Sequence clustering for genetic mapping of binary traits

by

Charith Bhagya Karunarathna

M.Sc.(Statistics), Sam Houston State University, 2014
B.Sc.(Hons.), University of Peradeniya, 2011

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2021

Declaration of Committee

Name:	Charith Bhagya Karunarathna
Degree:	**Doctor of Philosophy**
Thesis title:	**Sequence clustering for genetic mapping of binary traits**
Committee:	**Chair: Joan Hu** Professor, Statistics and Actuarial Science
	Jinko Graham Supervisor Professor, Statistics and Actuarial Science
	Kelly Burkett Committee Member Associate Professor, Mathematics and Statistics University of Ottawa
	Brad McNeney Examiner Associate Professor, Statistics and Actuarial Science
	Marie-Hélène Roy-Gagnon External Examiner Associate Professor, Epidemiology and Public Health Faculty of Medicine University of Ottawa

Abstract

Sequence relatedness has potential application to fine-mapping genetic variants contributing to inherited traits. We investigate the utility of genealogical tree-based approaches to fine-map causal variants in three different projects. In the first project, through coalescent simulation, we compare the ability of several popular methods of association mapping to *localize* causal variants in a sub-region of a candidate genomic region. We consider four broad classes of association methods, which we describe as single-variant, pooled-variant, joint-modelling and tree-based, under an additive genetic-risk model. We also investigate whether differentiating case sequences based on their carrier status for a causal variant can improve fine-mapping. Our results lend support to the potential of tree-based methods for genetic fine-mapping of disease. In the second project, we develop an R package to dynamically cluster a set of single-nucleotide variant sequences. The resulting partition structures provide important insight into the sequence relatedness. In the third project, we investigate the ability of methods based on sequence relatedness to fine-map rare causal variants and compare it to genotypic association methods. Since the true gene genealogy is unknown in reality, we apply the methods developed in the second project to estimate the sequence relatedness. We also pursue the idea of reclassifying case sequences into their carrier status using the idea of genealogical nearest neighbours. We find that method based on sequence relatedness is competitive for fine-mapping rare causal variants. We propose some general recommendations for fine-mapping rare variants in case-control association studies.

Keywords: Fine mapping; gene genealogy; association methods; sequence relatedness; disease association; causal variants

Dedication

To my beloved parents and sister in Sri Lanka.

Acknowledgements

First and foremost, I would like to convey my gratitude to my senior supervisor Dr. Jinko Graham for bringing me to the field of Statistical Genetics. Thank you for your encouragement, guidance and patience throughout my PhD studies. I value everything you have done for me throughout my life at Simon Fraser University.

My sincere thanks to the examining committee: Dr. Jinko Graham, Dr. Kelly Burkett, Dr. Brad McNeney and Dr. Marie-Hélène Roy-Gagnon for their valuable inputs.

I take this opportunity to convey my gratitude to the faculty in the department of Statistics and Actuarial Science at Simon Fraser University. You provide an incredible learning environment in the department. I extend my gratitude to Sadika, Charlene, Kelly and Jay for their kind assistance.

Many thanks to Payman Nickchi for collaborating with my third project. I enjoyed working with you and thank you for your cooperation. I also thank Christina Nieuwouldt for having useful discussions throughout my master's and PhD studies, and being a good friend of mine. I would like to thank my Sri Lankan fellows Harsha Perera, Lasantha Premarathna, Rajith Silva and Pulindu Ratnasekara for supporting me in various occasions. I truly value our friendship.

I would also like to thank Dr. Ananda Manage, Dr. Cecil Hallum, Dr. Melinda Holt and Dr. Ferry Butar Butar for their enormous support and guidance during my master's studies at Sam Houston State University. I extend my thank to all my fellow students for their friendship during my masters. Without you, I would never have been connected with Simon Fraser University. I had great time with these wonderful people at Sam Houston.

Last but not least, I am indebted to my father, mother and sister for their guidance, patience and blessing throughout my many years at school. I appreciate everything you all have done for me.

Table of Contents

List of Tables

List of Figures

Glossary

coalescent A stochastic process that models the relationships within a sample of sequences from present sequences back to the most recent common ancestor.

diploid An organism that has paired chromosomes, one inherited from each of two parents.

fastsimcoal2 A C++ program to simulate genetic markers in a sample of sequences under complex evolutionary models (`http://cmpg.unibe.ch/software/fastsimcoal2`).

fine mapping Determining the genetic variant (or variants) that contribute to complex trait in a genomic region.

haploid An organism having a single set of chromosomes, inherited from a single parent.

haplotype A chromosome segment with genetic variation that is inherited from one parent.

msprime A Python program to generate coalescent trees for a sample of sequences under a range of evolutionary scenarios (`https://tskit.dev/msprime/docs/stable`).

phasing The process of assigning alleles to the paternal and maternal chromosomes to get a pair of haplotypes.

population effective size The number of individuals in a population that contribute offspring to the next generation.

recombination A process by which DNA inherited from both parents is broken and recombined to produce new combinations of alleles on a sequence.

single-nucleotide variant Variation at a given site along the DNA sequence.

Chapter 1

Introduction

Genetic-association methods aim to identify the association between a trait and genetic markers or between a trait and genetic relationships. These methods can be classified a number of ways. We consider methods that are based on (i) a single genetic marker, (ii) multiple genetic markers, or (iii) a genealogical tree or sequence relatedness that underlies the genetic data.

Single-marker methods test the association between a trait and genetic marker, such as a single-nucleotide variant (SNV), one at a time. Single-marker methods are typically used to identify common variants. Multiple-marker methods, by contrast, assess the association between a trait and multiple markers simultaneously. These methods often focus on the cumulative effects of variants in a genomic region of interest.

We can classify multiple-marker methods according to whether they are based on burden or variance-component tests. Burden tests collapse information for multiple genetic markers into one genetic score and assess the association between the score and a trait (e.g., [1]). Variance-component tests assess the association by evaluating the variance of the random effects for individual genetic markers. (e.g., [2]) Both these methods are powerful for rare-variant association testing.

Genealogical tree-based methods, inspired by the gene genealogy [3], assess the association between clustering of related DNA segments and clustering of trait values. The idea is that the genealogical tree connecting DNA segments clusters related segments to nested clades. Within a clade, segments may carry the same trait-influencing mutation, inherited from a common ancestor of the clade. Locally, along the genome, the DNA segments are grouped together in nested clades of the tree. Hence, an association between trait similarity and relatedness of DNA segments in a certain genomic region suggests the presence of a disease-predisposing variant or variants. Therefore, genealogical tree-based methods have potential to identify both common and rare causal variants.

In this thesis, we explore the fine-mapping ability of genealogical-tree approaches as three different projects. In Chapter 2, we compare the fine-mapping ability of several popular association methods using the true genealogical trees as a reference. This chapter has been

published in the journal of Human Heredity in 2018. Chapter 3 implements a method to reconstruct partitions of the underlying genealogical tree from SNV haplotypes data. This chapter has been published in the journal of BMC Bioinformatics in 2019. The chapter includes a simple example of grouping haplotypes into nested clades for use in association mapping. Chapter 4 applies the methods developed in Chapter 3 to the problem of *detecting* and *localizing* the disease-causal genomic region and introduces methods to reclassify the case haplotypes based on their estimated carrier status for a causal SNV. We next describe each of the chapters in more detail.

Many methods have been proposed to detect disease association with DNA sequence variants in candidate genomic regions. However, the literature lacks a comparison of these methods in terms of their ability to localize or fine-map the disease causal variants lying within the candidate region. In Chapter 2, through coalescent simulation, we compare the ability of several popular methods of association mapping to localize causal variants in a sub-region of a candidate genomic region. This work is an extension of an earlier comparison of methods for detecting disease association with genomic sequence variants in a population of haploid or single-parent organisms [4] in two key ways. First, the earlier investigation considered the ability of the methods to *detect* association in the candidate region. We extend the results by comparing the methods' ability to *localize* the association signal to the causal sub-region. Second, the earlier investigation considered a case-control sample from a haploid population. We extend the results by sampling cases and controls from a diploid population, such as humans. We present a case study of one simulated dataset for insight into the methods and describe simulation results to score which method best localizes the middle sub-region of interest where the disease-causal variants lie. Our results lend support to the potential of genealogical-based methods for genetic fine-mapping of disease.

The previous question used the true trees known from the simulation. In practice, however, true trees are unknown. When the true trees can be reconstructed with a high degree of accuracy from the available sequence data, we would expect to be able to localize the causal genomic region well. In Chapter 3, using the concept of perfect phylogeny [5], we pursue this idea of genealogical tree reconstruction from sequence data. We present an R package, `perfectphyloR`, which is available on the Comprehensive R Archive Network for R $\geq$ 3.4.0, to reconstruct perfect phylogenies underlying a sample of DNA sequences. A perfect phylogeny is a rooted binary tree that represents a recursive partitioning of a set of objects such as DNA sequences. Note that perfect phylogeny is a partition of the sequences

and not an actual genealogical tree. For example, consider the following figure showing the genealogical tree of six haplotypes on the left and their partition structure on the right.

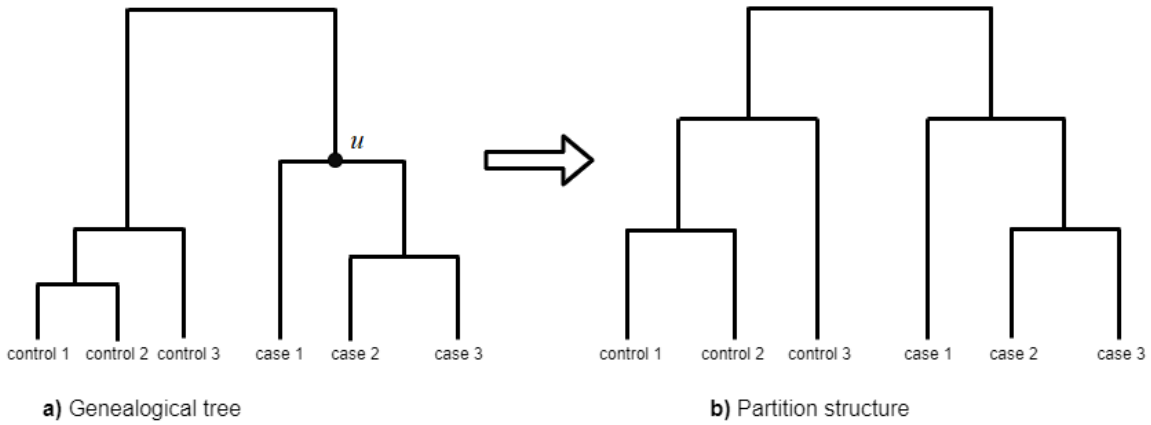

Figure 1.1: **a)** The true tree structure of the six haplotypes. **b)** Their partition structure.

The genealogical tree has information about time to the point where two haplotypes, or lines of descent, coalesce in their most recent common ancestor. When we make partitions from the true tree, we lose this information on time. Therefore, in perfect-phylogeny partitions, we do not have the information about ordering of all the coalescent events. For example, control 1 and control 2 coalesce more recently than case 2 and case 3 (Figure 1.1a) but this information is lost in the partition at right. Though perfect phylogenies are not ancestral trees, their nested partition structures provide insight into the pattern of ancestry of DNA sequence data as shown in figure 1b. The package `perfectphyloR` enables users to dynamically cluster a set of SNV sequences. The resulting partitions provide important insight into the local ancestral structure of the sequence data. In addition, `perfectphyloR` enables users to investigate the association between the reconstructed partitions and a user-specified partition in a variety of ways. We illustrate by example how users can reconstruct the partitions underlying a sample of DNA sequences and how such associations can be helpful for localizing trait-predisposing variants within a candidate genomic region.

In genetics, *identity-by-descent* or IBD refers to sequences being identical because they descend from a recent common ancestor. In Chapter 4, we compare IBD-based association methods to non-IBD based methods in terms of their ability to *detect* and *localize* disease causal variants. For fine mapping, IBD-based methods use the information about relatedness among segments of DNA sequences, whereas non-IBD based methods do not use such information. We also explore the idea of reclassifying the case haplotypes into carriers and non-carriers of causal variants based on some partition statistics obtained from genealogical nearest neighbors (GNN) [6]. GNN is a statistic based on the topological property of a

genealogical tree and can be applied to the partition structures as well. The statistic summarizes the identity of the nearest neighbors of a given haplotype on a tip of a genealogical tree. For example, to find the nearest neighbors of case 1 of the genealogical tree (Figure 1a), we first traverse upward from case 1 until we find the first internal node, u. Then the nearest neighbors of case 1 are all the sequences descending from node u. i.e. case 2 and case 3.

To perform IBD-based mapping, we need to identify IBD clusters on haplotype sequences or relatedness of sequences. However, in practice, we do not know true IBD clusters since we do not know the true genealogical trees. Therefore, using the method developed in Chapter 3, we reconstruct partitions underlying the sample haplotype data. With simulated case-control haplotypes, we compare the ability of proposed IBD-based methods with non-IBD based methods to *detect* the association, and to *localize* causal variants in a subregion of a candidate genomic region. We first work through an example dataset for insight into these methods. Using a simulation study, we then compare the detection and localization ability of IBD methods with non-IBD methods.

Chapter 2

Using gene genealogies to localize rare variants associated with complex traits in diploid populations

This chapter has been published in the journal of Human Heredity: C. B. Karunarathna and J. Graham, "Using gene genealogies to localize rare variants associated with complex traits in diploid populations," Human Heredity, vol. 83, no. 1, pp. 30-39, 2018.

2.1 Introduction

Most genetic association studies focus on common variants, but rare variants can play major roles in influencing complex traits [7, 8]. Rare causal variants identified through sequencing could thus explain some of the missing heritability of complex traits [9]. However, for rare variants, standard methods to test for association with single genetic variants are underpowered unless sample sizes are very large [10]. The lack of power of single-variant approaches holds in fine-mapping as well as genome-wide association studies.

In this report, we are concerned with fine-mapping a candidate genomic region that has been sequenced in cases and controls to identify a disease-risk locus. Our work extends an earlier comparison of methods for *detecting* disease association in a candidate genomic region [4] to a comparison of methods for *localizing* the association signal. Additionally, in the current investigation, we sample cases and controls from a diploid or two-parent population to mimic studies in humans. In the previous investigation, cases and controls were sampled from a haploid or one-parent population.

A number of methods have been developed to evaluate the disease association for both a single variant and multiple variants in a genomic region. Besides single-variant methods, we consider three broad classes of methods for analysing association: pooled-variant, joint-modelling and tree-based methods. Pooled-variant methods evaluate the cumulative effects

of multiple genetic variants in a genomic region. The score statistics from marginal models of the trait association with individual variants are collapsed into a single test statistic by combining the information for multiple variants into a single genetic score [10]. Joint-modeling methods model the joint effect of multiple genetic variants on the trait simultaneously. These methods can assess whether a variant carries any further information about the trait beyond what is explained by the other variants. When trait-influencing variants are in low linkage disequilibrium, this approach may be more powerful than pooling test statistics for marginal associations across variants [11]. Tree-based methods assess whether trait values co-cluster with the local genealogical tree for the haplotypes (e.g., [12], [13]). A local genealogical tree represents the ancestry of the sample of haplotypes at each locus. Haplotypes carrying the same causal alleles are expected to be related and cluster on the genealogical tree at a disease-risk locus.

In practice, true trees are unknown. However, clustering statistics based on true trees represent a best case for detecting or localizing association because tree uncertainty is eliminated. Burkett et al. [4] used known trees to assess the effectiveness of tree-based approaches for detection of disease-risk variants in a haploid population. They found that clustering statistics computed on the known trees outperform popular methods for detecting causal variants in a candidate genomic region. Following Burkett et al. [4], we use Mantel tests that associate phenotypic and genealogical distances as the clustering statistics. These statistics, which rely on known trees, serve as benchmarks against which to compare the popular association methods. However, unlike Burkett et al. [4], who focus on detection of disease-risk variants, we focus on localization of association signal in the candidate genomic region. Additionally, we use a diploid rather than a haploid disease model to mimic human populations.

In this report, we compare the ability of several popular methods of association mapping to localize causal variants in a subregion of a larger, candidate, genomic region. In our simulation study, we use sequence data generated under an approximation to the coalescent with recombination [3]. To illustrate ideas, we start by working through a particular example dataset as a case study for insight into the association methods. We next perform a simulation study involving 200 sequencing datasets and score which association method localizes the risk subregion most precisely. We conclude with a summary and discussion of our results. Our results confirm the earlier findings by Burkett et al. [4] indicating potential gains in performance from ancestral tree-based approaches. They also highlight some important differences between haploid and diploid populations when localizing causal variants.

2.2 Methods

In this section, we describe our data simulation, the association methods we considered and the way we assessed localization and detection of the association signal. Then, we describe the popular association methods we evaluated for fine mapping. Finally, we explain the simulation study involving 200 sequencing datasets to address the signal localization and detection.

2.2.1 Data simulation

First, we report how we simulated haplotype data from ancestral trees. Second, we describe how we assigned the disease status to individuals and sampled data for our case-control study. We used fastsimcoal2 [14] to simulate ancestral trees and 3000 haplotypes of 4000 equispaced single-nucleotide variants (SNVs) in a 2 million base-pair (Mbp) genomic region. We used a recombination rate of 1×10^{-8} per base-pair per generation [15], in a diploid population of constant effective size, $N_e = 6200$. To mimic random mating in a diploid population, we then randomly paired the 3000 haplotypes into 1500 diploid individuals. Next, disease status was assigned to the 1500 individuals based on randomly-sampled risk SNVs from the middle genomic region of $950 - 1050$ kbp. For risk SNVs, the number of copies of the derived (i.e. mutant) allele increased the risk of disease according to a logistic regression model,

$$\text{logit}\{P(D = 1|G)\} = \text{logit}(0.02) + \sum_{j=1}^{m} 2 \times G_j, \text{ where}$$

- $\text{logit}(p) = \log[p/(1-p)]$ for $0 < p < 1$,
- D is disease status ($D = 1$, case; $D = 0$, control),
- $G = (G_1, G_2, \ldots, G_m)$ is an individual's multi-locus genotype at m risk SNVs, with G_j being the number of copies of the derived allele at the j^{th} risk SNV, and
- the value of the intercept term is chosen to ensure that the probability of sporadic disease (i.e. $P(D = 1|G = \underset{\sim}{0})$) is approximately 2%.

To select risk SNVs in the model, we randomly sampled SNVs from the middle subregion one at a time, until the disease prevalence was between $9.5 - 10.5\%$ in the 1500 individuals. Our selection of risk SNVs is not restricted by the minor allele frequency (MAF) and therefore differs slightly from Burkett et al. [4], which allowed only SNVs with MAF $< 1\%$ to be risk SNVs. After assigning disease status to the 1500 individuals, we randomly sampled 50 cases (i.e. diseased) from the affected individuals and 50 controls (i.e. non-diseased) from the unaffected individuals. We then extracted the data for the variable SNVs in the resulting case-control sample to examine the patterns of disease association in subsequent analyses.

2.2.2 Association analysis

In this section, we review the methods for association mapping that we considered. These methods fall under four categories: single-variant method, pooled-variant methods, joint-modeling methods and tree-based methods.

2.2.2.1 Single-variant method

For the single-variant method, we used the standard Fisher's exact test of disease association with each of the SNVs in the case-control sample. In Fisher's exact test, each of the variant sites in the case-control sample was tested for an association with the disease outcome using a 2×3 table to compare genotype frequencies. This single-variant association was assessed with the p-value of Fisher's exact test on the contingency table. Each row in a table represents disease status of individuals, and a column represents the three possible genotypes. The -$\log_{10}$ p-value from the test was recorded as the association signal for each variant. However, single-variant tests are less powerful for rare variants than for common variants [16]. We therefore considered three other ways to assess the association signal based on pooled-variant, joint-modelling and tree-based methods.

2.2.2.2 Pooled-variant methods

For the pooled-variant methods, we evaluated the Variable Threshold (VT) and the C-alpha test. The variable threshold approach of Price et al. [1], uses a generalized linear model to relate the phenotypes to the counts of variants in the genomic region of interest which have MAFs below some user-defined threshold (e.g. 1% or 5%). The idea is that variants with MAF below the threshold have a higher prior probability of being functional than the variants with higher MAF, based on population-genetic arguments. For each possible MAF threshold, VT computes a score measuring the strength of association between the pheonotype and the genomic region, and uses the maximum of the score over all allele frequency thresholds. The statistical significance of the maximum score is then assessed by a permutation test. Price et al. [1] found that the VT approach had high power to detect the association between rare variants and disease traits when effects are in one direction. Unlike the VT test, the C-alpha test of Neale et al. [2] is a variance-components approach that assumes the effects of variants are random with mean zero. The C-alpha procedure tests the variance of genetic effects under the assumption that variants observed in cases and controls are a mixture of risk, protective or neutral variants. Neale et al. [2] found that the C-alpha test showed greater power than burden tests such as VT when the effects are bi-directional. We applied the VTWOD function in the R package RVtests [17] for the VT-test and the SKAT function in the R package SKAT [18] for the C-alpha test. We used sliding windows of 20 SNVs overlapping by 5 SNVs across the simulated region.

2.2.2.3 Joint-modeling methods

For the joint-modeling methods, we evaluated the CAVIARBF [19] and elastic-net [20] methods. CAVIARBF is a fine-mapping method that uses marginal test statistics for the SNVs and their pairwise association to approximate the Bayesian multiple regression of phenotypes onto variants that is implemented in BIMBAM [21]. However, CAVIARBF is much faster than BIMBAM because it computes Bayes factors using only the SNVs in each causal model rather than all SNVs. These Bayes factors can be used to calculate the posterior probability of SNVs in the region being causal; i.e., the posterior inclusion probability (PIP). To compute PIPs for SNVs, a set of models and their Bayes factors have to be considered. Let p be the total number of SNVs in a candidate region; then the number of possible causal models is 2^p. To reduce the number of causal models to evaluate and thus save computational time and effort, CAVIARBF imposes a limit, L, on the number of causal variants. This limitation reduces the number of models in the model space from 2^p to $\sum_{i=0}^{L} \binom{p}{i}$. Since there were 2747 SNVs in our example dataset, to keep the computational load down, we considered $L = 2$ throughout this investigation.

The elastic net [20] is a hybrid regularization and variable selection method that linearly combines the L1 and L2 regularization penalties of the lasso [22] and ridge (e.g., [23]) regression methods in multiple regression. This combination of lasso and ridge penalties provides a more precise prediction than using multiple regression, when SNVs are in high linkage disequilibrium [24]. In addition, the elastic net can accommodate situations in which the number of predictors exceeds the number of observations. We used the elastic net to select risk SNVs by considering only the main effects. The variable inclusion probability (VIP), a frequentist analog of the Bayesian posterior inclusion probability was used as a measure of the importance of a SNV for predicting disease risk [11]. To obtain the VIP for a SNV, we re-fitted the elastic-net model using 100 bootstrap samples and calculated the proportion of samples in which the SNV was included in the fitted model. In our analysis, we applied the elastic net using the R package glmnet [25].

2.2.2.4 Tree-based methods

We considered two tree-based methods to assess clustering of disease status on the gene genealogy connecting haplotypes at a putative risk variant: Blossoc (BLOck aSSOCiation; [13]), which uses reconstructed trees, and a Mantel test which uses the true trees. Blossoc aims to localize the risk variants by reconstructing genealogical trees at each SNV, using them to define clusters, and associating the cluster membership with disease status. The reconstructed trees approximate perfect phylogenies [5] for each SNV, assuming an infinite-sites model of mutation. These trees are scored according to the non-random clustering of affected individuals. The underlying idea is that genomic regions containing SNVs with high clustering scores are likely to harbour risk variants. Blossoc can be used for both phased and

unphased genotype data. However, the method is impractical to apply to unphased data with more than a few SNVs due to the computational burden associated with phasing. We therefore assumed the SNV data are phased, as might be done in advance with a fast-phasing algorithm such as fastPHASE [26], BEAGLE [27], IMPUTE2 [28] or MACH [29, 30]. We evaluated Blossoc with the phased haplotypes, using the probability-score criterion which is the recommended scoring scheme for small datasets [13].

In practice, the true trees are unknown but as the data were simulated we had access to this information. Also, the cluster statistics based on true trees represent a best case insofar as tree uncertainty is eliminated [4]. We therefore included two versions of the Mantel test as a benchmark for comparison. In the first version, the phenotype corresponding to a haplotype is scored according to whether or not the haplotype comes from a case. We refer to the first version as the *naive* Mantel test because all case haplotypes are treated the same, even those not carrying any risk variants. In the second version, the phenotype is scored according to whether or not the haplotype comes from a case and carries a risk variant. We refer to the second version as the *informed* Mantel test because it takes into account whether or not a case haplotype carries a risk variant. The informed Mantel test is a best-case scenario insofar as the uncertainty about the risk-variant-carrying status of the case haplotypes is eliminated. Both Mantel tests correlate the pairwise distance in the known ancestry with those in the phenotypes. Following Burkett et al. [4], we used pairwise distances calculated from the rank of the coalescent event rather than the actual times on the tree. To focus on the rare variants, the test statistic upweights the short branches close to present at the tip of the tree, by assigning a branch-length of one to all branches, even the relatively longer branches that are expected to occur close to the time to the most recent common ancestor. Pairwise distances between haplotypes on this re-scaled tree are then correlated to pairwise phenotypic distances. We determined the distance measures, $d_{ij} = 1 - s_{ij}$, where $s_{ij} = (y_i - \mu)(y_j - \mu)$ is the similarity score between haplotype i and j, y_i is the binary phenotype (coded as 0 or 1) and μ is the disease prevalence in the 1500 simulated individuals. We then used the Mantel statistic to compare the phenotype-based distance matrix, d, with the re-scaled tree-distance matrix. Note that we define a phenotype for each haplotype within an individual because we are interested in relatedness of sequences rather than individuals. Therefore, an individual has two phenotypes rather than one. (If we were interested in individual relatedness, another option would be to take the minimum distance of all four possible pairs of haplotypes between two individuals [31].)

2.2.3 Scoring localization and detection

To address the question of localization, we scored the distance of the peak association signal from the risk region based on the average absolute value of the distance of peak signals across the entire genomic region. The average distance was used when there were multiple peaks with the same maximum strength of association. Specifically, for each method, on each

dataset, we computed the average distance (in bases) of the peak association signals from the risk region and plotted the empirical cumulative distribution function (ECDF) of the average based on the 200 simulated samples. Thus, the ECDF at point x is the proportion of the 200 simulated samples with average distance less than or equal to x. A method with higher ECDF than another method localizes the signal better.

To detect association with a given method, we used a maximum score across all the SNVs in a dataset to obtain a global test of association across the entire genomic region. We determined the null distribution of the global test statistic for each method by permuting the case-control labels. For the global test statistic, we used either a maximum statistic or the maximum of - $\log_{10}$ of p-values across the genomic region. The global test statistics for the different association methods are not comparable since they are not on the same scale. To make these statistics comparable across methods, we considered their permutation p-values. We defined these p-values as the proportion of test statistics under the permutation-null distribution that are greater than or equal to the observed value. We then compared the distribution of the resulting p-values for the different methods by plotting their ECDFs.

2.3 Results

In this section, we first present the summaries of our example dataset and the resulting plots from the selected association methods. We then present our results from the simulation study for localizing and detecting the association signal.

2.3.1 Example dataset

2.3.1.1 Population and sample summaries

In the population of 1500 individuals that was simulated for the example dataset, we obtained 4000 SNVs, of which 16 were risk SNVs. Of the 4000 SNVs in the population, 2747 were polymorphic in the sample of 50 cases and 50 controls. Of the 16 risk SNVs in the population, 10 were polymorphic in the case-control sample. The linkage disequilibria between the polymorphic risk SNVs was low; all r^2 values were < 0.1 (results not shown). Table 2.1 summarizes the physical distances, effect sizes of risk SNVs, age of the most recent common ancestors of the derived alleles in generations, number of recombinations and MAFs of the 10 risk SNVs in the sample. Of these 10 risk SNVs, the fourth is the oldest but the seventh is the most frequent, owing to the neutral random variation of the simulated trees.

Figure 2.1 compares the distribution of risk haplotypes in cases and controls. We define a risk haplotype to be a haplotype that carries a risk SNV. Figure 2.2 shows the effect size of the polymorphic risk SNVs versus their location in the risk region (Figure 2.2a) and their age in generations, in the log-base-10 scale, respectively.

Table 2.1: Summaries for risk SNVs.

Risk SNV	Position (kbp)	Effect size[a]	Age[b]	NR [c]	MAF		
					Population	Cases	Controls
1	975.0	0.039	298,737	NA	0.001	0.01	0.00
2	990.0	0.079	330,135	18	0.009	0.02	0.00
3	990.5	0.307	802,458	0	0.013	0.07	0.01
4	991.5	0.380	20,690,057	2	0.039	0.07	0.03
5	993.0	0.380	1,498,566	0	0.031	0.08	0.02
6	997.5	0.156	438,663	0	0.006	0.02	0.02
7	1005.5	1.128	3,655,347	10	0.115	0.27	0.07
8	1012.5	0.232	2,115,335	5	0.013	0.05	0.01
9	1019.0	0.039	703,023	4	0.001	0.01	0.00
10	1038.5	0.195	405,567	14	0.013	0.04	0.01

[a] Effect sizes were computed from MAF in cases and controls and the regression effect, as described in [32].

[b] Age was measured by the number of generations back to the most recent common ancestor of the derived alleles.

[c] NR, the number of recombination events between the current and previous risk SNV in the sample.

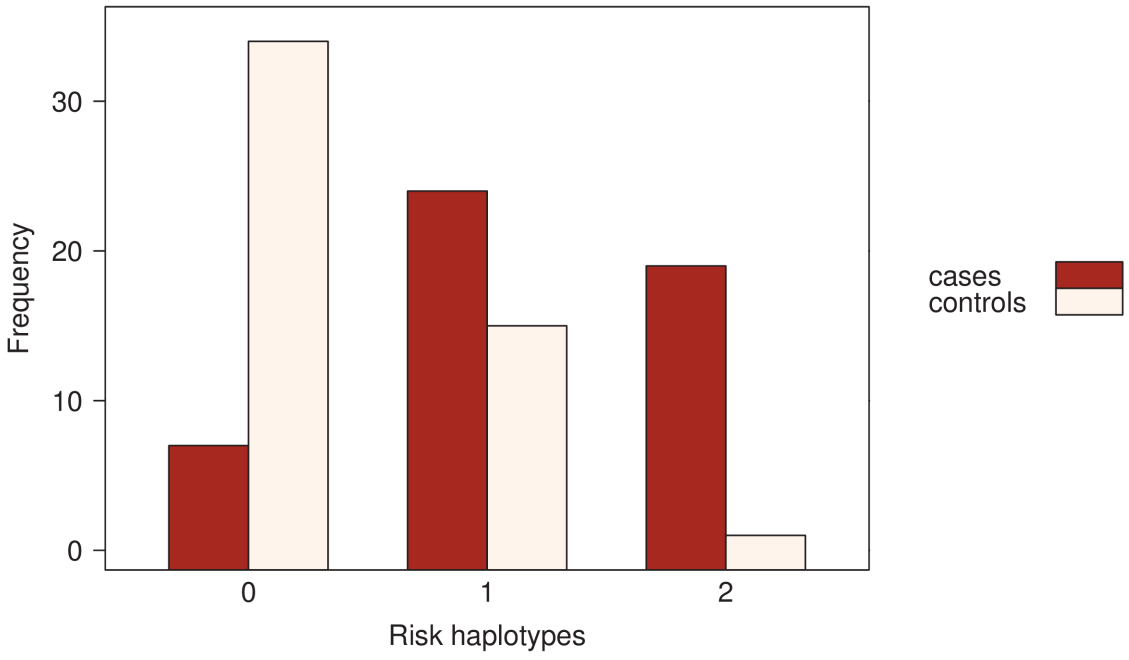

Figure 2.1: Number of haplotypes that carry risk variants for both cases and controls.

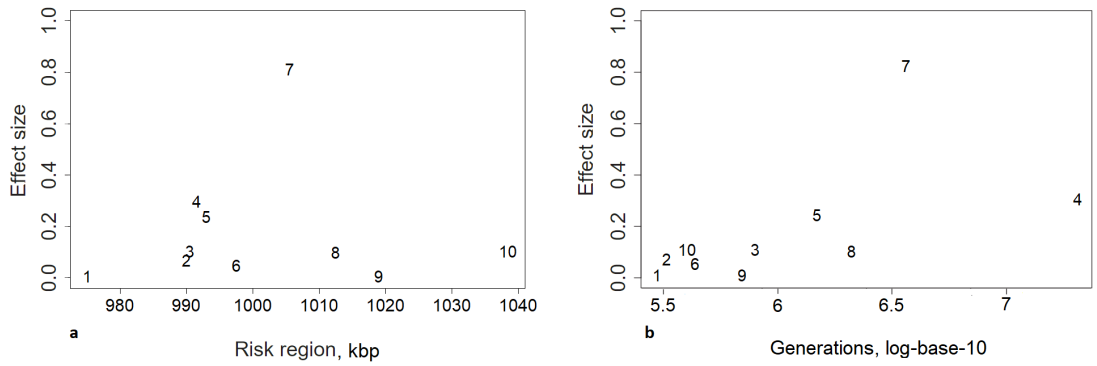

Figure 2.2: **a)** True effect size of polymorphic risk SNVs versus their positions in the risk region. **b)** True effect size of polymorphic risk SNVs versus their age in generations, in the log-base-10 scale. The risk SNVs are numbered according to their physical location in the risk region.

2.3.1.2 Association results

Figure 2.3 shows the resulting plots for each association method using the example dataset. The results from the single-variant method of Fisher's exact test are shown in panel (a). In our example dataset, Fisher's exact test does not localize the peak signal, which is distal to the disease-risk region.

Panels (b) and (c) of Figure 2.3 show the results from the pooled-variant methods. The C-alpha test in panel (c) has stronger associations than the VT test in panel (b), and the C-alpha test localizes the peak association signal to the disease-risk region whereas the VT test doesn't.

Panels (d) and (e) of Figure 2.3 show the results from the joint-modeling methods. The estimated PIP and VIP for the SNVs were computed from CAVIARBF (panel (d)) and elastic net (panel (e)), respectively. We used 100 bootstrap samples to estimate VIPs via elastic net. CAVIARBF provides estimates of the PIPs at each SNV. In our example dataset, both elastic net and CAVIARBF show peak signal outside the risk region, but CAVIARBF localizes the signal better than elastic net.

Panels (f), (g) and (h) of Figure 2.3 show the results from the tree-based methods: Blossoc and the two versions of the Mantel test, i.e. naive- and informed-Mantel. We applied Blossoc to the phased haplotypes, using the probability score criterion for each SNV across the region (panel (f)). In our example dataset, Blossoc shows relatively high association, but the peak signal is outside the risk region. Panel (g) shows the statistics computed from the naive-Mantel test. Our example dataset shows relatively high association signal within the risk region but the peak signal is outside of it. Panel (h) shows the statistics computed from the informed-Mantel test. This informed-Mantel test successfully localizes the peak signal to the risk region.

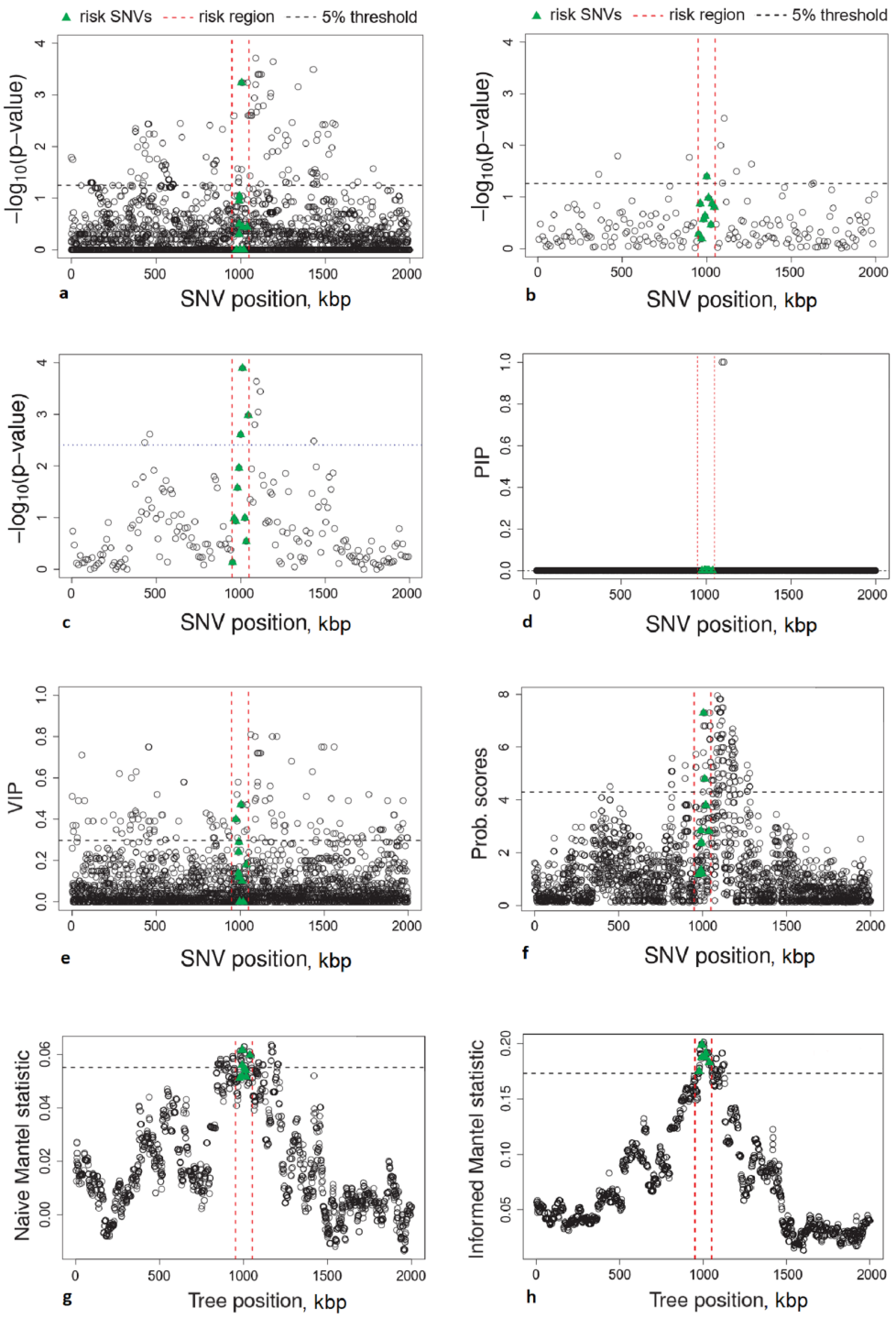

Figure 2.3: Plots of association results from the eight selected association methods in the 50 cases and 50 controls. **a)** Fisher's exact test. **b)** VT test. **c)** C-alpha test. **d)** Posterior inclusion probabilities (PIPs) computed from CAVIARBF. **e)** Variable-inclusion probabilities (VIPs) for SNVs computed from elastic net. **f)** Clustering scores for each SNV, using the probability scores criterion in Blossoc. **g)** Naive-Mantel statistics for each tree position (SNV). **h)** Informed-Mantel statistics for each tree position (SNV). The horizontal dashed line represents the 5% significant threshold based on permutation and adjusted for multiple testing across the entire genomic region.

2.3.2 Simulation study

We first present the simulation results for localizing the association signal, followed by the results for detecting the association signal.

2.3.2.1 Localizing the association signal

Figure 2.4 compares the ability of the different methods to localize the association signal. For each of the 200 simulated datasets, we considered the distance of the peak association signal from the risk region. As described in the Methods section, if there were ties in the peak signal, we took the average distance. The figure shows the ECDFs of these distances for the 200 datasets, for all eight methods. The informed Mantel test outperforms all the other methods; i.e., it has the highest proportion of simulated datasets at the lower distance values. Fisher's exact test, C-alpha test, CAVIARBF, and Blossoc perform comparably and relatively well for localizing signal. VT and naive Mantel test have the worst localization performance. As observed in the example dataset, VT has worse performance than C-alpha for localizing the signal.

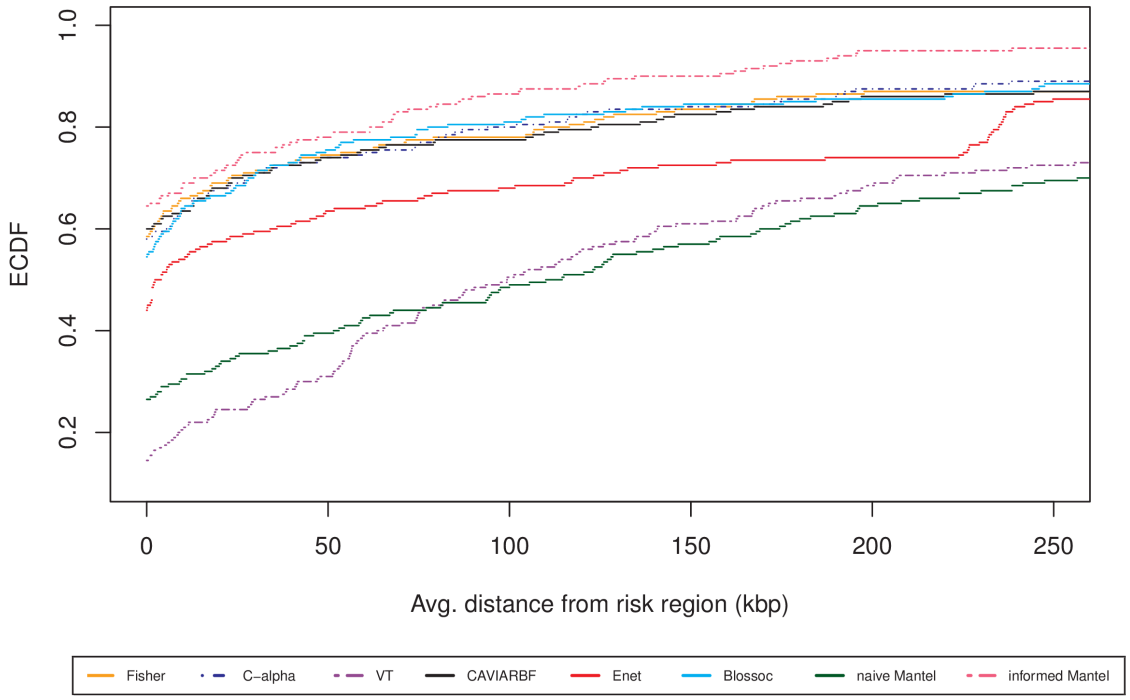

Figure 2.4: ECDFs of average distances of the peak association signals from the risk region for the 200 datasets. Eight methods are compared: Fisher's exact test, VT test, C-alpha test, CAVIARBF, elastic net (Enet), Blossoc, naive Mantel test and informed Mantel test. To better compare methods, the x-axis is shown only for distances $\leq$ 250kbp.

2.3.2.2 Detecting the association signal

Figure 2.5 compares the ability of the methods to detect any association with the disease across the entire genomic region that is being fine-mapped. For each method, we compare the ECDFs, over the 200 datasets, of the permutation p-values computed from the corresponding scores. As expected, the informed Mantel test performs better than all the other methods. The elastic net approach has the lowest power to detect association, followed by the VT approach.

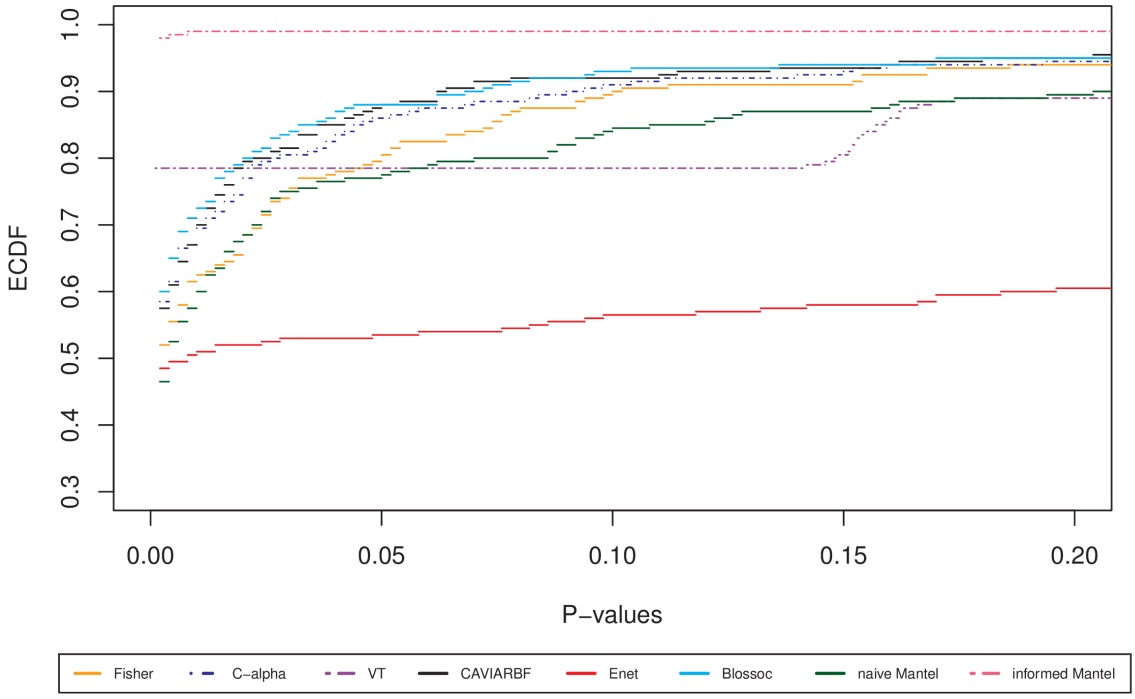

Figure 2.5: ECDFs of permutation p-values from a global test of association in the genomic region. Eight methods are compared: Fisher's exact, VT, C-alpha, CAVIARBF, elastic net (Enet), Blossoc, naive Mantel test and informed Mantel test. The x-axis is shown only for p-values ≤ 0.20 for a better resolution.

2.4 Discussion

In this study, through coalescent simulation, we have investigated the ability of several popular association methods to fine-map trait-influencing genetic variants in a candidate genomic region. While our simulation investigation can never replace an examination of true sequencing data in the framework where the actual variants are known, it can give some insight into the operating characteristics of these association methods under a popular and tractable model of sequence variation, the coalescent with mutation and recombination. As the first step, we worked through a particular example dataset as a case study for insight into the methods. We then performed a simulation study to score which method localizes the risk subregion most precisely.

In our simulations, the informed Mantel test localized the association signal most precisely among all the methods considered. By contrast, the naive Mantel test performed poorly relative to the other methods. In fact, the naive Mantel test localized the risk region more poorly than Blossoc, CAVIARBF, C-alpha, and Fisher's exact test. Our results for *localizing* risk variants in *diploid* populations therefore stand in contrast to previous results for *detecting* risk variants in *haploid* populations [4], which found that the naive Mantel test and related tree-based methods performed very well. The poor performance of the naive Mantel test in our simulations can be explained by the misclassification of haplotypes. In haploid populations, case haplotypes without a risk variant are rarer than in diploid populations, and so fewer would be misclassified by the naive case-control phenotypes than in diploid populations. In diploid populations, we do not know which of the two haplotypes in a case carries the disease-risk variant and score both as being "affected". When only one of the two haplotypes in a case carries a risk variant, defining both haplotypes as "affected" misclassifies one. Therefore, when the majority of cases carry only a single haplotype with risk variants, we expect the informed Mantel test to outperform the naive Mantel test because it defines only case haplotypes that carry the risk variant as "affected". The development of methods to identify which haplotypes carry risk variants would thus be an avenue for further research in genealogy-based approaches to fine-mapping risk variants.

When computing the naive and informed Mantel statistics, we have assumed that the true genealogical trees are known. In practice, however, these trees are not known. The accuracy of trees reconstructed from the sequence data is expected to affect the localization abilities of the Mantel statistics. When the true trees can be reconstructed with a high degree of accuracy from the available sequence data, the informed Mantel test applied to the reconstructed tree should localize the association signal well. To reconstruct the haplotype partitions implied by the genealogical trees, we applied the methods outlined in Mailund et al. [13] to the sequence data. To gain insight into the accuracy of the reconstruction, we computed the Rand index [33] between the true and reconstructed partitions at the genomic position of each risk SNV. The Rand index is a measure between 0 and 1 reflecting

the agreement of the partitions. For the ten risk SNVs labelled 1-10 in Table 2.1, the Rand index values based on ten clusters from each partition were 0.849, 0.814, 0.914, 0.900, 0.900, 0.900, 0.853, 0.885, 0.895, 0.882, respectively. These high values suggest good agreement and accuracy of reconstruction. However, candidate genomic regions with lower mutation and/or higher recombination rates than the rates we have used in our simulations would be expected to have less accurate reconstructions. In these cases, the performance of the Mantel procedures would be expected to be poorer than shown here. The nature and extent of this performance loss would be an interesting topic for future work.

Our simulation study also provides a comparison of the VT to the C-alpha test. Even though the effects are one directional, C-alpha showed higher localization signal in the risk region than VT. Our findings for localization with the VT and C-alpha tests in a diploid population are consistent with those of Burkett et al. [4] for detection in a haploid population. We would in fact expect better performance of the VT test than the C-alpha test since VT is for rare variants having the same direction of effect, which we simulated [1]. However, variance-component tests such as C-alpha have higher power than burden tests such as VT when the proportion of risk variants in the set of tested variants is low [34]. In our example dataset, the highest proportion of risk variants within moving windows of 20 SNVs was 20%. Therefore, a possible explanation for the better performance of C-alpha relative to VT is a relatively low proportion of risk variants within the moving windows. To examine the impact of larger window sizes in our example dataset, we experimented with windows of size 50 and 100 SNVs (overlapping by 5 SNVs) for both the VT and C-alpha tests. However, we could not see any improvement in localizing the association signal (results not shown).

In the example dataset, most risk variants were rare. Of the 10 risk SNVs that were polymorphic in the sample, four were rare with MAF $< 1\%$, five were low frequency with MAF of $1 - 5\%$, and one was common with MAF $> 5\%$. In addition, a majority of cases carried a single risk haplotype (see Figure 2.1) and most risk haplotypes contained a single risk SNV. These findings in the example dataset suggest that the results under a dominant model of genetic risk would be similar to our results under an additive model. Under a recessive model of disease risk, we would expect the naive Mantel test to perform as well as the informed Mantel test because both haplotypes in a case would tend to carry risk variants, and so misclassification of haplotypes would be minimized. In the example dataset, we found that the C-alpha test and the informed Mantel test were the only methods that successfully localized the association signal. However, the peak signals from all the other methods (Fisher's test, VT, CAVIARBF, elastic net, Blossoc and naive Mantel) were close to the disease-risk region.

There are a number of directions for future work. First, we have focused on a simple model of disease risk, with additive effects, no interactions, and no non-genetic covariates. Simulations with more complex risk models would be an area for further research. In the

approaches we have considered, the phenotypes can be adjusted for non-genetic covariates. Second, an examination of true sequencing data with known causal variants would be an interesting future direction once such data resources become more readily available to the public. Finally, for tree-based methods, differentiating between case haplotypes that carry or do not carry risk SNVs improves localization of the risk region. Preliminary work (not shown) suggests that carrier and non-carrier haplotypes can be differentiated based on their number of positively-associated alleles (at level 5%, uncorrected for multiple testing). In future work, we will pursue this idea with informed-Mantel tests on reconstructed trees.

2.5 Acknowledgements

We thank Kelly Burkett for helpful discussions and the Department of Statistics and Actuarial Science at Simon Fraser University for its generous support. This research was funded in part by the Natural Sciences and Engineering Research Council of Canada.

Chapter 3

perfectphyloR: An R package for reconstructing perfect phylogenies

This chapter is published in the journal of BMC Bioinformatics: C.B. Karunarathna and J. Graham, "perfectphyloR: An R package for reconstructing perfect phylogenies," BMC Bioinformatics, vol.20, no.1, pp 1-9, 2019.

3.1 Background

A perfect phylogeny is a rooted binary tree that represents a recursive partitioning of a set of objects such as deoxyribonucleic acid (DNA) sequences [35]. Though the perfect phylogenies are not ancestral trees, the structure of their nested partitions provides insight into the pattern of ancestry of DNA sequences. For example, the perfect phylogeny near a trait-influencing variant can provide useful information about trait association [36]. For instance, in a case-control study, case alleles may tend to cluster in a partition if the corresponding variant influences disease susceptibility. If a cluster has proportionally more case sequences than other clusters in the partition, there will be an association between the disease and cluster membership [12]. Thus, an R package to reconstruct perfect phylogenies from sequence data can be of use to researchers mapping the genetic location of trait-influencing variants.

We present an R package, `perfectphyloR`, to reconstruct perfect phylogenies underlying a sample of DNA sequences. The package uses a classic algorithm [35] together with heuristics [36] to partition sequences. Related software includes PerfectPhy [37] and BLOck aSSOCiation (BLOSSOC) [36].

PerfectPhy is a C++ program that implements efficient algorithms [38, 39] for reconstructing perfect phylogenies from multi-allelic DNA markers. The software comes with a collection of tools for importing/exporting files, handling missing data, filtering markers and drawing trees. PerfectPhy takes a given set of sequences and determines if it can be represented by a perfect phylogeny; if so, the partition is returned. The filtering tool can

be applied in advance to select a maximal subset of markers compatible with a perfect phylogeny.

BLOSSOC is a C++ program for genetic fine-mapping that returns association statistics computed on perfect phylogenies. The statistics are calculated for moving windows of DNA markers across a genomic region of interest. The statistics are returned but not the partitions used to construct them. Unfortunately, BLOSSOC is no longer actively maintained (T. Mailund, personal communication) and is challenging to install on up-to-date operating systems.

Our package `perfectphyloR`, like BLOSSOC, is intended for use with moving windows of markers along the genome. The window sizes should be large enough to allow relatively fine partitioning of the sample of input sequences. However, requiring all the DNA markers in the window to be compatible with a perfect phylogeny tends to be too restrictive and leads to crude partitions. To avoid this limitation, we have incorporated the heuristics implemented in the partitioning algorithm of BLOSSOC. Since `perfectphyloR` returns the sequence partitions, users can then leverage any of the statistical and phylogenetic tools available in R to understand them. In addition, as an R package, the software is easier to install and to maintain as operating systems change.

Throughout, we assume the infinite-sites model and account for diallelic DNA markers only. Since our package reconstructs partitions regardless of whether the variants are common or rare, we refer to markers as single-nucleotide variants (SNVs) instead of single-nucleotide polymorphisms. By SNV, we mean any strictly diallelic marker. Our package is primarily directed to applications at the population level, rather than the interspecies level. Briefly, a neighborhood of SNVs is determined about a focal SNV, as described below. Then, the perfect phylogeny is built by recursive partitioning on SNVs in this neighborhood.

We first discuss the implementation of the reconstruction of the partitions underlying a sample of DNA sequences. We then illustrate the major functionality of the package with worked examples.

3.2 Implementation

In this section, we describe the reconstruction process, which consists of three steps:

1. Create a `hapMat` data object.
2. Reconstruct the perfect phylogeny at a focal SNV.
3. Reconstruct perfect phylogenies across a genomic region.

We first create an object of (S3) class `hapMat` containing SNV sequences to be partitioned with the function `createHapMat()`. To construct a `hapMat` data object, users are required to specify:

- `hapmat`, a matrix of 0's and 1's, with rows representing sequences and columns representing SNVs,
- `snvNames`, a vector of names of SNVs labelling the columns of `hapmat`,
- `hapNames`, a vector of names labelling the sequences in the rows of `hapmat`,
- `posns`, a numeric vector specifying the physical locations along the chromosome (in base pairs) of SNVs in the columns of `hapmat`.

In principle, and as noted by a reviewer, the `hapMat` structure could be extended to accommodate multi-allelic variants, although we do not pursue this here.

With the main function `reconstructPP()`, the user can reconstruct the perfect phylogeny at a chosen focal SNV. The result is a `phylo` object to which the user may apply all the tools from the `ape` package [40] for summarizing the reconstructed partition of sequences.

The function `reconstructPP()` consists of three major steps:

1. Determine a neighborhood of SNVs around a given focal SNV.
2. Order the SNVs in the neighborhood.
3. Recursively partition sequences based on SNVs in the neighborhood.

For a given focal SNV, the algorithm finds a neighborhood of SNVs. Starting from the focal SNV, the neighborhood of SNVs that are compatible with the focal SNV is expanded as much as possible on either side of the focal SNV until an incompatible SNV is found. The compatibility of a pair of SNVs is determined by the Four-Gamete Test [41]. For example, under the infinite-sites mutation model and no recombination, if the patterns at two SNVs are 00, 01, 10 and 11, then a mutation must have occurred twice at the same SNV and the two SNVs are said to be incompatible. If the neighborhood of compatible SNVs is smaller than a user-defined minimum size, we include incompatible SNVs in order of their physical proximity to the focal SNV, until the minimum size is reached.

Once the neighborhood of SNVs is determined, we order the compatible SNVs in the neighborhood from the most ancient to the most recent based on the minor allele frequency. We use the minor allele frequency of an SNV as a proxy for its age. Our rationale is that, under the infinite-sites mutation model, the age of SNVs can be inferred from the derived allele frequency. Then, we order incompatible SNVs according to their physical proximity to the focal SNV.

The algorithm partitions sequences based on the most ancient compatible SNV in the neighborhood, and then recursively moves towards the most recent compatible SNV. When there are no further compatible SNVs in the neighborhood, the algorithm partitions sequences based on the incompatible SNVs, in order of their physical proximity to the focal

SNV. Starting with the most ancient compatible SNV in the neighborhood, the algorithm partitions the sequences based on their carrier status for its derived allele. Then the algorithm jumps to the next-oldest compatible SNV in the neighborhood based on allele frequency and continues partitioning. After considering the compatible SNVs, the algorithm moves to any incompatible SNVs in the neighborhood in order of their physical proximity to the focal SNV. This process is repeated until each cluster contains only one sequence or there are no more SNVs to consider in the neighborhood. Thus, the method requires phased data. If a user has unphased data, phasing can be done in advance with software such as fastPHASE [26], BEAGLE [27] , IMPUTE2 [28], or MACH [29, 30].

3.3 Examples

This section gives worked examples illustrating how to reconstruct the partitions underlying a sample of DNA sequences. In addition, we show how to investigate the association between the reconstructed partitions and a user-specified partition. The association statistics we consider include the Rand index [42], the distance correlation (dCor) statistic [43], the Heller-Heller-Gorfin (HHG) statistic [44], the Mantel statistic [45], and the R-Vector (RV) coefficient [46]. The Rand index quantifies the association between two partitions directly. The dCor statistic, HHG statistic, Mantel statistic, and RV coefficient quantify the association between two distance matrices derived from partitions.

We first illustrate how to create a `hapMat` data object of SNV sequences. We then reconstruct a perfect phylogeny at a focal SNV. Next, we reconstruct perfect phylogenies across a genomic region. Finally, we show how to visualize and test associations between these reconstructed partitions and

- a comparator partition or dendrogram,
- a comparator distance matrix, and
- a phenotypic distance matrix.

To illustrate, we consider a toy example with 4 sequences comprised of 4 SNVs at positions 1, 2, 3, and 4 kilo-base pairs (kbp). The required `hapMat` object is created by executing the following command:

```
R> ex_hapMat  <- createHapMat(hapmat = matrix(c(1,1,1,0,
                                                0,0,0,0,
                                                1,1,1,1,
                                                1,0,0,0),
                                                byrow = TRUE,
                                                 ncol = 4),
                                 snvNames = c(paste("SNV",1:4,sep = "")),
```

```
                                hapNames = c("h1","h2","h3","h4"),
                                   posns = c(1000, 2000, 3000, 4000))
```

The structure of the resulting object of class `hapMat` is as follows.

```
R> ex_hapMat
   $hapmat
      SNV1 SNV2 SNV3 SNV4
   h1    1    1    1    0
   h2    0    0    0    0
   h3    1    1    1    1
   h4    1    0    0    0

   $posns
   [1] 1000 2000 3000 4000

   attr(,"class")
   [1] "hapMat"
```

If a user has a variant call format (`vcf`) file that consists of SNV data with a single alternative allele and no missing values in the genotype field, the `hapMat` data object can be created by supplying the file path to the `vcf` file as follows:

```
R> # specify the file path
   vcf_file_path <- "C:/vcfData/vcfData.vcf.gz"

   # Create a hapMat object
   ex_vcf_hapMat <- perfectphyloR::vcftohapMat(vcf_file_path)
```

Once the `hapMat` object is created, the user can reconstruct a perfect phylogeny at a focal SNV with `reconstructPP()`, by specifying the following four arguments:

1. `hapMat`: A data structure of class `hapMat`, created by `createHapMat()`.

2. `focalSNV`: The column number of the focal SNV at which to reconstruct the perfect phylogeny.

3. `minWindow`: Minimum number of SNVs around the focal SNV in the neighborhood of SNVs used to reconstruct the perfect phylogeny (default is the maximum of one and 2% of the total number of the SNVs).

4. `sep`: Character string separator to separate sequence names for sequences that can not be distingiushed in the neighborhood around the focal point. For example, if sequences "h1" and "h3" can not be distinguished and `sep = "-"`, then they will be grouped together with the label "h1-h3". The default value is `"-"`.

For example, consider the dataset `ex_hapMatSmall_data` comprised of 10 sequences and 20 SNVs. This dataset is a subset of the larger example dataset, `ex_hapMat_data`, that comes with the package. The larger dataset has 200 sequences and 2747 SNVs, and was used in a previously published association association analysis [47]. We can reconstruct a perfect phylogeny at the first SNV of `ex_hapMatSmall_data` by executing the following commands:

```
R> # Load the example hapMat data object.
   data(ex_hapMatSmall_data)

   # Reconstruct dendrogram at the first SNV of ex_hapMatSmall_data.
   rdend <- reconstructPP(hapMat = ex_hapMatSmall_data,
                          focalSNV = 1,
                        minWindow = 1,
                              sep = "-")
```

Figure 3.1 shows the reconstructed dendrogram, `rdend`, at the first SNV of `ex_hapMatSmall_data`. The structure of `rdend` is as follows:

```
R> str(rdend)
   List of 6
    $ edge          : num [1:6, 1:2] 5 6 6 5 7 7 6 1 2 7 ...
    $ Nnode         : int 3
    $ tip.label     : chr [1:4] "1249" "354-1009-2818" "2909"
    "1904-454-2931-2994-370"
    $ edge.length   : num [1:6] 6 3 3 4 5 5
    $ node.label    : NULL
    $ snvWinIndices: int [1:2] 1 5
    - attr(*, "class")= chr "phylo"
    - attr(*, "order")= chr "cladewise"
```

The user can extract the positions of the lower and upper limits of the neighborhood of SNVs used to reconstruct `rdend` as follows:

```
R> ex_hapMatSmall_data$posns[rdend$snvWinIndices]

   [1] 1500 7000
```

To see the sequences in the neighborhood of SNVs used for the reconstruction, the user can execute the following command:

```
R> ex_hapMatSmall_data$hapmat[, rdend$snvWinIndices[1]:
                                rdend$snvWinIndices[2]]
```

```
      SNV3 SNV4 SNV7 SNV9 SNV14
1904     0    0    0    0     0
454      0    0    0    0     0
1249     1    1    1    1     0
2931     0    0    0    0     0
2994     0    0    0    0     0
2909     0    0    0    0     1
354      1    1    1    0     0
1009     1    1    1    0     0
370      0    0    0    0     0
2818     1    1    1    0     0
```

As can be seen in the above output, there are two groups of sequences that have the same ancestral and derived alleles at each SNV position: sequences 354, 1009 and 2818, and sequences 1904, 454, 2931, 2994 and 370. These two groups of sequences therefore cannot be distinguished in the reconstructed partition. In Figure 3.1, we can verify that two tips of the partition are comprised of these two groups of sequences.

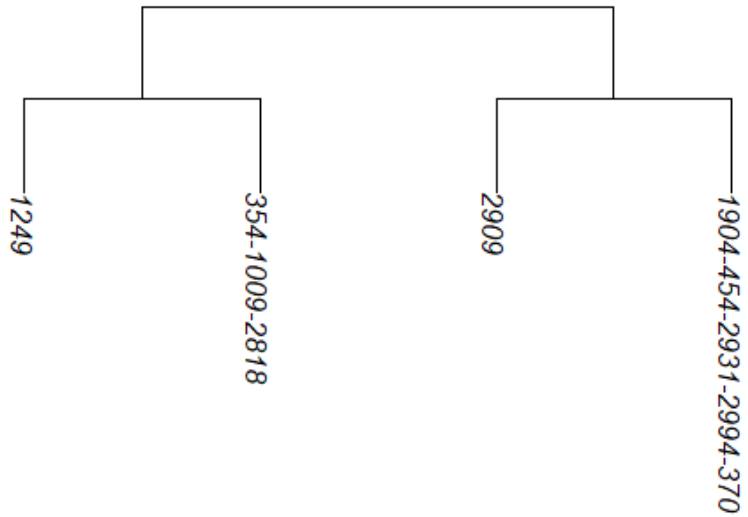

Figure 3.1: The reconstructed partition at the first SNV of `ex_hapMatSmall_data`.

With `reconstructPPregion()`, the user can reconstruct perfect phylogenies at each possible focal SNV in a `hapMat` data object. In the following example, we consider the

10 sequences with 20 SNVs in `ex_hapMatSmall_data`. We reconstruct perfect phylogenies across the 20 SNVs.

```
R> # Reconstruct partitions across the region.
   rdends <- reconstructPPregion(hapMat = ex_hapMatSmall_data,
                                 minWindow = 1)
```

`rdends` is an `ape` multiphylo object. The reconstructed partition at the first focal SNV in `ex_hapMatSmall_data` is the first `phylo` object in `rdends`:

```
R> str(rdends[[1]])
   List of 6
    $ edge          : num [1:6, 1:2] 5 6 6 5 7 7 6 1 2 7 ...
    $ Nnode         : int 3
    $ tip.label     : chr [1:4] "1249" "354-1009-2818" "2909"
                                "1904-454-2931-2994-370"
    $ edge.length   : num [1:6] 6 3 3 4 5 5
    $ node.label    : NULL
    $ snvWinIndices: int [1:2] 1 5
    - attr(*, "class")= chr "phylo"
    - attr(*, "order")= chr "cladewise"
```

If a user wants to reconstruct perfect phylogenies within a user-provided subregion of a `hapMat` object, they may specify the lower and upper values of the subregion in base pairs as follows:

```
# Reconstruct partitions between a given range SNV positions.
R> rdends_range <- reconstructPPregion(hapMat = ex_hapMatSmall_data,
                                       minWindow = 1,
                                         posn.lb = 500,
                                         posn.ub = 2000)
```

The function `testDendAssoRI()` uses the Rand Index to investigate the association between a comparator dendrogram or partition and multiple reconstructed dendrograms or partitions across a genomic region. `testDendAssoRI()` has five key arguments:

1. `rdend`: An `ape` multiphylo object of reconstructed dendrograms at each focal SNV.
2. `cdend`: An `ape` phylo object of the comparator dendrogram.
3. `hapMat`: An object of class `hapMat` containing SNV sequences.
4. `k`: An integer that specifies the number of clusters that the dendrogram should be cut into. The default is `k = 2`. Clusters are defined by starting from the root of the dendrogram, moving towards the tips and cutting across when the appropriate number of clusters is reached.

5. `nperm`: Number of permutations for the test of any association across the genomic region. The default is `nperm = 0`; i.e., association will not be tested.

To illustrate, we use the example dataset `ex_hapMat_data` with 200 sequences and 2747 SNVs. We plot the Rand index values summarizing the association between the comparator dendrogram at SNV position 975 kilobase pairs and the reconstructed dendrogram at each SNV position across the 2 Mbp genomic region (Figure 3.2a).

```
R> # Comparator true dendrogram at 975 kbp.
   data(tdend)

   # hapMat data object.
   data(ex_hapMat_data)

   # Reconstruct dendrograms across the region.
   allrdends <- reconstructPPregion(hapMat = ex_hapMat_data,
                                    minWindow = 55)

   # Rand index profile based on 6 clusters.
   RI_profile <- testDendAssoRI(rdend = allrdends,
                                cdend = tdend,
                                    k = 6,
                               hapMat = ex_hapMat_data,
                                nperm = 1000,
                                 xlab = "SNV positions (bp)",
                                 ylab = "Rand indices",
                                 main = "Association Profile")

    # Omnibus P value for overall associaion.
    RI_profile$OmPval
    [1] 0.000999001
```

Figure 3.2 shows the association profile between a comparator true dendrogram, `tdend`, at position 975 kbp, and a list of reconstructed dendrograms across the genomic region of `ex_hapMat_data`. In the two panels of the figure, the Rand indices are based on six and 24 clusters. Since we use simulated data, we know the true dendrogram at position 975 kbp. In Figure 3.2, using the Rand index, we investigate how the true dendrogram at position 975 kbp associates with the reconstructed dendrograms across the genomic region. As can be seen, the highest point for six clusters lies at position 975 kbp, and for 24 clusters is very close to position 975 kbp. According to the omnibus p-value, returned by `testDendAssoRI()`, the association across the genomic region is significant ($P \approx 0.001$) for both six and 24 clusters.

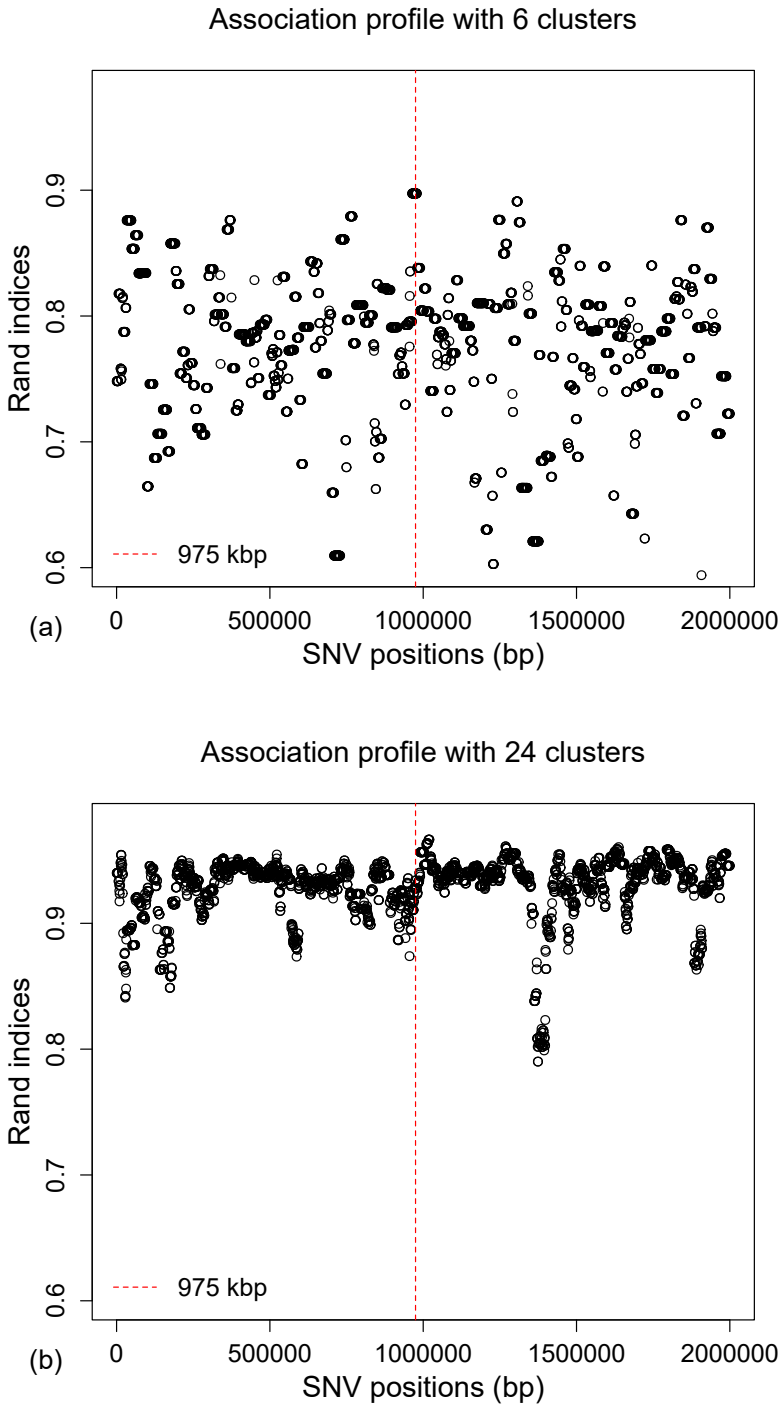

Figure 3.2: Rand indices associating a comparator true dendrogram at position 975 kbp and reconstructed dendrograms across the genomic region. **a)** Based on the six clusters. **b)** Based on 24 clusters. Red vertical dashed lines represent the position of the comparator dendrogram at 975 kbp.

The function `testAssoDist()` investigates the association between a comparator distance matrix and multiple reconstructed dendrograms across a genomic region. The association statistics available in the function are the dCor statistic, HHG statistic, Mantel statistic, and RV coefficient. The function has the following five key arguments:

1. `rdend`: An `ape` multiphylo object of reconstructed dendrograms at each focal SNV.
2. `cdmat`: A comparator matrix of pairwise distances (e.g. pairwise distances between sequences of a comparator dendrogram).
3. `method`: A character string specifying one of `"dCor"`, `"HHG"`, `"Mantel"` or `"RV"` for the dCor, HHG, Mantel or RV statistics, respectively.
4. `hapMat`: An object of class `hapMat` containing SNV sequences.
5. `nperm`: Number of permutations for the omnibus test of any association across the genomic region. The default is `nperm = 0`; i.e., association will not be tested.

To illustrate, we plot the dCor statistics summarizing the association between a comparator distance matrix, `cdmat`, and the reconstructed dendrograms across the genomic region of the example dataset `ex_hapMat_data`.

First, we compute the pairwise distances between sequences based on the comparator true dendrogram at SNV position 975 kbp. These pairwise distances are computed with the function `rdistMatrix()`, available in the package. The `rdistMatrix()` function uses the rankings of the nested partitions in the dendrogram to calculate rank-based distances between the sequences. However, users can provide any distance measures of interest for `cdmat`. We then plot the dCor statistic summarizing the association between the rank-based distance matrix for the reconstructed dendrograms at each SNV position and the comparator distance matrix at SNV position 975 kbp (Figure 3.3).

```
R> # Comparator true dendrogram at SNV position 975 kbp.
   data(tdend)

   # hapMat data object.
   data(ex_hapMat_data)

   # Compute rank-based distances between sequences based on
   # the comparator true dendrogram (tdend) using the function,
   # rdistMatrix().

   tdendDmat = perfectphyloR::rdistMatrix(tdend)
```

```
# Reconstruct dendrograms across the region.
allrdends <- reconstructPPregion(hapMat = ex_hapMat_data,
                                 minWindow = 55)

# dCor profile comparing the association between distance
# matrix of true dendrogram (comparator dendrogram)
# and all reconstructed dendrogram across the genomic region.

dCor_profile <- testAssoDist(cdmat = tdendDmat,
                             rdend = allrdends,
                            method = "dCor",
                            hapMat = ex_hapMat_data,
                             nperm = 1000,
                              xlab = "SNV positions(bp)",
                              ylab = "dCor Statistics",
                              main = "Association Profile")

# Omnibus p-value for overall association.
dCor_profile$OmPval
[1] 0.000999001
```

In Figure 3.3, we can clearly see the strongest association around the SNV position 975 kbp, and the association across the genomic region is significant ($P \approx 0.001$), as expected. The association signal is much clearer than for the Rand index plotted in Figure 3.2 because dCor uses the full information from the pairwise distance matrices whereas the Rand index is based on a discrete number of clusters.

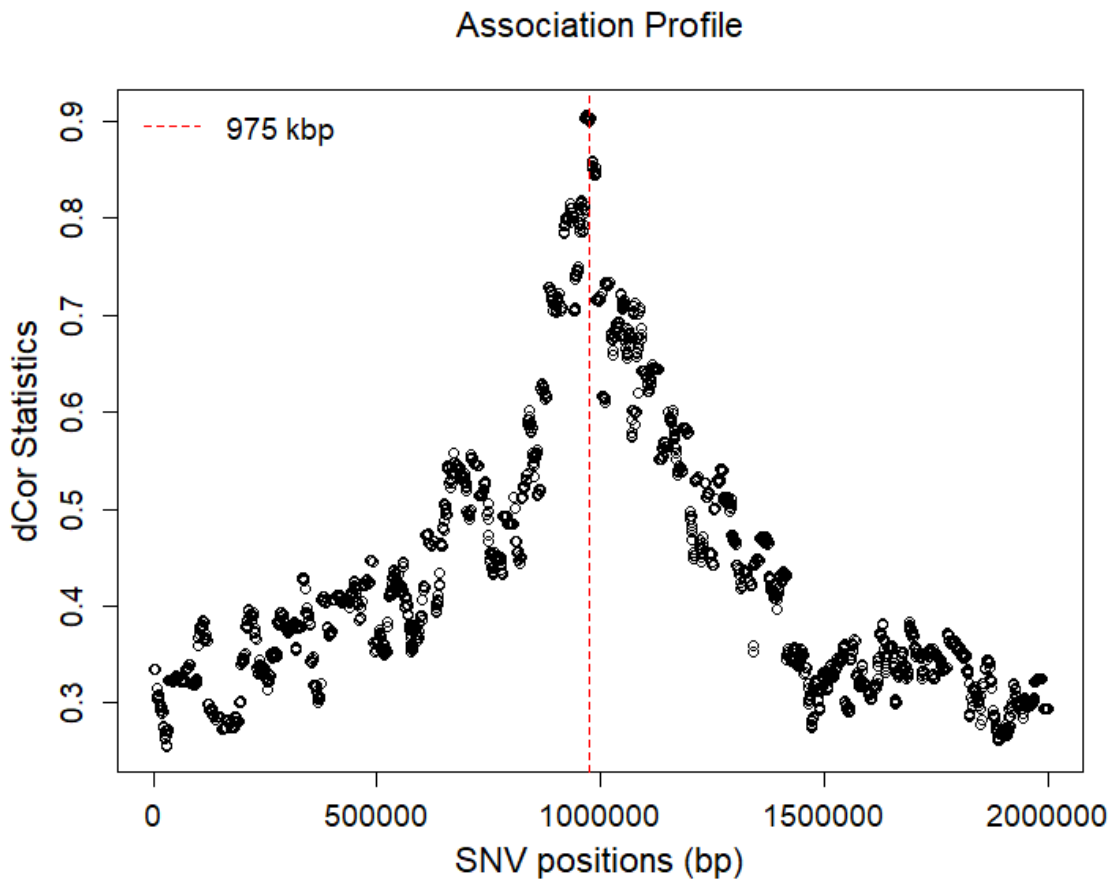

Figure 3.3: Associations between a comparator distance matrix from the true dendrogram at position 975 kbp and the reconstructed dendrograms across the genomic region. Red vertical dashed line represents the position of the comparator dendrogram at 975 kbp.

To illustrate another application of the function `testAssoDist()`, we perform the RV test of association between a phenotypic distance matrix as the `cdmat` argument and the reconstructed dendrograms across the genomic region of `ex_hapMat_data`. The phenotype data and distances are described in [47] and are contained in the data object `phenoDist`. Binary phenotype status was assigned based on causal SNVs from a causal subregion defined from 950 - 1050 kbp within the 2-Mbp genomic region.

```
R> # Phenotypic distances.
   data(phenoDist)
   # RV profile.
   RV_profile <- testAssoDist(cdmat = phenoDist,
                              rdend = allrdends,
                             method = "RV",
                             hapMat = ex_hapMat_data,
                              nperm = 1000,
                               xlab = "SNV positions (bp)",
                               ylab = "RV coefficients",
                               main = "Association Profile")

   # Indicate the region containing the causal SNVs.
   abline(v = 950000); abline(v = 1050000)

   # Omnibus P value for overall association.
   RV_profile$OmPval
   [1] 0.118
```

Figure 3.4 shows the resulting association profile between the phenotypic distances and the reconstructed dendrograms across the genomic region in `ex_hapMat_data`. The vertical lines indicate the causal subregion of 950 - 1050 kbp. The strongest association is close to the causal subregion. However, in this example, the association across the genomic region is not significant ($P \approx 0.1$).

3.4 Timing

Table 3.1 shows the computation times of the package's major functions. These computation times are for the 200 sequences comprised of 2747 SNVs in the example data `ex_hapMat_data` that is included in the package. Table 3.2 compares computation times of the function `reconstructPPregion()` for different numbers of sequences and numbers of SNVs. These times scale approximately linearly in the number of SNVs and quadratically in the number of sequences. Computation times are measured on an Intel E5-2683 v4 at 2.1 GHz with 20 GB of RAM.

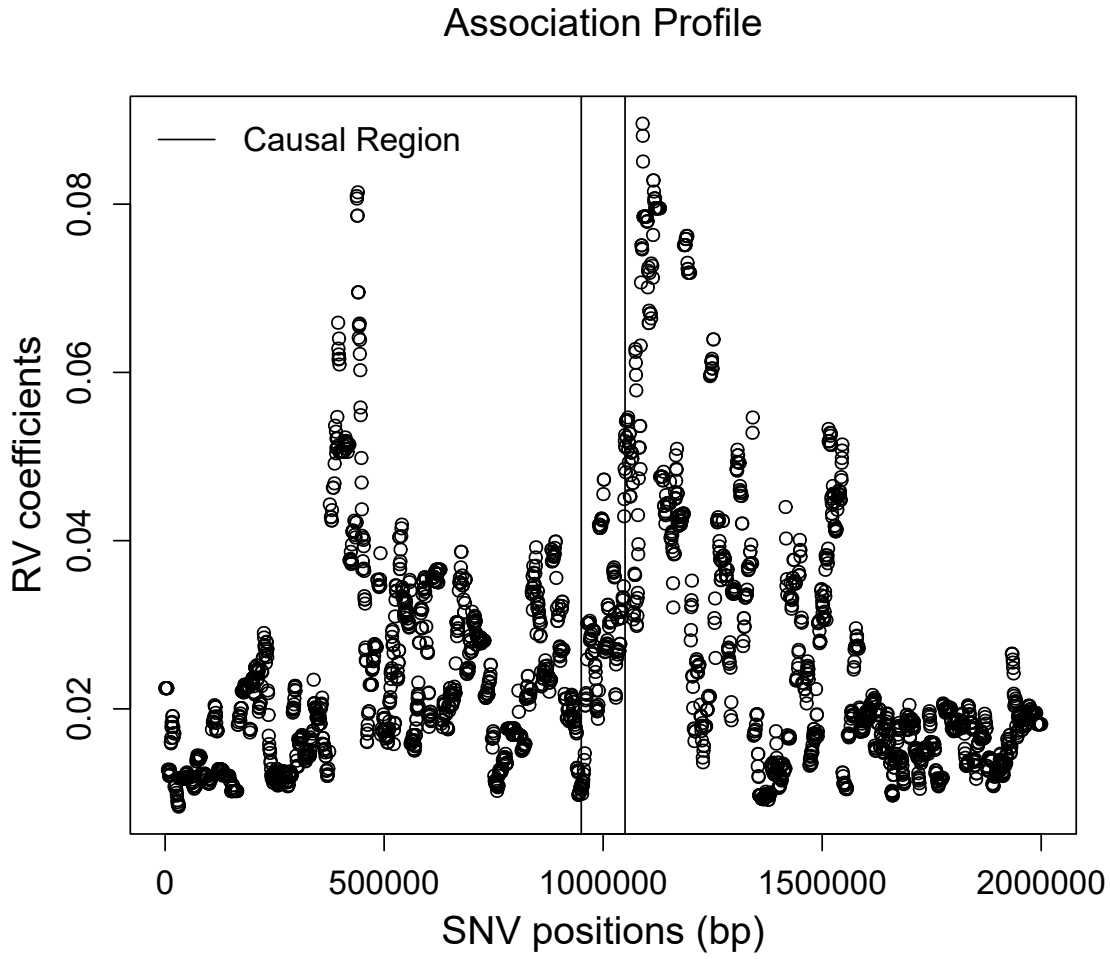

Figure 3.4: Associations between the phenotypic distance matrix and the reconstructed dendrograms across the genomic region. Black vertical lines indicate the limits of the genomic region containing trait-influencing SNVs.

Table 3.1: Computation times of the major functions of the package `perfectphyloR` for 200 sequences comprised of 2747 SNVs.

Function	**Computation Time (minutes)**	
	No permutation	**1000 permutations**
`reconstructPP()`	Few seconds	NA
`reconstructPPregion()`	12.00	NA
`testDendAssoRI()`	1.12	70.00
`testAssoDist()`	0.28	193.71

Table 3.2: `reconstructPPregion()` timing results (in minutes) for different number of sequences and SNVs.

	Number of sequences			
Number of SNVs	**200**	**300**	**400**	**500**
3000	17.66	21.47	26.07	26.38
4000	23.38	28.26	33.41	33.80
5000	29.99	34.82	41.40	41.87
6000	36.58	43.41	49.86	50.67
7000	43.17	51.68	58.72	59.42
8000	49.25	59.11	67.89	68.65
9000	55.63	66.75	77.00	77.92
10000	61.56	75.53	86.11	86.99

3.5 Discussion

We note that the computation time of `reconstructPPregion()` can vary a lot based on the size of the `hapMat` object (Table 3.2). Starting from the first SNV of the `hapMat` object, this function continues the reconstruction process until the last SNV. At each focal SNV, the function starts from ground level to construct a surrounding window of SNVs and rebuilds the partition, without utilizing the information from previously constructed partitions at nearby SNVs. As a result, many of the same computations may be done several times for similar focal SNVs. As noted by a reviewer, there may be ways to make `reconstructPPregion()` faster. For example, clustering similar successive SNVs before starting the reconstruction could lead to computational efficiencies and would be an avenue for future work.

Although we know of no software that is directly comparable to `perfectphyloR`, the PerfectPhy suite of tools is also set up to return sequence partitions. We therefore explored the use of PerfectPhy in a moving-window approach similar to that of `perfectphyloR`. Briefly, for each placement of the moving window, the following two steps were repeated: (i) filter out incompatible SNVs in the window and (ii) reconstruct the perfect phylogeny using the remaining compatible SNVs. We applied this approach to the 200 sequences in the example dataset, `ex_hapMat_data`, using the default minimum-window size of 55 for 2747 SNVs. For the first few window placements, we compared the computational time of steps (i) and (ii) in the PerfectPhy-based approach to that of `reconstructPP()` in `perfectphyloR`. For the PerfectPhy approach, the filtering step is the bottleneck, with computation times in excess of 600 minutes. By contrast, `reconstructPP()` took no more than 0.18 seconds.

3.6 Conclusion

The R package `perfectphyloR` provides functions to reconstruct a perfect phylogeny at a user-given focal SNV and perfect phylogenies across a genomic region of interest. The package also computes, tests and displays association measures based on the reconstructed partitions in a genomic region. The reconstructed partitions are useful to researchers seeking insight into the ancestral structure of DNA sequences. For example, associating the reconstructed partitions with a trait can help to localise trait-influencing variants in association studies. `perfectphyloR` can be freely downloaded from the Comprehensive R Archive Network (CRAN) or from `https://github.com/cbhagya/perfectphyloR/`.

3.7 Availability and requirements

Project name: perfectphyloR
Project home page: https://CRAN.R-project.org/package=perfectphyloR
Operating system(s): Windows, Linux, OS X
Programming language: R
Other requirements: R 3.4.0 or newer
License: GPL-2, GPL-3
Any restrictions to use by non-academics: none

The package `perfectphyloR` can be installed from CRAN using `install.packages ("perfectphyloR")`. The local zip file can be installed using R Studio by selecting the install package(s) from local zip files.

3.8 List of abbreviations

DNA: Deoxyribonucleic acid; BLOSSOC: BLOck aSSOCiation;
SNV: Single Nucleotide Variant; dCor: Distance Correlation;
RI: Rand Index; HHG: Heller-Heller-Gorfin;
RV: R-Vector, a vector version of standard r correlation;
GHz: Giga Hertz; GB: Gigabyte; RAM: Random Access Memory;
CRAN: Comprehensive R Archive Network

3.9 Acknowledgements

We thank the anonymous reviewers for constructive comments that improved and clarified the manuscript. We also thank Christina Nieuwoudt and Kelly Burkett for helpful discussions and comments, and the Department of Statistics and Actuarial Science at Simon

Fraser University for its generous support. This research was funded in part by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

Chapter 4

Fine-mapping rare variants by gene genealogies in case-control studies

4.1 Introduction

Different association methods are available to fine-map genetic variants, including single-variant and aggregation based approaches under the genotypic-association mapping. However, these genotypic-association mapping approaches do not consider the sequence-relatedness which can be useful for fine-mapping genetic variants with inherited traits.

As an alternative approach to the genotypic-association mapping, we can group sequences based on their relatedness by considering their gene genealogy which describes relationships among sequences sampled from a population (e.g., [4], [47]). The case sequences carrying a rare causal variant are assumed to be descendant from a common ancestral sequence. As a result, they are identical-by-descent (IBD) around a causal variant and will cluster together on the gene genealogy. Note that, two DNA sequences are IBD, if they are inherited from the same ancestral sequence in a reference population without an intervening mutation. This clustering behavior is useful to fine-map causal variants for methods based on sequence relatedness or IBD (e.g., [27], [4], [47]). In IBD mapping of a disease trait, causal variants are localized to a genomic region by associating the sharing of IBD DNA segments with the sharing of trait values. The intuition is that sequences that carry the same disease-predisposing variant should share the same segment inherited IBD from an ancestral sequence on which the disease-predisposing variant arose. By contrast, genotypic-association mapping associates individual genotypes directly with trait values. IBD and genotypic-association mapping are thus conceptually complementary approaches to finding causal variants for a trait. Because genotypic-association methods associate individual genotypes directly with trait values, they lose power when a disease locus harbours several causal variants. IBD methods on the other hand are robust to such allelic heterogenity because they associate the sharing of IBD DNA segments with the sharing of trait values.

Generally, the more closely related two sequences, are the more sharing of IBD segments they have. Two sequences with a recent common ancestor will tend to share longer DNA segments IBD than two sequences with a distance common ancestor. At the genomic location of a causal variant, the ancestral tree will tend to cluster the sequences that carry the causal variant, and their cluster membership will be correlated with the disease trait. Burkett et al. [4] investigated the utility of ancestral tree-based methods to detect multiple rare variants that contribute to complex disease in haploid populations. Karunarathna and Graham [47] extended the investigation to consider the localization ability of tree-based methods in diploid populations. They found that classiftying case sequences into carriers and non-carriers of causal variants improved the fine-mapping ability of IBD methods. However, both these studies relied on IBD information from ancestral trees for the DNA sequences, information that is not available in practice. In other work, Browning and Thompson [48] investigated the power of IBD mapping to detect a complex-disease locus in case-control studies. These authors contrasted the rates of IBD in case/case and non-case/case pairs of individuals at each single-nucleotide variant (SNV) and found that IBD mapping had higher power than genotypic-association mapping under alleclic heterogeneity.

In this work, we investigate the ability of different association methods to *detect* and *localize* rare causal variants in an allelically heterogenous disease with high genetic penetrance ratio. We view the sequence-relatedness method as a linkage method because it associates similarity in trait values to similarity in relatedness. However, the sequence-relatedness method we consider is not a traditional genetic linkage method in families because it uses a case-control sample from a population rather than families. As discussed in Ott et al. [49], linkage methods work better than traditional association methods for fine-mapping rare causal variants in an allelically heterogenous disease with high genetic penetrance ratio. We therefore set up the parameters in the model so that the penetrance ratio is large and multiple rare causal variants are influencing the trait. Thus, sufficiently powered linkage studies preceding association fine-mapping would be expected to produce the candidate region currently under investigation. We consider rare causal variants with very low frequency in the population since they are hard to detect by standard single-variant association methods. These rare causal variants are simulated to lie in a 100 kilo-base pair (kbp) subregion of a 2Mbp candidate genomic region. Since the ancestral trees of the DNA sequences are not known in practice, we do not know true relatedness of sequences. Therefore, we estimate the relatedness of sequences by using the methods developed in Chapter 3. Through coalescent simulation, we compare the ability of the proposed IBD-based methods with two popular genotypic-association (non-IBD) methods to *detect* and *localize* causal variants. Under non-IBD methods, we consider Fisher's exact test as a classical approach and SKAT-O which is a powerful regression approach for detecting rare variants. Under IBD methods, we consider the distance correlation and the Mantel test. However, the Mantel test has been criticized to be a biased test in the presence of nonexchangable units [50]. We therefore investigate

whether the Mantel test is biased in our context. To illustrate the ideas, we start by working through a particular example dataset as a case study. We then perform a simulation study inloving 500 datasets to compare the ability of the methods to *detect* and *localize* the causal variants. Moreover, we introduce two different post-hoc analyses after the *detection* and *localization*. We consider a method to help with discerning true-positive from false-positive results of SKAT-O. We also explore the idea of classifying the case sequences into carriers versus non-carriers of causal variants and then grouping the non-carrier case sequences with the control sequences.

4.2 Methods

In this section, we first describe how we simulated sequence and trait data, the association methods we considered, and the way we evaluated the ability of association methods to *detect* and to *localize* causal variants. We then present two post-hoc analyses: discerning true-positive from false-positive signal of SKAT-O and classifying case sequences into carriers and non-carriers of causal variants.

4.2.1 Data simulations

4.2.1.1 Simulating sequences

We simulated 500 datasets of 50 affected individuals (cases) and 50 unaffected individuals (controls) as follows. We used msprime [51] to simulate ancestral trees and 3000 haplotypes of around 8400 single-nucleotide variants (SNVs) in a candidate genomic region of length 2 million base pairs (Mbp). We set recombination rate to 1×10^{-8} per base-pair per generation [15] with the mutation rate of 2×10^{-8} per base-pair per generation [52] in a diploid population of constant effective size, $N_e = 6200$ [47]. To mimic random mating in a human population, we then randomly paired the 3000 haplotypes into 1500 individuals. We used the 1500 individuals as our population in what follows.

4.2.1.2 Disease trait model

We assigned disease status to the 1500 individuals based on randomly sampled causal SNVs from the middle subregion of 950-1050 kilobase pairs (kbp). For causal SNVs, the number of copies of the derived allele increased the risk of disease according to a logistic regression model,

$$logit(P(D = 1|G)) = \beta_0 + \beta_1 \sum_{j=1}^{m} G_j$$

where

- $logit(p) = log[p/(1-p)]$ for $0 < p < 1$,
- D is disease status (D = 1, case; D = 0, control),
- $G = (G_1, G_2, \cdots, G_m)$ is an individual's multi-locus genotype at m causal SNVs, with G_j being the number of copies of the derived allele at the j^{th} causal SNV,
- β_0 is the intercept term that represents the sporadic disease, $P(D = 1|G = 0)$ and
- β_1 is a penetrance parameter that measures the effect of causal variants on the disease.

4.2.1.3 Simulations under null hypothesis of no association

Without using the disease-trait model above, we randomly assigned disease status to the 1500 individuals in the population. Out of 1500 individuals in the population, 75 were randomly assigned as disease affected and the rest as unaffected, to ensure a population disease prevalence of $75/1500 = 5\%$. After assigning the disease status, we sampled 50 cases (i.e. diseased) from 75 affected individuals and 50 controls (i.e. non-diseased) from 1425 unaffected individuals. We then extracted the SNV data from the case-control sample for the analysis.

4.2.1.4 Simulations under alternative hypothesis of association

We assigned disease status to all individuals in the population according to the disease trait model above. We set up this model to have high penetrance ratio to ensure successful linkage analysis [49]. The penetrance ratio is g/f, where g is disease penetrance and f is phenocopy or sporadic rate. We set the value of β_0 in the linear model so that the $f = P(D = 1|\sum_{j=1}^{m} G_j = 0) = 4.5 \times 10^{-5} \approx 0$. We set $\beta_1 = 16$ so that $g \approx P(D = 1|\sum_{j=1}^{m} G_j = 1) = 0.9975$, as would be expected for rare variants of high penetrance. In our study, the penetrance ratio is about 22,167, and the log odds ratio is $\log(\frac{g}{1-g}\frac{1-f}{f}) = \beta_1 = 16$. We aimed for an allelically heterogeneous disease with 20 major causal variants of approximately equal population frequency (MAF $\sim$ 0.6%). These 20 major variants accounted for as much of the population prevalence as possible. When necessary, additional rare variants were chosen to be causal to achieve the targeted disease prevalence of 5%. Further details about the selection procedure for causal variants can be found in Appendix A. After assigning disease status to the 1500 individuals, we randomly sampled 50 cases from the affected individuals and 50 controls from the unaffected individuals. We then extracted the SNV data from the case-control sample for the analysis.

4.2.2 Association methods

In the next two sections, we review the association analysis we consider for fine mapping. The methods we consider fall into two categories: non-IBD and IBD-based. Non-IBD association methods associate trait values to genotypic values. By contrast, IBD-based association methods associate similarity in trait values to similarity in relatedness or identity-by-descent among segments of DNA sequences.

4.2.2.1 Non-IBD association methods

We considered Fisher's exact test and an optimal sequence kernel association test [53]. These methods do not consider relatedness among sequence segments but rather focus on the association between DNA variants and the trait.

Fisher's Exact Test

We tested the disease association with genotypes for each SNV using a standard Fisher's exact test as implemented in the `stats` package in base R. Specifically, each of the SNV sites was tested for an association with the disease outcome using a 2×3 table to compare the genotype frequencies. We recorded the -$\log_{10}$ exact p-value as the association signal from each SNV.

SKAT-O

For rare variants, the power of classical association tests, such as Fisher's exact test, is limited. We therefore considered an optimal test in an extended family of sequence kernel association tests (SKAT), known as SKAT-O [53]. SKAT is an efficient regression method to test for association between genetic variants in a region and a continuous or dichotomous trait. SKAT is more powerful than burden tests (eg., [54], [55]) when a large fraction of the variants in a region are noncausal, or the effects of causal variants on the trait are in different directions [53]. In contrast, burden tests are more powerful when the target region has many causal variants and the effects of the causal variants are in the same direction . By using the data, SKAT-O finds the optimal linear combination of the burden test and SKAT to maximize the power. Thus, SKAT-O maintains power in both scenarios. We applied the SKAT-O test using the `SKAT()` function in the R package `SKAT` [18] with window size of 21 SNVs. Windows were centered at a target SNV and extended up to 10 SNVs to the left and to the right. Therefore, at the edges of the genomic region, the window sizes are smaller than 21 SNVs.

4.2.2.2 IBD-based association methods

IBD-based association methods take into account the information on the identity-by-descent of DNA segments in their analysis. Specifically, these methods associate the clustering of DNA segments with the clustering of trait values. First, sequence variation is used to reconstruct how the sampled sequences cluster together, or are partitioned. Next, distances between sequences in a partition are computed. Then, the clustering of trait values for the sequences are evaluated. Finally, the clustering of DNA sequences is associated with the clustering of trait values. We next describe these steps.

Reconstructing sequence partitions

To reconstruct partitions from the sampled sequence data, we applied methods implemented in Chapter 3 (see Section 3.3). Using the function `reconstructPPregion()` described in Section 3.3, and a minimum window size of 500 variants, we reconstructed partitions across the 2-Mbp genomic region. A window size of 500 appeared to adequately capture the correlation between the reconstructed and true partitions. As noted in Chapter 1, partitions of sequences are not genealogical trees because we do not have information on coalescent times. However, the partitions provide information on the nested structure of the sequence clusters and therefore on relationships. This IBD information can be useful in the detection and localization of causal rare variants for a disease.

Sequence distances on partitions

To compute the pairwise distances between sequences in the reconstructed partition, we used the function `rdistMatrix()` in the package `perfectphyloR`. The function `rdistMatrix()` computes the pairwise distances between the sequences based on the rankings of the reconstructed nested partitions. In a reconstructed partition, this function assigns a distance of one between a node and its descendant. For example, Figure 4.1 shows a reconstructed partition of four sequences after assigning the distances. In the figure, the distance between sequences 1 and 2 is 2.

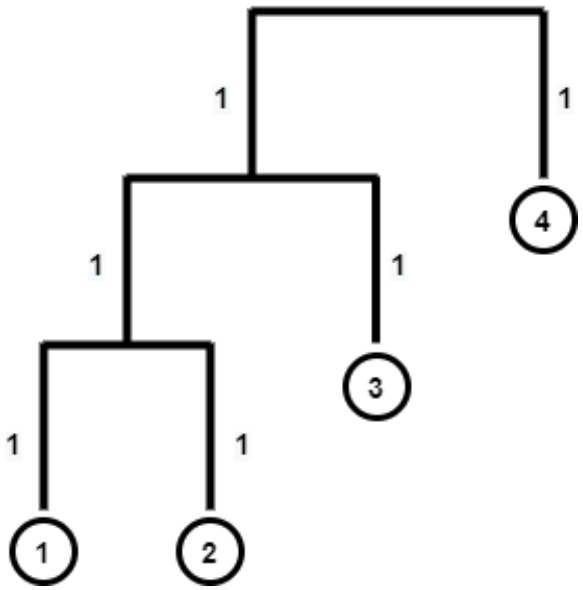

Figure 4.1: Partition showing the distances assigned to four sequences in the function `rdistMatrix()`.

Phenotypic distances

To compute the phenotypic distances, we used the method described in [4]. Briefly, following [56], the phenotypic distance between sequence i and j is $d_{ij} = 1 - s_{ij}$, where: $s_{ij} = (y_i - \mu)(y_j - \mu)$, a phenotypic similarity score between sequence i and j; y_i is the binary phenotype (0 or 1); and μ is the disease prevalence in the population.

Distance associations

To associate sequence distances and phenotypic distances, we use the distance correlation [43] and the Mantel statistics [45]. The distance correlation measures (possibly nonlinear) dependence between two random vectors of any dimension. We used the distance formulation of the statistic in [57]. The Mantel statistic measures the Pearson correlation coefficient between two distance matrices by considering the two vectors of distances. We compute the distance correlation and the Mantel statistic between these two distance matrices at each SNV position in the genomic region.

4.2.3 Scoring Detection

In this section, we describe how we evaluate the ability of both non-IBD and IBD-based association methods to *detect* the association signal. We first explain the global test of detecting association across the genomic region. We then explain how we compare the association methods. Finally, we discuss about the type-I error rate and power.

4.2.3.1 Global tests

To obtain a global test of association across the 2Mbp genomic region, we used a maximum test score across all the SNVs in a dataset with a given method. The nominal level of all tests was 5%. For the global test statistic, we used either a maximum distance correlation statistic or maximum of Mantel statistic from IBD-based methods, or the maximum of -$\log_{10}$ of the p-values from the non-IBD-based methods, across the genomic region. We determined the null distribution of the global test statistic for each method by permuting the case-control labels 1000 times. To make these statistics comparable across the methods, we considered their permutation p-values. We defined these p-values as the proportion of test statistics under the permutation null distribution that are greater than or equal to the observed value.

4.2.3.2 ECDF

To compare the distribution of the resulting p-values for the different methods, we plotted their empirical cumulative distribution function (ECDF). The ECDF at any point x is the proportion of the 500 simulated datasets with a p-value less than or equal to x. Therefore, any method with a higher value of the ECDF than the other methods, has a larger proportion of datasets with p-value less than x. Specifically, any method with a higher value of the ECDF for $x = 0.05$ detects the association signal more readily than the other methods at level 0.05.

4.2.3.3 Type-I error rate and power

For each method, we estimate type-I error rate and power by taking the proportion of 500 datasets that are rejected at level 5%, when these datasets are simulated under null hypothesis of no association and the alternative hypothesis of association, respectively. In particular, we are interested in the type-I error rate of the Mantel test because it has been criticized to be biased (i.e. to have inflated type-I error rate) in the presence of nonexchangable units [50]. We do not anticipate bias in our context because we permute the phenotype labels of individuals and individuals are exchangeable. We also report the approximate 95% confidence interval for the true type-I error rate of each method.

4.2.4 Scoring localization

To evaluate the localization, we scored the distance of the maximum association signal from the causal region in base pairs using the 500 datesets simulated under the alternative hypothesis. When there were multiple peaks, we used the average absolute value of the distance of all maxima across the 2Mbp genomic region. For each method, on each dataset, we computed the average distance of the peak association signals from the disease causal

region and plotted the ECDF of the average based on the 500 simulated datasets. The ECDF at point x is the proportion of the 500 simulated datasets with average distance less than or equal to x. A method with higher ECDF than another method localizes the signal better.

4.2.5 Post-hoc analysis

In this section, we introduce two different post-hoc analyses after detecting the association signal and localizing the rare causal variants. We first consider signal-to-noise ratio for discerning between true- and false-positive signals after SKAT-O detects association. We then introduce a labelling method to classify case sequences into carrier and non-carriers of causal variants.

4.2.5.1 Diagnostic for SKAT-O detection

To localize causal variants, we consider the association profile. By association profile, we mean the collection of association statistics at SNVs across the entire genomic region. In an association profile, a peak value or "signal" that stands out clearly from the background variation or "noise" draws more attention than a value that is indistinguishable from the background. When we focus on the datasets for which SKAT-O detects association, those simulated under association have a clearer signal in their Mantel-statistic profiles than those simulated under no association. We therefore explore the idea of using the Mantel-statistic profiles to distinguish true-positive from false-positive SKAT-O association. To contrast the size of the signal relative to the noise (i.e. signal-to-noise ratio) in the Mantel profile, we use the 95^{th} percentile of the absolute value divided by the interquartile range (IQR). We choose the 95^{th} percentile and the IQR because they are robust measures of the extreme and spread of a distribution, respectively. We compare these signal-to-noise ratios between two groups of datasets for which SKAT-O detects association: datasets simulated under the null and alternative hypothesis. We refer to the datasets simulated under the null hypothesis and rejected by SKAT-O as incorrectly rejected, and the datasets simulated under alternative hypothesis and rejected as correctly rejected. If datasets from the incorrectly rejected group have lower signal-to-noise ratios than datasets from the correctly rejected group, the Mantel localization profiles should have practical use for distinguishing true and false patterns of association.

4.2.5.2 Classifying case sequences

In this section, we describe how the case sequences can be classified into carriers and non-carriers of causal variants. Initially, we assume that all the case sequences are carriers of causal variants. We then classify the case sequences as non-carriers of causal variants using

the genealogical nearest neighbor [6] as described below. All control sequences are considered to be non-carriers.

Genealogical nearest neighbors

In practice, we do not know which of the case sequences carry causal variants and which do not. We therefore classify the case sequences into carriers and non-carriers of causal variants, using the idea of the genealogical nearest neighbor or GNN. The idea is that case sequences carrying a rare variant are descended from a common ancestral mutation that arose relatively recently back in the time. We therefore expect these case sequences to cluster together in the reconstructed partition and to be genealogical nearest neighbors. Our GNN statistic is based on the topological properties of the genealogical trees, as summarized by the reconstructed partitions of sequences. We take the average proportion of a sequence's nearest neighbours that are case sequences, where the average is weighted by the genomic length of the sharing. To illustrate the computation, we consider a toy example of four sequences of length 10 kbp, as shown in Figure 4.2.

In the figure, the subregion from 0 kbp to 6 kbp is spanned by the partition A, and the rest of the region by the partition B. In other words, when we reconstruct the partition at each SNV position within the first 6 kbp, we have only one partition structure (partition A), and the rest of the region has the structure of partition B. Then the GNN proportion for each sequence can be computed as follows. Suppose we choose sequence 1 as our target sequence. In partition A, we go upward from sequence 1 until we find the first internal node, u. All the sequences that descend from this u, excluding the target sequence 1, are the genealogical nearest neighbors of sequence 1. The GNN statistic for sequence 1 is the proportion of these neighbours (excepting the target sequence) that are case sequences within the clade below u. Considering all possible target sequences, we obtain a vector, G_A, of GNN statistics for partition A, with elements indexed by the target sequence. For example, in partition A, the GNN proportions for the four target sequences are: $1/1 = 1$, for sequence 1, $1/1 = 1$, for sequence 2, $2/2 = 1$, for sequence 3, and $2/3 = 0.67$, for sequence 4. Once we compute G_A, this vector is weighted by the proportion of genomic region spanned by partition A. Since partition A spans 60% of the total region of 10 kbp, the corresponding weight, $W_A = 0.6$. We repeat this process for the partition B. After that, we compute the average GNN proportion by taking the weighted average of all these proportions in both partitions. By taking the weighted average, we assign more weight to the partitions corresponding to long physical lengths of sequence than partitions corresponding to short lengths. The resulting vector,

$$GNN_{prop} = \frac{G_A W_A + G_B W_B}{W_A + W_B},$$

summarizes the proportion of nearest neighbours that are case sequences, along the genomic region. The intuition is that any case sequence with a lower value of the average GNN

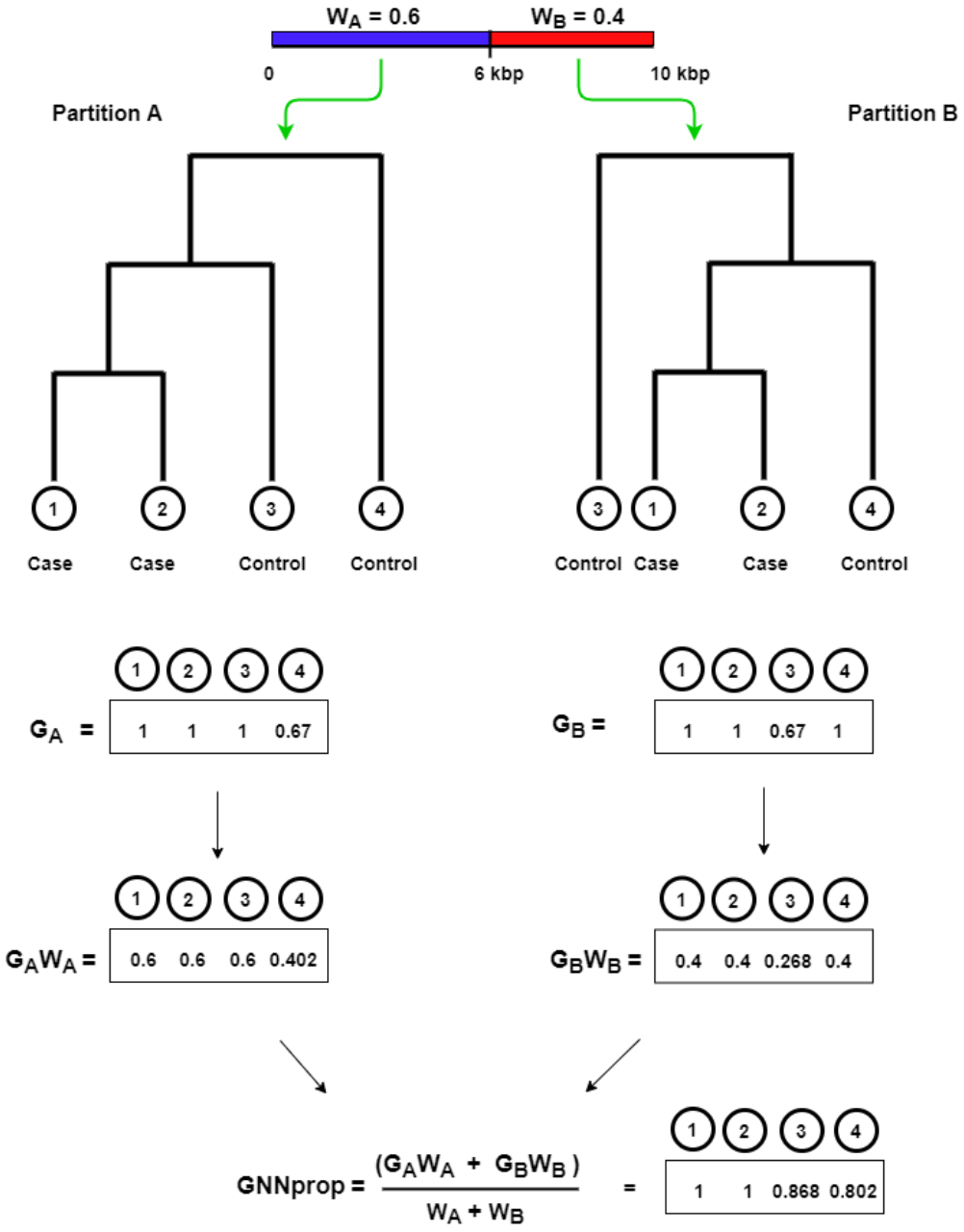

Figure 4.2: A toy example showing the computation of GNN proportions for the sequences.

proportion is more closely related to controls than to cases. We assume case sequences are carriers of causal variants unless the GNN proportion suggests otherwise. We compare the average GNN proportion of each case sequence to the distribution of the average GNN proportion in control sequences. Specifically, if a case sequence has an average GNN proportion that is lower than 25^{th} percentile in control sequences, we classify it as a noncarrier.

Misclassification error

Since we simulated the data, we know which of the case sequences are the carriers of causal variants and which are not. For one dataset, we therefore create the confusion matrix for 100 case sequences by what we refer to as naive and GNN labelling as shown in Table 4.1. In naive labelling, all the case sequences are assumed to be carriers of causal variants. In GNN labelling, certain case sequences are declared to be carriers of a causal variant based on their average GNN proportion as described above.

Table 4.1: Confusion matrix for case sequences

N=100		Declared carrier	
		No	Yes
True carrier	No	a	b
	Yes	c	d

We define the misclassification error for the above confusion matrix as:

$$me = \frac{b + c}{N}$$

We then compare the missclassification error rates of case sequences by naive and GNN labelling methods for all 500 datasets.

4.3 Results

In this section, we first present the results of an example dataset for insight into the methods. We then present results of a simulation study of 500 datasets. In the simulation study, we investigate the ability of the methods to *detect* and *localize* causal variants for an allelically heterogeneous disease with low sporadic rate and high penetrance. We further present the results of our post-hoc analysis.

4.3.1 Example dataset

4.3.1.1 Population and sample summaries

In the example population of 1,500 individuals, we obtain 8,394 SNVs of which 20 are causal variants. Of 8,394 SNVs in the population, 5,574 are polymorphic in the case-control sample of 50 affected and 50 unaffected individuals. Of the 20 causal variants in the population, 19 are polymorphic in the case-control sample. The linkage disequilibria between the polymorphic causal variants in the sample is low; all pairwise r^2 values are ≤ 0.1 (results not

shown). Table 4.2 summarizes the number of case and control sequences that carry each causal variant, in the population and in the case-control sample, respectively. The table also shows the derived allele frequency (DAF) of each causal variant in the population. All 20 causal variants are rare with population DAF $\leq 0.17\%$, and none are carried by control sequences in the population. The population of 1,500 individuals has 79 cases. All 79 cases in the population carry at lease one causal variant: 78 cases carry one variant, and one case carriers two.

Table 4.2: Summaries of causal variants.

Labels	Position (kbp)	Population counts			Sample counts	
		Case sequences	Control sequences	DAF(%)	Case sequences	Control sequences
SNV 3906	950.355	4	0	0.13	4	0
SNV 3912	951.583	5	0	0.17	2	0
SNV 3916	953.211	5	0	0.17	0	0
SNV 3927	956.339	4	0	0.13	3	0
SNV 3942	958.856	4	0	0.13	3	0
SNV 3953	961.845	4	0	0.13	2	0
SNV 3961	964.689	4	0	0.13	2	0
SNV 3975	967.138	4	0	0.13	1	0
SNV 3989	970.697	4	0	0.13	3	0
SNV 3994	972.600	5	0	0.17	5	0
SNV 4116	1000.587	4	0	0.13	2	0
SNV 4128	1005.069	4	0	0.13	2	0
SNV 4188	1018.687	4	0	0.13	3	0
SNV 4276	1040.663	3	0	0.10	3	0
SNV 4291	1042.789	3	0	0.10	2	0
SNV 4297	1043.136	4	0	0.13	2	0
SNV 4300	1043.694	3	0	0.10	3	0
SNV 4301	1044.044	4	0	0.10	4	0
SNV 4304	1044.809	4	0	0.13	2	0
SNV 4308	1045.703	4	0	0.13	3	0

4.3.1.2 Association profiles

Figure 4.3 shows the association profiles for the example dataset. In panels a, b and c, the horizontal dashed line shows the 5% significance threshold for the association test based on permutation, after adjusting for multiple testing across the entire genomic region. Panels a and b show results from Fisher's exact test and SKAT-O, respectively. In our example dataset, these non-IBD association methods do not localize the peak signal to the disease causal region. Panels c and d show the results from IBD-based association methods distance

correlation and Mantel, respectively. As can be seen in panels c and d, the Mantel test successfully localizes the peak signal to the disease causal region but not distance correlation.

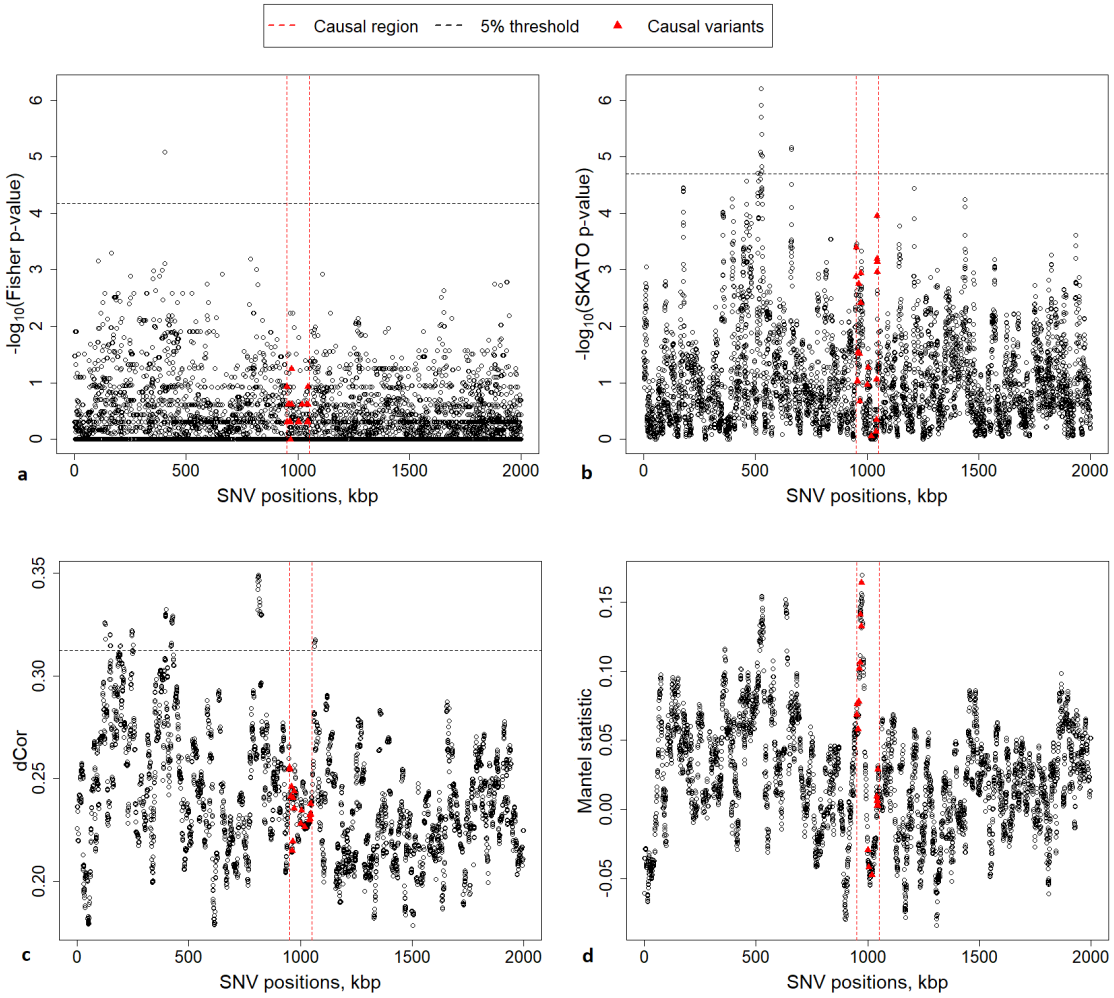

Figure 4.3: Association profiles for: a) Fisher's exact test, b) SKAT-O, c) Distance correlation (dCor), and d) Mantel statistic. The maximum value of variant-specific statistcs over the entire genomic region is used in a permutation test for the presence of any association. The horizontal dashed line shows the 5% significance threshold based on 1000 permutations and adjusted for multiple testing across the entire genomic region. The p-values for detecting any association are 0.007, 0.004, and 0.002, for Fisher's exact test, SKAT-O, and dCor, respectively. Note that we do not report the p-value for the Mantel test because of the concerns about the type-I error rate of the Mantel test (See section 4.3.2.1).

4.3.1.3 Performance of GNN labelling

In this section, we assess the performance of GNN labelling for classifying case sequences into carriers and non-carriers of causal variants in our example dataset. Figure 4.4 shows boxplots of the average GNN proportions for each sequence in the example dataset, grouped by their status as case carriers of causal variants, case non-carriers of causal variants or control sequences. The horizontal red-dashed line shows the 25^{th} percentile of the average GNN proportion in control sequences.

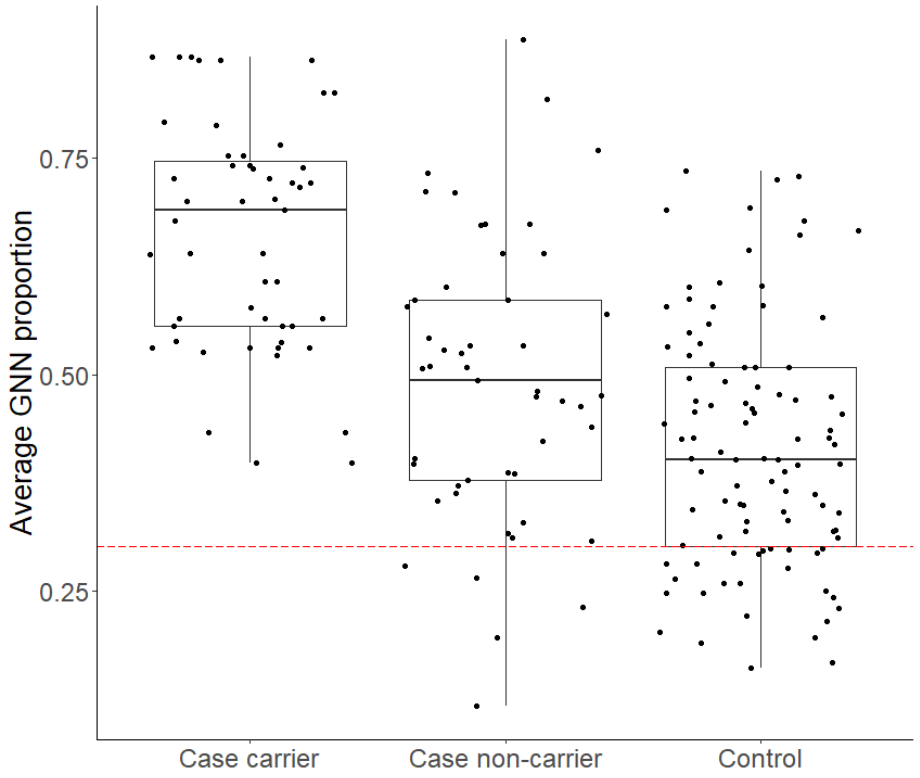

Figure 4.4: Average GNN proportions of sequences grouped by their status as case carriers or case non-carriers of causal variants and controls. The horizontal red-dashed line shows the 25^{th} percentile of average GNN proportion in control sequences.

For the GNN labelling, we classify all the case sequences with lower average GNN proportions than the 25^{th} percentile in controls as non-carriers, and the rest of the case sequences as carriers. The results are shown in Table 4.3a, with the naive labelling for comparison in 4.3b. In the table, none of the 51 carriers are misclassified as noncarriers by either the GNN or naive labelling. Also, of the 49 noncarriers, 44 are misclassified as carriers by GNN labelling and 49 by naive labelling. Thus, the misclassification rate of case sequences under GNN labelling $\left(\frac{44+0}{100} = 0.44\right)$ is less than the rate under naive labelling $\left(\frac{49+0}{100} = 0.49\right)$.

Table 4.3: Confusion matrices for 100 case sequences.

(a) Naive labelling

		declared carrier	
		No	Yes
true carrier	No	0	49
	Yes	0	51

(b) GNN labelling

		declared carrier	
		No	Yes
true carrier	No	5	44
	Yes	0	51

4.3.2 Simulation study

In this section, we present the results for detecting the association signal, followed by the results for localizing the association signal to the causal region.

4.3.2.1 Detection

We present the detection results for simulation under the null hypothesis of no association, followed by the results for simulation under the alternative hypothesis of association. Note that, we use simulations under the null hypothesis to investigate whether the Mantel test is biased or not in our context.

Simulations under null hypothesis

Figure 4.5 compares the empirical cumulative distribution functions (ECDFs) of permutation p-values from a global test of association across the genomic region. For each method, we compare the ECDFs over the 500 datasets, simulated under the null hypothesis of no association. Panel a shows all the ECDFs. We plotted y-axis up to 0.30 for better resolution. Panel b denotes zoomed part of the plot around the 5% significance threshold. As can be seen, the Mantel test inflates from the targeted type-I error rate at nominal level 5% but not other methods. The permutation based p-values would be susceptible to the same bias.

In table 4.4, we show the proportion of 500 null datasets that are rejected at level 5% (type-I error rate) and associated approximate 95% confidence interval by each method. Figure 4.6 shows the graphical representation of the point and approximate 95% confidence interval estimators for type-I error rate by each method. As can be seen, the approximate 95% confidence interval for the type-I error rate of the Mantel test barely covers 0.05, in contrast to the other test statistics.

Given the previously mentioned concerns about the Mantel test, however, we do not pursue it further as a test for detecting association. We will return to this point in the discussion.

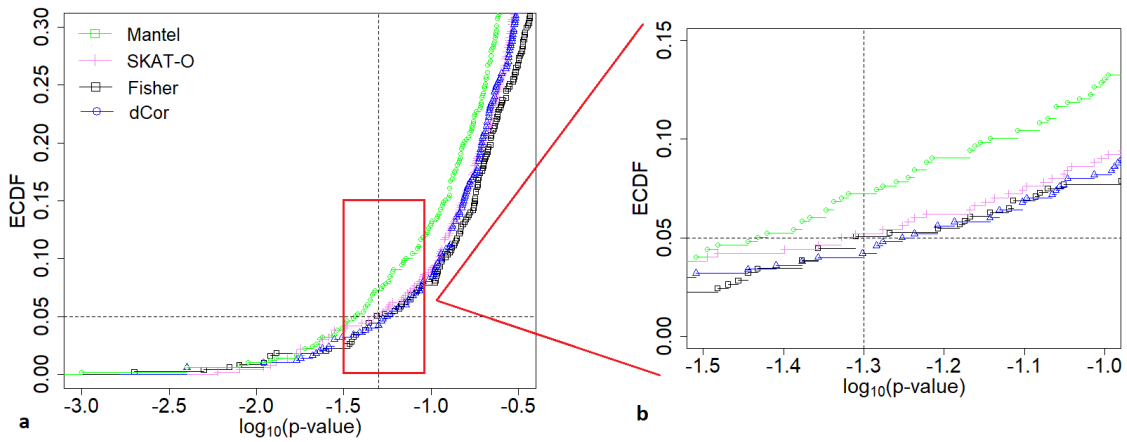

Figure 4.5: Both panels show the ECDFs of permutation p-values from a global test of association across the genomic region. Four methods are compared: Fisher's exact test, SKAT-O, distance correlation (dCor) and Mantel. a) Plot in full scale but shown up to 0.3 on y-axis for better resolution. b) Zoomed panel showing inflated estimate of type-I error rate from the Mantel test (in green). On the x-axis, p-values are converted to the log-10 scale for better resolution. Vertical line represents the 5% significance threshold (log_{10}(p-value) = -1.3).

Method	No of datasets	$\hat{p}$	Approx. LB (95% CI)	Approx. UB (95% CI)
Fisher	500	0.0506	0.0310	0.0702
SKAT-O	500	0.0520	0.0321	0.0719
dCor	500	0.0420	0.0245	0.0606
Mantel	500	0.0722	0.0491	0.0955

Table 4.4: The estimated type-I error rate or proportion of 500 null datasets that reject the null hypothesis ($\hat{p}$) and associated approximate 95% confidence interval.

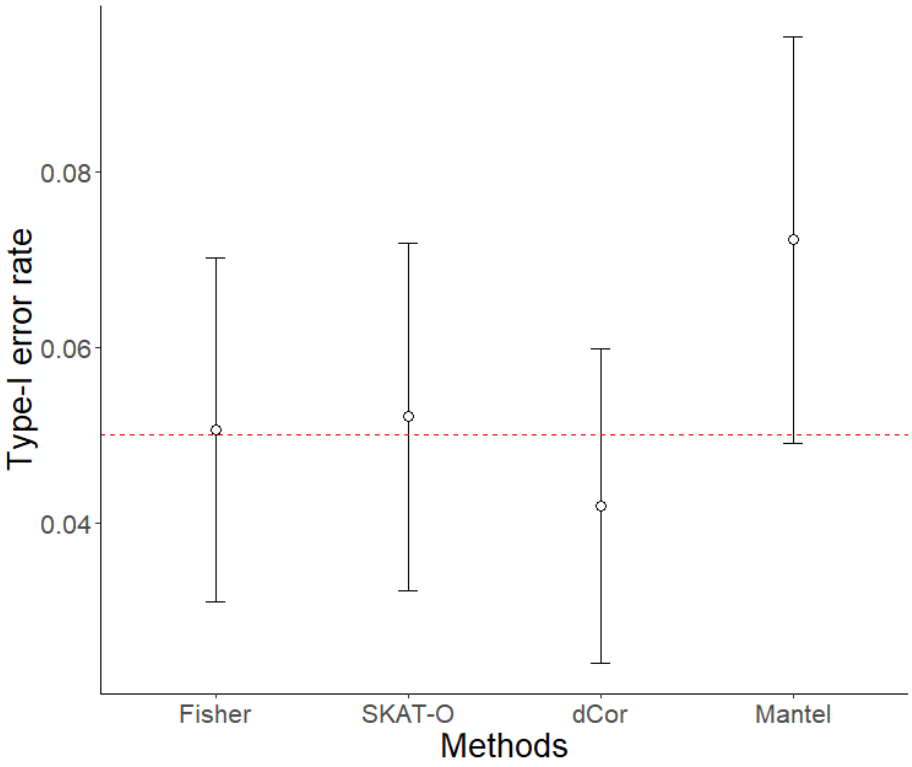

Figure 4.6: Point and approximate 95% confidence interval estimates for type I error rate. The horizontal dashed line represents the nominal 5% level.

Simulations under alternative hypothesis

Figure 4.7 compares the ability of the methods to detect any association with the disease across the entire genomic region. Over 500 datasets that are simulated under the alternative hypothesis, we compare the ECDFs of the permutation p-values computed from the corresponding statistics for all methods except the Mantel test. The resulting ECDFs are plotted against the $\log_{10}(p\text{-values})$. We convert the p-values to the log-base-10 scale for better resolution. As can be seen, SKAT-O performs considerably better than the distance correlation and Fisher's exact test; i.e., it has the highest proportion of simulated datasets with p-values below 5% ($\log_{10}(p\text{-values}) < -1.3$).

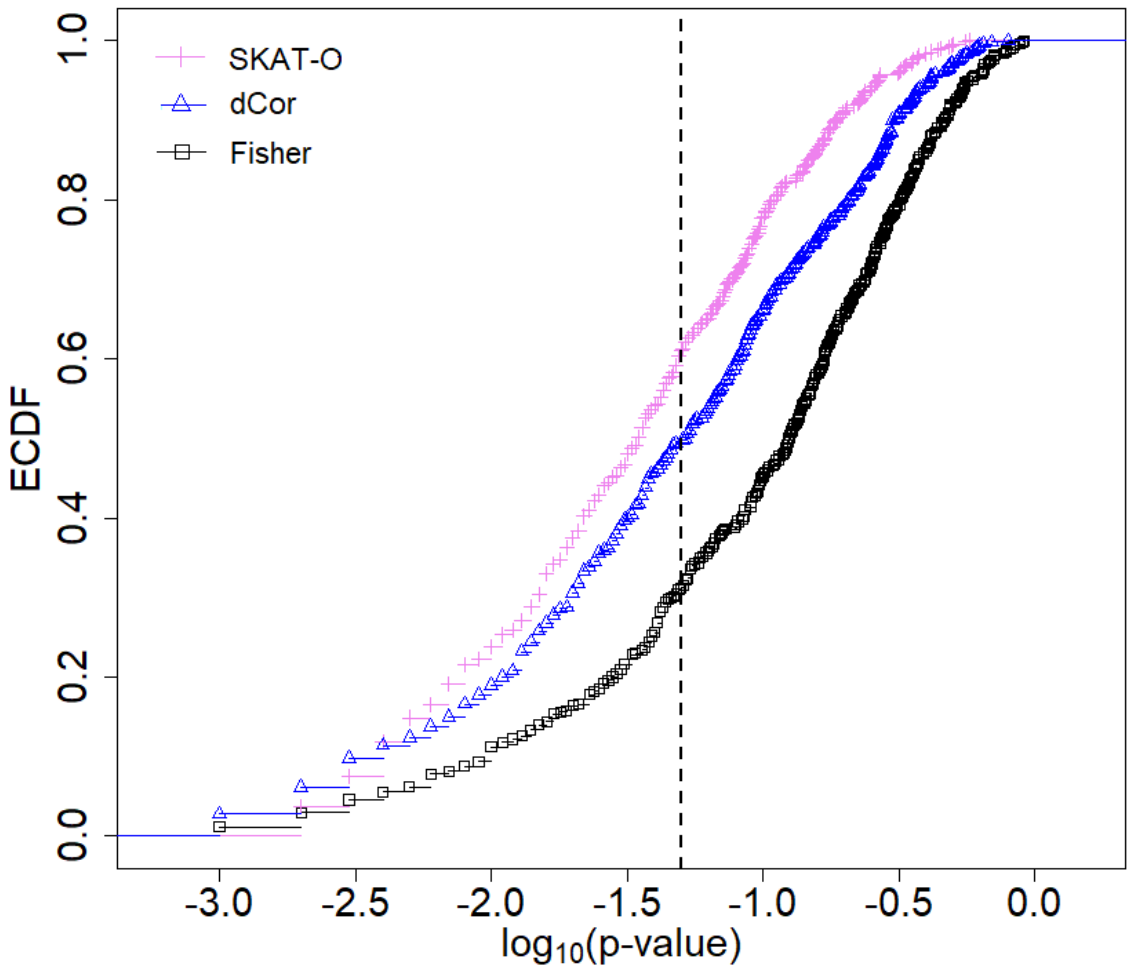

Figure 4.7: The ECDFs of permutation p-values from a global test of association across the genomic region. On the x-axis, p-values are converted to the log-10 scale for better resolution. Vertical line represents the 5% significance threshold ($\log_{10}(p\text{-value}) = -1.3$).

4.3.2.2 Localization

Figure 4.8 compares the ability of the methods to localize the association signal to the causal region. The figure shows the ECDFs of the distance of the peak association signal from the causal region, for all methods. The Mantel test outperforms all the other methods; it has the highest proportion of the 500 simulated datasets at the lower distance values. SKAT-O performs comparably better than distance correlation and Fisher's exact test. Among the 500 simulated datasets, the proportion of association signals that are localized to the causal region by the Mantel test ($\hat{p_1} = 0.352$) is significantly higher than the proportion of association signals that are localized by SKAT-O ($\hat{p_2} = 0.168$; McNemar's test, p-value ≈ 0). Distance correlation and Fisher's exact test perform similarly and only slightly better than random localization. For example, both methods locate the signal within the causal region

about 5% of the time, as would be expected if localization occurred completely randomly. Of all the methods, distance correlation and the Fisher's exact test have the most tendency to co-localize the signal (Figure A.2).

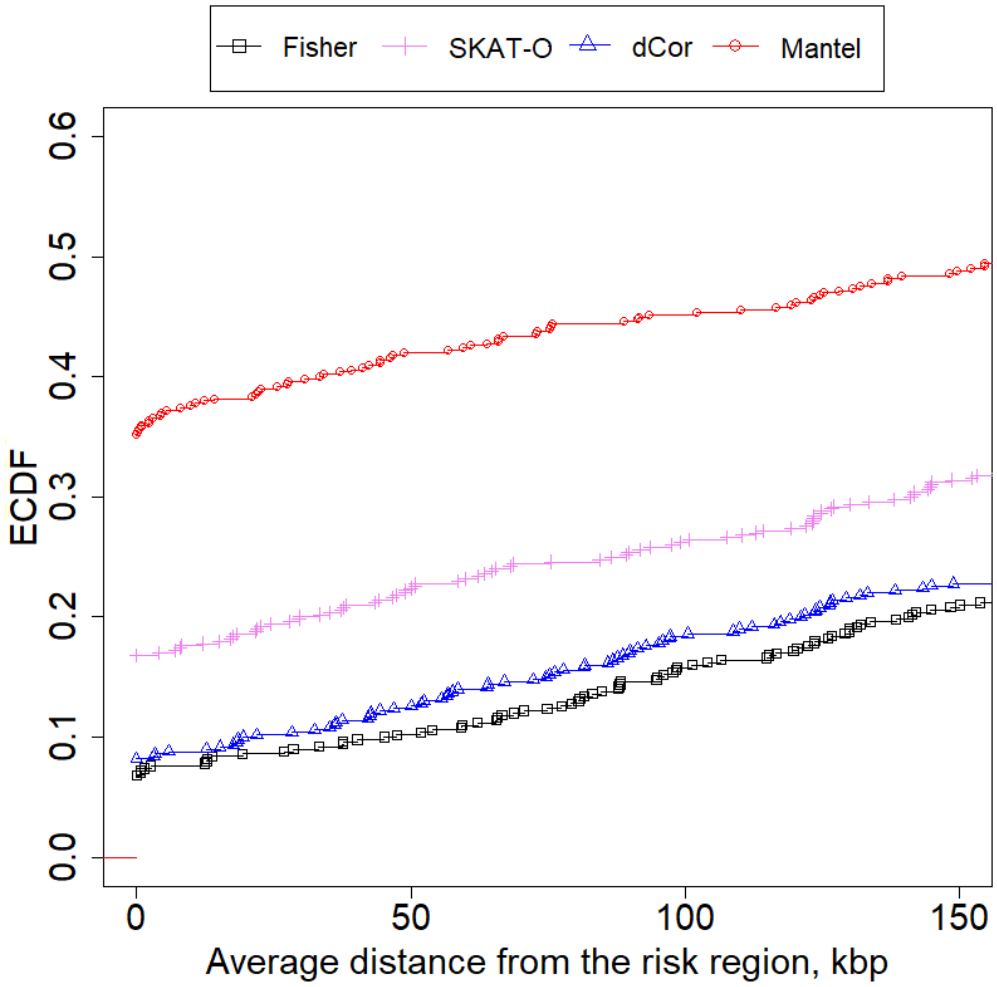

Figure 4.8: The ECDFs for the average distance of the peak association signal from the causal region, for 500 datasets simulated under the alternative hypothesis of association. Four methods are compared: Fisher's exact test, SKAT-O, distance correlation (dCor) and Mantel.

4.3.3 Post-hoc analysis

4.3.3.1 Diagnostic for SKAT-O detection

Figure 4.9 compares the Mantel localization profiles in two datasets, a positive (panel a) and negative control (panel b). Both datasets have been detected to be significantly associated by SKAT-O. However, the positive-control dataset is simulated under the alternative hypothesis of association while the negative-control dataset is simulated under the null hypothesis of no association. As can be seen in Figure 4.9, the peak signal for association in the positive-control dataset stands out clearly from the background variation, whereas the peak signal in the negative-control dataset is hard to distinguish from the background.

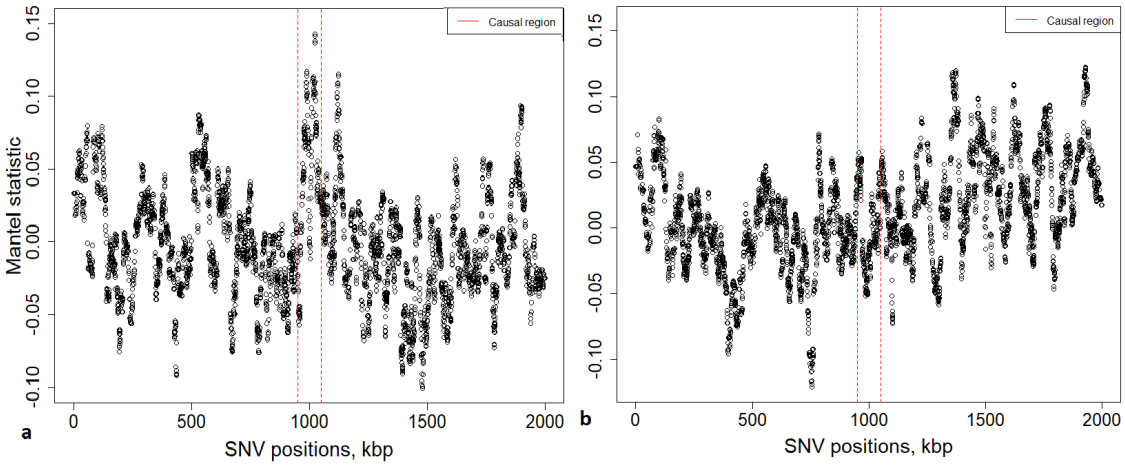

Figure 4.9: Comparison of Mantel localization profile in two datasets: a) Positive control where the dataset was simulated under the alternative hypothesis of association and rejected (i.e. detected to be significantly associated) by SKAT-O (p-value = 0.035) and b) Negative control where the dataset was simulated under the null hypothesis of no association and rejected by SKAT-O (p-value = 0.032).

Figure 4.10 compares the signal-to-noise ratio (as defined in the Method section) of the Mantel profiles in the two groups of datasets with sample sizes of 304 and 26 datasets for the alternative and null hypotheses, respectively. The mean signal-to-noise ratio of the Mantel profile for the datasets simulated under the alternative hypothesis is significantly greater than the mean for the datasets simulated under the null hypothesis (two-sample t-test with unequal variances, p-value $= 1.19 \times 10^{-5}$). Therefore, the Mantel profiles appear to have practical use for distinguishing between true and false positive association detected by SKAT-O.

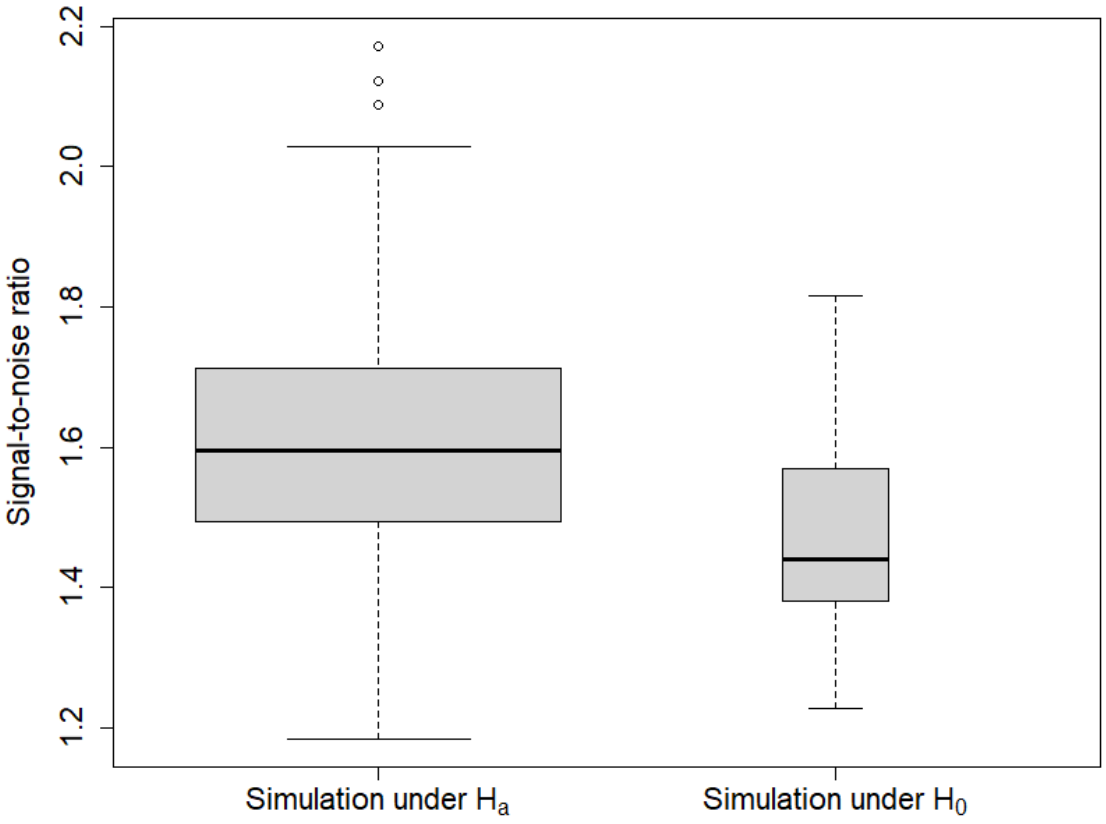

Figure 4.10: Comparison of signal-to-noise ratio statistic between positive- and negative-control datasets, where the datasets have been simulated under alternative and null hypotheses and rejected (i.e. detected to be significantly associated) by SKAT-O. Boxplots for the two groups with sample sizes of 304 and 26 datasets for the alternative and null hypotheses, respectively. Boxplots widths are adjusted to their sample sizes. The two distributions are significantly different (p-value $= 1.19 \times 10^{-5}$; two-sample t-test with unequal variances).

4.3.3.2 Overall performance of GNN labelling

We assess the performance of classifying case sequences as carriers or non-carriers of causal variants, using the 500 datasets that have been simulated under the alternative hypothesis. As described in the Methods, we compute the misclassification rate of case sequences under GNN and naive labeling. We then plot the misclassification rate of case sequences by GNN versus naive labelling, across 500 datasets, as shown in Figure 4.11. Each data point in the figure indicates the misclassification rate by GNN and naive labelling for a given dataset.

The naive misclassification rate has only three values (0.49, 0.50 and 0.51) which have been jittered on the x-axis to make viewing easier. Of 500 datasets, 496 have lower misclassification rates by GNN than naive labelling. Thus, most of the data points lie below the red-dashed line of $y = x$. Thus, GNN labelling of carriers improves naive labelling.

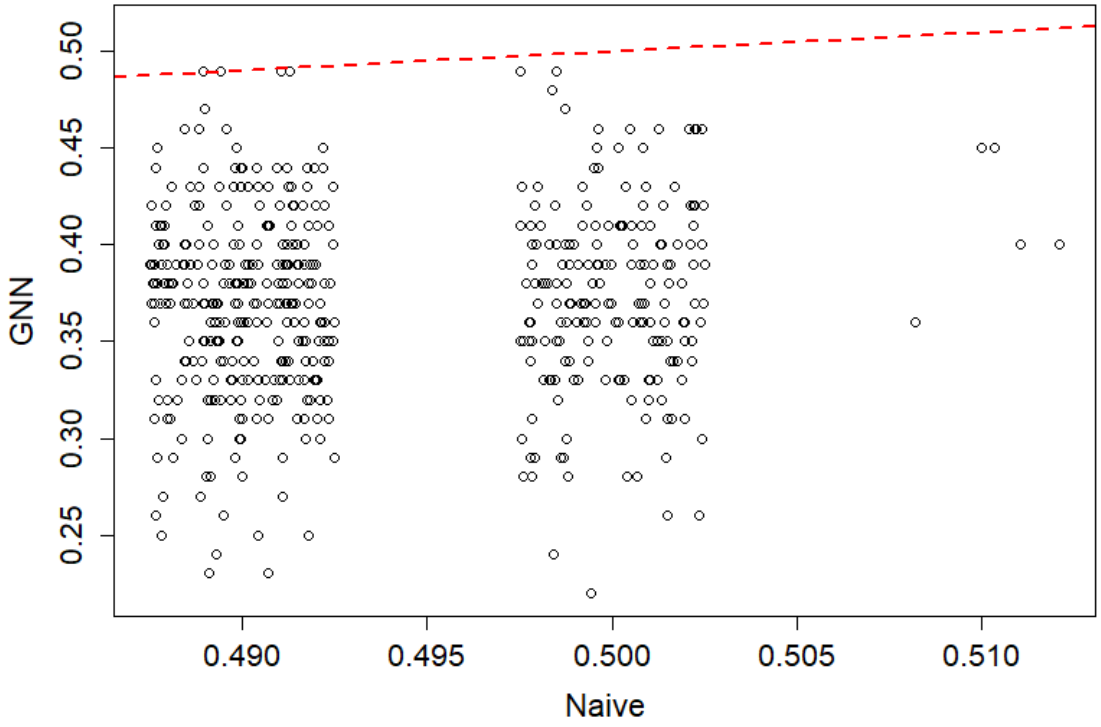

Figure 4.11: Misclassification rate of causal variant carrier status in case sequences, for GNN versus naive labelling in 500 simulated datasets. The red-dashed line is $y = x$.

4.4 Discussion

Sequence relatedness or IBD-based association methods have potential for fine-mapping rare variants when several occur in the same gene or locus. In this work, we evaluate the ability of non-IBD and IBD-based association methods to fine-map rare variants with high genetic penetrance ratio. As mentioned in the Introduction of this chapter, we view the sequence-relatedness method as a linkage method because it associates similarity in trait values to similarity in relatedness. However, our sequence-relatedness method is slightly different than a traditional linkage method for pedigrees because it considers case-control sequences in a population. Since linkage methods are powerful in fine-mapping rare causal variants in an allelically heterogenous disease with high genetic penetrance ratio [49], our focus was to set up the parameters in the model to showcase linkage analysis. Our logistic-regression model for disease penetrance has a low phenocopy rate of disease ($f = 4.5 \times 10^{-5}$) and high genetic penetrance ($g = 0.9975$). The disease-penetrance ratio, $g/f = 22167$, is therefore high. Moreover, the disease is allelically heterogeneous, with 20 causal variants in the disease locus defined by the middle genomic subregion of the candidate region. The high disease-penetrance ratio together with the allelic heterogenity make the disease particularly well-suited for linkage analysis or other methods based on IBD [49]. This genetic architecture is different than the one we used in the Chapter 2. In Chapter 2, our focus was to identify the potential of sequence-relatedness methods in fine-mapping without restricting the selection of causal variants by the minor allele frequency when the true genealogy is known. Therefore, our causal variants consist of collection of both rare and common variants. However, in this project, we only considered a genetic architecture for rare or low frequency causal variants.

As the first step, we work through an example dataset for insight. We then broaden our scope to many datesets using a simualtion study. In the simulation study, we compare the ability of these methods to *detect* and *localize* disease causal variants. Specifically, we are interested in the ability to detect *any* association in the candidate genomic region and to localize the association to a causal subregion within the candidate region. We also introduce a method to classify case sequences into carriers and non-carriers of causal variants using genealogical nearest neighbors that have been estimated from sequence data.

We evaluated the type-I error rate of all the association detection methods using the datasets simulated under the null hypothesis of no association. As described in Guillot and Rousset [50], the Mantel test has been shown to be biased in the presence of nonexchangable units. Under the null hypothesis, our sample sequences are not exchangeable owing to their underlying ancestry. However, the phenotype labels of individuals are exchangeable. Thus, the Mantel test should be unbiased in our context. However, in our simulation study, the approximate 95% confidence interval for the type-I error rate of the Mantel test barely covers 0.05. Given the previous concerns about the bias of the Mantel test, we did not

consider the Mantel test further for detecting association signal. In future work, we would like to investigate this further.

In detecting the association, our simulation study shows that the SKAT-O performs better than the distance correlation and Fisher's exact tests. Our findings for SKAT-O are consistent with previous studies which find it to be powerful in detecting rare causal variants. The distance correlation test shows the second best result in detecting the association signal. Fisher's exact test has the worst performance.

For localizing causal variants, our simulation results indicate that the Mantel test outperforms all the other methods. SKAT-O performs better than the distance correlation and Fisher's exact test. However, the distance correlation and Fisher's exact test perform only slightly better than the random localization. For SKAT-O, we used window size of 21 SNVs. We also tried localization with different window sizes of 11, 41, 63 and 101 SNVs. Though the larger window sizes localize better none outperform the Mantel test as shown in Figure A.1.

From the ECDFs for localization, the distance correlation and Fisher's exact test perform similarly. Of all the methods distance correlation and Fisher's exact test localize the most similarly across the 500 simulated datasets; see appendix A.3. Thus the distance correlation has a higher tendency than other methods to localize the peak signal in the same place as Fisher's exact test in individual datasets.

Turning to post-hoc analysis, we first consider the idea to discern true-positive from false-positive SKAT-O detection. For the Mantel-statistics profiles, the mean of the signal-to-noise ratios is significantly larger in positive-control than in negative-control datasets. This result suggests that the Mantel-statistic profile can be useful for distinguishing true- and false- positive patterns of association detected by SKAT-O. Next, we consider classifying case sequences into carriers and non-carriers of causal variants.

Based on our results in Chapter 2, we propose a data-based scheme to classify case sequences into carriers and non-carriers of causal variants. Our classification uses a GNN labelling method [6] based on the topological properties of the reconstructed trees. The resulting misclassification rates of case sequences indicate that the GNN labelling method classifies case sequences into their carrier status better than the original naive labelling of case sequences. Therefore, GNN labeling would appear to be useful for post-hoc analysis once an association is detected.

In our current study, we have used the default simulation model in *msprime* which is Hudson's coalescent with recombination (CwR). The CwR is well known to break down when the number of sampled sequences is large relative to the population effective size, N_e (e.g., [58]) In our simulations, N_e of 6200 individuals (12400 sequences) was not much larger than the sample size, n, of 3000 sequences (i.e., $\sqrt{12400} \approx 111$ is not greater than n). In future simulations, we plan to use a hybrid strategy implemented in *msprime* [59]. This hybrid strategy involves running a discrete-time Wright-Fisher model for the most recent

generations back and then, once the number of lineages decreases to a reasonably small number, switching over to Hudson's coalescent to complete the rest of the gene genealogy. With the hybrid simulation, we also plan to extend our investigation to quantitative traits.

Chapter 5

Summary and Future Work

In genetic association studies, the goal is to map genes that contribute to a disease trait in a candidate genomic region. Throughout this dissertation, we have explored the potential of tree-based or sequence-relatedness association to fine map such genes.

In Chapter 2, through coalescent simulation of a diploid population, we compare the ability of several popular association methods to localize causal variants in a subregion of a candidate genomic region. Under an additive genetic-risk model, we consider four broad classes of association methods which we describe as single-variant, pooled-variant, joint-modeling and tree-based. Our findings lend support to the potential of genealogical tree-based methods for genetic fine-mapping of disease.

In practice, true genealogical trees are unknown. However, reconstructing the genealogy of DNA sequences allows us to apply tree-based or sequence-relatedness methods for fine mapping. In Chapter 3, we pursue this reconstruction idea using the concept of perfect phylogeny. As described in Chapter 1, these perfect phylogeny partitions are not genealogical trees. But we can still use them to cluster the DNA sequences. Therefore, in Chapter 3, we present an R package, `perfectphyloR`, to reconstruct the perfect phylogeny for DNA sequences at a focal point in the genomic region. To our knowledge, `perfectphyloR` is currently the only R package that enables users to dynamically cluster a set of single-nucleotide variant sequences based on the underlying perfect phylogeny. Our implementation first partitions the DNA sequences using a classic partitioning algorithm of [35] and then uses heuristics introduced by [36] to refine them further. The resulting reconstructed partitions can provide important insight into the local ancestral structure of sequence data.

In Chapter 4, we investigate methods based on the concept of sequence relatedness to fine-map genetic variants with inherited traits. We compare the fine-mapping ability of methods based on sequence relatedness with methods that include single-variant testing and aggregation of multiple variants. Specifically, we focus on detecting and localizing rare causal variants for an allelically heterogenous disease with high disease-penetrance ratio, in case-control studies. Since the true gene genealogy is unknown, we used the methods developed in Chapter 3 to estimate sequence relatedness. We find the concept of sequence relatedness

to be useful for improving the localization of rare causal variants. We also pursue the idea of classifying the case sequences into carriers and non-carriers of causal variants, using the idea of genealogical nearest neighbor. Our proposed classification shows an improvement in classifying case sequences into carriers and non-carriers of causal variants.

In summary, we have explored the potential of tree-based methods or methods based on sequence relatedness to fine-map rare causal variants in case-control studies for dichotomous traits. These tree-based methods would be of practical use for fine-mapping of causal variants, for example, *after* a genome-wide "non-parametric" linkage analysis of sibling or other relative pairs (e.g., [60]). One could try all the methods we have considered in the candidate genomic region for detecting and localizing association signal. In future work, it would be interesting to extend our investigation to quantitative traits and streamline tree reconstruction. Our R package, `perfectphyloR`, has been developed for use with moving windows of SNVs along the genomic region of interest by considering one SNV at a time. Therefore, computation time tends to be high for a large dataset. Future work on speeding up the reconstruction would be beneficial for researchers seeking sequence relatedness underlying genetic data. Another possible direction would be to expand the input file format to include the PED/MAP file formats of PLINK. As mentioned in Chapter 4, we had a concern about bias in the Mantel test for detecting the association signal [50]. This behavior may be due to the simulation error but we would like to investigate about it further in future.

Bibliography

[1] A. L. Price, G. V. Kryukov, P. I. de Bakker, S. M. Purcell, J. Staples, L.-J. Wei, and S. R. Sunyaev, "Pooled association tests for rare variants in exon-resequencing studies," *The American Journal of Human Genetics*, vol. 86, no. 6, pp. 832–838, 2010.

[2] B. M. Neale, M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S. M. Purcell, K. Roeder, and M. J. Daly, "Testing for an unusual distribution of rare variants," *PLoS genetics*, vol. 7, no. 3, p. e1001322, 2011.

[3] R. R. Hudson *et al.*, "Gene genealogies and the coalescent process," *Oxford surveys in evolutionary biology*, vol. 7, no. 1, p. 44, 1990.

[4] K. M. Burkett, B. McNeney, J. Graham, and C. M. Greenwood, "Using gene genealogies to detect rare variants associated with complex traits," *Human heredity*, vol. 78, no. 3-4, pp. 117–130, 2014.

[5] D. Gusfield, "Efficient algorithms for inferring evolutionary trees," *Networks*, vol. 21, no. 1, pp. 19–28, 1991.

[6] J. Kelleher, Y. Wong, A. W. Wohns, C. Fadil, P. K. Albers, and G. McVean, "Inferring whole-genome histories in large population datasets," *Nature genetics*, vol. 51, no. 9, pp. 1330–1338, 2019.

[7] J. K. Pritchard, "Are rare variants responsible for susceptibility to complex diseases?," *The American Journal of Human Genetics*, vol. 69, no. 1, pp. 124–137, 2001.

[8] N. J. Schork, S. S. Murray, K. A. Frazer, and E. J. Topol, "Common vs. rare allele hypotheses for complex diseases," *Current opinion in genetics & development*, vol. 19, no. 3, pp. 212–219, 2009.

[9] E. E. Eichler, J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau, "Missing heritability and strategies for finding the underlying causes of complex disease," *Nature Reviews Genetics*, vol. 11, no. 6, p. 446, 2010.

[10] S. Lee, G. R. Abecasis, M. Boehnke, and X. Lin, "Rare-variant association analysis: study designs and statistical tests," *The American Journal of Human Genetics*, vol. 95, no. 1, pp. 5–23, 2014.

[11] S. Cho, K. Kim, Y. J. Kim, J.-K. Lee, Y. S. Cho, J.-Y. Lee, B.-G. Han, H. Kim, J. Ott, and T. Park, "Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis," *Annals of human genetics*, vol. 74, no. 5, pp. 416–428, 2010.

[12] C. Bardel, V. Danjean, J.-P. Hugot, P. Darlu, and E. Génin, "On the use of haplotype phylogeny to detect disease susceptibility loci," *BMC genetics*, vol. 6, no. 1, p. 24, 2005.

[13] T. Mailund, S. Besenbacher, and M. H. Schierup, "Whole genome association mapping by incompatibilities and local perfect phylogenies," *BMC bioinformatics*, vol. 7, no. 1, p. 454, 2006.

[14] L. Excoffier, I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, "Robust demographic inference from genomic and snp data," *PLoS genetics*, vol. 9, no. 10, p. e1003905, 2013.

[15] H. R. Johnston and D. J. Cutler, "Population demographic history can cause the appearance of recombination hotspots," *The American Journal of Human Genetics*, vol. 90, no. 5, pp. 774–783, 2012.

[16] J. Asimit and E. Zeggini, "Rare variant association analysis methods for complex traits," *Annual review of genetics*, vol. 44, pp. 293–308, 2010.

[17] C. Xu, M. Ladouceur, Z. Dastani, J. B. Richards, A. Ciampi, and C. M. Greenwood, "Multiple regression methods show great potential for rare variant association tests," *PloS one*, vol. 7, no. 8, p. e41694, 2012.

[18] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-variant association testing for sequencing data with the sequence kernel association test," *The American Journal of Human Genetics*, vol. 89, no. 1, pp. 82–93, 2011.

[19] W. Chen, B. R. Larrabee, I. G. Ovsyannikova, R. B. Kennedy, I. H. Haralambieva, G. A. Poland, and D. J. Schaid, "Fine mapping causal variants with an approximate bayesian method using marginal test statistics," *Genetics*, vol. 200, no. 3, pp. 719–736, 2015.

[20] H. Zou and T. Hastie, "Addendum: regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 5, pp. 768–768, 2005.

[21] B. Servin and M. Stephens, "Imputation-based analysis of association studies: candidate regions and quantitative traits," *PLoS genetics*, vol. 3, no. 7, p. e114, 2007.

[22] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 273–282, 2011.

[23] S. Le Cessie and J. C. Van Houwelingen, "Ridge estimators in logistic regression," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 41, no. 1, pp. 191–201, 1992.

[24] S. Cho, H. Kim, S. Oh, K. Kim, and T. Park, "Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis," in *BMC proceedings*, vol. 3, p. S25, BioMed Central, 2009.

[25] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[26] P. Scheet and M. Stephens, "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase," *The American Journal of Human Genetics*, vol. 78, no. 4, pp. 629–644, 2006.

[27] B. L. Browning and S. R. Browning, "A fast, powerful method for detecting identity by descent," *The American Journal of Human Genetics*, vol. 88, no. 2, pp. 173–182, 2011.

[28] B. N. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS genetics*, vol. 5, no. 6, p. e1000529, 2009.

[29] Y. Li, C. Willer, S. Sanna, and G. Abecasis, "Genotype imputation," *Annual review of genomics and human genetics*, vol. 10, pp. 387–406, 2009.

[30] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, "Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes," *Genetic epidemiology*, vol. 34, no. 8, pp. 816–834, 2010.

[31] J. Wessel and N. J. Schork, "Generalized genomic distance–based regression methodology for multilocus association analysis," *The American Journal of Human Genetics*, vol. 79, no. 5, pp. 792–806, 2006.

[32] J.-H. Park, S. Wacholder, M. H. Gail, U. Peters, K. B. Jacobs, S. J. Chanock, and N. Chatterjee, "Estimation of effect size distribution from genome-wide association studies and implications for future discoveries," *Nature genetics*, vol. 42, no. 7, p. 570, 2010.

[33] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[34] T. A. Thornton, "Statistical methods for genome-wide and sequencing association studies of complex traits in related samples," *Current protocols in human genetics*, vol. 84, no. 1, pp. 1–28, 2015.

[35] D. Gusfield, "Efficient algorithms for inferring evolutionary trees," *Networks*, vol. 21, no. 1, pp. 19–28, 1991.

[36] T. Mailund, S. Besenbacher, and M. H. Schierup, "Whole genome association mapping by incompatibilities and local perfect phylogenies," *BMC Bioinformatics*, vol. 7, no. 1, p. 454, 2006.

[37] M. Coulombe, K. Stevens, and D. Gusfield, "Construction, enumeration, and optimization of perfect phylogenies on multi-state data," in *2015 IEEE 5th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, pp. 1–6, IEEE, 2015.

[38] R. Agarwala and D. Fernández-Baca, "A polynomial-time algorithm for the perfect phylogeny problem when the number of character states is fixed," in *Proceedings of 1993 IEEE 34th Annual Foundations of Computer Science*, pp. 140–147, IEEE, 1993.

[39] S. Kannan and T. Warnow, "A fast algorithm for the computation and enumeration of perfect phylogenies," in *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, Citeseer, 1995.

[40] E. Paradis, J. Claude, and K. Strimmer, "APE: Analyses of Phylogenetics and Evolution in R language," *Bioinformatics*, vol. 20, no. 2, pp. 289–290, 2004.

[41] R. R. Hudson and N. L. Kaplan, "Statistical properties of the number of recombination events in the history of a sample of DNA sequences," *Genetics*, vol. 111, pp. 147–164, 1985.

[42] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[43] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[44] R. Heller, Y. Heller, and M. Gorfine, "A consistent multivariate test of association based on ranks of distances," *Biometrika*, vol. 100, no. 2, pp. 503–510, 2012.

[45] N. Mantel, "The detection of disease clustering and a generalized regression approach," *Cancer research*, vol. 27, no. 2 Part 1, pp. 209–220, 1967.

[46] Y. Escoufier, "Le traitement des variables vectorielles," *Biometrics*, pp. 751–760, 1973.

[47] C. B. Karunarathna and J. Graham, "Using gene genealogies to localize rare variants associated with complex traits in diploid populations," *Human Heredity*, vol. 83, no. 1, pp. 30–39, 2018.

[48] S. R. Browning and E. A. Thompson, "Detecting rare variant associations by identity-by-descent mapping in case-control studies," *Genetics*, vol. 190, no. 4, pp. 1521–1531, 2012.

[49] J. Ott, J. Wang, and S. M. Leal, "Genetic linkage analysis in the age of whole-genome sequencing," *Nature Reviews Genetics*, vol. 16, no. 5, pp. 275–284, 2015.

[50] G. Guillot and F. Rousset, "Dismantling the mantel tests," *Methods in Ecology and Evolution*, vol. 4, no. 4, pp. 336–344, 2013.

[51] J. Kelleher, A. M. Etheridge, and G. McVean, "Efficient coalescent simulation and genealogical analysis for large sample sizes," *PLoS computational biology*, vol. 12, no. 5, 2016.

[52] C. D. Campbell, J. X. Chong, M. Malig, A. Ko, B. L. Dumont, L. Han, L. Vives, B. J. O'Roak, P. H. Sudmant, J. Shendure, M. Abney, C. Ober, and E. E. Eichler, "Estimating the human mutation rate using autozygosity in a founder population," *Nature Genetics*, vol. 44, no. 11, pp. 1277–1281, 2012.

[53] S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, E. L. P. Team, D. C. Christiani, M. M. Wurfel, X. Lin, *et al.*, "Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies," *The American Journal of Human Genetics*, vol. 91, no. 2, pp. 224–237, 2012.

[54] B. E. Madsen and S. R. Browning, "A groupwise association test for rare mutations using a weighted sum statistic," *PLoS genetics*, vol. 5, no. 2, 2009.

[55] B. Li and S. M. Leal, "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data," *The American Journal of Human Genetics*, vol. 83, no. 3, pp. 311–321, 2008.

[56] L. Beckmann, D. Thomas, C. Fischer, and J. Chang-Claude, "Haplotype sharing analysis using mantel statistics," *Human heredity*, vol. 59, no. 2, pp. 67–78, 2005.

[57] J. Josse and S. Holmes, "Measuring multivariate association and beyond," *Statistics surveys*, vol. 10, p. 132, 2016.

[58] A. Bhaskar, A. G. Clark, and Y. S. Song, "Distortion of genealogical properties when the sample is very large," *Proceedings of the National Academy of Sciences*, vol. 111, no. 6, pp. 2385–2390, 2014.

[59] D. Nelson, J. Kelleher, A. P. Ragsdale, C. Moreau, G. McVean, and S. Gravel, "Accounting for long-range correlations in genome-wide simulations of large cohorts," *PLoS genetics*, vol. 16, no. 5, p. e1008619, 2020.

[60] C. M. Greenwood and S. B. Bull, "Analysis of affected sib pairs, with covariates–with and without constraints," *The American Journal of Human Genetics*, vol. 64, no. 3, pp. 871–885, 1999.

Appendix A

Supplementary materials for Chapter 3

A.1 Selecting causal variants

This appendix describes the selection procedure for causal variants in the additive logistic regression model of Section 4.2.1.2. We select variants to be causal assuming that a diseased individual in the population inherits only one copy of a causal variant from their parents. We make the simplifying assumption that each person carries at most one copy of a causal variant because our causal variants are rare and the chance of inheriting two copies is therefore negligible. Given the values of the logistic-regression parameters defined in Section 4.2.1.2, we first need to compute the number of individuals, N_1, that carry one copy of a causal variant, and the number of individuals, N_0, that carry no copies of a causal variant. To compute N_0 and N_1, we follow these steps:

- From the additive logistic-regression model in Section 4.2.1.2,

$$logit(P(D=1|G)) = \beta_0 + \beta_1 \sum_{j=1}^{m} G_j,$$

  we can obtain:

$$logit(P(D|G)) = \begin{cases} \beta_0; & \text{when } \sum_{j=1}^{m} G_j = 0 \\ \beta_0 + \beta_1; & \text{when } \sum_{j=1}^{m} G_j = 1, \text{ and} \end{cases}$$

$$P(D|G)) = \begin{cases} \frac{\exp(\beta_0)}{1+\exp(\beta_0)}; & \sum_{j=1}^{m} G_j = 0 \\ \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}; & \sum_{j=1}^{m} G_j = 1 \end{cases} \tag{A.1}$$

- We can also write the probability of disease as follows:

$$P(D) \approx P(D|\sum_{j=1}^{m} G = 0)P(\sum_{j=1}^{m} G = 0) + P(D|\sum_{j=1}^{m} G = 1)P(\sum_{j=1}^{m} G = 1) \qquad \text{(A.2)}$$

- Since the disease prevalence is about 0.05, by equation A.1 and A.2, we can obtain:

$$0.05 \approx \frac{1}{N}\left\{\frac{\exp(\beta_0)}{1+\exp(\beta_0)}N_0 + \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}N_1\right\},$$

  where N is the total number of individuals in the population.

- By setting $\beta_0 = -10$, $\beta_1 = 16$, and $N = 1500$ in the above, we obtain:

$$\begin{aligned} 0.05 &\approx \frac{1}{1500}\left\{N_0\frac{\exp(-10)}{1+\exp(-10)} + N_1\frac{\exp(-10+16)}{1+\exp(-10+16)}\right\} \\ &= \frac{1}{1500}\left\{N_0\frac{\exp(-10)}{1+\exp(-10)} + N_1\frac{\exp(6)}{1+\exp(6)}\right\} \end{aligned}$$

- Thus, re-arranging terms, we get $75 \approx (4.54 \times 10^{-5})N_0 + 0.99N_1$.

- Moreover, since $N_0 + N_1 \approx 1500$, we may solve for N_1 and N_0 to get $N_1 \approx 75$, and $N_0 \approx 1425$.

We select 20 approximately-equal-frequent variants from the population such that their total number of copies is around 75.

A.2 Localization by SKAT-O

This appendix shows Figure A.1 mentioned in Section 4.4. We consider different window sizes for SKAT-O to see if they can improve on the localization of the Mantel test. For localization, we consider the average distance of the peak association signal from the causal region. We plot ECDFs of these average distnaces for the Mantel test and for SKAT-O tests under different window sizes. SKAT-O window sizes are 11, 21, 41, 63 and 101 SNVs. As can be seen, the Mantel test outperforms the SKAT-O with different window sizes. For example, about 36% of the 500 datasets are localized signal exactly to the causal region with the Mantel test. By contrast, with the SKAT-O test having window size 11 SNVs, about 18% of the 500 datasets localize signal exactly to the causal region.

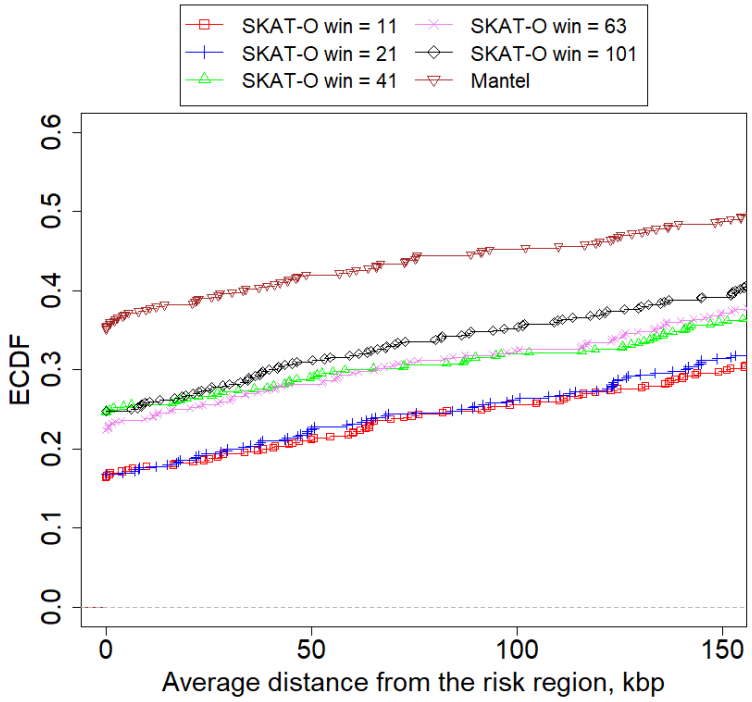

Figure A.1: The ECDFs for the average distance of the peak association signal from the causal region, for 500 datasets simulated under the alternative hypothesis of association. The results show SKAT-O with different window sizes of 11, 21, 41, 63 and 101 SNVs, as well as the Mantel test.

A.3 Correlation of localization distances

This appendix shows the Figure A.2 mentioned in Section 4.4. For localization, we consider the average distance of the peak association signal from the causal region. The figure shows the correlation of the localization distances between all possible pairs of the methods. As can be seen, the distance correlation and Fisher's exact test are the most correlated.

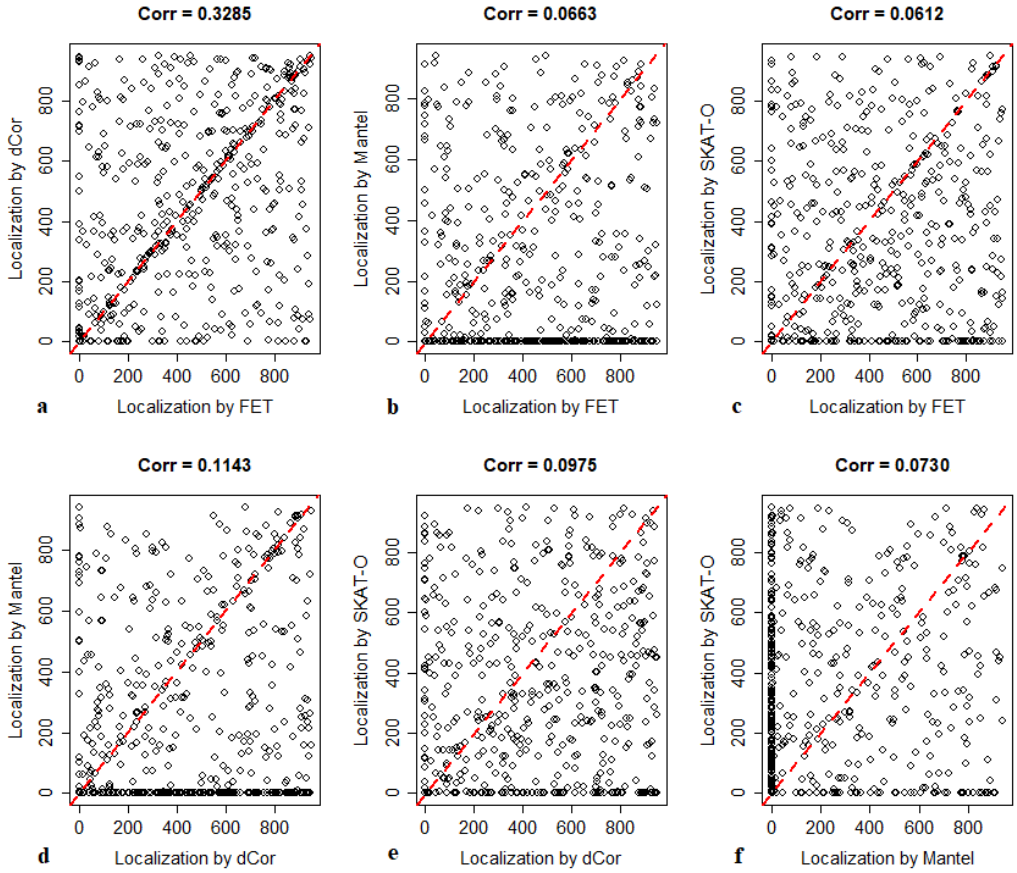

Figure A.2: Correlation of the average distances from the causal region between all possible pairs of the methods: a) Distance correlation (dCor) and Fisher's exact test (FET), b) Mantel and FET, c) SKAT-O and FET, d) Mantel and dCor, e) SKAT-O and dCor, f) SKAT-O and Mantel. The red-dashed line is $y = x$.