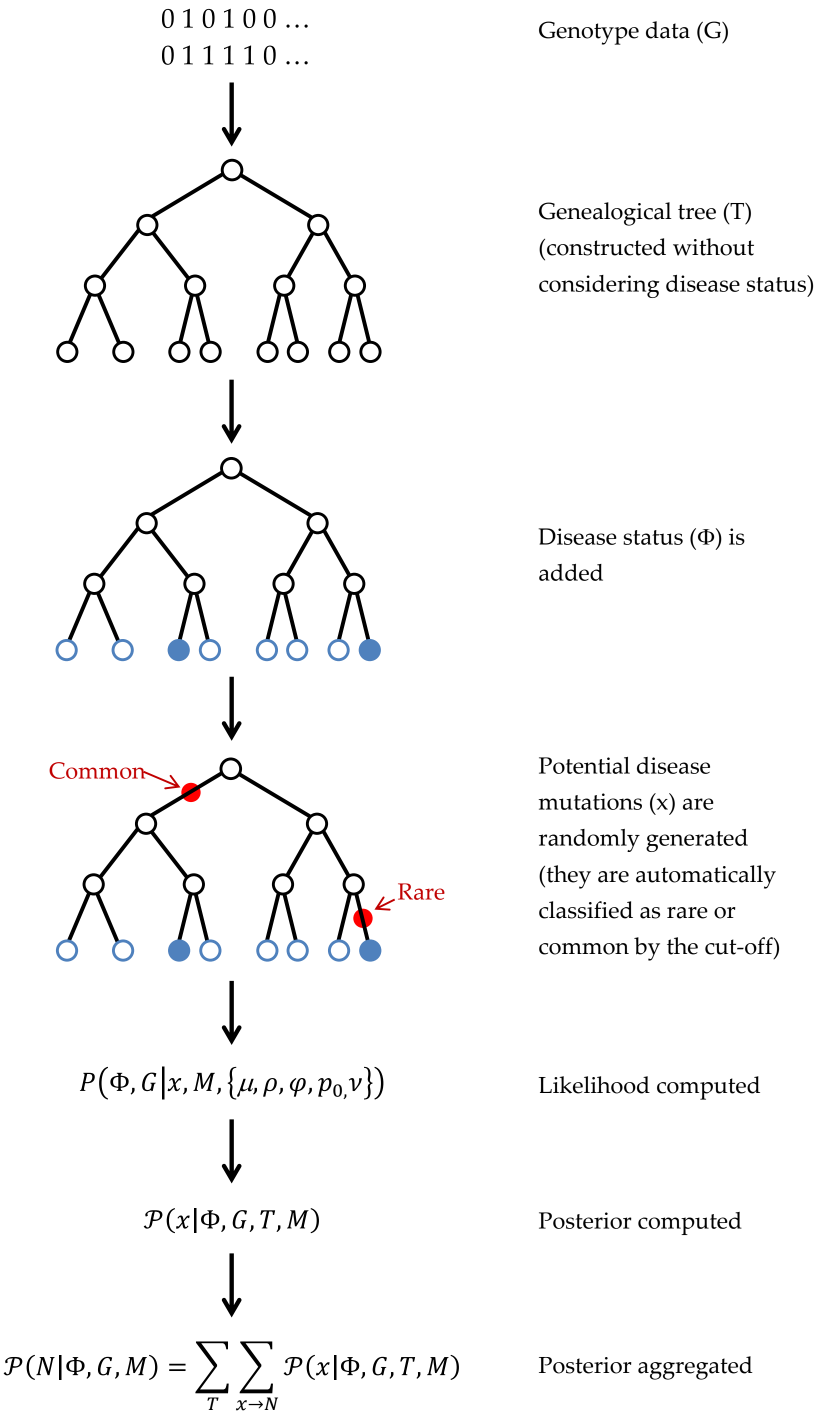

0 1 0 1 0 0 …
0 1 1 1 1 0 …
Genotype data (G)
Genealogical tree (T) (constructed without considering disease status)
Disease status (Φ) is added
Common
Rare
Potential disease mutations (x) are randomly generated (they are automatically classified as rare or common by the cut-off)
$P\left(\Phi, G \middle| x, M, \{\mu, \rho, \varphi, p_{0,} \nu\}\right)$
Likelihood computed
$\mathcal{P}(x|\Phi, G, T, M)$
Posterior computed
$\mathcal{P}(N|\Phi, G, M) = \sum_{T} \sum_{x \to N} \mathcal{P}(x|\Phi, G, T, M)$
Posterior aggregated

Figure 1.2: Flowchart for the algorithm.

1.2.2 Bayesian Analysis

[Morris et al., 2002], [Zöllner and Pritchard, 2005] perform Bayesian analysis to obtain posterior probabilities of of the SNPs being (in LD with) the true DSLs. In the first step, they use the SNP genotype data (G) to generate possible ARGs (T) from the posterior $P(T \mid G)$. In the next step, they evaluate posterior probabilities of the disease loci (x) given observed disease phenotypes (Φ) and the tree structure, i.e. $P(x \mid \Phi, T)$. The locus with the highest posterior probability can then be reported.

Our Bayesian analysis, while along the lines of [Zöllner and Pritchard, 2005], extends to complex diseases by allowing multiple DSLs with different penetrance. So, instead of a single DSL, we evaluate the posterior probability of a particular configuration of DSLs. Then, we evaluate posterior probabilities corresponding to counts of rare and common variants by aggregating over such posteriors.

Thus, instead of a single location, the vector x now contains all the information about (simulated) disease-susceptible mutations in the gene — the locations of the mutations, type of the mutation, as well as the allele at that locus. This x leads to a count of common and rare variants — denoted by the tuple $N = (N_c, N_r)$. We can obtain $P(N)$ by aggregating over $P(x)$, and the posterior probability of the counts — $P(N \mid \Phi, T)$ is what we will be interested in.

1.2.3 The Likelihood

The likelihood has three terms in all — first, the probabilities of the minARGs given the SNP genotypes, $P(T \mid G)$, secondly, the probabilities of the disease mutations occurring on a tree, $P(x \mid T)$, and finally, the disease probabilities given the disease mutations, $P(\Phi \mid x, M)$. It will also include the modeling parameters, which we discuss later.

We will explicitly mention the disease model M, which includes models for the disease probabilities given the mutations, and involves models for penetrance, epistasis, phenocopy etc. In the initial stages of our calculations we keep the model M, as a conditional term, to clearly identify situations where those disease modeling assumptions play a role. We can later omit the term, as we consider it fixed.

Modeling the Tree

While modeling the tree, a simple top-down approach is taken. The root node has probability 1. Conditional on the parental node, a mutated offspring has probability μ, so a direct descendant, whose genotypes are same as his parent, has probability $(1 - \mu)$. Mutations are assumed to be independent. As each node only allows one mutation, if there are $\mathcal{N}$ nodes in the tree, and m mutations, then the likelihood term for mutation is $\mu^m(1 - \mu)^{\mathcal{N}-m}$.

We also assume that mutation and recombinations are independent. If the probability of a recombination is ρ_j at locus j, at one particular node, the contribution is $\rho_j^r(1 - \rho_j)^{1-r}$, where r indicates if there is a recombination or not.

Modeling the Mutations

Our model extends the approach of [Zöllner and Pritchard, 2005], where we had only one possible DSL, and since we do not know beforehand which SNP-interval it will belong to, it was assumed that without phenotype information the tree does not contain information about the DSL.

Now, since we have multiple mutations possible, one simple extension would be to model $P(x \mid T, \mu) = \mu^k$, where k is the total number of mutations in the tree that contribute to the disease. We still preserve the basic assumption that the tree topology by itself does not provide any information on the causal DSLs; this will be explained in detail in the discussion on priors (section 1.2.4).

Modeling the Disease Probability

As we have seen, the disease loci variable x contains the locations of the mutations, its type — rare/common, and the zygosity — 0/1/2 if common, 0/1 if it is a rare variant (as a person having two copies of the same rare variant is extremely low, without high inbreeding). These disease mutations are distinct from the marker SNP mutations, which are only used in making the trees.

Given a tree, a variant can be determined to be rare or common based on the threshold — if we take MAF $< 1\%$ to be the definition of rare variants, then by looking at where the mutation occurs in a tree, we obtain the proportion of people who have that mutation, and simply compare it with 1%.

While modeling the rare variants, we consider the mathematically simplifying assumption that if a subject has inherited any of the hypothetical (simulated) rare causal mutations, that individual will be diseased, i.e. complete penetrance. With completely penetrant rare variants, having a second rare variant in addition to the first one does not change the likelihood — that is the mathematical simplification we aim for. This is partly motivated by the standard infinite-sites model [Balding et al., 2007]. Note that this is not same as assuming that a person can have only one rare variant.

On the other hand, a person can have multiple common variants, each with small to moderate effect. Let c be the number of common variants a person carries, adding over all loci in the two chromosomes. The penetrance is modeled as a function of the common variant count c — it is easy to see that this function should be a positive non-decreasing concave function. That is so because, it is desirable that the gain in penetrance while moving from 20 to 21 common variants should be rather small than while moving from 1 to 2 common variants; if we used an additive or multiplicative instead, the change would have remained same or even increased, something which is not desirable. It also means that we do not enforce the common variants to have constant effect sizes. A suitable model for the penetrance is thus the power law, where $p = p_0 \cdot c^{\nu}$, where $p_0 \geq 1$ is a multiplying factor (does not have the 'base rate' interpretation), $c \in (0, 1)$ is the standardized count, and $\nu \in [0, 1]$ is the shape parameter. In this expression, the common variant count c is transformed to be in $(0, 1)$ by dividing with the total number of loci.

We also allow for phenocopy, i.e. an individual can be affected without having any causal variants. It is modeled as $P(\text{disease} \mid \text{no variants}) = \varphi$, which is a small but non-zero quantity.

These modeling assumptions constitute our disease model M, and the results will certainly vary to some extent if a different model is used. To note, M comes only in the $P(\Phi \mid x, M)$ term.

1.2.4 Priors

We have denoted likelihoods by $P(\cdot)$, and we denote priors by $\pi(\cdot)$. Following the standard practice, we take $\pi(T)$ and $\pi(x)$ to be uniform. As mentioned in section 1.2.3, having a uniform prior on the tree topology on absence of any specific information is reasonable, and then our prior on the number of recombinations or mutations compensate for structures which have more mutations or recombinations, because they are rare events in reality. Also, if we do not have specific information about increased or decreased mutation rate in a specific part of the candidate region, it would be reasonable to assume that mutation is equally likely in any particular locus. One important point to consider is that these are proper uniform priors. This is so because the spaces of T and x are finite, as the number of possible minARGs is finite, and also any tree being of finite size, the number of possible mutations is also finite.

Now we put priors on the parameters $\mu, \rho, \varphi, p_0, \nu$. We do not go on to use hyperpriors on the prior parameters, but instead choose them carefully, e.g. the mean recombination rate from HAPMAP.

Mutation rate: $\pi(\mu) = \text{beta}(\alpha_\mu, \beta_\mu)$,

recombination: $\pi(\rho) = \text{beta}(\alpha_\rho, \beta_\rho)$.

Phenocopy: $\pi(\varphi) = \text{beta}(\alpha_\varphi, \beta_\varphi)$.

Penetrance: $\pi(p_0) =$gamma(α_p, β_p); $\pi(\nu) = \frac{\log \delta}{\delta - 1} \cdot \delta^\nu,\ \delta \in (0, \infty)$.

The mutation, recombination and phenocopy rates are probabilities, and so it is standard to use a beta prior for them, as beta distribution is generally a conjugate prior for probabilities. For our real data analysis, we could estimate the rates from HAPMAP data extracted about the same region, and compute the prior parameters. In general, the program uses standard values provided in the literature, but since the probabilities can vary across the genomic region and the phenocopy rate might vary based on various environmental factors, the program allows for updated parameter values to better suit the dataset at hand. In Bayesian methods, the priors have a larger effect when the sample size is small, and the effect of the prior 'washes away' as the sample size tends to infinity. So, especially for smaller sample sizes, the user can choose to vary the prior parameters themselves to see to what extent the posterior distribution is affected.

The parameters p_0 and ν have prior distributions chosen in a way such that they provide conjugate priors for the distribution of common variant penetrance described in the previous section and achieve the intended ‘positive non-decreasing concave’ shape.

1.2.5 Posterior

Because the parameters here are setwise independent, i.e. the three likelihood terms have different parameters, we can simplify the likelihood as (steps follow):

$$P(\Phi, G \mid x, M, \{\mu, \rho, \varphi, p_0, \nu\}) = \sum_T P(\Phi \mid x, M, \{\varphi, p_0, \nu\}) \cdot P(x \mid T, \{\mu\}) \cdot P(T \mid G, \{\rho\}),$$

which can be written concisely, by integrating out the parameters, as

$$P(\Phi, G \mid x, M) = \sum_T P(\Phi \mid x, M) \cdot P(x \mid T) \cdot P(T \mid G).$$

Simplification steps: The parameters here are setwise independent, i.e. the the three likelihood terms have different parameters. We will later see why this is true. Then, we can simplify the likelihood as:

$P(\Phi, G \mid x, M) = \sum_T P(\Phi, G \mid x, T, M) P(T \mid x, M)$

$\approx \sum_T P(\Phi \mid x, T, M) \cdot P(G \mid x, T, M) \cdot P(T \mid x, M)$ [assumption A]

$= \sum_T P(\Phi \mid x, M) \cdot P(G \mid T) \cdot P(T \mid x)$

$= \sum_T P(\Phi \mid x, M) \cdot P(x \mid T) \cdot P(T \mid G).$

Assumption A says that given the complete ancestral history of both marker and disease loci, and given the disease model, the marker SNPs and disease phenotypes are independent. Which is reasonable given that the DSLs are not the markers, and we are conditioning on the DSLs directly.

The next step follows as the model M only controls the disease probabilities.

Then we use that $P(G \mid T) \propto P(T \mid G)/P(T)$, and $P(T \mid x) \propto P(x \mid T) \cdot P(T)$.

-

As with likelihoods $P(\cdot)$ and priors $\pi(\cdot)$, we denote posteriors by $\mathcal{P}(\cdot)$. We are interested in $\mathcal{P}(x \mid \Phi, G, M)$.

By Bayes rule, (note that $\pi(x)$ is an uniform prior):

$$\mathcal{P}(x \mid \Phi, G, M) \propto P(\Phi, G \mid x, M) \cdot \pi(x) \propto P(\Phi, G \mid x, M).$$

Note that $P(\Phi, G \mid x, M)$ can be easily obtained from the likelihood after integrating out the model parameters. That step is much simplified by observing again that the parameters are setwise independent, and therefore the three terms in the likelihood can be integrated independently. (Actually, μ contributes to both tree and mutation terms. But by the construction of coalescent trees, each SNP locus can mutate exactly once, and therefore the term involving μ is same for all trees. So we take it out of the calculations. Hence, μ remains only in the mutation term.) The details are shown next:

As mentioned in the deduction of posterior distributions, we integrate out the model parameters from the three likelihood components. This is facilitated by the parameters being setwise independent, which can be easily seen by looking at the three terms.

$$P(T \mid G) = \int P(T \mid \rho, G) \pi(\rho) \, d\rho,$$

$$P(x \mid T) = \int P(x \mid \mu, T)\pi(\mu)\, d\mu,$$

$$P(\Phi \mid x, M) = \int P(\Phi \mid p_0, \nu, \varphi, x, M)\pi(p_0)\pi(\nu)\pi(\varphi)\, dp_0\, d\nu\, d\varphi. \qquad \bullet$$

Hence, we can write,

$$\mathcal{P}(x \mid \Phi, G, M) = \sum_T P(T \mid G) \cdot P(x \mid T) \cdot P(\Phi \mid x, M).$$

In the final stage, we summarize the information on mutations to the count vector $N = (N_c, N_r)$, which stores the number of common and rare variants present in the case-control sample. Since are actually interested in is N, not x, so we aggregate over x to get the posterior distribution of N.

$$\mathcal{P}(N \mid \Phi, G, M) = \sum_T \sum_{x \to N} P(T \mid G) \cdot P(x \mid T) \cdot P(\Phi \mid x, M)$$

$$= \sum_T P(T \mid G) \left\{ \sum_{x \to N} P(x \mid T) \cdot P(\Phi \mid x, M) \right\}.$$

1.2.6 The Steps for Computation

After describing the model in the previous sections, we now outline the steps for computing the posterior $\mathcal{P}(N \mid \Phi, G, M)$, which is to be used for making inferences. These steps are also illustrated on the flowchart (figure 2.1) in the beginning of this section.

1. We use SNP haplotype data (G) to generate possible minARGs (T).

2. Given a tree (T), we model the likelihood $P(x \mid T)$ of the putative DSL configurations (x), which depends on the probability of mutation and

recombination at each site.

3. Next, we model the disease probabilities, $P(\Phi \mid x, T, M)$, where Φ is the disease status, M is the disease model.

4. They are used to simulate configurations of mutations (x) at probable DSLs. Each configuration corresponds to a particular count of common and rare variants, $N = (N_r, N_c)$.

5. So the terms in the likelihood are: $P(\Phi \mid x, T, M), P(x \mid T), P(T \mid G)$. The complete likelihood $P(\Phi, G \mid x, M)$ aggregates the previous terms by summing over all possible T's.
   $P(\Phi, G \mid x, M) = \sum_T P(\Phi \mid x, M) \cdot P(x \mid T) \cdot P(T \mid G)$.

6. We use appropriate priors, e.g. uniform prior on trees (T), prior on recombination rate obtained from HAPMAP, etc.

7. We evaluate posterior probabilities $\mathcal{P}(x \mid \Phi, G, T, M)$ of such configurations (x), given the observed phenotypes (Φ), SNP data (G), and the tree (T).

8. Finally, we get posteriors $\mathcal{P}(N \mid \Phi, G, M)$ for variant configurations (N), by aggregating over corresponding configurations, and over simulated trees.
   $\mathcal{P}(N \mid \Phi, G, M) = \sum_T \sum_{x \to N} \mathcal{P}(x \mid \Phi, G, T, M)$.

The trees are generated in step 1 by Wu's algorithm of generating minARGs uniformly from a given haplotype data. The mutations in step 4 are generated randomly on the branches of a given tree. Both these simulations are used for Monte Carlo estimates of probabilities by averaging, and therefore our computed posterior depends on the accuracy of the drawn samples — the number of draws and how well they span the sample space. As the sample space is finite in both cases, ensuring these criteria are much more straight-forward.

1.2.7 A toy example

In the following simple example, we show how this method compares different tree structures based on their posterior (computed using likelihood and prior as disucessed below). we fix a tree and compute the posterior probabilities for different mutation configurations. Our data has 4 affected and 5 unaffected subjects. For simplicity, let

$p = \mathrm{P}(\text{disease}|1 \text{ common variant}) = 0.4$,

$r = \mathrm{P}(\text{disease}|\text{rare variant}) = 1$,

$\varphi = \mathrm{P}(\text{disease}|\text{wildtype allele}) = \mathrm{P}(\text{phenocopy}) = 0.1$,

$\mu = \mathrm{P}(\text{mutation at any node}) = 0.05$.

Given the tree structure, there are various possible configurations of DSLs. In figure 1.3, we show four such scenarios.

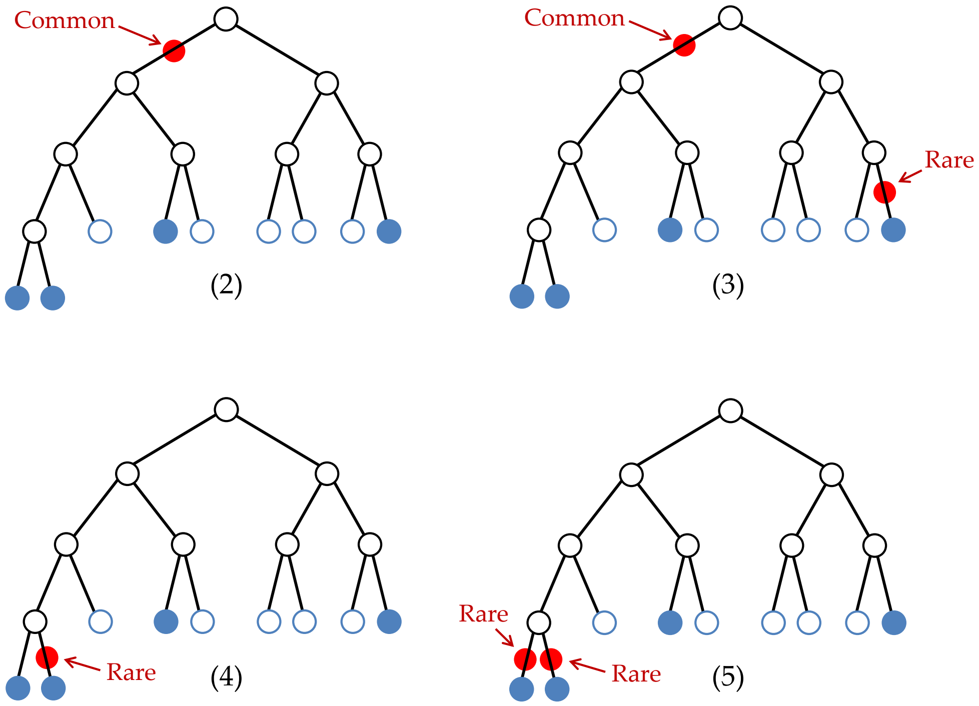


Figure 1.3: Some possible DSL configurations corresponding to the toy example.

First, we compute $P(\Phi \mid x, T, M) = \mathrm{P}(\text{phenotype} \mid \text{mutations, tree, disease model})$.

1. If we had no mutations, i.e. only phenocopies;

$$l = [\varphi^4 \cdot (1-\varphi)^5] = [0.1^4 \cdot 0.9^5], \log l = -9.7.$$

2. 1 common (MAF = 5/9), no rare variants;

$$l = [p^3 \cdot (1-p)^2] \times [\varphi^1 \cdot (1-\varphi)^3] = [0.4^3 \cdot 0.6^2] \times [0.1^1 \cdot 0.9^3], \log l = -6.4.$$

3. 1 common (MAF = 5/9), 1 rare variant (MAF = 1/9);

$$l = [p^3 \cdot (1-p)^2] \times [r^1] \times [(1-\varphi)^3] = [0.4^3 \cdot 0.6^2] \times [1^1] \times [0.9^3], \log l = -4.1.$$

4. 1 rare (MAF = 1/9), no common variants;

$$l = [r^1] \times [\varphi^3 \cdot (1-\varphi)^5] = [1^1] \times [0.1^3 \cdot 0.9^5], \log l = -7.4.$$

5. 2 rare (MAF = 1/9), no common variants;

$$l = [r^2] \times [\varphi^2 \cdot (1-\varphi)^5] = [1^2] \times [0.1^2 \cdot 0.9^5], \log l = -5.1.$$

Next, we compute $P(x \mid T) = \text{P(mutations|tree)}$.

$$\log l \propto \begin{cases} 2\log(1-\mu) = -0.1, & \text{no mutation} \\ \log\mu + \log(1-\mu) = -3, & \text{1 mutation} \\ 2\log\mu = -6, & \text{2 mutations.} \end{cases}$$

Combining these together, the total log-likelihood is:

$$\log l = \begin{cases} -9.8, & \text{no mutation (1)} \\ -9.4, & \text{1 common (2)} \\ -11.1, & \text{1common,1 rare (3)} \\ -10.4, & \text{1 rare (4)} \\ -11.1, & \text{2 rare (5).} \end{cases}$$

Here, we have not evaluated all possible configurations (the ones with higher number of variants will have even smaller probability). Among those considered, case (2) with one common mutation seems most likely.

1.3 Results

1.3.1 Real Data Analysis

Nonsyndromic cleft lip with or without cleft palate (NSCL/P) is a common congenital malformation that is caused by an interplay of multiple genetic and environmental factors [Mossey et al., 2009]. Our dataset comprises 96 NSCL/P cases and 96 controls of Central European ethnicity in whom the exonic and adjacent intronic regions of one candidate gene for NSCL/P has been sequenced. The gene was among the candidate regions in an independent GWA study [Mangold et al., 2010]. Moreover, this gene has functional importance, as it codes for a protein which is involved in bone development, and is therefore relevant for further analysis.

The genotypes were obtained as unphased, and the software PHASE [Stephens et al., 2001], [Stephens and Donnelly, 2003] was used to infer the haplotypes. Phasing probability estimates were high in general (i.e. posterior probability computed by the software for the phase calls were mostly 100%, and in occasional cases going down to 85%, but never below). There was a single missing SNP in a person, and it was imputed using PHASE. The recombination rates estimated [Li and Stephens, 2003], [Crawford et al., 2004] from the data by PHASE tally with those from the CEU population of HAPMAP (phase II) [The International HapMap Consortium, 2007], which comments favorably on our data quality.

On a cursory comparison of the allele frequencies between cases and controls, it seems that lower-frequency SNPs have comparatively higher relative difference in terms of allele frequencies between the two groups, as compared to higher-frequency i.e. common variants. This implies that there should be some effect from rare variants. If we graphically display the generated ARGs for the data (a tree is shown in figure 1.4), they also imply that the top-level (i.e. higher MAF) SNPs do not provide a good partition of the cases and controls, whereas the lower-level SNPs provide small clusters of mostly cases and controls. These are scenarios where our algorithm, as expected, provide higher evidence for rare variants.

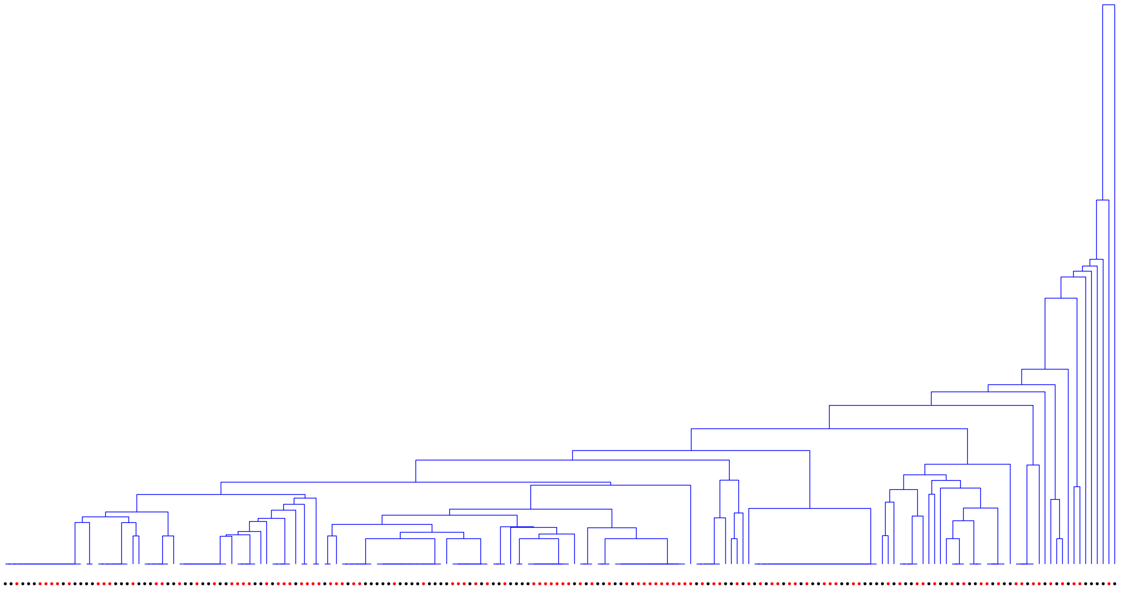


Figure 1.4: A simulated dendrogram for the dataset with cases and controls shown corresponding to the leaves, as red and black dots respectively. Note that some subjects have identical genotypes and therefore are grouped together.

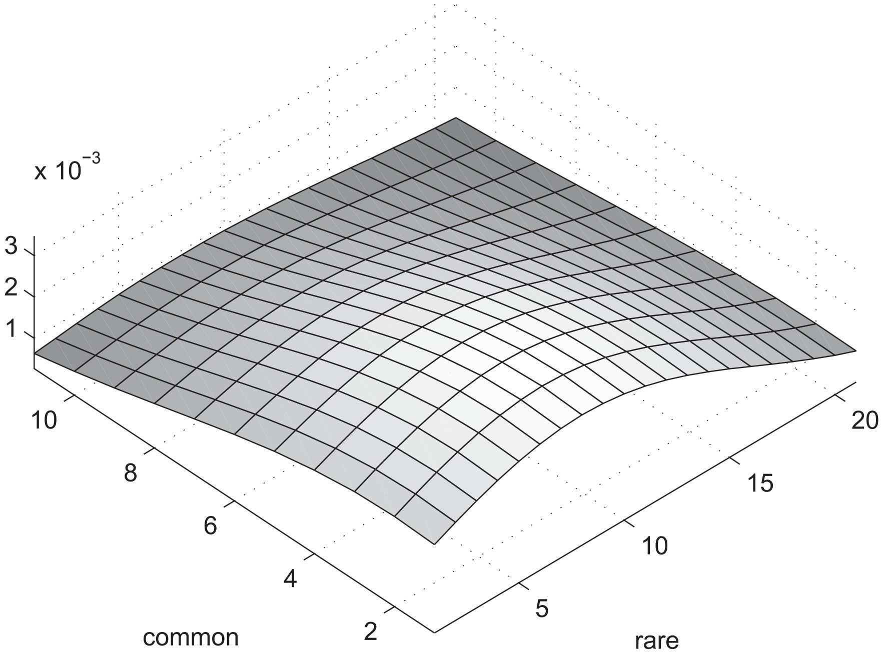


Figure 1.5: Smoothed posterior for real dataset. The X and Y axes denote counts of rare and common variants, and the Z axis shows the (un-normalized) posterior density. The density is smoothed with a Gaussian kernel.

When we look at the obtained posterior distribution (figure 1.5) for the joint distribution of count of common and rare variants, we observe that the posterior mode is positioned along the axis of rare variants. The posterior dies out with increase in the number of common variants, which is expected to be the case when there are few or no common variants. Albeit, the posterior is not highly peaked, something to be expected given the small number of cases.

In fact, as the number of cases is only 96, the possible number of haplotypes is only 192, therefore for rare variants with MAF $< 1\%$, on average we expect to see them in < 2 people. Such situations make it difficult to distinguish between real variant and phenocopy, therefore we can expect the detection efficiency to be low.

But the posterior does indicate a skew towards rare variants, which is something we expect, based on our knowledge of the data. So we can conclude that the method works reasonably with this small sample too. In practice, as verified in our extensive simulations, for usual case-control studies with hundreds or even thousands of subjects, this method will have reasonably good performance.

Since we allow for phenocopy in our model, the posterior distribution also allows for the case with no causal variants in the genomic region, i.e. the (0,0) point. Thus, a posterior distribution comparing the presence or absence of causal variants in that region can also be derived from the current posterior distribution. That can lead to a statistical test for presence of causal variants, something which can complement the objective of this current paper, which works on a genomic region already identified as a potential candidate by a GWAS. It can be an interesting future project.

1.3.2 Simulated Data Analysis

We conduct a simulation study for a number of different scenarios - samples with no underlying variants, samples with only rare variants, with only common variants, and with both types of variants. Some (smoothed) plots of the bivariate posterior densities are produced as examples.

The haplotype distribution is generated by drawn the haplotypes from a coalescent genealogy via MaCS [Chen et al., 2009]. The loci are selected at random to be the DSLs with equal probability. The disease phenotypes are generated under various disease models (e.g. rare DSLs, common DSLs, both, or none), from which a specified number of cases and controls are selected without replacement. The causal loci are then excluded, following the assumption that the marker loci are not the DSLs. The number of SNPs is varied from 30 to 100, and similarly the number of causal rare and common variants. The sample size is also varied — we examine scenarios with total sample size ranging from 300 to 3000. Although we have taken equal number of cases and controls, it is not mandatory for our program. 1000 replications are typically used for calculating the posterior distribution, though this number can be changed easily.

For all the figures presented here as examples, there is more spread in the posterior distribution on the axis of rare variants, as expected; both when there are true rare variants and when not. When there are true rare variants (figure 1.6), the posterior is shifted towards an increased count of rare variants. Similarly, for common variants as well (figure 1.7), the posterior is shifted more along the common variants axis as more common variants are added while simulating the dataset. For the scenarios presented here, we use 1000 cases and controls each, and 1000 replicates for the simulation. The data is generated as phase known.

We observe that the peakedness of the posterior increases with the increase with sample size, which is natural, given that we get more information about the true number of variants. This is especially true for rare variants. As mentioned in the last section, when we have sample sizes of just a couple hundred, it is hard to distinguish rare variants from phenocopies. Then the posterior is largely dictated by the priors. But as we increase our sample size, given a true rare variant, we should see a cluster of affected people under some particular leaf of the genealogical tree. That is how we get increased efficiency to detect the presence of rare variants.

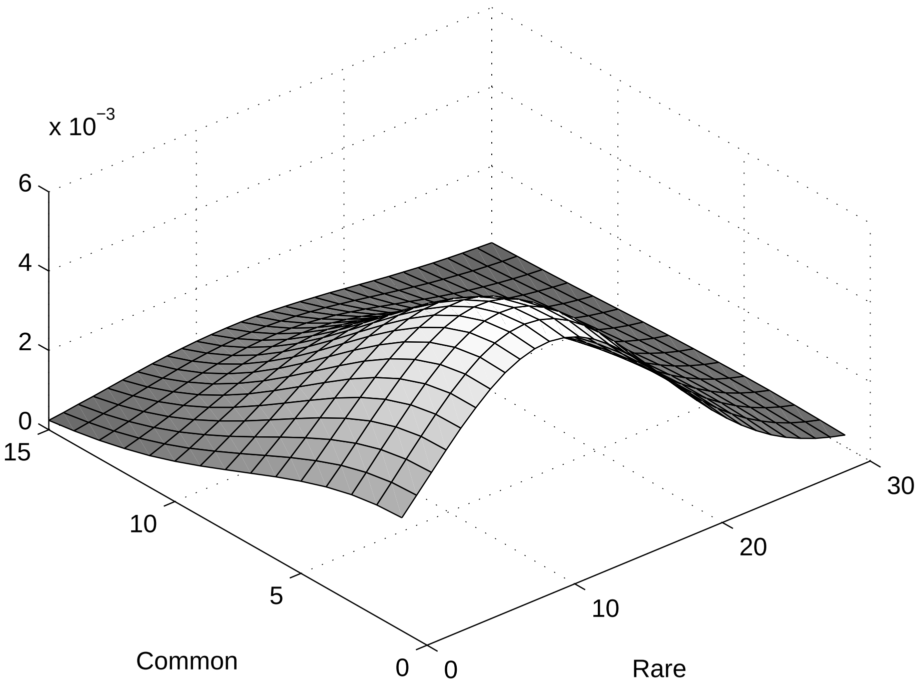


Figure 1.6: Posterior for simulated dataset with rare variants.

The posterior mode gives a rough idea about how many rare and common variants there are. This can be considered as an estimate, though is bound to have some variability, particularly in rare variants, as illustrated by the higher spread in that co-ordinate. We can see in figure 1.8, for example, that even under no true variant, the mode may be close to zero but not exactly zero.

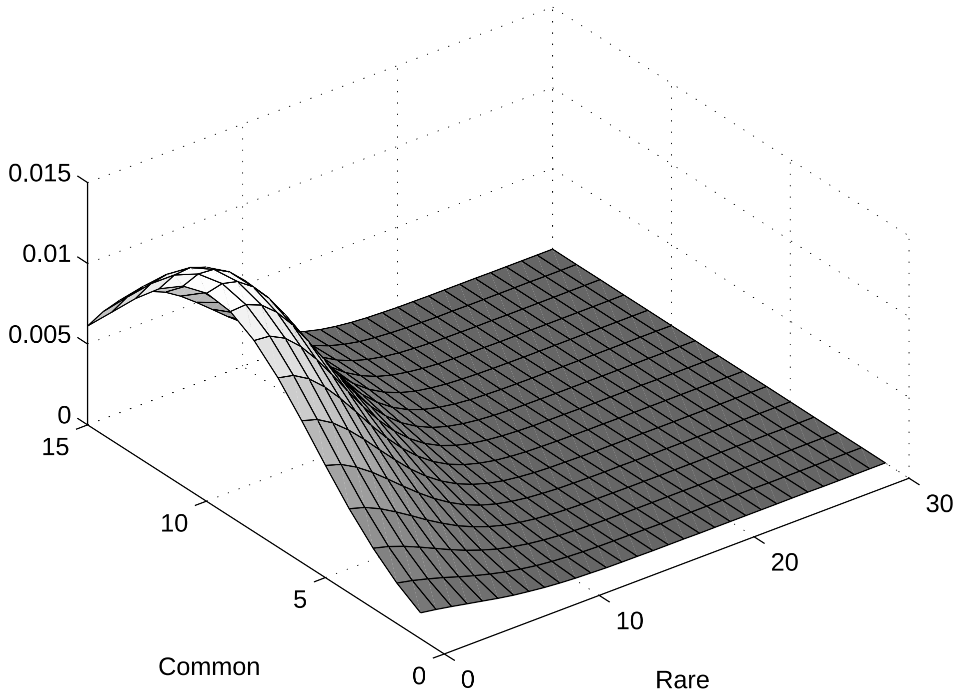


Figure 1.7: Posterior for simulated dataset with common variants.

The simulated scenario with no true variants (figure 1.8) shows what can be considered of the null behavior for this method, where as the other cases reflect its efficiency. While that is somewhat model and parameter-dependent, it appears that the method does well in large samples; in the GWAS era, a sample of about a few thousand is not unexpected.

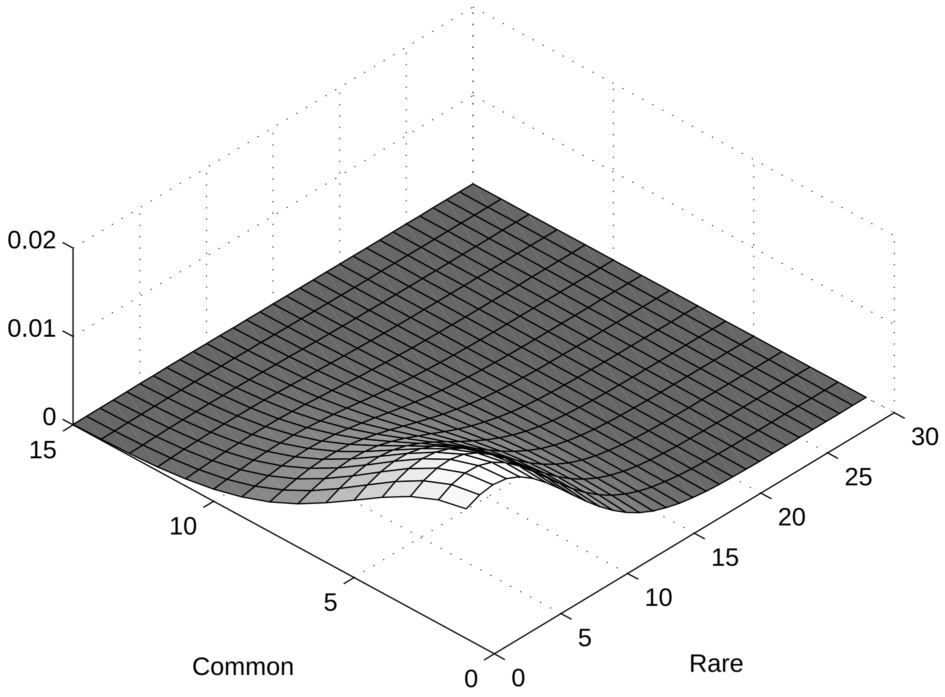


Figure 1.8: Posterior for simulated dataset with no variants.

Although the tree sizes increases with the sample size, the tree nodes work with the haplotypes instead of the genotypes. And the number of possible haplotypes will be much less that the number of people, in particular if there is LD, which we expect to see in the small candidate regions that we work with. Thus, a moderate increase in sample size does not incur an unreasonable increase in tree computation.

1.4 Discussion

In this paper, we presented a method to predict the number of rare and common variants in a genomic region underlying a complex disease. While still based on

SNP data, we are able to obtain information on this by utilizing the genealogical history inherent in the sample, by the use of genealogical trees (more specifically, ancestral recombination graphs). With a Bayesian approach, we provide a bivariate posterior distribution for the counts.

While being a Bayesian analysis, we avoid the use of Markov Chain Monte Carlo (MCMC) or Gibbs sampling to obtain posterior distributions, by taking simple conjugate likelihood and prior models. The choice of priors always has a scope of debate, and might be improved if more information is available from the studies of real datasets. For now, we try to incorporate as much available information as possible, e.g. using the mutation and recombination rates obtained from the HAPMAP. We think that this method can be extended by considering better models for rare variants, which can improve upon some simplifying assumptions.

By excluding such useful but computation-intensive methods, we are able to cut down on runtime, and the program runs under a few minutes for moderate sized datasets, even on personal computers. To be specific, with a few hundreds of subjects, and SNP counts around 30 to 100, the program runs in less that a couple minutes on a laptop (2.5GHz, 3 GB RAM), when we perform 500 simulations for each scenario. The program has fast computation speed for two reasons — first, we mostly use conjugate priors and are able to integrate mathematically, so we avoid Monte Carlo integrations for many variables (and we avoid MCMC altogether) — the computation time is mostly spent on simulating the trees and doing Monte Carlo integration. Secondly, the total number of unique haplotypes after phasing is much smaller than 2 × the total number of subjects, so the number of tree nodes in the ARG is much smaller than the expected number of nodes in a standard genealogical tree. However, as the computation required for trees increases rapidly with large number of SNPs, this method might not be well-suited for large datasets, e.g. a whole-genome scan, at this point; as the capability of computing infrastructure is increasing rapidly, and sequencing costs are also going down very fast, such extensions could become possible in the near future.

Another interesting way of extending this method would be to include covariates. Using covariates in order to better model the environmental effects, in addition to the genetic effects modeling, is becoming increasingly popular, and we could easily extend our method to allow for environmental factors by incorporating covariates into the phenocopy rate parameter.

At this point, we follow the standard assumption ([Wu, 2008], [Zöllner and Pritchard, 2005]) that that haplotype phase is known or readily available. But this method can be readily extended to include haplotype uncertainty. As the package PHASE provides haplotype estimates along with posterior probability estimates corresponding to those phase calls, those can easily be incorporated into our likelihood calculations, and simulations based on different phase configurations can be aggregated with phasing probabilities as weights, to produce the final posterior probability.

The method is fairly robust to population stratification, as it employs genealogical trees to model the population. It will be interesting to see how this method can be extended to family-based studies, which already contains some useful structural information, and is believed to be more powerful for studying rare variants.

2

Rare or common? The clue's in cluster

While analyzing complex diseases for genetic association signals via GWAS (genome-wide association studies), it is of importance to know whether the association signal is caused by some underlying common or rare disease variants. In this paper, we present a new cluster-based method to analyze SNP (single nucleotide polymorphism) data in case-control studies to obtain some insight on this fundamental question. We extend the discovery-based cluster ensemble methodology by Grimmer and King [Grimmer and King, 2011] to genetic data, since clusters can be seen to encompass genealogical trees as a particular form of hierarchical clustering. Utilizing the structure provided by the clusters in a Bayesian framework, we compute posterior configurations of common and/or rare variants to predict roughly how many variants there might be underlying this complex disease.

2.1 Introduction

It is commonly understood that complex diseases, which is the main focus of study for genetic association studies, is driven by multiple disease variants, each conferring small to moderate excess risk. In the early days of SNP studies, only common polymorphisms ($> 5\%$ frequency) could be sequenced, and so the common-disease-common variants (CDCV) hypothesis [Balding et al., 2007] was popular. And indeed, many common variants were found to be linked with complex diseases via GWA studies. However, as common variants could not explain a large part of the heritability for complex diseases, rare variants were put forward as one alternative [Dickson et al., 2010]. Coupled with the improvement in quality in SNP data, so that rarer ($> 1\%$ frequency) polymorphisms could be sequenced, this brought forward the interest in common-disease-rare-variants (CDRV) hypothesis, which says that rare disease variants with higher penetrance might underlie complex diseases.

While it has been increasingly easier to sequence and study thousands of common variants, and to incorporate them into SNP-chips to facilitate disease risk prediction, analyzing rare variants is still harder, as there are much fewer subjects who carry such variants, and the error probability is higher, so detection has low power, and replication is difficult. Thus, while analyzing a segment of the genome for association with a complex disease, it is of importance to be able to have some insight on whether this region contains rare or common disease variants. Such information is useful for planning the analysis strategy — for example, if we know that rare variants is the main factor in this region, we will need to have this region sequenced entirely to obtain all the polymorphisms that this region carry.

Thus, we aim to answer this fundamental question, that whether a candidate region for disease susceptibility contain rare or common variants that cause the disease? To be able to answer this, without having to use sequence-level data for the region, means that we need to utilize the structured information contained in the SNP data. It has been an established technique to construct genealogical trees to model the samples' ancestry from the SNP data, which depicts the transmission of mutations through generations. We observe that, a genealogical tree is simply a particular hierarchical clustering on the set of subjects. Therefore, it is reasonable that we seek to employ the various other clustering techniques available on this data, and each will provide a somewhat different way of looking into the same data, thereby providing new angles to explore the inherent structure provided by the SNP markers.

The question is most naturally answered by Bayesian methodology, which, by providing a posterior on the counts of rare and/or common variants, explicitly indicates how many variants there might be, and with how much certainty can we make such statements. For example, we might report the posterior mode which is $P(\text{the region contains 2 rare variants and 1 common variant}) = 0.3$.

In short, our algorithm is as follows — we apply the Grimmer & King clustering methodology [Grimmer and King, 2011] on the SNP data to produce a cluster space. We randomly select points from this space, where each point represents a new clustering matrix formed by local cluster ensemble of the different clustering methods. This point is converted into a genealogical tree by hierarchical clustering. So, essentially, we generate a set of trees using clustering methods. To perform a Monte Carlo integration, we then generate different configurations of disease mutations on these trees, and calculate the likelihood of the observed disease status of all the subjects. The likelihood is aggregated with priors to obtain the posterior probabilities. In the Monte Carlo step, these are aggregated to obtain the final posterior distribution. A flowchart denoting these steps is shown in the next section.

2.2 Material and methods

2.2.1 The cluster-selection tool of Grimmer & King

Grimmer and King develop a general-purpose clustering methodology in which users can pick new clusters from the cluster space. Their approach is to use all the standard clustering algorithms available and their variants (around 150), to create a space of clustering results, where each algorithm produces its own clustering output as a distance matrix. Therefore each output is a point in the n^2-dimensional space of all possible clusterings. The distance matrix among these points are calculated, based on counting the number of pairs two clustering methods assign to the same clusters.

Then a multidimensional scaling method is used to project this space into a 2-dimensional metric space. Their objective is to let the user click on any point in this space, to generate a new cluster. Such clusters are called ‘local cluster ensembles'. Given any selected point in the two-dimensional space, the distance of this point to all computed cluster points are obtained.

In the next step, an aggregated distance matrix is produced, by averaging each pre-computed distance matrix, with weights proportional to a function of the distance (in the 2-dim space). Usually Gaussian or Epanechnikov kernel is used.

Therefore, their method produces, given any point in the clustering space, a similarity matrix for all the subjects, based on a weighted average of the similarity matrices (or clusterings) given by clustering methods that are 'close' to the given point.

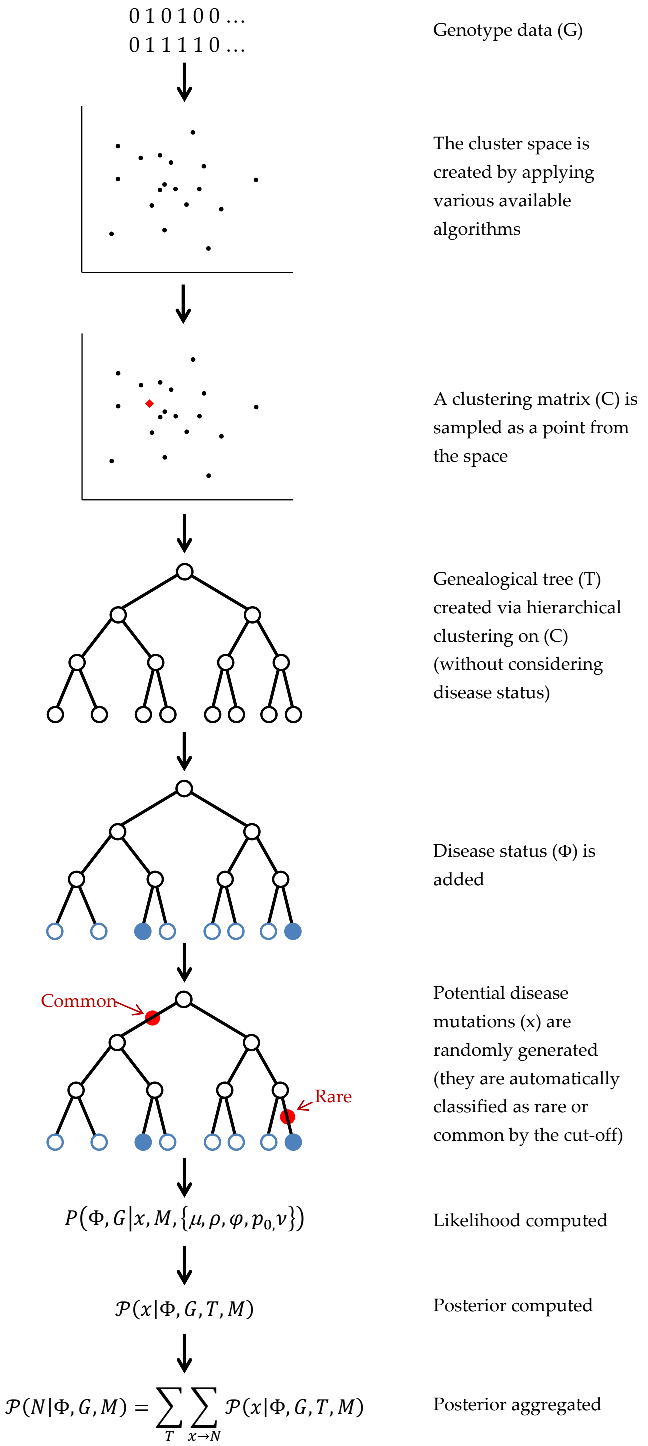
In the end, any clustering method applied on this similarity matrix will give a clustering on the subjects, corresponding to the chosen point in the 2-dim space. Thus, any point in the clustering space corresponds to a particular clustering of the subjects.

2.2.2 Adopting their method for our purpose

In the tree-based approach, we randomly generated ARGs [Griffiths and Marjoram, 1996] for all the subjects, and used it in our calculations. But as we said earlier, a genealogical tree is only a special case of providing a clustering. The space in the Grimmer-King algorithm from which we can randomly pick our clusters is a richer space, and we can use such clusters directly in our algorithm. This is so because, each clustering matrix, when run through any hierarchical clustering method, produces a dendrogram, which is equivalent to a bifurcating genealogical tree. Such trees will not have recombination events marked into them. Therefore, these trees are simpler than the ARGs.

Instead of an user clicking manually into the (visualized) 2-dimensional space, we pick points randomly from that space. Therefore, we need to put a probability distribution on the (projected) space of pre-computed clusterings. It seems that the bivariate uniform distribution, restricted to a large rectangle containing the convex hull of all the points, would be a reasonable choice. As we move inside the hull, we get closer to one or the other point, and our newly selected cluster will be give more weight to those points. So, if our selected point is close to a clustering method, the produced cluster will have similar clustering as that. On the other hand, if we move away from the points, towards the boundary, the weights will become nearly equal, and therefore the new clustering matrix will be a simple average of all the clustering matrices.

Even when we pick points randomly from this space, the space of all possible generated clusterings is a subspace of all possible clusterings. This is because, we are taking an weighted average of the clustering decisions made by the standard clustering algorithms, and not from random clusters. For example, if we have four data points $\{1100, 1100, 0011, 0011\}$, any standard clustering method will put the first two points in one cluster and the last two points in another. But the complete clustering space will also have points like $\{(1, 3), (2, 4)\}$ or $\{(1, 4), (2, 3)\}$. Such points will not come into our randomly selected clusters, as the clustering algorithms we use make 'informed decisions'. Therefore, we can avoid picking 'bad' clusters.



Genotype data (G)

The cluster space is created by applying various available algorithms

A clustering matrix (C) is sampled as a point from the space

Genealogical tree (T) created via hierarchical clustering on (C) (without considering disease status)

Disease status (Φ) is added

Potential disease mutations (x) are randomly generated (they are automatically classified as rare or common by the cut-off)

$$P\left(\Phi, G \middle| x, M, \{\mu, \rho, \varphi, p_{0,} v\}\right)$$

Likelihood computed

$$\mathcal{P}(x|\Phi, G, T, M)$$

Posterior computed

$$\mathcal{P}(N|\Phi, G, M) = \sum_{T} \sum_{x \to N} \mathcal{P}(x|\Phi, G, T, M)$$

Posterior aggregated

Figure 2.1: Flowchart for the algorithm.

2.2.3 Computing the Posterior

The posterior computation procedure is exactly the same as the previous paper, and is not discussed in detail here. The only difference is that, we are generating coalescent trees (T) directly from the simulated clusters (C), and not using ARGs as before. Therefore, we do not need to consider recombinations any more, and the recombination rate parameter ρ is out of the calculations.

2.2.4 The Steps for Computation

Here we outline the steps for computing the posterior $\mathcal{P}(N \mid \Phi, G, M)$, which is to be used for making inferences. These steps are also illustrated on the flowchart (figure 2.1) in the beginning of this section.

1. We use SNP haplotype data (G) to generate the clustering space by applying the various clustering algorithms.

2. We randomly generate clustering matrices (C) by sampling points from the clustering space.

3. Each simulated clustering matrix (C) corresponds to a genealogical tree (T) via applying any hierarchical clustering method.

4. Given a tree (T), we model the likelihood $P(x \mid T)$ of the putative DSL configurations (x), which depends on the probability of mutation at each site.

5. Next, we model the disease probabilities, $P(\Phi \mid x, T, M)$, where Φ is the disease status, M is the disease model.

6. They are used to simulate configurations of mutations (x) at probable DSLs.

Each configuration corresponds to a particular count of common and rare variants, $N = (N_r, N_c)$.

7. So the terms in the likelihood are: $P(\Phi \mid x, T, M), P(x \mid T), P(T \mid G)$. The complete likelihood $P(\Phi, G \mid x, M)$ aggregates the previous terms by summing over all possible T's.
   $P(\Phi, G \mid x, M) = \sum_T P(\Phi \mid x, M) \cdot P(x \mid T) \cdot P(T \mid G)$.

8. We use appropriate priors, e.g. uniform prior on trees (T), prior on recombination rate obtained from HAPMAP, etc.

9. We evaluate posterior probabilities $\mathcal{P}(x \mid \Phi, G, T, M)$ of such configurations (x), given the observed phenotypes (Φ), SNP data (G), and the tree (T).

10. Finally, we get posteriors $\mathcal{P}(N \mid \Phi, G, M)$ for variant configurations (N), by aggregating over corresponding configurations, and over simulated trees.
    $\mathcal{P}(N \mid \Phi, G, M) = \sum_T \sum_{x \to N} \mathcal{P}(x \mid \Phi, G, T, M)$.

The clustering space and generation of clustering matrices in steps 2–3 are performed by the Grimmer-King algorithm [Grimmer and King, 2011]. The mutations in step 6 are generated randomly on the branches of a given tree. Both these simulations are used for Monte Carlo estimates of probabilities by averaging, and therefore our computed posterior depends on the accuracy of the drawn samples — the number of draws and how well they span the sample space. As the sample space is finite in both cases, ensuring these criteria are much more straight-forward.

2.3 Data Analysis

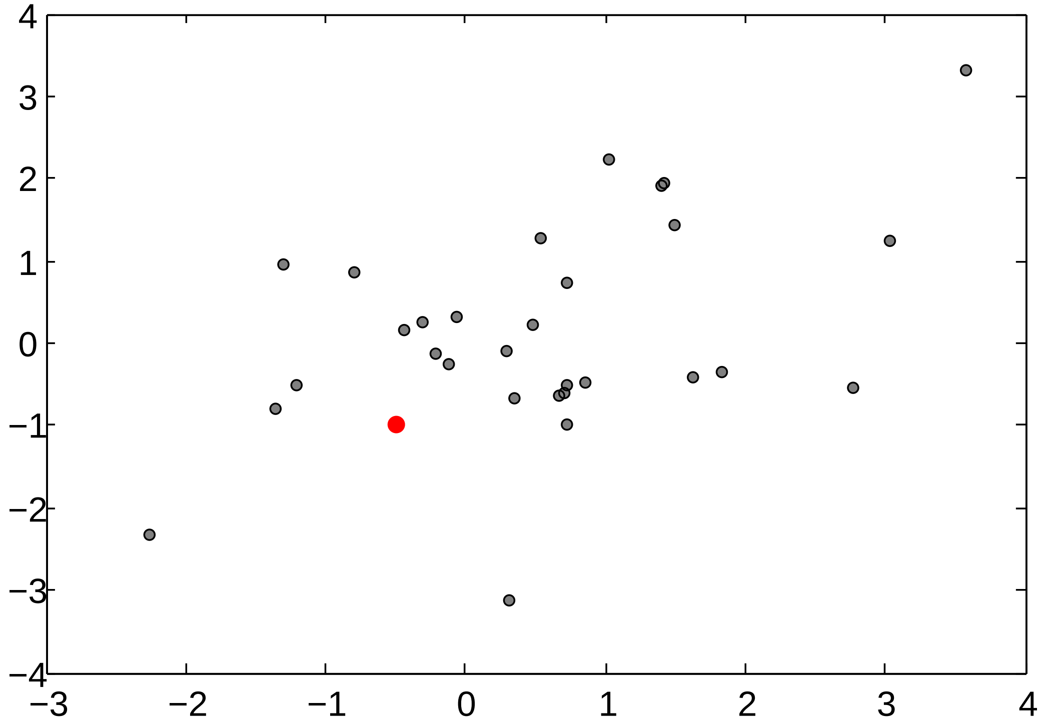


Figure 2.2: The clustering space generated by the different clustering methods, projected into a two-dimensional space. The grey points correspond to the different methods, and the red dot is a new sampled point, corresponding to a new clustering matrix.

We perform simulations with set-ups similar to paper 1, with similar parameter and sample size set-ups. The basic behavior of the results is similar, as expected — the bivariate posterior modes are able to distinguish between the null scenario, presence of rare variants, common variants, or both, with performance being increasingly better as sample size increases. In the following two plots, we compare two scenarios similar to the previous paper — one where the disease is caused by rare variants, and in the other case, by common variants. The behavior of the outputs is similar to the previous method, which isn't surprising, given that the probability models are the same — the tree structure differs due to the new the tree generation method. It seems that the posterior distribution has less variance compared to the previous method, which could be a consequence of the clustering algorithms picking 'good' clusters and therefore 'good' trees only (by 'good', we mean trees that better approximate the underlying structure). This is discussed in more detail in the next section.

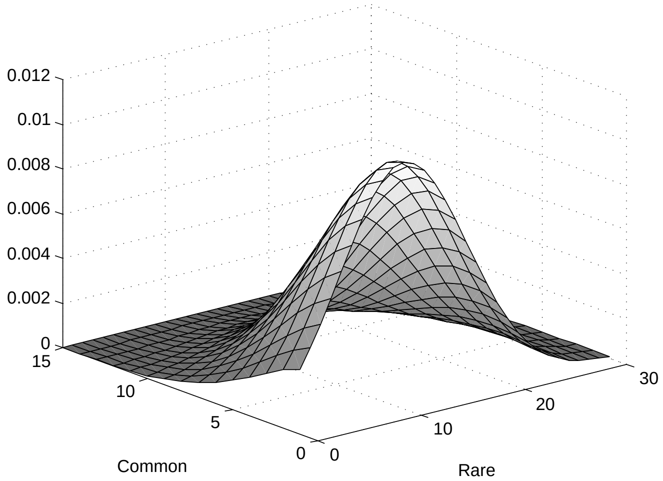


Figure 2.3: Posterior for simulated dataset with rare variants.

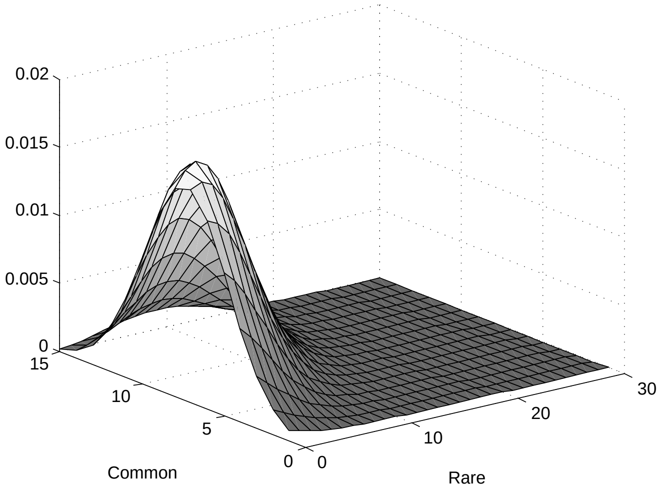


Figure 2.4: Posterior for simulated dataset with common variants.

2.4 Discussions

As it was remarked before, the sampled clusters are weighted combinations of existing clustering methods, which are generally considered to be ‘good’ clustering methods, that is, they somewhat accurately represent the underlying clustering structure, and are not haphazard in their output. Therefore, the sampled trees are also not exactly a uniform sample from the space of all trees, since the cluster sampling method does not strive to enforce that — as the topology of the clustering space is difficult to express, enforcing any distribution would be difficult too. However, aggregating the posterior over these random samples would still remain a valid procedure.

One advantage of using the clustering method over the ARGs is that we do not need to infer the haplotypes — the genotype vectors are directly usable as inputs for clustering. Therefore, the additional level of uncertainty introduced by the phasing is not an issue here. On the other hand, the clustering matrix size scales with n^2 where n is the sample size, whereas the size for the previous method scaled with the number of unique haplotypes, which is $\leq 2n$ and usually smaller. In contrast, clustering only weakly depends on the number of SNPs m, while the ARG method is much more affected when m increases. So this clustering approach will probably be more efficient when applied to rare variant (i.e. sequence) data.

3

Rare Variants? Let's go Nonparametric.

3.1 Introduction

Genetic sequencing for complex diseases is becoming increasingly feasible, with rapidly decreasing costs and increasing data quality. Rare variant testing using sequencing data has therefore gained momentum in the past few years, and a variety of different methods has been proposed for association testing of rare variants with complex diseases. As testing individual variants have very low power, as compared to Genome-Wide Association Studies (GWAS) using Single Nucleotide Polymorphism (SNP) data, almost every test involves some sort of collapsing of the rare variants or the test statistics into an univariate test statistic. We propose here a new non-parametric test statistic aimed at case-control studies, which is motivated by the well-known Kolmogorov-Smirnov non-parametric tests. While employing a non-parametric test implies robustness, i.e. no need to to model the data with distributional assumptions such as Gaussian or logistic models, simulation studies show that this test does not sacrifice on power — it is comparable to the standard rare variant tests currently used, in the sense that it has similar power to test like RBT ([Ionita-Laza et al., 2011]) or SKAT ([Wu et al., 2011]), and has better power than some other test such as C-alpha ([Neale et al., 2011]). On the other hand, another useful corollary of using the Kolmogorov-Smirnov structure is that it implicitly utilizes the natural ordering of the variants on the genome. This test is designed to handle the situations when both deleterious and protective variants are present. analytical results show some useful features of the test statistic, and the p-value can be calculated either mathematically or by permutation. Coupled with the ability of analytically obtaining the p-value, the procedure is computationally extremely fast.

3.2 Materials and Methods

We start by describing the notations used in the derivation of the test. Then we will describe the Kolmogorov-Smirnov test statistic, and present our statistic which is a modification of the K-S statistic aimed particularly at genetic data. In the end, we will show some asymptotics results which enable us to compute the p-value analytically.

3.2.1 Notation

Suppose we have a genomic region of length N base pairs. As N is very large, it can also be thought of as the interval $[0, 1]$ being discretized with a grid of size N. Therefore, the location i corresponds to a point $u = i/N$ in the interval.

For an individual j, the variable x_{ji} denotes if a variant is present at location i for the j^{th} individual. We can say that this follows a distribution function $F(u)$, which characterizes the probability of observing a variant at location u. We do not make any distributional assumptions regarding the form of $F(\cdot)$, since mutation rates can vary across the genome.

If F_1 is the distribution function for an affected individual, and F_0 for an unaffected individual, then the basic idea is to compare F_1 versus F_0. Under the null hypothesis that variants in that genomic region is not associated with the complex trait, $F_1(\cdot) \equiv F_0(\cdot) \equiv F(\cdot)$. This can be performed with a two-sample Kolmogorov-Smirnov test, for which asymptotic distributions are known and p-values can be computed analytically using numerical methods.

Suppose we have n_X subjects as cases and n_Y subjects as controls. Let us use the variable x for cases and y for controls. We define the variant counts at each location by

$$x_i = \sum_{j=1}^{n_X} x_{ji}, \qquad y_i = \sum_{j=1}^{n_Y} y_{ji}.$$

And we denote the cumulative counts as

$$X_i = \sum_{l=1}^{i} x_l, \qquad Y_i = \sum_{l=1}^{i} y_l.$$

Note that X_N denotes the total number of variants across all the subjects in cases, and Y_N denotes the same in controls. Using this, the empirical CDFs obtained over the grid are

$$\hat{F}_X(u) \equiv \hat{F}_X(i) = X_i/X_N, \hat{F}_Y(i) = Y_i/Y_N.$$

And the standard two-sample Kolmogorov-Smirnov test statistic would be

$$D_1 = \sup_{1 \leq i \leq N} \left| \hat{F}_X(i) - \hat{F}_Y(i) \right|.$$

3.2.2 The Proposed Test Statistic

Consider the situation where the mutation rate is the same across the entire genomic region, but the cases have a higher variant proportion, and consequently higher mutation rate, for each loci than the controls. This implies that the genomic region is associated with the disease trait. However, the normalized rate will be the same in both groups, and therefore the test statistic will not show any difference. To be able to reflect this situation in the test statistic we propose the following modification — instead of dividing by X_N in the cases, we only divide by the total number of subjects n_X, and similarly divide by n_Y instead of Y_N in the controls. The test statistic then becomes $\sup_i |X_i/n_X - Y_i/n_Y|$. But unlike D_1, it is not restricted to be in $[0, 1]$. So we define the new test statistic to be

$$D = \sup_i \left| \frac{X_i/n_X}{K} - \frac{Y_i/n_Y}{K} \right|, \quad \text{where } K = \max(X_N/n_X, Y_N/n_Y).$$

Corresponding to this, we can construct empirical CDFs

$$\hat{F}^*_X(i) = \frac{X_i/n_X}{K}, \quad \hat{F}^*_Y(i) = \frac{Y_i/n_Y}{K}, \quad 1 \leq i \leq N.$$

Actually, only one of them is a proper CDF, in the sense that it achieves probability 1 at the end of its domain, and the other one is an incomplete CDF or a subdistribution function. But we can easily extend them to become proper CDFs by adding an extra point $N+1$ to their domains and defining

$$\hat{F}_X^*(i) = \begin{cases} X_i/n_X K, & 1 \le i \le N, \\ 1, & i = N+1. \end{cases}, \quad \hat{F}_Y^*(i) = \begin{cases} Y_i/n_Y K, & 1 \le i \le N, \\ 1, & i = N+1. \end{cases}$$

Then we can use the standard definition of a K-S statistic to define

$$D = \sup_{1 \le i \le N+1} |\hat{F}_X^*(i) - \hat{F}_Y^*(i)| \equiv \sup_{1 \le i \le N} |\hat{F}_X^*(i) - \hat{F}_Y^*(i)|.$$

The previous step implies the following, which will be important later:

$$D \ge D_2 = \left| \frac{X_N/n_X}{K} - \frac{Y_N/n_Y}{K} \right|.$$

3.2.3 Decomposition

$$D = \sup_i |\hat{F}_X^*(i) - \hat{F}_Y^*(i)| = \sup_i \left| \frac{X_i/n_X}{K} - \frac{Y_i/n_Y}{K} \right| = \sup_i \left| \hat{F}_X(i) \frac{X_N/n_X}{K} - \hat{F}_Y(i) \frac{Y_N/n_Y}{K} \right|$$

As the expression is symmetric in X and Y, without loss of generality assume that K attains the maximum via X. Then,

$$D = \sup_i \left| \hat{F}_X(i) - \hat{F}_Y(i) \frac{Y_N/n_Y}{X_N/n_X} \right| = \sup_i \left| \left(\hat{F}_X(i) - \hat{F}_Y(i) \right) + \hat{F}_Y(i) \left(1 - \frac{Y_N/n_Y}{X_N/n_X} \right) \right|.$$

$$\therefore D \le \sup_i \left| \hat{F}_X(i) - \hat{F}_Y(i) \right| + D_2 \cdot \sup_i \hat{F}_Y(i) = D_1 + D_2.$$

So we obtain two-sided strict bounds for D:

$$D_2 \leq D \leq D_1 + D_2.$$

As one can see, the previous inequality shows that while D is similar to the K-S statistic D_1, they are not exactly equal.

3.2.4 Asymptotics for D_1

To do asymptotics, we consider X_N, Y_N to be large, which will be so when n and N are large. Furthermore, we require that X_N and Y_N are of the same order. We use the Kolmogorov-Smirnov statistic ([Kolmogorov, 1933]) — from the asymptotics derived by Smirnov ([Smirnov, 1948], [Doob, 1949]), we know that (conditional on X_N, Y_N)

$$\sqrt{\frac{X_N Y_N}{X_N + Y_N}} \cdot D_1 \to \sup |B(\cdot)|,$$

where $B(\cdot)$ is the Brownian bridge, and the limiting distribution is called the Kolmogorov distribution ([Donsker, 1952]). Let's denote $C = \sup |B(\cdot)|$; the probability distribution for C is easily computable numerically, and it is known that for this distribution, mean $\mu_C = \sqrt{\pi/2} \cdot \ln 2 = 0.8687$, variance $\sigma_C^2 = \pi^2/12 - \mu_C^2 = 0.26^2$ ([Wang et al., 2003]).

We start by deriving the distributions of X_N and Y_N under the null. Let $p_i = P(\text{an individual has a variant at location } i)$. The individual p_i's are assumed to be very small, that is, $p_i \approx 0$. And since N is large, we assume $P = \sum_{i=1}^{N} p_i$ is substantial. Then an appropriate model would be to take the total variant counts of an individual j, $x_j = \sum_i x_{ji} \sim Poisson(P)$.

The above definition of P assumes that variants at different locations are independent. Even if not, as long as the individual p_i's are small, a $Poisson(P)$ model is a valid model for some general P estimated from the data.

Then, under the null, every individual is i.i.d., and by CLT,

$$\sqrt{n_X}(X_N/n_X - P) \xrightarrow{d} N(0, P),$$

$$\sqrt{n_Y}(Y_N/n_Y - P) \xrightarrow{d} N(0, P).$$

Let us define $n = n_X + n_Y$. Let $c = 2n_X/n$, $\bar{c} = 2n_Y/n$, and define $\mu = P \cdot n/2$. If the sample sizes are equal, $c = \bar{c} = 1$. With this, we can approximately write,

$$\bar{X}_N = X_N/\mu \sim N(c, c/\mu), \quad \bar{Y}_N = Y_N/\mu \sim N(\bar{c}, \bar{c}/\mu).$$

Or, with the CLT notation, as cases and controls are independent, we have

$$\sqrt{\mu}\left(\begin{pmatrix} \bar{X}_N \\ \bar{Y}_N \end{pmatrix} - \begin{pmatrix} c \\ \bar{c} \end{pmatrix}\right) \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} c & 0 \\ 0 & \bar{c} \end{pmatrix}\right).$$

Now we work with D_1. Conditional on X_N, Y_N, we have seen that $D_1 \to A \cdot C$, where

$$A = \sqrt{\frac{X_N + Y_N}{X_N Y_N}} = \frac{1}{\sqrt{\mu}}\sqrt{\frac{\bar{X}_N + \bar{Y}_N}{\bar{X}_N \bar{Y}_N}} = \frac{1}{\sqrt{\mu}}\sqrt{\frac{1}{\bar{X}_N} + \frac{1}{\bar{Y}_N}}.$$

Applying the multivariate delta theorem on the previous expression, we get

$$\sqrt{\mu}\left(\sqrt{\frac{1}{\bar{X}_N} + \frac{1}{\bar{Y}_N}} - \sqrt{2k}\right) \xrightarrow{d} N\left(0, \frac{1}{4}\left(\frac{1}{c^2} + \frac{1}{\bar{c}^2} - \frac{1}{c\bar{c}}\right)\right),$$

where $k = \frac{1}{2}\left(\frac{1}{c} + \frac{1}{\bar{c}}\right)$. Note that k is the inverse of the harmonic mean of c and $\bar{c}$, and therefore by the A.M. $\geq$ G.M. $\geq$ H.M. inequality, $k \geq 1$. When we have equal sample sizes for both cases and controls, $k = 1$, and

$$\sqrt{\mu}\left(\sqrt{\frac{1}{\bar{X}_N} + \frac{1}{\bar{Y}_N}} - \sqrt{2}\right) \xrightarrow{d} N\left(0, \frac{1}{4}\right).$$

This implies, $\sqrt{\mu}\left(A - \dfrac{\sqrt{2k}}{\sqrt{\mu}}\right) \xrightarrow{P} 0$. Recall that, as n is large, we consider μ to be large as well.

So, comparable to the one-sample notation of $\sqrt{n}D_n \to C$, we get
$\sqrt{\mu}D_1 \cdot \dfrac{1}{\sqrt{2k}} \to C$.

3.2.5 Asymptotics for D_2

$$D_2 = \left|\frac{X_N/n_X - Y_N/n_Y}{\max(X_N/n_X, Y_N/n_Y)}\right|.$$

We define $\tilde{X}_N = X_N/(n_X \cdot P) = X_N/(c \cdot \mu) = \bar{X}_N/c$. Similarly
$\tilde{Y}_N = Y_N/(n_Y \cdot P) = \bar{Y}_N/\bar{c}$. Then, using the previous asymptotic expression,

$$\sqrt{\mu}\left(\begin{pmatrix} \tilde{X}_N \\ \tilde{Y}_N \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/c & 0 \\ 0 & 1/\bar{c} \end{pmatrix}\right).$$

From the denominator of D_2, we can see that we need the joint distribution of order statistics. Suppose X_1, X_2 are two independent Gaussian variables with mean 0, and their variances are a, b respectively. Let Z_1, Z_2 be the two order statistics, Z_1 the minimum and Z_2 the maximum. Then, we know that, ([Nadarajah and Kotz, 2008], [Jones, 1948]),

$$E(Z_2) = -E(Z_1) = \frac{\sqrt{a+b}}{\sqrt{2\pi}}, \mathrm{Var}(Z_2) = \mathrm{Var}(Z_1) = \frac{1}{2}\left(1 - \frac{1}{\pi}\right)(a{+}b), \mathrm{Cov}(Z_1, Z_2) = \frac{1}{2\pi}(a{+}b).$$

Here, if we define Z_1, Z_2 to be the two order statistics corresponding to $\tilde{X}_N, \tilde{Y}_N$, we have

$$\sqrt{\mu}\left(\begin{pmatrix} Z_2 \\ Z_1 \end{pmatrix} - \begin{pmatrix} 1 + \frac{\sqrt{k}}{\sqrt{\pi\mu}} \\ 1 - \frac{\sqrt{k}}{\sqrt{\pi\mu}} \end{pmatrix}\right) \xrightarrow{d} N_+\left(0, k\begin{pmatrix} 1 - \frac{1}{\pi} & \frac{1}{\pi} \\ \frac{1}{\pi} & 1 - \frac{1}{\pi} \end{pmatrix}\right).$$

Note: here we are using the notation N_+ because the distribution only has support in the region $Z_2 \geq Z_1$.

In this set-up, $D_2 = 1 - \frac{Z_1}{Z_2}$. So, we use the multivariate delta theorem to obtain

$$\sqrt{\mu}\left(\frac{Z_1}{Z_2} - \frac{\sqrt{\pi\mu} - \sqrt{k}}{\sqrt{\pi\mu} + \sqrt{k}}\right) \xrightarrow{d}$$

$$N_+\left(0, \frac{k\mu}{(\sqrt{\pi\mu} + \sqrt{k})^2}\left[(\pi - 1) - 2\frac{\sqrt{\pi\mu} - \sqrt{k}}{\sqrt{\pi\mu} + \sqrt{k}} + (\pi - 1)\left(\frac{\sqrt{\pi\mu} - \sqrt{k}}{\sqrt{\pi\mu} + \sqrt{k}}\right)^2\right]\right).$$

As μ is large, we can ignore the higher order terms in the variance to simplify the expression. This leads to

$$D_2 \approx N_+\left(2\frac{\sqrt{k}}{\sqrt{\pi\mu} + \sqrt{k}}, (2\pi - 4)\frac{k}{(\sqrt{\pi\mu} + \sqrt{k})^2}\right)$$

or

$$\sqrt{\mu}D_2 \approx N_+\left(2\sqrt{\frac{k}{\pi}}, (2\pi - 4)\frac{k}{\pi}\right).$$

Here too we use N_+ to denote the fact that this distribution is truncated at 0.

3.2.6 Testing from a Truncated Normal

An important point we need to remember is that, both components of D are truncated at 0, and in particular, the distribution of D_2 is a truncated normal distribution ([Barr and Sherrill, 1999]). So, its derived expectation and variance are under truncation as well. For example, if $X \sim N(\mu, \sigma^2)$, and Y is X truncated at a, then

$$E(Y) = \mu_* = \mu + \beta\sigma, \text{Var}(Y) = \sigma_*^2 = \sigma^2(1 + \alpha\beta - \beta^2), \text{ where } \alpha = \frac{a - \mu}{\sigma}, \beta = \frac{\phi(\alpha)}{\Phi(-\alpha)}.$$

The mean and variance derived in the asymptotics above are actually μ_* and σ_*. From these, the original μ and σ can be obtained numerically.

Then, we can do a one-sided test by obtaining a cut-off τ for the upper tail, such that

$$\tau = \mu + \sigma\Phi^{-1}\left(0.95 + 0.05\Phi(\alpha)\right).$$

3.2.7 Behavior of the Test Statistic

The derivation shows that the test statistic is controlled by two different components — the Kolmogorov-Smirnov statistic D_1 depending on the difference in the way variant proportions are distributed among cases and controls, and the statistic D_2 taking care of the overall difference in variant counts. Under the null both terms have an equal effect — $\sqrt{\mu}D_1$ and $\sqrt{\mu}D_2$ both have distributions of same order. On the other hand, different alternative scenarios will lead to different components dominating the test statistic. If there is a difference in the number of variants among the two groups, D_2 will dominate, and the test asymptotically will have power 1 as sample size goes to infinity. Even if not, as long as there is a difference between variant proportion distribution in the two groups, the presence of D_1 will still lead the asymptotic power towards 1.

For example, consider the situation of $n = 2, N = 5$, and the observed x vector is $(0, 1, 0, 0, 0)$, and $y = (0, 0, 1, 0, 0)$. Then, $X_N = Y_N = 1$, so that $D_2 = 0$. But $F^*_X = (0, 1, 1, 1, 1)$ and $F^*_Y = (0, 0, 1, 1, 1)$, so that $D = 1$. Here, the entire contribution comes from the Kolmogorov-Smirnov part. In this case, therefore, the statistic successfully distinguishes between the cases, which have a deleterious variant at location 2, and controls, which have a protective variant at location 3.

In the following two diagrams (3.1, 3.2), we illustrate the two situations where the test statistic will be powerful. The null situation is presented in the third diagram (3.3) for comparison.

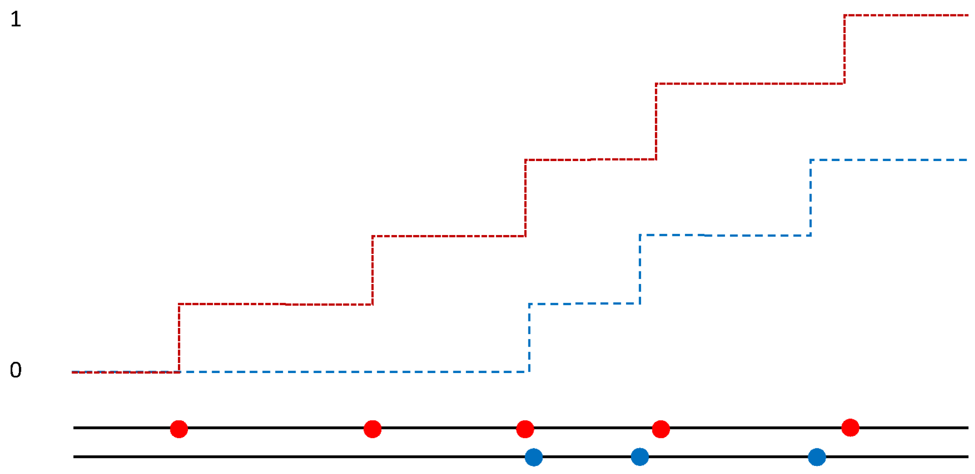


Figure 3.1: Example 1: Cases and controls have different mutation rates.

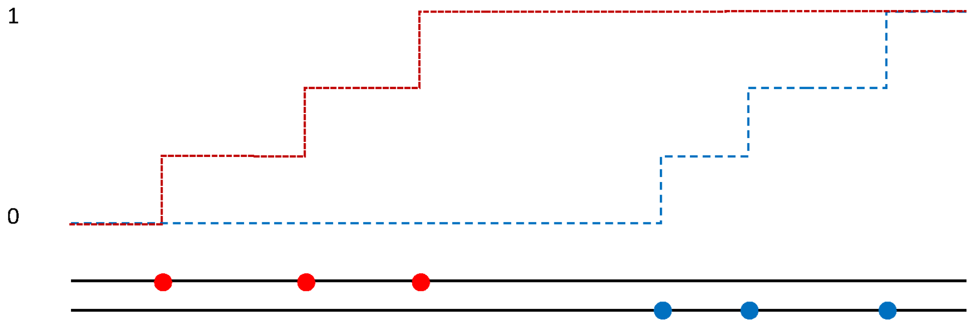


Figure 3.2: Example 2: Cases and controls have similar mutation rates overall, but the mutation regions are different.

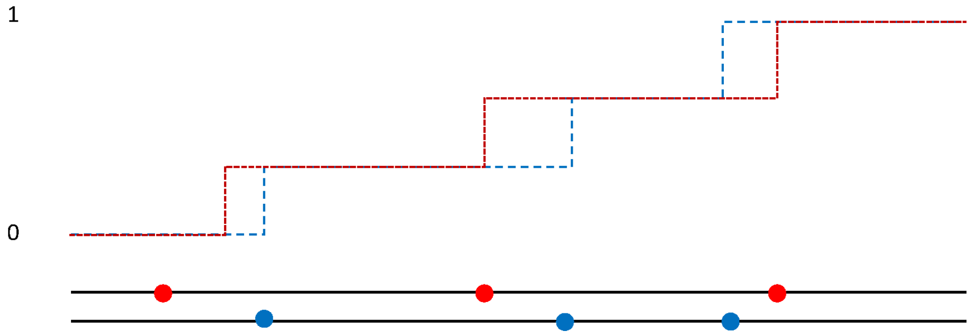


Figure 3.3: Example 3: No difference between mutation rates in cases and controls.

This is also shown in the next two simulations results — the first scenario (3.4) is where the mutation rates are different, and in the second scenario (3.5), cases and controls have different mutation rates in clusters. On the left, the case and control CDF's are shown, and on the right, the permutation distribution of the test statistic is shown. The red star denotes the observed value of the test statistic, and the magenta circle denotes the 5% cut-off.

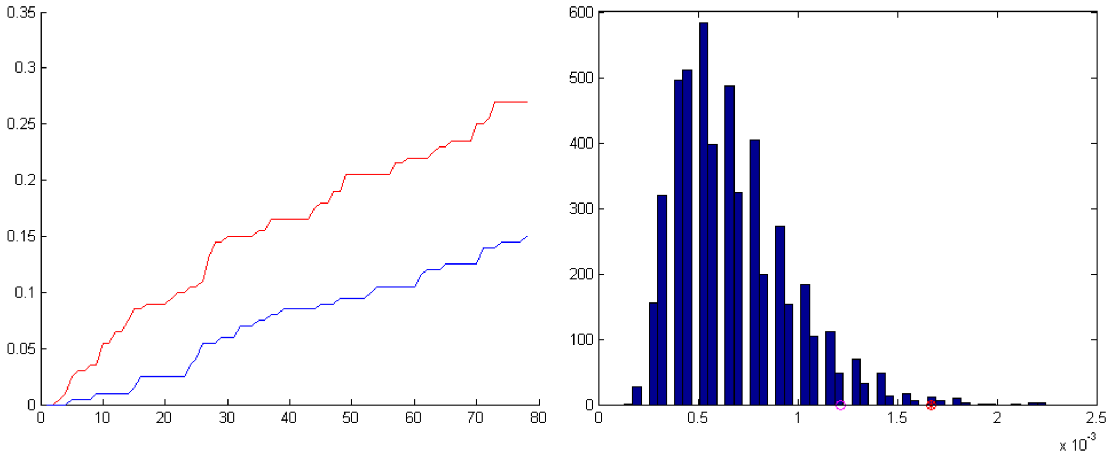


Figure 3.4: Example 1: Case and control mutation rates are different.

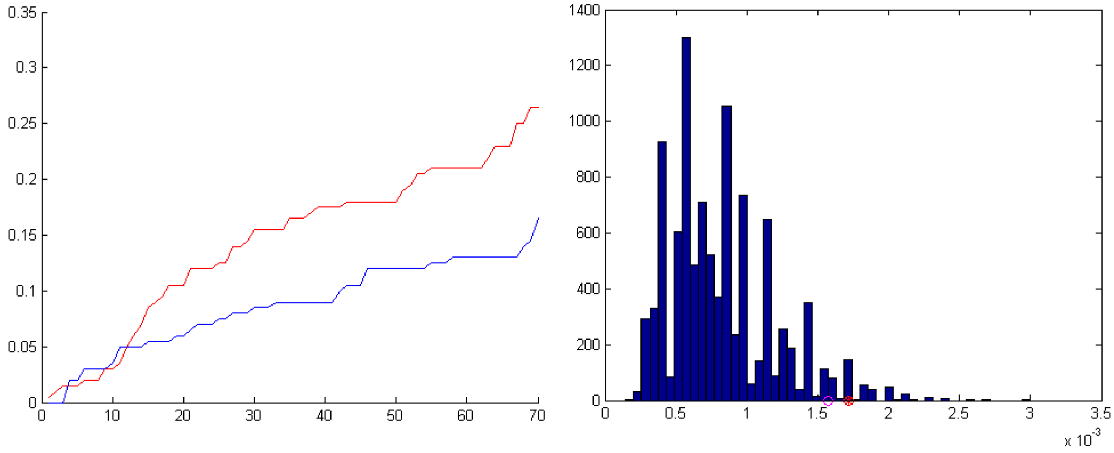


Figure 3.5: Example 2: Cases and controls have mutation rates varying in two separate clusters.

The asymptotics for D_1 assumes continuity for the underlying probability distribution, i.e. p_i's are very small. This is ensured by the rare variant assumption. Moreover, common variants lead to a large jump in the empirical CDF, irrespective of whether it is causal or not. So we have decided to remove all the variants with MAF $> 1\%$ from the data. Although having only rare variants does not exactly imply the continuity assumption — every variant does not occur in an unique location — but that is not a serious issue. Even if the underlying mutation distribution has discrete mass at some points, it has been shown that the Kolmogorov-Smirnov test is still a valid test under such circumstances, while being conservative ([Conover, 1972]). It is also known that for large data size, the loss in power is not very large ([Horn, 1977]), especially when the individual probabilities are not large, which will be the case when we remove the common variants.

Due to the presence of μ in the asymptotics, which involve the mutation rate P, this test is no longer an exactly nonparametric test. But that is acceptable, because mutation rate is an important parameter, and difference in mutation rates is precisely the reason why we modified our statistic to be different from the standard Kolmogorov-Smirnov statistic. Even though it is not exactly nonparametric anymore, the test still retains most of the robustness properties of the Kolmogorov-Smirnov test — it still does not make any distributional assumptions about the mutation rate across the genomic region, neither about the LD pattern, nor about effect sizes/directions.

3.3 Comparison to Other Rare Variant Tests

In order to avoid the burden of high dimensionality, most rare variant tests apply pooling of some sort. The collapsing-based Cohort Allelic Sums Test (CAST) ([Morgenthaler and Thilly, 2007]) and Combined Multivariate and Collapsing (CMC) method ([Li and Leal, 2008]) were among the first to address the rare variants issue specifically, and there have been many other proposed tests, each designed to take advantage of a particular situation, with particular models and assumptions. For example, the Replication-Based Test (RBT) method ([Ionita-Laza et al., 2011]) assumes Poisson probability models for variant counts at each loci; the Sequential Kernel Adaptive Test (SKAT) method ([Wu et al., 2011]) uses logistic regression to model the disease probability given each variant (the weighted combination of individual logistic score tests is equivalent to a overall variance component score test). [Asimit and Zeggini, 2010], [Bansal et al., 2010], [Basu and Pan, 2011] are some good reviews of the various rare variant methods proposed over time. It is worthwhile to mention that there are also some Bayesian approaches to rare variant analysis ([Shen et al., 2011]), even if those are not as popular, and are generally not covered by these review papers.

While tests like CMC combine the variant genotypes directly, other tests such as SKAT combine individual test statistics (scores or variances) corresponding to the variants. According to a review by [Pongpanich et al., 2012], neither method is uniformly the best — genotype collapsing is more sensitive to presence of noise (non-causal loci), while score collapsing is more robust to variations in effect sizes, association directions, MAFs, and LD.

Recently, it was highlighted that rare variants can have effects in both directions — some could be harmful and some could be protective towards the disease. Therefore, newer tests were designed which would remain powerful under such scenarios. The test we propose is also motivated by that fact, and this is where the robustness of the Kolmogorov-Smirnov method is effective — no specification of the underlying variant distribution is necessary.

Frequently, weights are used in the test statistics to up-weight the rare variant contributions — e.g. the Weighted-Sum method ([Madsen and Browning, 2009]) uses weights inversely proportional to the standard deviation under the binomial model, i.e. $w = 1/\sqrt{p(1-p)}$ where the MAF p is generally estimated from the controls. [Price et al., 2010] showed that using such an weight function implies this assumption on the log odds ratio of the disease — $\log(OR) \propto 1/\sqrt{p(1-p)}$. The SKAT method uses heuristic weights derived from beta distributions with MAF as the parameter, and the RBT method uses Poisson probability of the rare variant counts as weights.

Some tests are generalized to include covariates or interaction terms, and some are applicable to quantitative traits as well. At this point, we do not extend our test to those directions.

In this paper, we compare our method to the C-alpha test ([Neale et al., 2011]), RBT ([Ionita-Laza et al., 2011]), and SKAT ([Wu et al., 2011]). Among the more well-known tests, we do not consider the CMC ([Li and Leal, 2008]), Kernel Machine Regression (KMR) method ([Liu et al., 2008]), Kernel-Based Adaptive Cluster (KBAC) method ([Liu and Leal, 2010]), Step-up method ([Hoffmann et al., 2010]), Variable Threshold (VT) test ([Price et al., 2010]). We chose to compare to those tests because they allow for variant effects in both directions.

In the review by [Basu and Pan, 2011], they provide simulation results showing power calculations for various test statistics. We are only interested in the part where effects in both directions are considered — otherwise, as expected, simple genotype collapsing methods perform the best. They did not consider the SKAT and VT methods; among the rest, KBAC, KMR, C-alpha and RBT were the best performers, and had similar powers even though they are based on different approaches — genotype vs. score collapsing. Since they had similar performances, we chose to only consider C-alpha and RBT, each representing a different collapsing approach.

Method	Author	Year	+/−	Collapsing	Model	Weight	P-value
CAST	Morgenthaler & Thilly	2007	×	Genotype	Logistic	None	Analyt.
CMC	Li & Leal	2008	×	Genotype	Linear	None	Analyt.
KMR	Liu et al.	2008	Yes	Genotype	Logistic	None	Analyt.
Weighted-Sum	Madsen and Browning	2009	×	Genotype	Rank	Yes	Perm.
KBAC	Liu & Leal	2010	×	Genotype	Hyper-geom.	Yes	Perm.
Step-up	Hoffmann et al.	2010	Yes	Genotype	Logistic	Yes	Perm.
VT	Price et al.	2010	×	Genotype	Linear	Yes	Perm.
RBT	Ionita-Laza et al.	2011	Yes	Genotype	Poisson	Yes	Perm.
SKAT	Wu et al.	2011	Yes	Score	Logistic	Yes	Analyt.
C-alpha	Neale et al.	2011	Yes	Score	Binomial	None	Perm.
wSSU	Basu & Pan	2011	Yes	Genotype	Logistic	Yes	Perm.
Seq-Sum-VS	Basu & Pan	2011	Yes	Genotype	Logistic	None	Perm.
EREC	Lin & Tang	2011	Yes	Genotype	Logistic	Yes	Perm.
Bayesian	Shen et al.	2011	Yes	Genotype	Lognormal	None	Perm.
K-S Test	—	2012	Yes	Genotype	None	None	Analyt.

3.4 Application to Real Data

Nonsyndromic cleft lip with or without cleft palate (NSCL/P) is a common congenital malformation that is caused by an interplay of multiple genetic and environmental factors [Mossey et al., 2009]. Our dataset comprises 96 NSCL/P cases and 96 controls of Central European ethnicity in whom the exonic and adjacent intronic regions of one candidate gene for NSCL/P has been sequenced. The gene was among the candidate regions in an independent GWA study [Mangold et al., 2009]. Moreover, this gene has functional importance, as it codes for a protein which is involved in bone development, and is therefore relevant for further analysis.

There was a single missing SNP in a person, and it was imputed using PHASE [Stephens et al., 2001], [Stephens and Donnelly, 2003]. The recombination rates estimated [Li and Stephens, 2003], [Crawford et al., 2004] from the data by PHASE tally with those from the CEU population of HAPMAP (phase II) [The International HapMap Consortium, 2007], which comments favorably on our data quality.

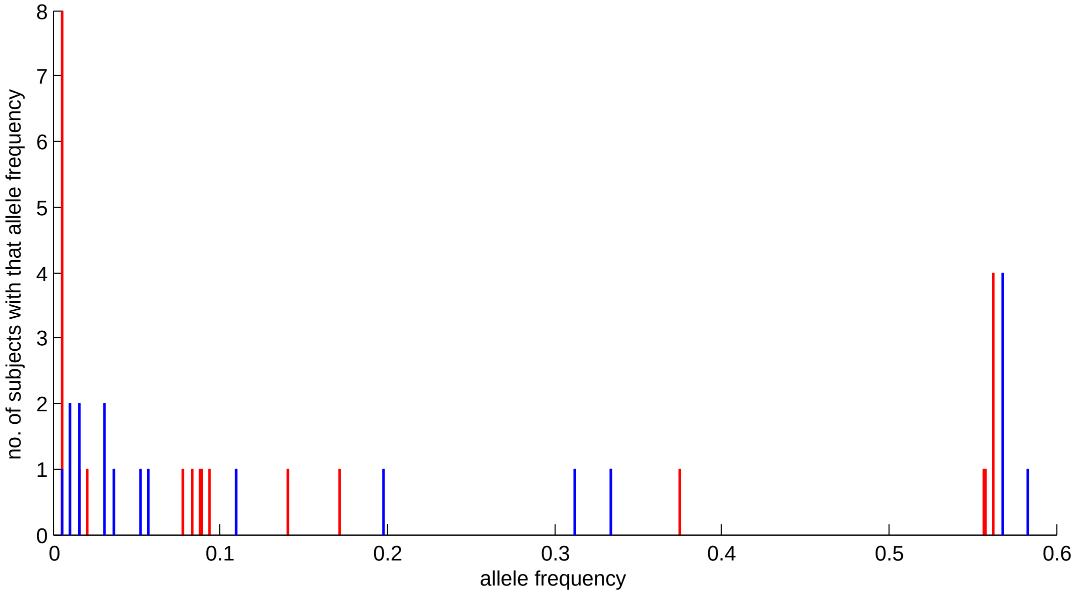


Figure 3.6: The case and control allele frequency distributions for the real data. The bin locations for this histogram are the specifici allele frequencies, and for each particular allele frequency the height of the bar denotes the number of subjects who have a variant with that allele frequency. Red denote cases and blue denote controls. The lowest allele frequency corresponds to variants which are present only once in 96 subjects (0.5%) — we have 8 such loci in cases, and 1 in controls. In the middle region, we can see common variants which differ in allele frequencies in cases and controls — the location of the red and blue bars are different.

On a cursory comparison of the allele frequencies between cases and controls (3.6), it seems that lower-frequency SNPs have comparatively higher relative difference in terms of allele frequencies between the two groups, as compared to higher-frequency i.e. common variants. This implies that there should be some effect from rare variants. But as the number of cases is only 96, for rare variants with MAF $< 1\%$, on average we expect to see them in ≤ 2 people. On the other hand, with only 27 SNPs, the number of rare variant loci is low as well. Overall, this implies a high signal-to-noise-ratio, and therefore low power for association detection. Still, our test statistic shows significant association between the disease status and rare variants, which is something we expect to see, based on our knowledge of the data. So we can conclude that the method works reasonably with this small sample too.

With only the rare variants selected, we compare our method to the three different methods mentioned in the previous section and show the corresponding p-values in the table below. As this test has the lowest p-value, it might be the case that this test has more power for this situation, although we can not make a general statement with just one dataset.

Method	P-value
K-S test	0.04
RBT	0.20
SKAT	0.43
C-alpha	0.50

3.5 Simulation Study

3.5.1 Simulation Set-up

We use detailed simulations to compare the power of our test statistic to the C-alpha, RBT and SKAT methods. There are a huge variety of simulation methods in practice, with almost every paper having their own simulation method, and in general the simulation set-ups correspond to the model assumptions in the test statistic, e.g. logistic link function for disease probabilities when logistic regression model is used in the test. To accommodate this, we use not one but two simulation scenarios:

1. In the set-up of [Basu and Pan, 2011], [Pan, 2009], they use a multivariate normal distribution $\underline{Z} = (Z_1, \ldots, Z_m)$ as latent variables for the genotypes. A first-order autoregressive (AR1) covariance structure is used: $\text{Corr}(Z_i, Z_j) = \rho^{|i-j|}$, with ρ varied from 0 to 0.9. For each loci, these are dichotomized using a quantile obtained from a desired MAF (randomly generated to be $\leq 1\%$) value to produce the genotypes $\underline{X} = (X_1, \ldots, X_m)$. With chosen parameters $\underline{\beta} = (\beta_1, \ldots, \beta_m)$, they used a logistic model $\text{logit } P(Y = 1) = \beta_0 + \underline{X}\underline{\beta}'$ to generate the disease status. Some components in β are set to 0 to represent non-causal loci, while some are positive and some negative, to indicate different association directions. As logistic and probit models are quite similar, we chose to use the probit link instead, as it makes simulations easier.

2. The previous simulation method is simplistic in the sense that it does not consider population genetics models like coalescent genealogy or the Wright-Fisher allele frequency distribution model [Balding et al., 2007]. As some simulation studies ([Ionita-Laza et al., 2011], [Price et al., 2010]) use such models to generate data, we considered it as an alternative simulation scenario. Specifically, we used MaCS (Markovian Coalescent Simulator, [Chen et al., 2008]) to simulate genomic sequences under coalescent model with a Markov process. Standard demographic parameters, as discussed in ([Price et al., 2010], [Chen et al., 2008]) were used. Then, with randomly generated effect sizes for selected causal loci, the overall disease risk of an individual is calculated by summing over all the loci, and then the disease status is generated by a Bernoulli model.

Sample size is varied from 300 to 1500 cases, and an equal number of controls. Number of loci is varied from 10 to 100, and the number of causal loci is varied accordingly. Effect sizes and other simulation parameters were also varied. We consider five different parameter set-ups to include the different situations:

1. Simulation under the null,
2. Under the alternative of only harmful variants,
3. Under the alternative of both protective and deleterious variants,
4. Protective and deleterious variants in separate zones on the genomic region,
5. With causal common variants in addition to the causal rare variants.

3.5.2 Simulation Results

Firstly, we verified that the test maintains the proper alpha level — when tested at 5% level under the null in our simulations, the rejection rate never exceeded 5%. This is also shown and explained in figure 3.8.

Now we show the behavior of this K-S test statistic while varying different simulation parameters. First, we show how the power increases as sample size is increased. In 3.7, the X-axis shows the sample size in cases (we took equal sample size in controls) versus the power of the test, for a fixed number of variant loci – 15 – out of which 3 are causal. The standard Neyman-Pearson testing scenario has power growing with sample size as the order of $\sqrt{n}$, and we see a similar relationship here.

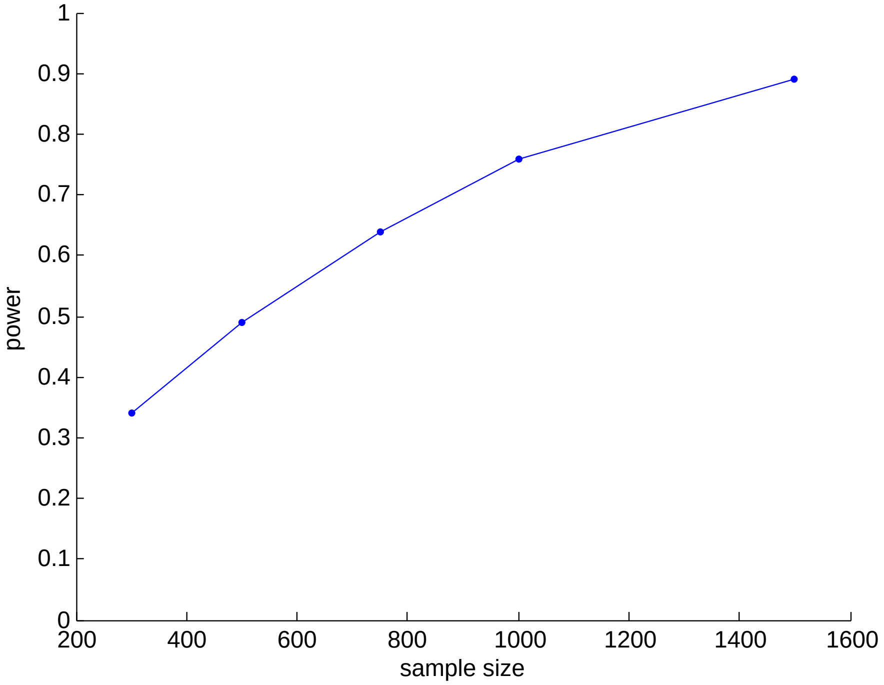


Figure 3.7: Plot of power versus sample size.

Next, we show how the power increases as the number of causal variants increase. In the next figure (3.8), we consider the scenario with 500 cases and 500 controls, and 15 loci. The number of causal variants among these 15 is varied from 0 to 7. At 0, this represents the level of the test, which is 5%. And when we have around half of the variants to be causal, the power reaches 1.

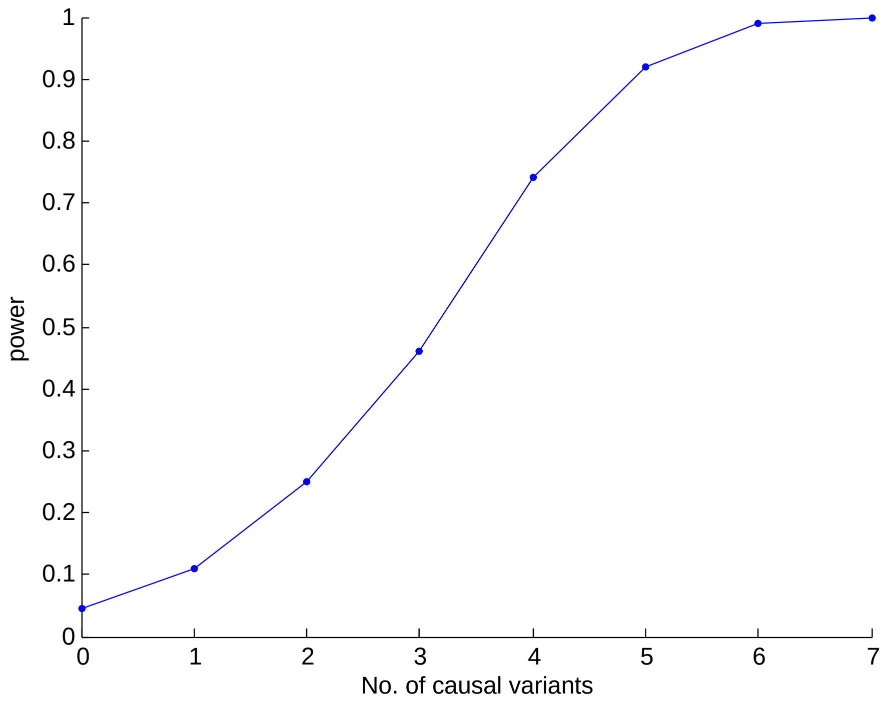


Figure 3.8: Plot of power versus number of causal variants.

It is intuitive that if we vary it the other way — if we keep the number of causal loci fixed but increase the total number of loci, then we will lose power, because the increasing number of non-causal loci will wash away the contribution of the causal loci to the test statistic. This is also discussed in [Pongpanich et al., 2012] as a common issue in rare variant testing, and it is observed that this is more of a problem in genotype-based collapsing methods. In figure 3.9, we again consider 500 cases and 500 controls, and fix 5 rare variants to be causal. Then, we vary the total number of loci from 5 (i.e. all variants are causal) to 100. The decrease in power is shown in the graph.

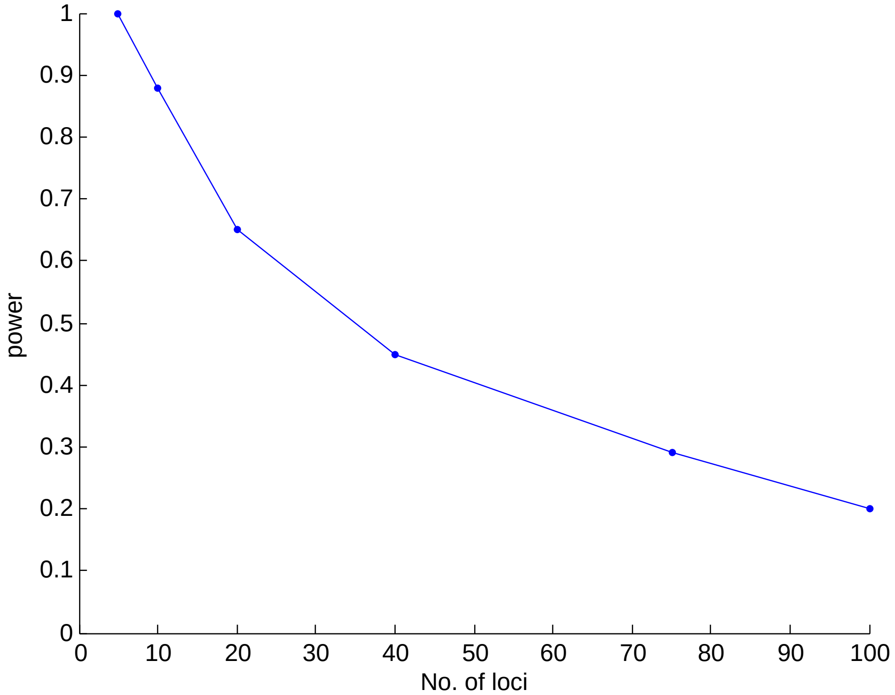


Figure 3.9: Loss of power as number of non-causal variants increase.

It was discussed earlier that, being location-based, this test will gain power when harmful and protective variants are clustered together on the genome, rather than having random locations. In the next figure (3.10), we show the power gain (via ratio) of the test when simulations are performed with random locations of causal variants versus when variants of same effect directions are clustered. The sample size is varied from 300 to 1500. As expected, the power in both situations tend to 1 as sample size increases, and so the ratio decreases with increase in sample size.

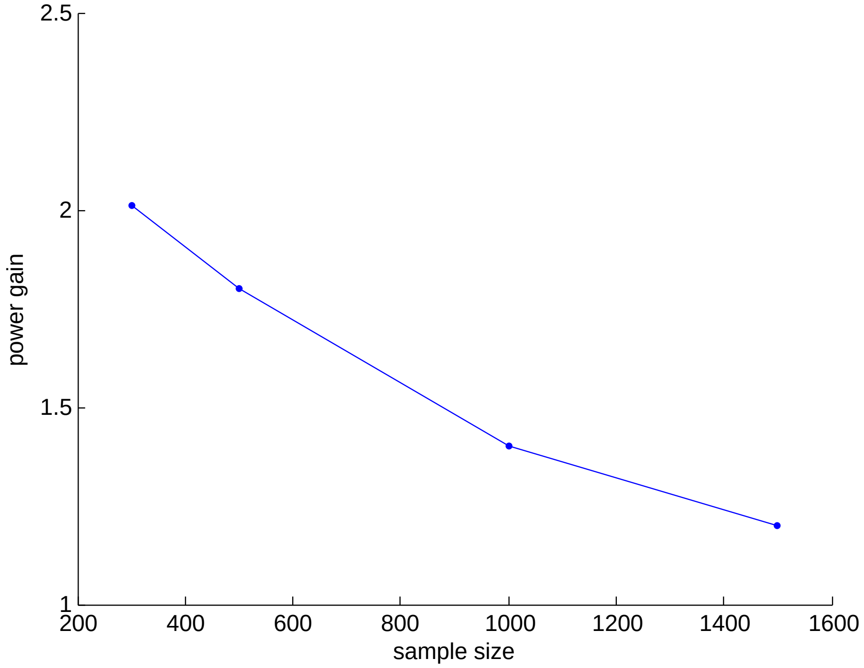


Figure 3.10: Gain in power when harmful and protective variants are cluster in separate regions on the genome.

Lastly, we compare the performance of this statistic with the three other rare variant testing methods (RBT, C-alpha and SKAT) via simulation study. We fix 500 cases and 500 controls, and 10 variant loci at random. Then, we vary the number of causal variant loci from 1 to 7, and as expected, the power increases for all the tests. In the graph (3.11), we see that the power of the K-S test is always comparable to the RBT and SKAT tests, which were found to be among the more powerful tests in the review papers. The power of the C-alpha test does not seem to grow as fast as the other three tests.

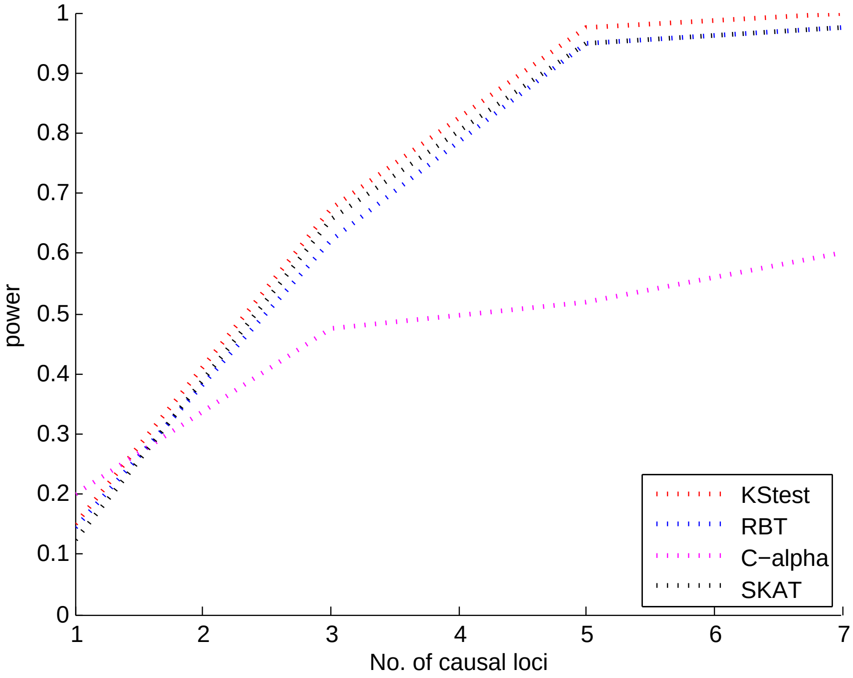


Figure 3.11: Comparing power for the different rare variant tests.

3.6 Discussion

We have presented a modification of the Kolmogorov-Smirnov test suitable for rare variant association testing. It is a robust, nonparametric test, able to work with variant effects in both directions, and any mutation distribution. We compared this method to three well-known rare variant tests, covering both genotype-based and score-based collapsing methods. It is seen that the method has similar power compared to the other rare variant tests in various situations, and as expected from the structure of the test, does well when variants of different effect directions are clustered separately.

As mentioned in the review by [Pongpanich et al., 2012], genotype collapsing methods such as this one is more sensitive to increasing number of non-causal loci. However, as we do not up-weight the rarer variants, this means that most noise is not up-weighted, and therefore the power loss is not very high. The robustness of the Kolmogorov-Smirnov method also helps in this regard. Overall, this method compares favorably to the RBT or SKAT method, and often performs better than the C-alpha test.

Due to the structure of the Kolmogorov-Smirnov test, this method would be most powerful when variants with same effect directions occur together on the genomic region. This is an unique feature of this test, in contrast to the previous test statistics, which are independent of the variant locations. As expected, this test gains power under such scenarios. And that might have a genetic basis, especially when functional information form bioinformatics is borrowed. Even when the location ordering of harmful and protective variants are mixed, the test still has decent power.

Finally, it is computationally very fast, and analytical p-values are available.

References

Asimit, J. and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. *Annual Review of Genetics*, 44(1):293–308.

Balding, D. J., Bishop, M., and Cannings, C. (2007). *Handbook of Statistical Genetics*. Wiley-Interscience, 3 edition.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, 11(11):773.

Barr, D. R. and Sherrill, E. T. (1999). Mean and variance of truncated normal distributions. *The American Statistician*, 53(4):357–361. ArticleType: research-article / Full publication date: Nov., 1999 / Copyright © 1999 American Statistical Association.

Basu, S. and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology*, 35(7):606–619.

Bodmer, W. and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*, 40(6):695–701.

Chen, G. K., Marjoram, P., and Wall, J. D. (2008). Fast and flexible simulation of DNA sequence data. *Genome Research*, 19(1):136–142.

Chen, G. K., Marjoram, P., and Wall, J. D. (2009). Fast and flexible simulation of dna sequence data. *Genome Research*, 19(1):136–142.

Conover, W. (1972). A kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association*, page 591–596.

Crawford, D. C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M. J., Nickerson, D. A., and Stephens, M. (2004). Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genetics*, 36(7):700–706. PMID: 15184900.

Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D. B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biology*, 8(1):e1000294.

Donsker, M. (1952). Justification and extension of doob's heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of mathematical statistics*, 23(2):277–281.

Doob, J. L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of Mathematical Statistics*, 20(3):393–403. Mathematical Reviews number (MathSciNet): MR30732; Zentralblatt MATH identifier: 0035.08901.

Griffiths, R. C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 3(4):479–502. PMID: 9018600.

Grimmer, J. and King, G. (2011). General purpose Computer-Assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, 108(7):2643–2650.

Gusfield, D., Eddhu, S., and Langley, C. (2004). Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *Journal of Bioinformatics and Computational Biology*, 2(1):173–213. PMID: 15272438.

Hoffmann, T., Marini, N., and Witte, J. (2010). Comprehensive approach to analyzing rare genetic variants. *PLoS One*, 5(11):e13584.

Horn, S. D. (1977). Goodness-of-Fit tests for discrete data: A review and an application to a health impairment scale. *Biometrics*, 33(1):237–247. ArticleType: research-article / Full publication date: Mar., 1977 / Copyright © 1977 International Biometric Society.

Ionita-Laza, I., Buxbaum, J. D., Laird, N. M., and Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genetics*, 7(2):e1001289.

Jones, H. L. (1948). Exact lower moments of order statistics in small samples from a normal distribution. *The Annals of Mathematical Statistics*, 19(2):270–273. Mathematical Reviews number (MathSciNet): MR25121; Zentralblatt MATH identifier: 0031.37104.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *G. Inst. Ital. Attuari*, 4:83.

Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics*, 83(3):311–321. PMID: 18691683.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using Single-Nucleotide polymorphism data. *Genetics*, 165(4):2213–2233.

Liu, D., Ghosh, D., and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 9(1):292.

Liu, D. J. and Leal, S. M. (2010). A novel adaptive method for the analysis of Next-Generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genetics*, 6(10):e1001156.

Madsen, B. and Browning, S. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, 5(2):e1000384.

Mangold, E., Ludwig, K. U., Birnbaum, S., Baluardo, C., Ferrian, M., Herms, S., Reutter, H., Assis, N. A. d., Chawa, T. A., Mattheisen, M., Steffens, M., Barth, S., Kluck, N., Paul, A., Becker, J., Lauster, C., Schmidt, G., Braumann, B., Scheer, M., Reich, R. H., Hemprich, A., Pötzsch, S., Blaumeiser, B., Moebus, S., Krawczak, M., Schreiber, S., Meitinger, T., Wichmann, H., Steegers-Theunissen, R. P., Kramer, F., Cichon, S., Propping, P., Wienker, T. F., Knapp, M., Rubini, M., Mossey, P. A., Hoffmann, P., and Nöthen, M. M. (2009). Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nature Genetics*, 42(1):24–26.

Mangold, E., Ludwig, K. U., Birnbaum, S., Baluardo, C., Ferrian, M., Herms, S., Reutter, H., de Assis, N. A., Chawa, T. A., Mattheisen, M., Steffens, M., Barth, S., Kluck, N., Paul, A., Becker, J., Lauster, C., Schmidt, G., Braumann, B., Scheer, M., Reich, R. H., Hemprich, A., Potzsch, S., Blaumeiser, B., Moebus, S., Krawczak, M., Schreiber, S., Meitinger, T., Wichmann, H. E., Steegers-Theunissen, R. P., Kramer, F. J., Cichon, S., Propping, P., Wienker, T. F., Knapp, M., Rubini, M., Mossey, P. A., Hoffmann, P., and Nothen, M. M. (2010). Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nature Genetics*, 42:24–26. PMID: 20023658.

Morgenthaler, S. and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1–2):28–56.

Morris, A. P., Whittaker, J. C., and Balding, D. J. (2002). Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *American Journal of Human Genetics*, 70(3):686–707. PMID: 11836651.

Mossey, P. A., Little, J., Munger, R. G., Dixon, M. J., and Shaw, W. C. (2009). Cleft lip and palate. *Lancet*, 374:1773–1785. PMID: 19747722.

Nadarajah, S. and Kotz, S. (2008). Exact distribution of the Max/Min of two gaussian random variables. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 16(2):210–212.

Neale, B., Rivas, M., Voight, B., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S., Roeder, K., and Daly, M. (2011). Testing for an unusual distribution of rare variants. *PLoS genetics*, 7(3):e1001322.

Pan, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology*, 33(6):497–507.

Pongpanich, M., Neely, M. L., and Tzeng, J. (2012). On the aggregation of multimarker information for Marker-Set and sequencing data analysis: Genotype collapsing vs. similarity collapsing. *Frontiers in Genetics*, 2.

Price, A., Kryukov, G., de Bakker, P., Purcell, S., Staples, J., Wei, L., and Sunyaev, S. (2010). Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86(6):832–838.

Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current Opinion in Genetics and Development*, 19(3):212 – 219. <ce:title>Genetics of disease</ce:title>.

Shen, Y., Cheung, Y., Wang, S., and Pe'er, I. (2011). A parametric bayesian method to test the association of rare variants. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*, page 137–143.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281. ArticleType: research-article / Full publication date: Jun., 1948 / Copyright © 1948 Institute of Mathematical Statistics.

Stephens, M. and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, 73(5):1162–1169. PMID: 14574645.

Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68(4):978–989. PMID: 11254454.

The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861.

Wang, J., Tsang, W. W., and Marsaglia, G. (2003). Evaluating kolmogorov's distribution. *Journal of Statistical Software*, 08(i18).

Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*.

Wu, Y. (2008). Association mapping of complex diseases with ancestral recombination graphs: models and efficient algorithms. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 15(7):667–684. PMID: 18651799.

Zöllner, S. and Pritchard, J. K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, 169(2):1071–1092. PMID: 15489534.