Methods for statistical and population genetics analyses

by

Shyam S. Gopalakrishnan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2011

Doctoral Committee:

Associate Professor Sebastian K. Zoellner, Chair
Professor Michael L. Boehnke
Associate Professor Zhaohui Qin
Associate Professor Noah A. Rosenberg
Assistant Professor Jun Li

To Mom and Dad

ACKNOWLEDGEMENTS

I want to express my gratitude to all the people who made this dissertation possible. I would like to thank my advisors, Drs. Steve Qin and Sebastian Zöllner for their support and advice throughout the course of my graduate studies. Steve made my transition from Engineering to Biostatistics a smooth one, teaching me the nuances of the field along the way. I have Sebastian's class to thank for my strong interest in population genetics. His mentorship during my dissertation work has been instrumental in the formation of my academic temperament. I would also like to thank my dissertation committee for providing me with feedback and suggestions. In particular, I would like to thank Dr. Michael Boehnke, for taking the time to advice me on various subjects.

Just as "it takes a village to raise a child", it takes an entire department to complete a dissertation. I have made several good friends during my stay in Ann Arbor. Matthew Zawistowski and Tanya Teslovich have been trusted friends without whom this journey would not have been half as interesting. I can only hope to find colleagues such as Rui, Jun, Weihua, Abby, Mark, Rob, Matthew, Anne, Wei and Heather. This thesis would not have come to fruition without the help of Sue Burke and Dawn Keene, who were there to help me with matter small and big.

I would like to thank my family for their steadfast support even when it looked like I would never finish. I draw my strength from my mother and father, the best

teachers that I could have asked for. My sister, Shubha, has been a confidante my entire life. Finally, I would like to thank my rock, Roshni, for all her support during all the ups and downs of my life.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES vii

LIST OF TABLES ix

CHAPTER

I. Introduction 1

1.1 Scope of this dissertation 2

II. An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria 5

2.1 Introduction 5
2.2 Methods 7
2.3 Results 14
2.4 Discussion 18

III. Framework for remapping multiply mapped short reads . . . 22

3.1 Introduction 22
3.2 Methods 24
3.3 Results 29
3.4 Discussion 31

IV. Feasibility of admixture mapping to identify rare susceptibility variants 34

4.1 Introduction 34
4.2 Methods 39

4.3 Results 52
4.4 Discussion 54

V. Estimating site frequency spectra from low coverage sequencing data 62

5.1 Introduction 62
5.2 Methods 65
5.3 Results 72
5.4 Discussion 74

VI. Conclusion 80

BIBLIOGRAPHY 84

LIST OF FIGURES

Figure

2.1 Example configuration where greedy approach does not pick the least number of tagSNPs 9

3.1 Proportion of multiply mapped reads aligned to their true location using our algorithm 31

4.1 Admixed chromosomes and admixture blocks 36

4.2 Population model used for simulating haplotypes using *ms* 49

4.3 Relative risk vs minor allele frequency. Power of single marker test is set at 10% for 20000 cases and controls, correcting for 1 million tests. 51

4.4 Contour plot showing the power of admixture mapping against the contribution to prevalences in the two founding populations. The power of admixture mapping is shown for two different mixing ratios of 80%-20% and 50%-50%. 57

4.5 The relationship between power of admixture mapping and contribution to prevalence in population A, plotted for various sample sizes. The contribution to prevalence in the second population was fixed at 1%. 58

4.6 Mean ratio of contribution to prevalences in the African population vs the European population plotted against the cumulative risk allele frequency. The contribution to prevalence in Europeans was fixed to 1%. 59

4.7 Power of admixture mapping compared to the power of single marker tests (blue lines) in Europeans. Two levels of European ancestry were considered, 20% (red lines) and 50% (black lines). The hollow symbols represent tests with 1000 cases and 1000 controls, whereas the filled symbols represent test with 10000 cases and controls each. 60

4.8 Power of admixture mapping compared to indirect association. Sample size was 5000 cases and controls each. TagSNPs were chosen with two minor allele frequency cutoffs, 1% and 0.5%. 61

5.1 Estimated SFS using individual based genotype calls using 200 samples sequenced at 30-fold average depth 73

5.2 Estimated SFS for low pass, 4-fold, short read sequencing data. The panels show the first 10 bins of the estimated SFS using genotypes from (a) an individual level caller, (b) population level caller and (c) population level LD aware caller. 76

5.3 MLE estimate of the SFS using simulated data 77

5.4 Mean and standard error of the ratio of true to estimated SFS for each bin 78

5.5 QPOC data: Comparison of MLE and counting estimate of SFS . . 79

LIST OF TABLES

Table

2.1 Summary of Chromosome 2: Number of tagSNPs with greedy approach and FESTA 18

2.2 TagSNP distribution on Chromosome 2: Number of tagSNPs per precinct. 19

2.3 Summary of tagSNP results for Encode regions: CEU samples . . . 19

2.4 Comparing different criteria for tagSNP selection: Effect on number of tagSNPs in denser SNP maps 20

3.1 Population parameters for simulating short read sequence data . . . 29

3.2 Read characteristics for chromosome 21 simulated data 30

3.3 Effect of multiply mapped reads on variant discovery 32

5.1 Expected SFS bin counts under neutral and our parameterization . 67

CHAPTER I

Introduction

One of the major goals in the field of genetics is to identify disease predisposing genetic variants, to better understand the underlying mechanism and ultimately identify targets for intervention. Aided by advances in technology, the tools used to accomplish these goals have rapidly evolved.

Risch and Merikangas [1] first advocated the use of large scale association studies to detect disease predisposing variants. Large scale association studies were not feasible at the time. Since then, the advances in high throughput low cost genotyping have propelled association studies, especially genome-wide association studies (GWAS), to become the instrument of choice in disease genetics.

GWAS have been highly successful in identifying variants associated with a wide array of common diseases and quantitative traits, e.g. type 2 diabetes[2], Parkinson's disease, LDL cholesterol etc[3]. GWAS are best suited to identify variants that fall under the Common Disease Common Variant (CDCV) hypothesis. The CDCV hypothesis states that the prevalence of common diseases can be attributed to a few common variants with moderate effect sizes. Though GWAS have identified more than 4900 associated loci for more than 200 traits, these common variants explain

only a small fraction of the heritability of the common diseases[4, 5].

The presence of rare causal variants has been suggested as a possible source for the missing heritability. GWAS are not well powered to detect rare variants. The Common Disease Rare Variant (CDRV) hypothesis suggests that common diseases can be explained by multiple rare variants with large effect sizes[6]. Several methods have been proposed to identify rare causal variants[7, 8]. Since testing individual rare variants is not statistically powerful, most methods collapse information across multiple rare variants. Development of new methods to identify rare susceptibility variants remains an area of much interest.

The search for rare susceptibility variants was further expedited by the emergence of short read sequencing[9]. Short read sequencing technologies allowed low cost large scale sequencing. Sequencing studies have been used to catalog variants in the human genome [10], identify rare causal variants for traits such as LDL and HDL cholesterol[11, 12], quantify gene expression [13], identify DNA protein interaction[14] and perform population genetics studies [15]. Analysis of short read sequence data present many interesting statistical challenges.

1.1 Scope of this dissertation

In this dissertation, I present novel methods aimed at tackling some of the statistical challenges in the field of genetics.

TagSNP Selection

Association studies rely on indirect association to maintain power. They test a representative set of markers, tagSNPs, in lieu of all the variants in the genome. The

power of association tests using tagSNPs is directly proportional to the linkage disequilibrium measure r^2 between the tagSNP and the causal variant. In chapter 2, we propose a graph-based method to select the optimal set of tagSNPs. We define optimality in terms of the number of tagSNP markers. We use a "divide and conquer" approach to identify the smallest set of tagSNPs that is highly correlated with all the variants in the region. As an example, we apply our method to chromosome 2 HapMap [16] data and ENCODE regions.

Remapping multiply mapped reads

The alignment of short reads is dependent on many factors like read length and error model. Further, the presence of repetitive elements and structural variants on the reference sequence results in a fraction of short reads being aligned to multiple locations in reference. These multiply mapped reads are often discarded.

In chapter 3, we present a Gibbs sampling approach to identify the most likely genomic location for multiply mapped reads. Additional information from the multiply mapped reads can improve the performance of downstream analyses. We illustrate the effect of including multiply mapped reads using variant discovery in a simulated sample.

Admixture mapping to identify rare variants

An admixed population derives its ancestry from multiple founding populations. Admixture mapping is a tool that identifies regions associated with the disease by testing the correlation of the ancestry of the region with the disease status.

Methods designed to detect rare causal variants combine information across multiple markers. Several strategies exist to collapse across markers, viz., presence of minor

allele, sum of minor alleles, sum of weighted allele frequencies etc. Ancestry across a testing unit can be construed as a way to summarize the information contained in the markers in the region. In chapter 4, we explore the power of admixture mapping to detect regions harboring multiple rare causal variants. We propose a disease model for the rare causal variants. In settings unsuitable for single marker association tests, we test the feasibility of admixture mapping to detect the rare susceptibility loci.

Site frequency spectrum estimation

The site frequency spectrum (SFS) is a population genetics statistic that contains information on the number of variant positions at each minor allele frequency in the sample. The SFS is an important summary statistic in population genetics, encompassing information on selection and demographic history. All population genetics statistics that do not include linkage disequilibrium information can be expressed as functions of the (SFS). Estimates of the SFS obtained from genotyping platforms suffer from ascertainment bias, since there exist potentially variable positions that are not included on the genotyping array. Since all positions are queried in a sequencing study, SFS estimated from sequencing studies do not suffer from ascertainment bias.

In chapter 5, we present a maximum likelihood estimation procedure to the estimate the SFS from low coverage short read sequence data. First, we show that estimates of the SFS obtained from genotype calling methods underestimate the number of rare variants, especially singletons and doubletons. We demonstrate that our method performs better than SFS obtained from genotype calling algorithms using both simulated and real data examples.

CHAPTER II

An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria

2.1 Introduction

Genome-wide association studies have emerged as the predominant approach to detect genetic variants that contribute to human diseases. Initially, genome-wide association studies focused on single nucleotide polymorphisms (SNPs) because of their high abundance in the human genome, their low mutation rates and their accessibility to high-throughput genotyping [17]. There are more than 10 million verified SNPs in dbSNP (build 124)[18], but typing all available SNP markers is inefficient and unnecessary since many will provide redundant information due to linkage disequilibrium (LD). A better strategy is to select a subset of representative SNPs (tagging SNPs or tagSNPs) and to remove the rest from consideration [19, 20]. The objective is to have little information overlap among the selected SNPs while retaining much of the signal contained in the original set.

The selection of tagSNPs is a well researched topic and many strategies have been proposed [21, 22, 23, 19, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34]. Zhang and Jin[35]

and Carlson et al. [36] introduced methods based on the LD measure r^2. These methods search for a small set of SNPs that are in strong LD (measured through pairwise r^2) with other SNPs that are not selected for genotyping. Pairwise r^2 is an attractive criterion for tagSNP selection since it is closely related to statistical power for case-control association studies, where a directly associated SNP is replaced with an indirectly associated tagSNP [6].

In this manuscript, we describe efficient algorithms for tagSNP selection based on pairwise LD measure r^2. The algorithms were implemented in a computer program named FESTA (fragmented exhaustive search for tagging SNPs). Essentially, we replace a greedy search, where markers are added sequentially to the tagSNP set, with an exhaustive search where all marker combinations are evaluated. To achieve this, we arrange the genome into precincts of markers in high LD, such that markers in different precincts show only low pairwise disequilibrium. TagSNP selection can then be performed within each precinct independently, greatly reducing computation cost. In most settings, our method is guaranteed to find the optimal tagSNP set(s) defined by the r^2 criterion. For a small proportion of precincts where exhaustive search is computationally too expensive to carry out, an efficient greedy-exhaustive hybrid search algorithm is described. Using data from the HapMap project [16], we show that the majority of these precincts contain relatively small numbers of SNPs, especially when a stringent r^2 criterion is used. Our algorithm readily identifies equivalent tagSNP sets, so that additional selection criteria can be incorporated. Other useful extensions are also discussed in this manuscript, such as the inclusion/exclusion of certain SNPs and double coverage, which can increase robustness of tagSNP sets against sporadic genotyping failures or errors.

2.2 Methods

Consider a set $\mathbb{S}$ which contains M bi-allelic SNP markers $a_1, a_2, \ldots, a_M$. Further assume that all these markers have minor allele frequency (MAF) above a certain threshold (0.05 was used in this study). First, two-SNP haplotype frequencies were estimated [37], and then the pairwise LD measure r^2 (also referred to as D^2) [38] was calculated for each pair of markers using the inferred haplotype frequencies [39]. Two markers a_i and a_j are said to be in strong LD if the r^2 between them is greater than a pre-specified threshold value r_0, denoted as $r^2(a_i, a_j) \geq r_0$ ($r_0 = 0.5$ or 0.8 in in this study). Both are considered tagSNPs for each other; i.e. a_i can be used as a surrogate for a_j, and vice versa.

Our aim is to a find tagSNP set, denoted by T, a subset of $\mathbb{S}$ such that $\forall a_i \in \mathbb{S}\backslash T$, $\exists a_j \in T$ that satisfies $r^2(a_i, a_j) \geq r_0$. In our presentation, we introduce two intermediate SNP sets, P and Q. P is called the candidate set which contains all the markers that are eligible to be chosen as tagSNPs and Q is named the target set which contains all the markers that are yet to be tagged, i.e. no marker in Q is in LD with any tagSNP in T. For each marker a_m in P, let $C(a_m) := \{a : a \in Q \& r^2(a, a_m) \geq r_0\}$ represent the subset of Q which contains markers that are in strong LD with a_m, and let $|C(a_m)|$ be the number of the elements in the set $C(a_m)$. Typically, the candidate set P is the complement of the tagSNP set T, $P = \mathbb{S}\backslash T$ and $P = Q$. One exception occurs when some SNPs are excluded as tagSNPs because they cannot be easily genotyped, but they still should be tagged by other markers if possible. In this case, the candidate set is a subset of target set. We describe several different algorithms for updating P, Q and T starting with a greedy approach [36]. We then outline successive refinements and extensions of a partition and exhaustive search algorithm, designed to handle various scenarios encountered when planning association studies.

2.2.1 Greedy Approach

The detailed algorithm is as follows [36].

Algorithm 1 (greedy approach):

1. Set $T = \emptyset$ and $P = Q = \mathbb{S}$.
2. For each marker $a_m \in P$, calculate $|C(a_m)|$.
3. For every marker a_m where $|C(a_m)| = 0$, add a_m into T, and remove it from Q.
4. Find the marker in P that has the highest $|C(a_m)|$ value, denoted as a_{max}, and add a_{max} into T, removing it and all connected SNPs, i.e. $C(a_m)$ from Q. (5) Repeat Steps 2-4 until $Q = \emptyset$.

In Step 4, by removing associated markers from consideration, the coverage overlap among tagSNPs is greatly reduced. Although it is simple to implement, the greedy procedure may miss more efficient solutions. Figure 2.1 gives a simple example, where markers A and B each tag half of all markers and together can tag all the markers. However, marker C is connected to more than half of all markers, and it is the first marker selected by the greedy algorithm. In this example, the greedy algorithm produced a set with three tagSNPs, despite the fact that the optimal solution contains only A and B.

2.2.2 Exhaustive search

An exhaustive search guarantees the minimum tagSNP set. Therefore, theoretically, the exhaustive search solves the tagSNP selection problem. But in practice, genome-wide tagSNP selection requires consideration of hundreds of thousands of SNP markers. For problem of this scale, exhaustive searches cannot be directly applied due to

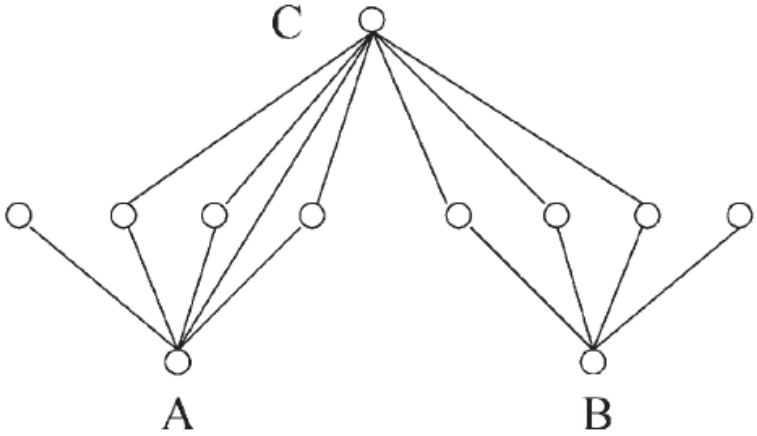


Figure 2.1: Example configuration where greedy approach does not pick the least number of tagSNPs

prohibitive computation costs.

Since appreciable LD only occurs within clusters of nearby markers along chromosomes, a practical solution is to first decompose the set of markers into disjoint precincts, such that markers in different precincts are never in strong LD. Then, selecting tagSNPs using the r^2 criterion in the whole set is equivalent to selecting tagSNPs in each precinct and then combining all the tagSNPs together. Here the concept of precinct is defined based on pairwise LD measure. It is therefore closely related to haplotype blocks [40, 21, 41, 42, 23, 43], which are regions where historical recombination events are rare. The main difference is that the precincts of markers in high r^2 are determined purely based on satistical correlation. Unlike haplotype block, markers within each precinct may not be consecutive markers sitting next to each other.

Partitioning the markers into precincts can be achieved using standard algorithms

in graph theory. We applied the Breadth First Search (BFS) algorithm [44]. Starting from any node (a marker) in a new precinct, this algorithm adds all neighboring nodes (markers in LD) and all neighbors of the newly added nodes to the precinct, until there are no neighbors to be added to the precinct. This process is restarted from different nodes until all the nodes are assigned to a precinct.

After the partitioning step, we perform the tagSNP selection within each precinct. Starting with $K = 1$, all K-marker combinations are searched to see if they cover all markers within this precinct. If not, K is increased by one and the search is repeated until a tagSNP set is found or a pre-specified search limit is reached.

When evaluating all K-marker combinations, the computation cost required for an exhaustive search might be too great in some precincts. In such cases, we propose a hybrid solution which reduces the computation cost and retains a good chance of finding optimal tagSNP sets. For each precinct i with N_i markers (Here on, all parameters with subscript i indicate parameters within the i-th precincts, such as K_i, J_i, P_i, Q_i, T_i and N_i.), we decide whether an exhaustive search is feasible by comparing the computation cost required for evaluating all K-marker combinations within a precinct, $\binom{N_i}{K}$, with a computation cost limit L specified a priori, determined based on available computing resources. Larger limits allow a more comprehensive search, which may result in fewer tagSNPs being selected, but require additional computational effort. In this study, we set this limit at 1 million. When this limit is exceeded, we apply the following hybrid algorithm. Specify K_i^* such that it is the largest K possible that satisfies $\binom{N_i}{K} \leq L_0$, where L_0 is a pre-specified computation cost limit (less than L, set at 10000 in studies conducted here). Subsequently, for each K_i^*-marker combinations, denoted as $\{a_1 \dots a_{K_i^*}\}$, assume that these markers have already been selected, remove a_m together with all the markers in $C(a_m)$ from candidate set P_i and

target set Q_i, $m = 1 \ldots K_i^*$, i.e. $P_i = Q_i = \mathbb{S}_i \bigcup_{m=1}^{K_i^*}(a_m \cup C(a_m))$, then apply the greedy approach to identify a subset of P_i that is able to cover Q_i, which contains the remaining untagged markers.

The tagSNP set obtained in the reduced set plus the previous K_i^* markers together form a complete tagSNP set for the i-th precinct. The detailed algorithm is as follows:

Algorithm 2 (FESTA: greedy-exhaustive hybrid search):

1. Apply the Breadth First Search to decompose the entire set of markers into precincts such that high LD can only be observed within precincts. $\mathbb{S} = \bigcup_n \mathbb{S}_i, and S_i \cap S_j = \emptyset for all i \neq j$

2. Within each precinct $\mathbb{S}_i$, set $K = 1$,

   a If $\binom{N_i}{K} \geq L$, move to (b), otherwise conduct an exhaustive search over all possible K-marker combinations. Both the candidate set P_i and the target set Q_i are $\mathbb{S}_i$. If no combination of K SNPs can cover the entire precinct, set $K = K + 1$, and repeat this step.

   b Find K_i^* such that $\binom{N_i}{K_i^*} \leq L_0$ and $\binom{N_i}{K_i^*+1} > L_0$. For every K_i^* marker combination in $\mathbb{S}_i$, denoted as $\{a_1 \ldots a_{K_i^*}\}$, let $T_i = \cup_m \{a_m\}$, $P_i = Q_i = \mathbb{S}_i \backslash \bigcup_{m=1}^{K_i^*}(\{a_m\} \cup C(a_m)$, and apply the greedy approach to identify a subset of P_i that is able to cover the remaining untagged markers Q_i. Among all the resulting tagSNP sets, we choose the smallest set.

3. Record all minimum tagSNP sets that cover the precinct. These form the complete minimum tagSNP sets $\{T_i^j : j = 1 \ldots J_i\}$, where J_i is the total number of such minimum tagSNP sets.

4. Any combination of tagSNP sets identified from all disjoint precincts forms a

tagSNP set for the whole set $\mathbb{S}$. Suppose the size of the minimum tagSNP set(s) in each precinct is K_i, then the overall size of such minimum tagSNP sets is $\sum_{i=1}^{n} K_i$, and the total number of such minimum tagSNP sets is $\prod_{i=1}^{n} J_i$.

FESTA executes either a pure exhaustive search or a greedy-exhaustive hybrid search in each precinct depending on the computational cost. Exhaustive search is first attempted, and if the computation cost becomes too high, the hybrid algorithm is used as a fall back. Typically, only a small proportion of the precincts require the greedy-exhaustive hybrid search.

2.2.3 Additional features

Mandatory tagSNP markers

Our algorithm readily allows users to force certain mandatory SNP markers to be included or excluded in the tagSNP set. There are several scenarios where such functionality is important. First, in candidate gene studies, previous knowledge may be available as to which SNPs are functionally important. These might include non-synonymous coding region SNPs (cSNPs) as well as SNPs located in regulatory regions. Second, in genome wide studies, one might carry out multiple rounds of genotyping and tagSNP selection. In such cases, additional tagSNPs could be selected at each round to cover the markers not tagged by tagSNPs successfully genotyped in the previous round. We provide an example of this in the results section. In other settings, it may be useful to exclude certain SNPs from consideration as tags. For example, some SNP markers may be difficult to genotype using a particular platform.

When there are mandatory markers $\{t_1, \ldots, t_r\}$ to be included, add these markers into the tagSNP set T, and remove them from the candidate set, i.e. $P = P \backslash \bigcup_{i=1}^{r} \{t_i\}$. The target set $Q = Q \backslash \bigcup_{i=1}^{r} (\{t_i\} \cup C(t_i))$. If there are SNPs $\{u_1, \ldots, u_s\}$ that need

to be excluded from the tagSNP set, we simply remove them from the candidate set, the target set Q is unchanged.

Choosing between alternative solutions

Within a densely typed SNP set, redundant tagSNPs are common, which results in multiple tagSNP sets of the same size. All of these sets are equal in the sense of minimizing the number of tagSNPs. In order to choose one set for genotyping, additional criteria can be employed. Here we evaluate several alternative criteria:

1. Maximize average r^2 between tagSNPs and untagged SNPs they represent
2. Maximize the lowest r^2 between tagSNPs and the untagged SNPs they cover
3. Minimize the average r^2 among all pairs of tagSNPs within a precinct
4. Maximize the average r^2 among all pairs of tagSNPs within a precinct
5. Maximize the average minor allele frequencies among all tagSNPs

In criteria 1 and 2, we try to identify the tagSNP sets that have the strongest connections with those untagged SNPs, which should increase power on average and in the worst case respectively. The purpose of using criteria 3 is to find a tagSNP set whose members are as independent as possible which minimizes overlap between tagSNPs and potentially increases the chance of linking to untyped SNPs. Criteria 4 may increase redundancy and robustness to genotype failure; and criteria 5 may improve genotyping success for some assays.

To evaluate the relationship between each tagSNP set identified by the aforementioned criteria, and more importantly, their potential of uncovering the disease causing mutations in association studies, we conducted some empirical evaluations, summarized in the Result section.

Other types of criteria may be of even greater interest in practice. For example, in many genotyping technologies, some SNPs are harder to genotype than others due to characteristics of surrounding genome sequence. We can use this information to select tagSNPs that are likely to have a high success rate, and to avoid SNPs that are prone to genotyping failure.

Double Coverage

So far, both the greedy approach and our FESTA algorithm focus on finding a tagSNP set such that each SNP is either a tagSNP itself or is in LD with at least one of the tagSNPs. This is a criterion aimed at minimizing the number of tagSNPs selected. In reality, random genotyping failure or genotyping error on these tagSNPs can result in loss of power to identify the true signal. To be more robust against such adverse events, we evaluated a more stringent criterion requiring that every untyped SNP be in LD with at least two tagSNPs.

Our FESTA algorithm can be extended to find tagSNP sets that will have double coverage on the SNP markers considered. As always, an exhaustive search is able to find such tagSNP sets when the marker set considered is not too large. When exhaustive search is not feasible, the same greedy-exhaustive hybrid search strategy can be applied. In practice, it may be useful to consider double coverage only for large precincts, where the cost of losing an SNP to genotyping failure might be higher.

2.3 Results

To illustrate our proposed piecewise exhaustive search strategy, compare it with the greedy approach and explore the various characteristics of the tagSNP sets selected by our method, we applied both methods to two sets of data, the entire Chromosome

2 and five ENCODE regions (ENr112, ENr131, ENr113, ENm010 and ENm013) genotyped by the HapMap project (release 16c, June 2005). All three populations: CEU (European), YRI (Yoruban) and JPT + CHB (Japanese and Chinese) were studied. The first is in the context of a genome-wide association study and the second is similar to the situation of a candidate region study.

2.3.1 Chromosome wide tagging

We have applied the greedy algorithm and FESTA to Chromosome 2 using HapMap Phase 1 genotype data (release 16c, June 2005). Tables 2.1 and 2.2 summarizes the results. FESTA produces fewer tagSNPs compared with the greedy approach in all three populations. When compared across populations, the YRI samples have about twice the amount of tagSNPs as the CEU or the JPT+CHB samples. The JPT+CHB samples have slightly less tagSNPs identified than the CEU samples. With r^2 threshold 0.5, the percentages of tagSNPs identified by our algorithm are 21.6% in CEU, 39.3% in YRI and 20.9% in JPT+CHB samples, respectively.

The size of the tagSNP set is optimal for precincts where the greedy approach indicates that one or two tagSNPs are enough to cover all the SNPs in it. Improvements over the greedy approach is only possible for the remaining precincts. In the CEU samples, there are 599 of such precincts, in which the greedy approach identified 2423 tagSNPs, and FESTA identified 2022, a 16.5% reduction. When the r^2 threshold is 0.8, 154 precincts require more than two tagSNPs, as identified by the greedy approach. Among them, the greedy approach and FESTA identified 526 and 402 tagSNPs, respectively, a reduction of 23.6% in tagSNPs chosen by FESTA. When double coverage is required, 69.1% and 45.9% more tagSNPs are needed with r^2 thresholds of 0.5 and 0.8, respectively. Similar results were obtained from the YRI

and JPT + CHB samples.

Among all the non-singleton precincts in the CEU samples (6545 for r^2 threshold of 0.5 and 10196 for r^2 threshold of 0.8), most require only a small number of tagSNPs, so that the exhaustive search can be applied directly. With r^2 threshold of 0.5, the greedy-exhaustive hybrid approach was required for only 98 precincts or 1.5% of all precincts (11 precincts (0.1%) with r^2 threshold of 0.8).

2.3.2 Densely typed region

A dense SNP map was released by the HapMap project on the ENCODE regions. We used five such regions (ENr112, ENr131, ENr113, ENm010 and ENm013) to evaluate the performance of our algorithm. Each ENCODE regions is 500 kb in length, for the CEU samples, the average number of SNPs in these regions is 832 (ranges from 551 to 1126), corresponding to an SNP density about 1 SNP per 601 bps (1 SNP per 907 bps to 1 SNP per 444 bps for individual regions). The detailed results were summarized in Table 2.3. In this set of densely typed SNPs, using our method with r^2 threshold of 0.5, the average percentage of tagSNPs required to cover each of the five regions is 8.3% of all markers (ranges from 5.4 to 11.3%). For double coverage, on average, 76.7% more tagSNPs are required (ranges from 70.7 to 83.6%). With a more stringent r^2 threshold of 0.8, the average percentage of tagSNPs required increased to 16.6% of all markers (ranges from 11.4 to 24.1%). To double cover these regions, 62.9% more tagSNPs are required (ranges from 56.9 to 71.6%). For precincts where improvement over greedy search is possible, using FESTA, the improvement in the number of tagSNPs is 17.9 and 23.0% on average for the five ENCODE regions with r^2 thresholds of 0.5 and 0.8 respectively. Using our method on YRI and JPT + CHB samples results in similar trends (data not shown).

2.3.3 Additional TagSNPs for denser SNP map

With the improvement in genotyping technologies and discovery of rarer variants, progressively denser SNP maps will become available. As more refined association studies are carried out, it will be useful to select new tagSNPs to 'fill holes' in the initial sparse maps. With a good picking strategy for the first round of tagging, this staged approach should result in only a small-to-moderate increase in the total number of tagSNPs compared to a one-stage strategy.

To evaluate this strategy, we constructed an artificial sparse SNP map for each of the five ENCODE regions (using the CEU samples only). Specifically, we selected one in every five consecutive SNP markers. The density of this sparse map is about 1 SNP per 3kb, close to the density of the phase I HapMap. Then, three different tagSNP sets are identified using the three criteria described previously, denoted by $T_i, i = 1, 2, 3$. Finally, we applied our approach to the full ENCODE SNP set, using each of these tagSNP sets as a seed, to search for additional tagSNPs to cover the previously 'hidden' SNP markers. The effectiveness of these tagSNP sets is evaluated by comparing the number of new tagSNPs needed to cover the 'newly found' SNPs. In addition to the three criteria, we also compared three other tagSNP selection strategies: Z random SNPs, assume Z is the number of tagSNPs for the sparse map, a picket fence strategy with Z equally spaced SNPs, where we place equally spaced grid points along the interval and then select markers that are closest to these grid points or using all original SNPs as tagSNPs. The results are summarized in tables 2.4. When the r^2 threshold is 0.5, 14.4% more tagSNPs (range from 7.0 to 20.9%) are needed to fill holes in the original map and that number is only 5.4% (range from 3.8 to 7.0%) with an r^2 threshold of 0.8. The three tagSNP sets require

	CEU	YRI	CHB+JPT
No. of SNPs	64801	69630	57810
$r^2 \geq 0.5$			
No. of precincts	11786	24752	10248
No. of tagSNPs (Greedy)	14384	27804	12454
No. of tagSNPs (FESTA)	13983	27379	12108
No. of tagSNPs (FESTA, double cover)	23644	41668	20644
$r^2 \geq 0.8$			
No. of precincts	23426	41079	20178
No. of tagSNPs (Greedy)	24300	41729	21044
No. of tagSNPs (FESTA)	24176	41664	20963
No. of tagSNPs (FESTA, double cover)	35824	54101	31463

Table 2.1: Summary of Chromosome 2: Number of tagSNPs with greedy approach and FESTA

fewer tagSNPs to cover the holes, compared with tagSNPs picked using a picket fence strategy (31.6% difference for r^2 threshold of 0.5 and 21.6% difference for r^2 threshold of 0.8) or picked at random (33.8% difference for r^2 threshold of 0.5 and 21.0% difference for r^2 threshold of 0.8).

2.4 Discussion

In this manuscript, we developed an efficient computational framework for tagSNP selection using the r^2 criteria. Our algorithm can handle 100,000s of linked markers and can identify smaller tagSNP sets than the greedy approach [36]. Using both chromosome wide data and densely typed ENCODE data from HapMap, we illustrated the utility of our approach and showed savings increase in more densely typed regions and inside large LD "blocks". Computational effort required by our method can be tailored to available computing resources. Another important feature is the ability of our method to identify multiple equivalent tagSNP sets and use additional criteria, such as assay design scores, to choose an optimal tagSNP set for genotyping. This feature offers flexibility in picking tagSNPs which is desirable when designing real association studies.

	CEU		YRI		CHB+JPT	
	Greedy	FESTA	Greedy	FESTA	Greedy	FESTA
Singleton	5241	5241	15079	15079	4416	4416
1	5172	5172	8096	8096	4660	4660
2	774	911	924	1070	634	770
3	318	278	355	291	312	250
4	144	99	127	100	113	90
5	59	42	73	53	60	30
6	27	18	36	28	16	15
7	17	17	21	10	14	6
8	16	4	10	8	11	6
9	11	1	6	3	4	4
10+	7	3	25	14	8	1
Total	14384	13983	27804	27379	12454	12108

Table 2.2: TagSNP distribution on Chromosome 2: Number of tagSNPs per precinct.

Region	ENr112	ENr131	ENr113	ENm010	ENm013
No. of SNPs	863	988	1061	539	708
	$r^2 \geq 0.5$				
No. of precincts	55	78	43	44	26
No. of singletons	23	31	16	16	11
No. of tagSNPs (Greedy)	81	110	71	66	41
No. of tagSNPs (FESTA)	75	105	67	61	38
No. of tagSNPs (double cover)	128	183	123	109	67
	$r^2 \geq 0.5$				
No. of precincts	134	184	131	125	72
No. of singletons	63	81	62	61	25
No. of tagSNPs (Greedy)	152	197	142	131	83
No. of tagSNPs (FESTA)	146	193	141	130	81
No. of tagSNPs (double cover)	237	311	229	204	139

Table 2.3: Summary of tagSNP results for Encode regions: CEU samples

Region	ENr112	ENr131	ENr113	ENm010	ENm013
# SNPs in dense map	863	988	1061	539	708
# SNPs in sparse map	173	198	213	108	142
	One-stage picking				
TagSNPs in dense map	75	105	67	61	38
	Two-stage picking				
Max average r^2 b/w tags and non-tags	85	114	81	72	40
Min lowest r^2 b/w tags and non-tags	85	115	80	75	40
Min average r^2 among	85	117	82	74	42
	Other Strategies				
Random picking	103.2	137.7	91.4	71.0	52.0
Picket fence	103	136	94	78	52
Use all sparse	200	241	239	134	153

Table 2.4: Comparing different criteria for tagSNP selection: Effect on number of tagSNPs in denser SNP maps

The key improvement of FESTA over the greedy approach is the 'precinct partitioning' step which enables the exhaustive search to be carried out very rapidly in most of the partitioned precincts. This is similar in spirit to the idea of 'partition-ligation' algorithm proposed by Niu et al. [45] for haplotype inference.

Many of the existing tagSNP picking algorithms aim to capture haplotype diversity using the reduced set of markers (called haplotype tagging SNPs, htSNPs) such as BEST [25]. They work well when a small number of common haplotypes exist (typically true in the vicinity of a candidate gene) but these approaches often require the knowledge of complete haplotype phase and the boundary of the haplotype blocks. On the other hand, tagSNP selection using r^2 criteria does not require knowledge of block boundaries and can easily be applied to cover the whole chromosome. Multiple-marker tagging strategies [46, 47] in which multiple tagSNPs can be used to represent each untagged SNPs have been proposed. While these methods further

reduce the number of tagSNPs selected, this approach may be sensitive to random genotyping failures.

Our approach is amenable to further computational improvements. For example, parallel programming could be used to search for tagSNPs in separate precincts, further speeding up the computation.

CHAPTER III

Framework for remapping multiply mapped short reads

3.1 Introduction

Short read sequencing has been deployed in many areas, such as, cataloging population variation [10], identifying disease susceptibility loci [48], differential gene expression analysis [49, 50] and epigenetics [51, 14].

Most sequencing study designs consist of three main steps, short read sequencing, alignment to the reference sequence and finally downstream analyses. The alignment step is affected by many factors, such as read length, error rate and model, repetitive elements and sequence homology in the reference sequence. These factors, combined with the parameters of the alignment algorithm, can result in the ambiguous alignment of reads, i.e., these reads can be mapped, with similar confidence, to multiple locations in the reference sequence within the constraints placed on the alignment process.

Several algorithms have been proposed for aligning short sequences to a reference sequence[52, 7, 53]. Most of the alignment algorithms only report reads that can be mapped to a unique location in the reference sequence. As a consequence, the *multiply*

mapped reads are commonly excluded from downstream analyses. The exclusion of multiply mapped reads can result in information loss leading to decreased sensitivity and specificity in downstream analyses. It can also produce biased results, especially in quantitative analyses [54].

Several methods have been proposed to incorporate multiply mapped reads in downstream analyses [54, 55]. Most approaches assign multiply mapped reads proportionally to their mapping positions, either based on the coverage at the mapping locations[55, 56] or a model based assignment[54, 57]. Many of the approaches were developed specifically for RNA-seq and ChIP-seq datasets. The quantitative nature of the downstream analysis allows for proportional assignment. The proportional assignment approaches cannot be utilized for DNA sequencing projects where downstream analysis is predicated upon a single accurate alignment of each read.

We present a model based approach designed to identify the most probable mapping location for the multiply mapped reads. We model the the abundance of the individual bases at each location and use a Gibbs sampler to identify a single most probable mapping for each multiply mapped read. We perform a simulation study to test accuracy in identifying the true alignment of multiply mapped reads using our method. In a subsequent simulation, we use variant discovery as the analysis of interest to quantify the improvement in downstream analysis when adding multiply mapped reads. Our algorithm was able to align upto 87% of correctly mappable short reads back to their true location. The inclusion of multiply mapped reads in variant discovery resulted in a 3% increase in the number of variants detected.

3.2 Methods

Let R be the set of all reads mapped successfully to the reference sequence L. R can be partitioned into two mutually exclusive sets, R_1 and R_{2+}, based on the number of mappings returned by the alignment algorithm. R_1 is the set of reads for which the alignment algorithm found exactly one mapping and R_{2+} is the set of reads with multiple mappings.

$$R_1 := \{r : r \in R, |M_r| = 1\} \tag{3.1}$$

$$R_{2+} := \{r : r \in R, |M_r| \geq 2\} \tag{3.2}$$

where M_r is the set of mappings for the read r. Each mapping is a location in the reference sequence to which the read can be aligned.

3.2.1 Count matrix

Consider a single location i. Let $C_i = \{C_i^A, C_i^C, C_i^G, C_i^T\}$ be the counts of bases A, C, G and T aligned to the location i. C_i can be obtained by counting the number of A, C, G and T bases present in all the reads that contain location i in their alignment, i.e. their selected mapping covers location i. Conditional on the underlying true genotype at location i, G_i, we assume that the counts follow a multinomial distribution.

$$(C_i|G_i = g) \sim Multinomial(N_i, p_g) \tag{3.3}$$

$$\begin{aligned} P(C_i|G_i = g) &= \binom{N_i}{C_i}(p_g)^{C_i} \\ &= \binom{N_i}{C_i^A, C_i^C, C_i^G, C_i^T}(p_g^A)^{C_i^A}(p_g^C)^{C_i^C}(p_g^G)^{C_i^G}(p_g^T)^{C_i^T} \end{aligned} \tag{3.4}$$

where $N_i = C_i^A + C_i^C + C_i^G + C_i^T$ is the total number of reads covering the location and $p_g = \{p_g^A, p_g^C, p_g^G, p_g^T\}$ is the vector of probabilities of observing A, C, G and T

conditional on the true genotype g.

The probability vector p_g depends on the error model for the sequencing process. In this work, we assume a uniform error model. If the underlying genotype is homozygote, i.e. $g = b_1/b_1$, the probability of observing the different bases can be written as

$$P(o|g = (b_1, b_1)) = \begin{cases} 1 - \epsilon & , \quad o = b_1 \\ \frac{\epsilon}{3} & , \quad o \neq b_1 \end{cases} \tag{3.5}$$

where ϵ is the sequencing error rate per base. In case of a heterozygote genotype, i.e. $g = b_1/b_2$, we can compute the probabilities to be

$$P(o|g = (b_1, b_2)) = \begin{cases} 0.5 - \frac{\epsilon}{3} & , \quad o \in \{b_1, b_2\} \\ \frac{\epsilon}{3} & , \quad o \notin \{b_1, b_2\} \end{cases} \tag{3.6}$$

Since the underlying genotype is unobserved, we compute the probability of observing count configuration C_i by integrating over all the possible genotypes.

$$P(C_i) = \sum_{g \in \{AA,\ldots,TT\}} P(C_i|G_i = g)P(G_i = g) \tag{3.7}$$

We use the reference sequence information to construct the genotype probabilities, $P(G_i)$. We assume the probability of a base different from the reference sequence to be 0.001, equal to the expected sequence difference for human sequences. We assign equal probabilities to all three non-reference bases. We use Hardy-Weinberg equilibrium (HWE) to obtain the genotype probabilities. In the absence of reference sequence information, we can assign equal probabilities to all bases and use HWE to get genotype probabilities.

Let $\mathbf{C} = \{C_i : i \in L\}$ be the count matrix for the entire reference sequence. Un-

der the assumption that the counts at different locations are independent, we can compute the probability of observing the count matrix $\mathbf{C}$.

$$\begin{aligned} P(\mathbf{C}) &= \prod_{i \in L} P(C_i) \\ &= \prod_{i \in L} \sum_{g \in \{AA,\ldots,TT\}} P(C_i|G_i = g)P(G_i = g) \end{aligned} \tag{3.8}$$

Consider a single read r. Let a_r be the alignment of read r, i.e. a_r is the mapping selected from M_r as the estimate of the true location of r.

3.2.2 Alignment of uniquely mapped reads

If $r \in R_1$, r has exactly one mapping returned by the aligner, i.e. $|M_r| = 1$. Thus, there is no ambiguity in the alignment of reads in R_1. Let $A_1 := \{a_k : k \in R_1\}$ be the set of alignments for the reads in R_1. We initialize the count matrix using the sequence of reads in R_1 at positions covered by A_1.

3.2.3 Alignment of multiply mapped reads

Consider the set of multiply mapped reads, R_{2+}. For each read $r \in R_{2+}$, assume that a single mapping from M_r has been selected as the current alignment. Let $A_2 := \{a_k : k \in R_{2+}\}$ be the set of alignments for the multiply mapped reads. Using Bayes rule, we can write the posterior probability of the alignments in A_2 as

$$P(A_2|\mathbf{C}, A_1) = \frac{P(\mathbf{C}|A_2, A_1)P(A_2|A_1)}{P(\mathbf{C}|A_1)} \tag{3.9}$$

We propose a Gibbs Sampling scheme to compute the posterior distribution of the multiply mapped reads. Let $r \in R_{2+}$ be a single multiply mapped read with map-

pings M_r. Additionally, let $A_{2,-r}$ be the set of alignments for all multiply mapped reads excluding r. We can write the posterior distribution of the alignment, a_r, of r conditional on the current alignment of the other reads.

$$P(a_r = m|\mathbf{C}, A_{2,-r}, A_1) \propto P(\mathbf{C}|(A_{2,-r} \cup m), A_1)P(a_r = m|A_1, A_{2,-r}) \tag{3.10}$$

where m is an element of M_r. We assume a uniform prior distribution over the mappings in M_r, i.e. $P(a_r = m|A_1, A_{2,-r}) = |M_r|^{-1}\ \forall m \in M_r$.

$$P(a_r = m|\mathbf{C}, A_{2,-r}, A_1) \propto \frac{P(\mathbf{C}|(A_{2,-r} \cup m), A_1)}{|M_r|} \tag{3.11}$$

$$P(a_r = m|\mathbf{C}, A_{2,-r}, A_1) \propto P(\mathbf{C}|(A_{2,-r} \cup m), A_1) \tag{3.12}$$

Since the probability of the count matrix using only the alignments in $A_{2,-r}$ and A_1 is independent of the mappings in M_r, we can simplify the computation of the conditional distribution as follows

$$P(a_r = m|\mathbf{C}, A_{2,-r}, A_1) \propto \frac{P(\mathbf{C}|(A_{2,-r} \cup m), A_1)}{P(\mathbf{C}|A_{2,-r}, A_1)} \tag{3.13}$$

Using equation (3.8), we get

$$\begin{aligned} P(a_r = m|\mathbf{C}, A_{2,-r}, A_1) &\propto \frac{P(\mathbf{C}|(A_{2,-r} \cup m), A_1)}{P(\mathbf{C}|A_{2,-r}, A_1)} \\ &\propto \prod_{i \in m} \frac{P(C_i|(A_{2,-r} \cup m), A_1)}{P(C_i|A_{2,-r}, A_1)} \end{aligned} \tag{3.14}$$

We limit our computation to the locations covered by the mapping m. The read contributes one additional count at each location covered by the mapping. Assume that $C_{r,l}$ is the contribution of the read r to the counts at location l. As an example, if the reads contains the base A at location l, $C_{r,l} = (1, 0, 0, 0)$; if it contains the base C at location l, $C_{r,l} = (0, 1, 0, 0)$ and so on. Using the read counts and substituting

equation (3.8) into (3.14), we get

$$\begin{aligned} P(a_r = m|\mathbf{C}, A_{2,-r}, A_1) &\propto \prod_{i \in m} \frac{P(C_i|(A_{2,-r} \cup M), A_1)}{P(C_l|A_{2,-r}, A_1)} \\ &\propto \prod_{i \in m} \frac{\sum_{g \in \{AA,...TT\}} \binom{N_i}{C_i + C_{r,i}} (p_g)^{(C_i + C_{r,i})}}{\sum_{g \in \{AA,...TT\}} \binom{N_i}{C_i} (p_g)^{C_i}} \end{aligned} \tag{3.15}$$

We iteratively align each multiply mapped read using the conditional distribution derived above. After allowing for burn-in, we sample the alignments of the reads to obtain the joint posterior distribution of the alignments of the multiply mapped reads. We obtain the maximum *a posteriori* (MAP) estimate of the alignments by selecting the marginal mode of the posterior alignment distribution for each multiply mapped read.

3.2.4 Simulations

We conduct two simulation studies to test the performance of our method. In the first study, we simulate a single individual using the coalescent simulator *ms* [58]. The population parameters for the coalescent simulation are given in table 3.1. We generate a region approximately the size of chromosome 21. Using the reference sequence of chromosome 21 as the ancestral state, we introduce variants using the sites obtained from the *ms* haplotypes. We randomly place 70 bp long reads on the simulated chromosomes. Assuming independent errors at each location on the read, we introduce errors using a uniform error model. We generate two datasets, with 10X and 30X average coverages. We use the Burrows-Wheeler aligner, *bwa* [52], to align these reads back to the reference sequence. Using our algorithm, we obtain the alignment for the multiply mapped reads. We measure the performance of our algorithm using the proportion of multiply mapped reads that were aligned back to their true location in 1000 replicates.

Parameter	Value
Effective population size (N_e)	10000
Mutation rate (μ)	1.5×10^{-8}
Recombination rate (r)	1×10^{-8}
Gene Conversion rate	4.5×10^{-9}

Table 3.1: Population parameters for simulating short read sequence data

In the second simulation study, we simulate 1Mb long haplotypes for 100 diploid individuals using the same population parameters as the first simulation. We use a 1Mb long sequence from chromosome 1 as the ancestral state. Using the mechanism described above, we generate short reads for each individual with 10X and 30X average coverage. After aligning the short reads using *bwa*, we use our algorithm to resolve the alignment of the multiply mapped reads. We use *glfMultiples* [59] to identify single nucleotide variants in the sample. We repeat the variant discovery step using only uniquely mapped reads. We compare the sensitivity and specificity of variant discovery between the two datasets.

3.3 Results

We present the results of our simulation study to characterize multiply mapped reads. Table 3.2 shows the numbers of uniquely mapped and multiply mapped reads with 10 and 30 fold average coverage on chromosome 21. More than 90% of all reads are uniquely mapped to the reference region and 5% of reads are multiply mapped. BWA could not align about 4% of reads. Greater than 95% of the uniquely mapped reads are aligned to their true location in the reference sequence. The proportion of multiply mapped reads that contain the true alignment in the list of mappings returned by BWA is approximately 80%. Since 20% of the multiply mapped reads do not contain the true alignment, our algorithm cannot remap them to their true genomic location. Fig. 3.1 shows the proportion of multiply mapped reads aligned to their true genomic

Coverage	Read type	Fraction	Correctly mappable
10X	Uniquely mapped	0.91	96.54%
	Multiply mapped	0.05	79.01%
	Unmappable	0.04	-
30X	Uniquely mapped	0.90	95.10%
	Multi-mapped	0.06	80.77%
	Unmappable	0.04	-

Table 3.2: Read characteristics for chromosome 21 simulated data

location using the Gibbs sampling approach. The proportion of multiply mapped read aligned correctly increases quickly before plateauing out at 70%. We see a slight increase in the proportion of correctly mapped read at 30 fold coverage compared to 10 fold coverage. Since only 80% of multiply mapped reads can be aligned back to their true location, our method is able to align approximately 87% of multiply mapped to their true location conditional on their true mapping being identified by the alignment algorithm.

The improvement in variant discovery with the inclusion of multiply mapped reads aligned by our method is given in table 3.3. For the 10 fold coverage data, using only uniquely mapped reads uncovers 87% of all variants present in the sample. Using the alignment provided by our method, we were able to increase the number of variants discovered from 1782 to 1842, a 3% increase. If all multiply mapped reads had been placed in their true genomic location, 93% of all variants would have been discovered. Similar trends can be seen for the dataset with 30 fold average coverage. Variant discovery with uniquely mapped reads resulted in 1897 variants. Including multiply mapped reads in the analysis led to a 3% increase in the number of variants discovered.

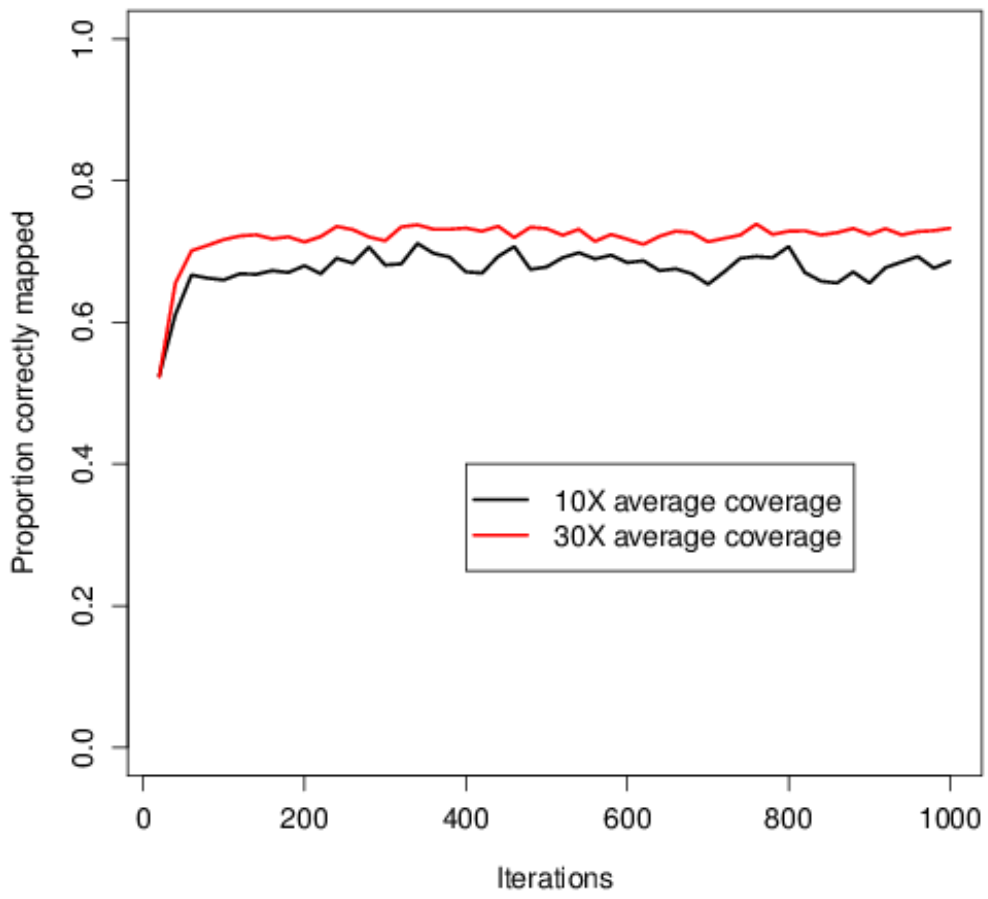


Figure 3.1: Proportion of multiply mapped reads aligned to their true location using our algorithm

3.4 Discussion

In this work, we developed a Bayesian framework to identify the true location of multiply mapped reads. In simulation studies, our method was able to resolve the mapping ambiguities of more than 70% of the multiply mapped reads. Restricting to multiply mapped reads which could be correctly assigned, our method was able to correctly align ∼90% of the multiply mapped reads. Incorporating the estimated alignment of multiply mapped reads into variant discovery resulted in a 3% increase

Coverage	Reads used	# SNPs	True SNPs discovered	False positives[a]
10X	Uniquely mapped		1782	0
	Multiply mapped		1842	5
	Mappable[b]	2057	1918	1
30X	Uniquely mapped		1897	0
	Multiply mapped		1951	3
	Mappable[b]	2057	2011	0

[a] Monomorphic positions identified as single nucleotide variants

[b] All reads that were mapped by the aligner were assigned to their true location

Table 3.3: Effect of multiply mapped reads on variant discovery

in the number of variants detected.

Since variant discovery is a computationally intensive process, we use the MAP estimate for downstream analysis. Since ∼20% of the multiply mapped reads are not aligned to their true location, variant calling on a single individual results in a high false discovery rate. Using information from multiple samples helps overcome the adverse effects of misaligned multiply mapped reads.

Most existing methods for resolving multiply mapped reads assign the reads proportionally to multiple locations. In contrast, our method can estimate a unique alignment for each multiply mapped read. Further, the Gibbs sampling approach allows the estimation of the joint posterior distribution of the alignment of all multiply mapped reads. Using this joint distribution, our method can be used for quantitative analyses by aligning multiply mapped reads proportional to their posterior probabilities.

Short read sequencing protocols are constantly evolving, resulting in ever increasing read lengths. As read length is inversely proportional to mapping ambiguity, this leads to a reduction in the proportion of multiply mapped reads. Paired end sequenc-

ing is another tool that can be used to improve the mapping fidelity. Although these advances result in fewer multiply mapped reads, repetitive sequences, structural variants and sequence homology in the human genome will ensure ambiguity in mapping short read sequences. Our method can be used as an additional tool to extract information from multiply mapped reads.

CHAPTER IV

Feasibility of admixture mapping to identify rare susceptibility variants

4.1 Introduction

Genome wide association studies (GWAS)[1] have been the instruments of choice in detecting genetic susceptibility loci. GWAS use genotype-phenotype correlation, computed using sets of affected and unaffected samples, to identify loci associated with the trait of interest. GWAS are best suited to detect variants that fall under the Common Disease-Common Variant (CDCV)[60] hypothesis. The CDCV hypothesis proposes that common diseases are caused by commonly occurring genetic variants with low to moderate effect sizes. GWAS have been successfully employed to identify common risk variants for common diseases such as diabetes[2], cardio-vascular diseases[61], Parkinson's disease [62, 63] and colorectal cancer [64, 65]. They have been used to identify genetic variants affecting quantitative traits such as lipid levels[66, 67] and BMI[68, 69]. While GWAS have been successful in identifying genetic loci influencing many heritable traits [3], in many cases, the variants so identified have not been able to adequately explain the heritability of common diseases [4, 5].

The unexplained heritability may be attributed to the presence of rare causal vari-

ants that GWAS are not well suited to detect. The Common Disease-Rare Variant (CDRV)[70] hypothesis surmises that the prevalence of common diseases can be attributed to the presence of several rare causal variants in the population, each with a moderate to high effect size. The two hypotheses, CDCV and CDRV, reflect the presence of causal variants at different parts of the allelic spectrum of susceptibility variants. These hypotheses fit as complementary pieces in our effort to unravel the underlying genetic architecture of complex common diseases.

Several methods have been proposed to identify rare causal variants [71, 72, 73, 8, 74]. Rare variants mapping methods have been used to find loci affecting LDL cholesterol[12, 75] and HDL cholesterol levels[11]. Since testing each rare variant individually is not statistically powerful, many rare variant mapping methods combine the rare variants across a larger unit, such as a gene or an exon. Subsequently, they test the burden or distribution of rare variants in each testing unit across affected and unaffected samples. Different strategies have been proposed to combine the rare variants across a testing unit, such as the presence or absence or rare variants[11, 7], sum of minor alleles across rare variants[8] and weighted rare allele counts[71].

In admixed populations, i.e. populations with multiple founding populations, we can use the ancestry of a testing unit as one way to collapse information across the variants present in the region. An admixed population derives its ancestry from two or more genetically different founder populations. The chromosomes of individuals from an admixed population consist of a mosaic of genetic material from the different founding populations. A block of shared ancestry is known as an admixture block. Genetic recombination reduces the size of admixture blocks over generations. Figure 4.1 shows the changes in admixed chromosomes over multiple generations.

Admixture mapping is a tool used to identify regions associated with a trait of

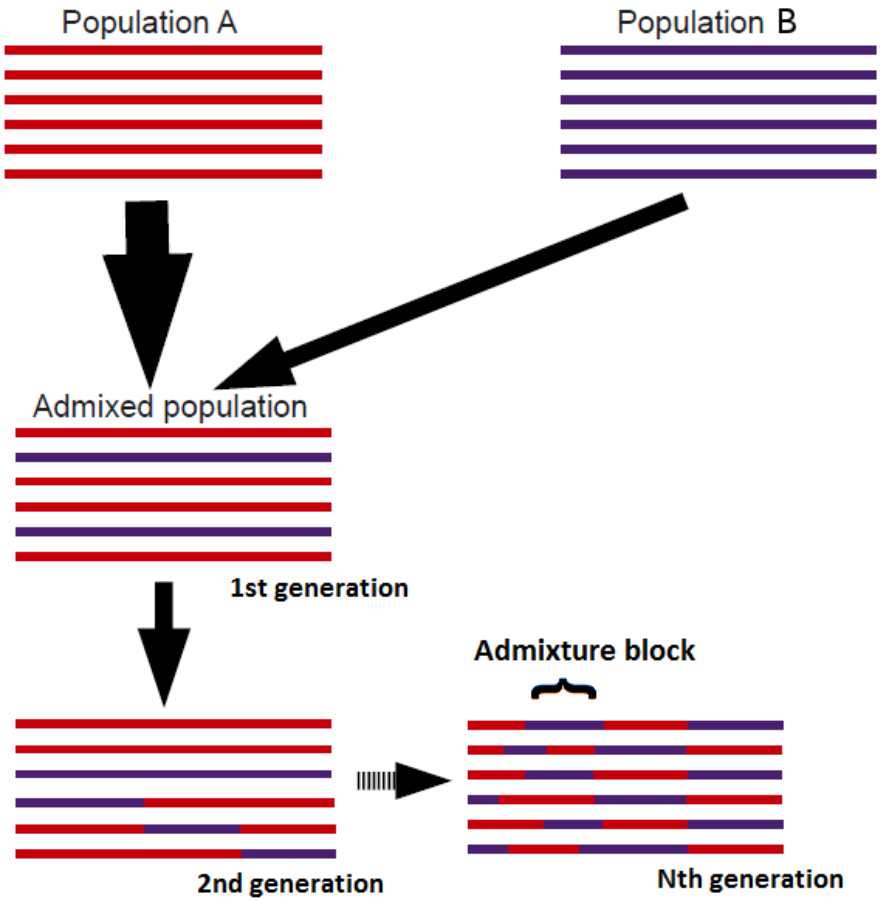


Figure 4.1: Admixed chromosomes and admixture blocks

interest by examining the differences in ancestry between affected and unaffected individuals. Consider an admixed population with two founding populations A and B. In the admixed population, the genomic regions that are not associated with the diseases should have ancestry equal to the mixing proportions, irrespective of the disease status of the individuals. For regions associated with the disease, the ancestry contribution is dependent on the disease status. For affected individuals, we expect to see excess ancestry from the founding population where the region's contribution to disease prevalence is higher. Conversely, for unaffected individuals, we expect to see diminished ancestry contribution from this founding population. This difference

in ancestry between affected and unaffected samples can be tested to map regions associated with the disease.

Admixture mapping can be used in two flavors. Case-control admixture mapping tests the difference in ancestry between affected and unaffected samples. A more powerful approach is cases only admixture mapping, which compares the ancestry of admixture blocks in cases to the estimated average genome-wide ancestry. Cases only admixture mapping can produce spurious results in the presence of deviations from average ancestry unrelated to the disease.

Admixture mapping has been shown to be powerful in identifying common causal variants [76]. Risk variants for end-stage renal disease[77] have been identified using admixture mapping. Admixture mapping has also yielded promising candidate regions associated with prostate cancer. Several diseases including hypertension, lung cancer, stroke have disparate prevalences in African and European populations [78], making them good candidates for admixture mapping. For many of these diseases, susceptibility loci have yet to be identified. Here, we explore the feasibility of admixture mapping in the context of mapping rare causal variants.

Data from HapMap [16] and 1000 Genomes[10] projects show that African populations contain more polymorphic sites than European or Asian populations. This difference is more pronounced for low frequency and rare variants. Many rare variants are private to one of the two populations and those that occur in both populations seldom have the similar frequencies [10]. This difference in the frequencies of rare variants among populations combined with the CDRV hypothesis, implies a difference in contribution of the locus to disease prevalence in the two populations. Since admixture mapping draws its power from the difference in prevalence contributions,

it is a natural choice to identify loci harboring rare variants.

The first step in admixture mapping is the estimation of the ancestry of admixture blocks. For the remainder of this work, we assume that ancestry of admixture blocks can be accurately estimated. This assumption is justified since modern genome-wide genotyping platforms include a dense set of markers spread across the genome. We can estimate ancestry of genomic locations by combining information from neighboring markers [79, 80]. We compare admixture mapping to single marker association studies using a simulation study. Using a multiplicative model for disease risk, we analytically compute the power of admixture mapping conditional on the contribution of the locus to prevalence in both populations. We simulate the founding populations using *ms*[58]. We calibrate the coalescent simulator using a model for population history of Africans and Europeans[81]. We compute the power of the single marker association tests using simulated cases and controls sampled from the founding populations. The power of the single marker association tests is computed as the proportion of datasets with a significant result using a 1 degree of freedom test for allelic association. We perform association tests using two strategies, viz., direct association where the causal variants are among the variants tested and indirect association, where a set of tagging SNPs are tested. We compare the performance of admixture mapping to single marker association analysis across different cumulative risk allele frequencies.

Under our simulation settings, we find that admixture mapping has moderate power to detect the susceptibility region. For our disease model, admixed populations with equal contributions from two founding populations yield the best power for admixture mapping. The power of admixture mapping is directly proportional to the cumulative risk allele frequency. When the cumulative risk allele frequency is held under 1% in the European population, admixture mapping has less than 10% power to detect as-

sociation of the region with the disease, even with sample sizes upto 10000 cases and 10000 controls. At a cumulative risk allele frequency of 5% in Europeans, with 10000 cases and controls each, admixture mapping has 60% and 80% power in populations with mixing ratios of 80-20 and 50-50 respectively.

4.2 Methods

In this section, we present the disease model that we use for rare variants. We also derive the relationship between the contribution of the locus to disease prevalence in the founding populations and the admixed populations. Using these results, we analytically compute the power of admixture mapping under the chosen disease model.

We need to set up some parameters to elaborate the disease model and compute the power of admixture mapping. Consider an admixed population, C, with two genetically different founding populations. Let the two founding populations be denoted by population A and population B. Let p_A and $p_B = 1 - p_A$ denote the contributions of population A and population B to the admixed population respectively. This implies that for any locus not associated with disease, an admixed individual is expected to derive their ancestry from population A with probability p_A. Additionally let F_A denote the contribution of the causal region to disease prevalence in population A and let F_B be the same in population B. We show that the power of admixture mapping is a monotonically increasing function of the ratio of the contributions to prevalences.

4.2.1 Disease Model

To compute the power of admixture mapping, we need to specify the underlying disease model. Following the model proposed by Zhu and Risch [82], we assume a multiplicative disease model at each susceptibility locus. In addition, we assume that

the different causal variants at the same locus interact in a multiplicative manner to affect the disease risk.

Consider the causal region in population A. Assume there are K causal variants in this region. We denote the relative risk of causal variant k by r_k. Given the relative risk of all the causal variants and a model of multiplicative interaction of causal variants, we can write the relative risk associated with a given haplotype h, R_h, as

$$R_h = \prod_{k=1}^{K} r_k^{I(h^k)} \tag{4.1}$$

where $I(h^k)$ is a variable indicating whether haplotype h carries the risk allele at the k-th causal variant. Given an individual from population A, carrying haplotypes h_1 and h_2, we can write the risk of disease in that individual as

$$\begin{aligned} P(D|(h_1, h_2), A) &= b^2 R_{h_1} \cdot R_{h_2} \\ &= b^2 \prod_{k=1}^{K} r_k^{I(h_1^k)} r_k^{I(h_2^k)} \end{aligned} \tag{4.2}$$

Here, b^2 is the baseline risk associated with the null genotypes, i.e. the risk conferred by carrying no causal variants.

4.2.2 Contribution to prevalence in the founding and admixed populations

Using the haplotype risks, we can compute the contribution of this region to the prevalence of the disease, F_A. Let $\mathcal{H}$ be the set of all possible haplotypes of the causal variants. Let f_h^A denote the frequency of haplotype h in population A. We can

write F_A in terms of haplotype risks and frequencies as

$$\begin{aligned} F_A &= \sum_{(h_1,h_2)\in\mathcal{H}^{\in}} P(D|(h_1,h_2)) f_{h_1}^A f_{h_2}^A \\ &= \sum_{(h_1,h_2)\in\mathcal{H}^{\in}} b^2 R_{h_1} R_{h_2} f_{h_1}^A f_{h_2}^A \\ &= \sum_{(h_1,h_2)\in\mathcal{H}^{\in}} (bR_{h_1} f_{h_1}^A)(bR_{h_2} f_{h_2}^A) \\ &= \left(\sum_{h\in\mathcal{H}} bR_h f_h^A\right)^2 \end{aligned} \tag{4.3}$$

We can derive the contribution of the locus to the prevalence in population B similarly. If we assume that identical haplotypes confer identical risks in both populations, we find that the risk of disease for an individual from population B carrying haplotypes h_1 and h_2 is exactly the same as in equation (4.2). Using an approach similar to equation (4.3), we can compute the contribution of the locus to disease prevalence in population B, F_B.

$$F_B = \left(\sum_{h\in\mathcal{H}} bR_h f_h^B\right)^2 \tag{4.4}$$

In the above equation, f_h^B is the frequency of haplotype h in population B. We note that the difference in contributions to prevalence of the region in the two founding populations is driven by the difference in the frequencies of the haplotypes carrying the causal variants.

We can calculate the contribution to prevalence of the region in the admixed population by combining the equations (4.3) and (4.4). First, we note that for an individual carrying a pair of haplotypes h_1 and h_2 admixed population, there are three possible ancestries. They can both trace their ancestry back to a single population, A or B. Alternatively, one haplotype each can be inherited from populations A and B. Let f_h^C

be the frequency of haplotype h in the admixed population. We can write f_h^C as the weighted mean of the frequencies of the haplotype in populations A and B.

$$f_h^C = p_A f_h^A + p_B f_h^B \tag{4.5}$$

Since we assume that the baseline and haplotype risks remain unchanged in the admixed population, we can explicitly write out the contribution of the region to disease prevalence in the admixed population, F_C, as follows

$$\begin{aligned} F_C &= \sum_{h_1 \in \mathcal{H}} \sum_{h_2 \in \mathcal{H}} b^2 R_{h_1} f_{h_1}^C R_{h_2} f_{h_2}^C \\ &= \left(\sum_{h \in \mathcal{H}} b R_h f_h^C \right)^2 \end{aligned} \tag{4.6}$$

We can use equations (4.3), (4.4) and (4.5) to simplify equation (4.6).

$$\begin{aligned} F_C &= \left\{ \sum_{h \in \mathcal{H}} b R_h \left(p_A f_h^A + p_B f_h^B \right) \right\}^2 \\ &= \left\{ p_A \sum_{h \in \mathcal{H}} b R_h f_h^A + p_B \sum_{h \in \mathcal{H}} b R_h f_h^B \right\}^2 \\ &= \left\{ p_A \sqrt{F_A} + p_B \sqrt{F_B} \right\}^2 \\ &= p_A^2 F_A + p_B^2 F_B + 2 p_A p_B \sqrt{F_A F_B} \end{aligned} \tag{4.7}$$

We can break the contribution to prevalence of the region in the admixed population into three parts.$p_A^2 F_A$ and $p_B^2 F_B$ are the contributions from individuals both of whose haplotypes are inherited from population A and B respectively, while $2 p_A p_B \sqrt{F_A F_B}$ is the contribution of individuals with one haplotype from each population. Conversely, we can view F_A, F_B and $\sqrt{F_A F_B}$ as the probability of disease given that the individual

is carrying two haplotype from population A, two haplotypes from population B and one haplotype each from either population, respectively.

4.2.3 Power of admixture mapping

We can calculate the power of admixture mapping analytically, conditional on knowing the contribution of the locus to disease prevalence in the two founding populations, viz., F_A and F_B. We will focus on case-control admixture mapping. Admixture mapping compares the ancestry proportions of cases and controls. It tests the difference in ancestries for cases versus controls using a one degree of freedom χ^2 test for independence between ancestry and disease status. If the contribution of the region to disease prevalence is the same in both founding populations, the ancestry proportions of cases and controls should be identically distributed and admixture mapping rejects the hypothesis that the region is associated with the disease. Let us consider an admixture mapping setup with n_C cases and $n_{\overline{C}}$ controls sampled from the admixed population. Let the counts of the 2x2 table be given by n_{AC}, n_{BC}, $n_{A\overline{C}}$ and $n_{B\overline{C}}$.

	Ancestry A	Ancestry B
Case	n_{AC}	n_{BC}
Control	$n_{A\overline{C}}$	$n_{B\overline{C}}$

Here n_{AC} and n_{BC} are the counts, in cases, of haplotypes inherited form population A and B respectively. Similarly, $n_{A\overline{C}}$ and $n_{B\overline{C}}$ are the counts of haplotypes in controls inherited from population A and B respectively. We can obtain the probability of a haplotype falling in each one of the four cells of the 2x2 table using the contributions of the region to disease prevalence in the founding and the admixed populations. Let $P(C, A)$, $P(C, B)$, $P(\overline{C}, A)$ and $P(\overline{C}, B)$ be the aforementioned probabilities. Consider the probability of a haplotype falling in the cell for cases with ancestry in population A.

$$P(C,A) = P(A|C)P(C) = \frac{n_C}{n_C + n_{\overline{C}}}\ P(A|C) \tag{4.8}$$

$P(A|C)$ is the probability of the haplotype being from population A if it was sampled from a case. We can expand equation (4.8) by explicitly calculating $P(A|C)$ using $P(A,A|C)$, the probability of carrying the second haplotypes also from population A conditional on being a case and $P(C|A,B)$, the probability of carrying the second haplotype from population B conditional on being a case.

$$\begin{aligned}
P(C,A) &= P(A|C)P(C) \\
&= \frac{n_C}{n_C + n_{\overline{C}}}\left(P(A,A|C) + P(A,B|C)\right) \\
&= \frac{n_C}{n_C + n_{\overline{C}}}\left(\frac{p_A^2 F_A}{F_C} + \frac{p_A p_B \sqrt{F_A F_B}}{F_C}\right) \\
&= \frac{n_C}{n_C + n_{\overline{C}}}\left(\frac{p_A\sqrt{F_A}}{F_C}(p_A\sqrt{F_A} + p_B\sqrt{F_B})\right) \\
&= \frac{n_C}{n_C + n_{\overline{C}}}\left(p_A\frac{\sqrt{F_A F_C}}{F_C}\right)
\end{aligned} \tag{4.9}$$

Similarly, we can obtain the other probabilities. They are given below.

$$\begin{aligned}
P(C,B) &= \frac{n_C}{n_C + n_{\overline{C}}}\left(p_B\frac{\sqrt{F_B F_C}}{F_C}\right) \\
P(\overline{C},A) &= \frac{n_{\overline{C}}}{n_C + n_{\overline{C}}}\left(p_A\frac{(1-\sqrt{F_A F_C})}{1-F_C}\right) \\
P(\overline{C},B) &= \frac{n_{\overline{C}}}{n_C + n_{\overline{C}}}\left(p_B\frac{(1-\sqrt{F_B F_C})}{1-F_C}\right)
\end{aligned} \tag{4.10}$$

Finally, conditional on the contribution of the region to prevalence in the founding populations, we can calculate the power of case-control admixture mapping using the

probabilities of each cell in the 2x2 table. We calculate the non-centrality parameter, λ, of the χ^2 statistic under the alternative hypothesis of non-independence of ancestry and disease status. The power of case-control admixture mapping, $P_{CC,adm}$, at level α, under our disease model for rare variants is given by

$$\begin{aligned} \lambda &= 4n\frac{(P(C,A)-P(\overline{C},A))^2}{P(A)P(B)} && (4.11) \\ P_{CC,adm} &= P(\chi^2_{1,\lambda} > \chi^2_1(1-\alpha)) && (4.12) \end{aligned}$$

where $P(A) = P(C,A)+P(\overline{C},A)$ and $P(B) = P(C,B)+P(\overline{C},B)$ are the probabilities of sampling a chromosome inherited from population A and B respectively, in a balanced case-control study. Note that these probabilities are not equal to the mixing proportions p_A and p_B.

4.2.4 Relationship between power and contributions to prevalence

We investigate the effect of the contribution to prevalences in the founding populations on the power of case-control admixture mapping. We can rewrite the non-centrality parameter, λ, by substituting the values of $P(C,A)$ and $P(\overline{C},A)$ from equations (4.9) and (4.10).

$$\begin{aligned} \lambda &= 4n\frac{(P(C,A)-P(\overline{C},A))^2}{P(A)P(B)} \\ &= 4n\frac{\left(\frac{n_C}{n_C+n_{\overline{C}}}\frac{\sqrt{F_A F_C}}{F_C} - \frac{n_{\overline{C}}}{n_C+n_{\overline{C}}}\frac{1-\sqrt{F_A F_C}}{1-F_C}\right)^2}{P(A)P(B)} && (4.13) \end{aligned}$$

Assuming a balanced case-control design, i.e. $n_C = n_{\overline{C}} = n$, we can further simplify the non-centrality parameter as,

$$\begin{aligned}\lambda &= 4n(0.5)^2 \frac{\left\{ p_A \frac{\sqrt{F_A F_C}}{F_C} - p_A \frac{(1-\sqrt{F_A F_C})}{1-F_C} \right\}^2}{P(A)P(B)} \\ &= n \frac{(p_A p_B)^2}{P(A)P(B)} \left\{ \frac{\sqrt{F_A} - \sqrt{F_B}}{\sqrt{F_C}(1-F_C)} \right\}^2 . \end{aligned} \tag{4.14}$$

For fixed mixing proportion p_A and contribution to prevalence in the admixed population, the non-centrality parameter, λ, depends on the relationship between the two populations solely through the term $(\sqrt{F_A} - \sqrt{F_B})^2$, which can be expressed in terms of the ratio of the contribution to prevalences. Any deviation of the ratio, F_A/F_B, from 1, results in an increase in λ, resulting in increased power for admixture mapping. Conversely, if the contributions to prevalence in the two populations are the same, i.e. $F_A = F_B$, we obtain a central χ^2 distribution under the alternative hypothesis, resulting in no power for admixture mapping. It is important to note that λ is not symmetrically related to the contributions to prevalence in the founding populations in case of non-equal mixing proportions, i.e. $p_A \neq 0.5$, since it depends on the contribution to prevalence in the admixed population.

Using the explicit form for F_A and F_B from eqns. (4.3) and (4.4), we can rewrite (4.14) as

$$\begin{aligned}\lambda &= n \frac{(p_A p_B)^2}{P(A)P(B)F_C(1-F_C)^2} \left\{ b \left(\sum_{h \in \mathcal{H}} R_h f_h^A - \sum_{h \in \mathcal{H}} R_h f_h^B \right) \right\}^2 \\ &= n \frac{(p_A p_B)^2}{P(A)P(B)F_C(1-F_C)^2} \left\{ b \sum_{h \in \mathcal{H}} R_h (f_h^A - f_h^B) \right\}^2 \end{aligned} \tag{4.15}$$

From equation (4.15), we can view the non-centrality parameter as being proportional to the squared difference between the mean haplotype risks in the two populations, weighted by the haplotype frequencies. If we make the simplifying assumptions that each risk variant lies on its own haplotype, we can compute the expected non-centrality parameter using equation (4.15) as

$$\begin{aligned} E(\lambda) &= E(n\frac{(p_A p_B)^2}{P(A)P(B)F_C(1-F_C)^2}\left\{bR_h\sum_{k=1}^{K}(f_k^A - f_k^B)\right\}^2) \\ &= n\frac{(p_A p_B)^2}{P(A)P(B)F_C(1-F_C)^2}b^2\Delta_{AB}^2 E(R_h^2) \\ &= n\frac{(p_A p_B)^2}{P(A)P(B)F_C(1-F_C)^2}b^2\Delta_{AB}^2(\mu_R^2+\sigma_R^2) \end{aligned} \tag{4.16}$$

where f_k^A and f_k^B are the frequencies of risk variant k in populations A and B respectively, $\Delta_{AB} = \left(\sum_{k=1}^{K} f_k^A - f_k^B\right)$ is the difference in the cumulative risk variant frequencies between the two populations and μ_R and σ_R^2 are the mean and variance of the haplotype risks. The power of case-control admixture mapping is a monotonically non-decreasing function of the difference in risk variant frequencies between the two populations.

4.2.5 Power of single marker association test

We compare the case-control admixture mapping test to the 1 degree of freedom single marker case-control test for allelic association. This test is more powerful that the 2 degrees of freedom genotype association test under the disease model considered here. We can analytically compute the power of a single marker test, assuming that the causal variants are tested directly and linkage equilibrium between the susceptibility loci. For illustration purposes, we consider single marker tests conducted in population A. Given n cases and n controls, we can compute the non-

centrality parameter of the 1-df χ^2 statistic testing allelic association for risk variant k in a similar fashion as (4.14).

$$\lambda_k^A = n\frac{(f_k^A(1-f_k^A)b_k)^2}{P(k)(1-P(k))}\left(\frac{r_k-1}{\sqrt{F_A}(1-F_A)}\right)^2 \tag{4.17}$$

where $P(k)$ is the probability of sampling the risk allele in a balanced case-control study and b_k is the baseline risk at causal variant k. We have the most power to detect the causal variant with the highest non-centrality parameter. Since we test multiple markers, we compare the case-control admixture mapping with the most extreme test statistic. We perform the single marker association test under two scenarios; direct association, when the causal markers are among the variants tested for association, and indirect association, where tagging SNPs are used as proxies for the causal variants. In either scenario, the distribution of the most extreme test statistic does not follow the non-central χ^2 distribution under the alternative hypothesis. Since we cannot compute power analytically, we use a simulation study to estimate the power of the single marker association test.

4.2.6 Simulations

The power of admixture mapping depends on the distribution of the contributions to prevalence in the founding populations, F_A and F_B. The joint distribution of these parameters is controlled by the relationship between the two populations, the disease model and the linkage disequilibrium (LD) structure in the locus harboring the causal variants. Similarly, the power of the single marker association test depends on the local LD structure and distribution of the frequencies of the risk variants. We present a simulation scheme to generate admixture blocks and case-control samples using a calibrated population history and the previously described rare causal variant disease model.

For admixture mapping, we are most interested in two admixed populations, the African-American and Latino populations. We use a model of population history[81] shown in fig. 4.2 to simulate haplotypes from the founding populations, viz., Africans and Europeans. We consider an admixture block 100 kb long. We use *ms*[58] to generate haplotypes in the admixture block from each of these two founding populations. We simulate 200 independent realizations of the coalescent process using the same population parameters. We create 50 replicates of case-control samples and contribution to prevalences in the two populations using each coalescent realization, for a total of 10000 data points.

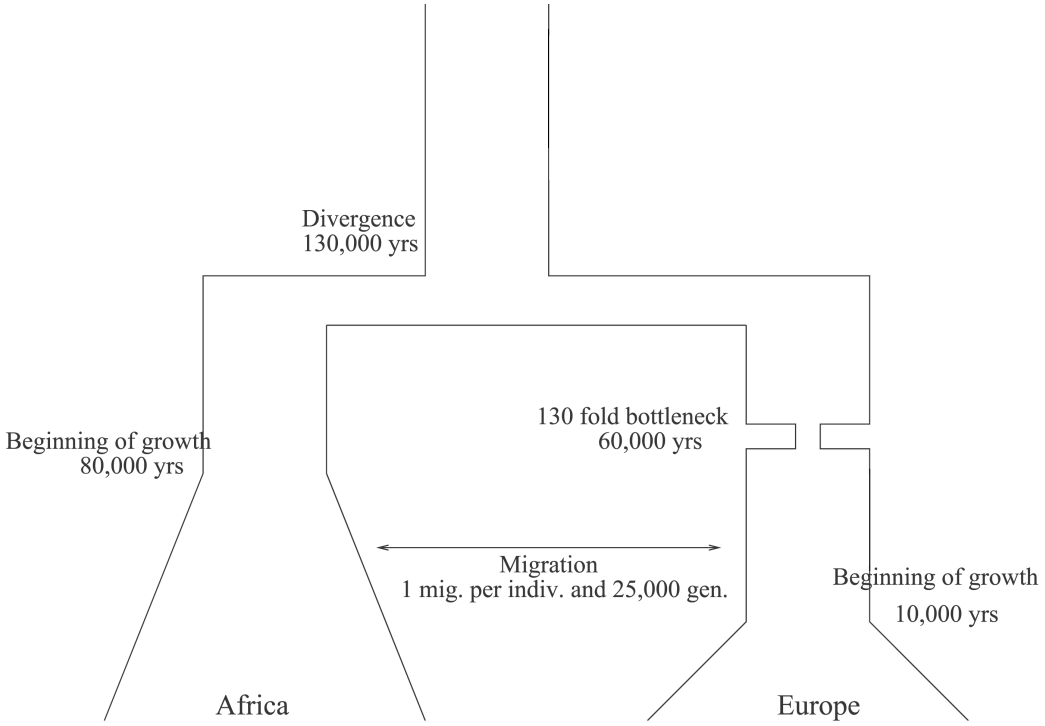


Figure 4.2: Population model used for simulating haplotypes using *ms*

4.2.6.1 Choosing risk variants

Using the simulated haplotypes, we select causal variants randomly from the set of variants with minor allele frequency (MAF) less than 10% in both populations. We select causal variants such that the cumulative MAF reaches a pre-specified threshold in the European population. We use different values of the cumulative MAF threshold in our simulation study. We assume that the minor allele is the risk allele. We also assume that the risk of the variant is inversely proportional to its MAF. Since we focus on variants that are unlikely to be found using a genome wide association study, we assign relative risks to causal variants such that the power of a genome wide association study, with 20000 cases and 20000 controls is fixed at 10% after correcting for 1 million tests. Fig. 4.3 shows the relative risk as a function of minor allele frequency. We use the higher of the two MAF from the two populations to assign the relative risk. We compute the risk of each haplotype using equation (4.1). We fix the baseline risk by setting the contribution to prevalence in the European population to be 1% in equation (4.3).

4.2.6.2 Admixture mapping

We generate an admixed individual by picking random haplotypes from the two founding populations, proportional to their contribution to the admixed population. For the African-American population, we use a mixing ratio of 80% African contribution and 20% European contribution. For the Latino population we simulate, we use an equal contribution from both populations. We note that all the markers in the admixture block share their ancestry, i.e. the boundaries of the admixture block are shared among the individuals. We compute the contribution to prevalence in the African population using haplotype risks and frequencies in that population and using the same baseline risk as the European population. We compute the power of

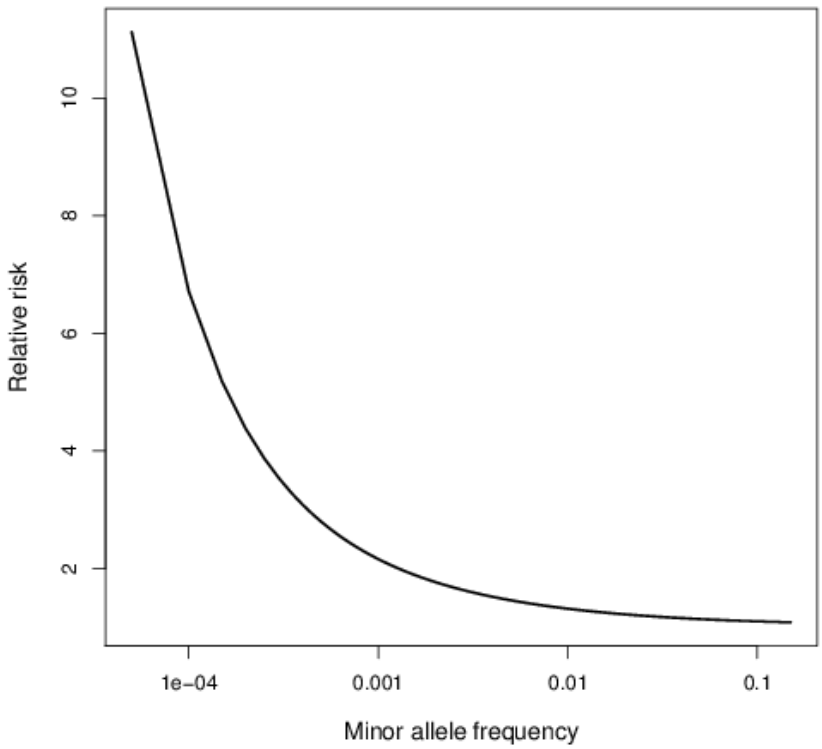


Figure 4.3: Relative risk vs minor allele frequency. Power of single marker test is set at 10% for 20000 cases and controls, correcting for 1 million tests.

admixture mapping by plugging the contribution of prevalences for both populations into equations (4.7), (4.12) and (4.14) and correcting for 3000 equivalent independent tests. Finally, we obtain the overall power of admixture mapping by averaging power across the 10000 datasets.

4.2.6.3 Single marker test

We perform the single marker association test on two different sets of markers. In the first set of markers, we perform the single marker association test on all the variants present in the region, including directly testing the risk variants. In the second set,

we limit ourselves to tagging SNPs, which are chosen to maximize information on common markers present in the region. We choose tagSNPs such that all the variants in the region with a minor allele frequency greater than 1% are either present in the tagging SNP set or have a pairwise LD measure, r^2, greater than 0.8 with at least one member of the tagging SNP set. We repeat the analysis with a tagging minor allele frequency threshold of 0.5%. We compute the power of the single marker test by checking the significance of the maximum test statistic in the region corrected for genome wide significance ($\alpha = 5 \times 10^{-8}$) in the 10000 replicates.

4.3 Results

We present the analytical results for the power of admixture mapping under our disease model. Figure 4.4 shows the power of admixture mapping over a range of contributions to prevalence in the two founding populations. Admixture mapping is more powerful when two founding populations contribute equally to the admixed population. In the case of unequal contributions, the contribution to prevalence in the population with lower ancestral contribution plays a larger role in determining the power of admixture mapping. Figure 4.5 shows the relationship between the power of admixture mapping, contribution to prevalence in one population and sample size. The contribution to prevalence in the other founding population is fixed at 1%. For both mixing ratios, viz., 50-50 and 80-20, admixture mapping power increases sharply as contribution to prevalence moves away from 1%. This increase is more pronounced with higher sample sizes.

Figure 4.6 shows the mean and standard error of the ratio of contribution to prevalence in Africans and Europeans, across a range of different cumulative risk allele frequencies in Europeans. We find a moderate increase in the ratio of prevalence

contributions with cumulative risk allele frequency. The ratio starts to plateau at the higher end of the allele frequency range. We obtain similar results across a range of contribution to prevalence in the European population, indicating that the ratio of prevalence contributions is solely dependent on the risk allele frequencies. We also note that this distribution is independent of the mixing ratios.

We compare the power of admixture mapping and direct association, where the causal variants are directly tested for correlation to the disease, in figure 4.7. For both sample sizes considered, admixture mapping performs better than single marker association tests. The powers of all three tests increase with cumulative risk allele frequency. The difference in power between the admixture mapping for the two mixing ratios is higher with the 10000 cases and controls compared to 1000 cases and controls. Admixture mapping with 10000 cases and controls each, on a population similar to the African-Americans, increases in power from 10% to 60% as the risk allele frequency increases from 1% to 5%. The power of admixture mapping under the same settings, in a population with a 50-50 mixing ratio, increases from 25% to 80%.

We compute the power of indirect association using tagSNPs and compare to admixture mapping. Using 5000 cases and controls each, the power of single marker test using tagSNPs is much lower than the power of both admixture mapping tests. For cumulative risk allele frequency of 0.5%, all tests have very low power with direct association being the best test at 3% power. Power improves for all tests with increasing risk allele frequency, with admixture mapping gaining the most. Single marker association tests using tagging SNPs have power less than 5% for all risk allele frequencies below 0.05.

4.4 Discussion

In this work, we explored the feasibility of admixture mapping as a tool to identify regions harboring rare causal variants. Under the multiplicative disease model for rare variants and relative risk settings resulting in very low power for single marker association studies, admixture mapping had moderate power to detect the susceptibility locus.

In our simulation studies, the power of admixture mapping was higher than 50% only for sample sizes larger than 5000 cases and controls each. Modern admixture mapping studies have much smaller sample size[77, 83], of the order of 1000 cases and controls each. Admixture mapping with current sample sizes would be underpowered to detect regions harboring rare susceptibility variants.

The power of admixture mapping is directly proportional to the cumulative frequency of the risk alleles. At cumulative risk allele frequencies less than 0.5%, admixture mapping had almost no power to detect the association of the region to the disease. Increasing the cumulative risk allele frequency to 1% significantly increased the power of admixture mapping. Admixture mapping draws its power from the difference in risk allele frequencies between the founding populations. Since very low risk allele frequencies preclude a large frequency difference between the founding populations under our disease model, it follows that admixture mapping is not well powered to detect an accumulation of really rare risk variants.

For the purposes of this study, we considered Africans and Europeans as the two founding populations. The disparity in the number and frequency of rare variants between the two populations makes African-Americans well suited for admixture mapping[16]. Equal ancestry contributions results in significantly higher power for

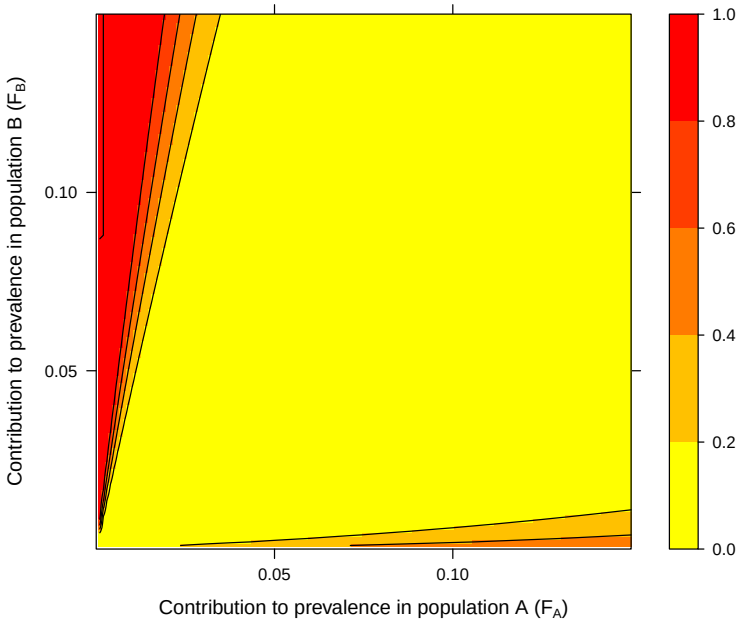
admixed population compared to a 80%-20% split in ancestry. We can find candidate admixed populations with equal contributions from Native Americans and Africans. We expect admixture mapping power to diminish significantly if the two founding populations are not as distant, e.g. Native Americans and Europeans.

Admixture mapping is predicated upon accurately identifying the ancestry of admixture blocks. In this study, we have assumed perfect ancestry estimation. Modern genotyping platforms contain a dense set of markers, with more than a million variants. The cumulative information from these markers can be used to identify admixture blocks and their ancestry accurately [84].

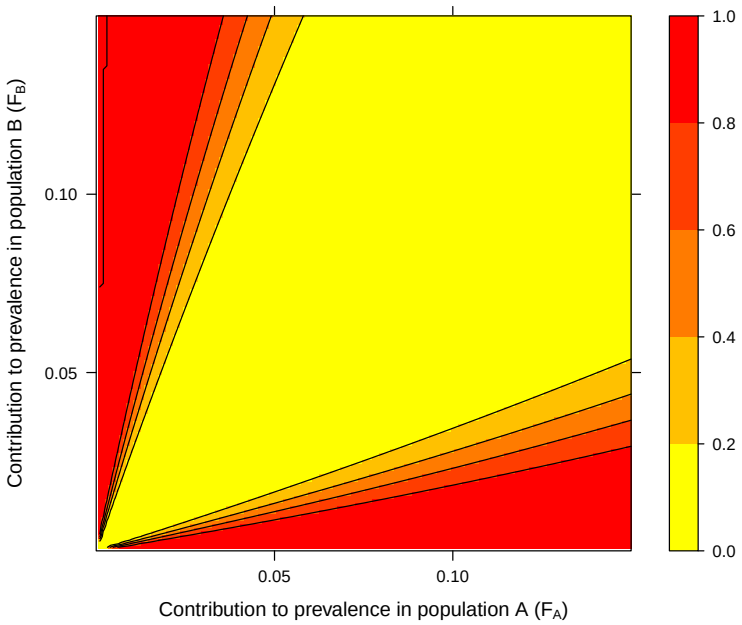
In this study, we have assumed that risk variants are equally likely to be more frequent in either of the two founding populations. Differential selection on the risk variants in the two founding populations can lead to scenarios where this assumption does not hold. Violation of this assumption can lead to a large difference in the cumulative frequency difference of risk variants, thus resulting in different contributions to prevalence in the two founding populations and higher power for admixture mapping.

Several methods have been proposed that test the cumulative burden of carrying multiple rare causal variants across a testing unit. These burden tests have been successful in detecting rare disease predisposing variants. As a follow up to this work, we would compare admixture mapping to burden tests. Admixture mapping has some advantages over the burden tests. The burden tests suffer from drawbacks such as non-robustness to variant misspecification and loss of power in the presence of both risk and protective variants. Admixture mapping combines information across all the markers across an admixture block; hence it does not suffer from the effects of variant misclassification. The presence of protective variants can work favorably

for admixture mapping by further skewing the ratio of contributions to prevalence. Additionally, burden tests are not a genome-wide testing strategy as they rely on accumulating variants across a testing unit such as a gene or an exon, whereas admixture mapping can be used to test across the entire genome. Admixture mapping can be an effective complementary tool to burden tests in the effort to identify rare risk variants.

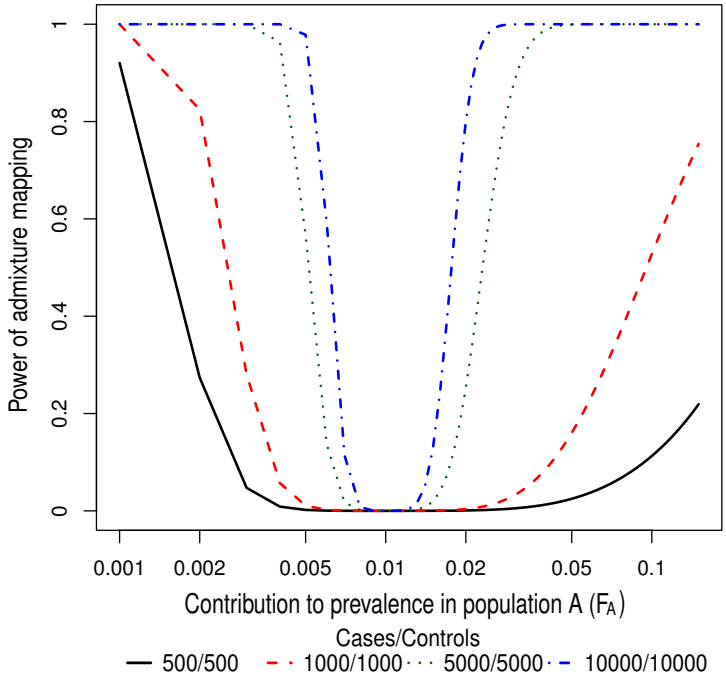


(a) Mixing ratio of 80%-20%

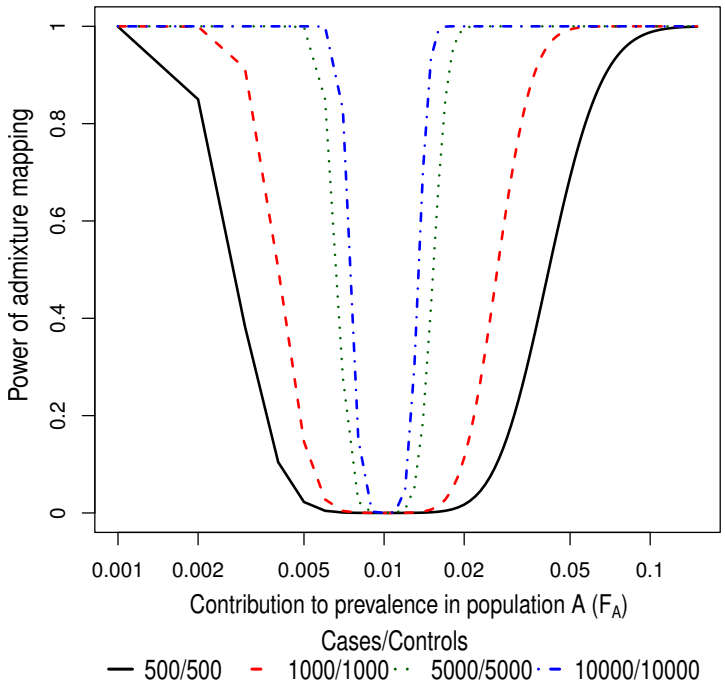


(b) Mixing ratio of 50%-50%

Figure 4.4: Contour plot showing the power of admixture mapping against the contribution to prevalences in the two founding populations. The power of admixture mapping is shown for two different mixing ratios of 80%-20% and 50%-50%.



(a) Mixing ratio of 80%-20%



(b) Mixing ratio of 50%-50%

Figure 4.5: The relationship between power of admixture mapping and contribution to prevalence in population A, plotted for various sample sizes. The contribution to prevalence in the second population was fixed at 1%.

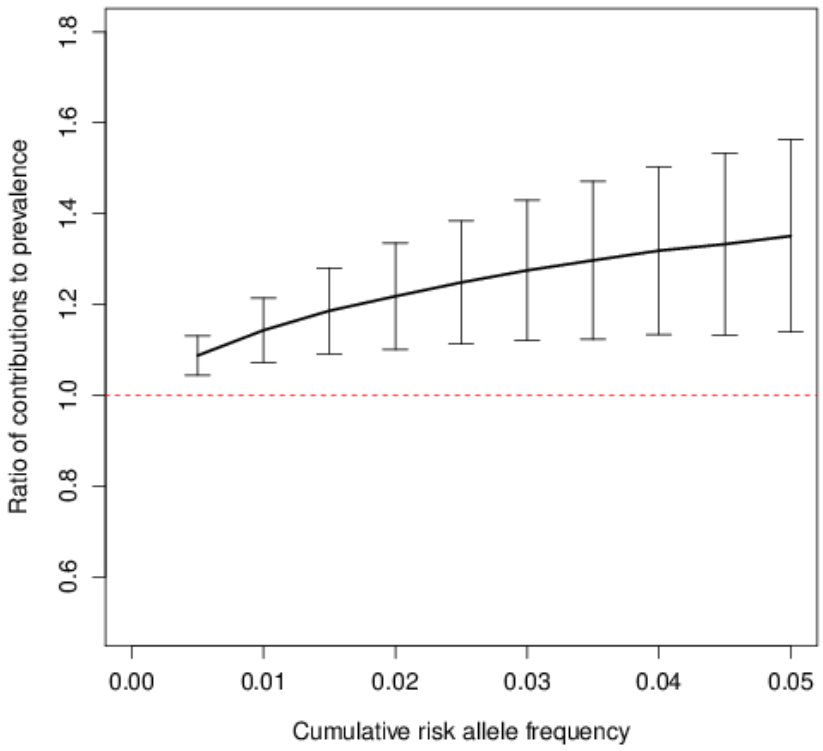


Figure 4.6: Mean ratio of contribution to prevalences in the African population vs the European population plotted against the cumulative risk allele frequency. The contribution to prevalence in Europeans was fixed to 1%.

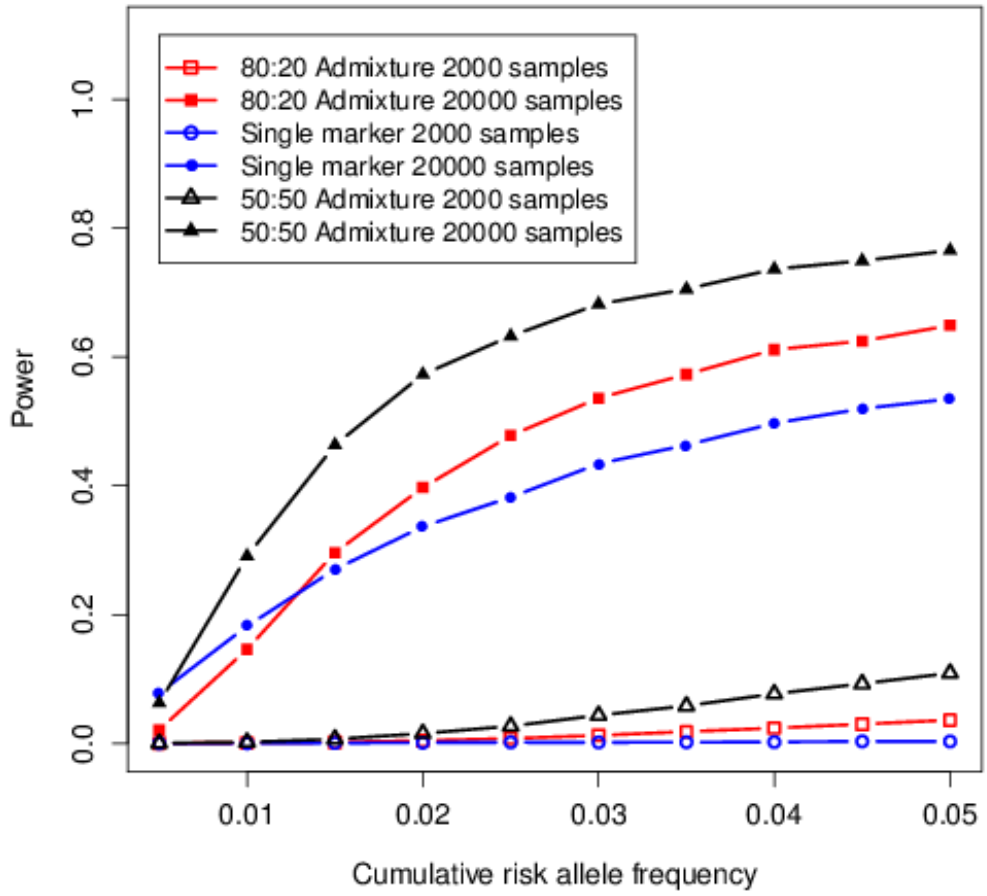


Figure 4.7: Power of admixture mapping compared to the power of single marker tests (blue lines) in Europeans. Two levels of European ancestry were considered, 20% (red lines) and 50% (black lines). The hollow symbols represent tests with 1000 cases and 1000 controls, whereas the filled symbols represent test with 10000 cases and controls each.

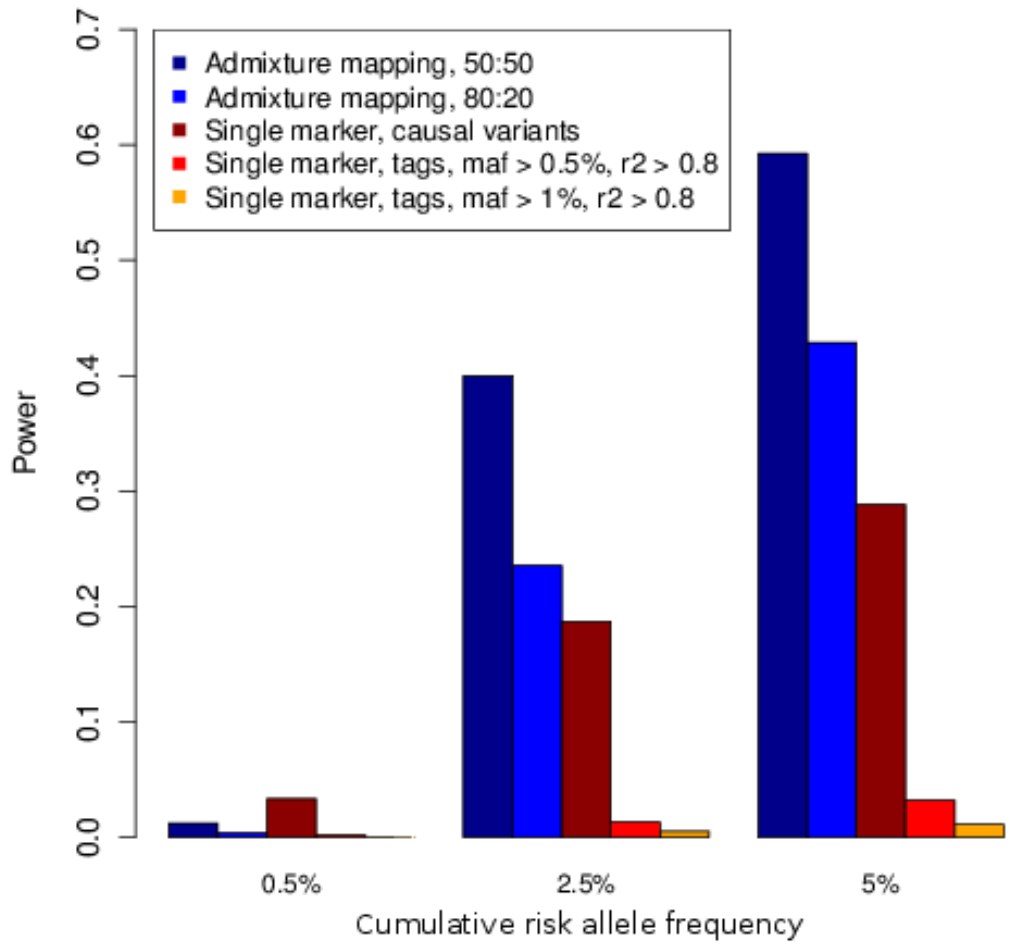


Figure 4.8: Power of admixture mapping compared to indirect association. Sample size was 5000 cases and controls each. TagSNPs were chosen with two minor allele frequency cutoffs, 1% and 0.5%.

CHAPTER V

Estimating site frequency spectra from low coverage sequencing data

5.1 Introduction

The Site Frequency Spectrum (SFS) is an important sample summary statistic in population genetics. The SFS encapsulates the abundance of variations occurring at different frequencies of the derived allele. Many demographic history and molecular evolution parameters can be calculated from the SFS. Estimators of the scaled mutation rate θ, such as Watterson's estimator, θ_W, [85] and the mean pairwise sequence difference, π are functions of the SFS. The SFS also contains information on the demographic history of the population like expansion, shrinkage and recent bottlenecks [86]. Adams and Hudson [87] used the SFS to obtain the maximum likelihood estimates (MLE) of demographic parameters. Statistics that test for deviation from a constant sized Wright-Fisher population with neutral molecular evolution, such as tajima's D [88] and Fay and Wu's H [89], can be derived from the SFS. Fu [90] presented several statistical tests of neutrality based on the SFS. All population genetics statistics that do not use linkage or linkage disequilibrium information can be represented as functions of the SFS.

Before the emergence of large scale sequencing technologies, the SFS was estimated using genotypes obtained using genotyping platforms such as Illumina 1M or the Affymetrix 600K SNP chips. Genotyping arrays do not include all variants present in the region; they are designed using variants identified using a discovery panel. This discovery scheme introduces an ascertainment bias in SFS estimation. Several methods [91] are available to correct for the ascertainment bias, but they depend on knowing the ascertainment scheme *a priori*.

With the advent of low cost high throughput sequencing platforms, SFS can be estimated using short read sequence data. Since short read sequencing queries every nucleotide in the region of interest, estimates of the SFS obtained using sequence data do not suffer from ascertainment bias. The SFS can be computed directly using the genotypes estimated from the short read sequences. The genotypes estimated from the short read sequence data are probabilistic. The uncertainty in genotype estimate is a function of many factors such as alignment fidelity, sequencing depth, i.e. the number of short reads covering the variant position and the error model of the sequencing technology.

Sequencing depth is an important determinant of genotype calling accuracy. It is a study design parameter that can be adjusted based on study goals. The trade off between sample size and sequencing depth has given rise to two competing study designs. The "small sample high coverage" strategy results in a high variant discovery rate with high confidence genotype calls. Under this design, the SFS can be accurately estimated. Conversely, the "large sample low coverage" strategy has low confidence genotype calls with a moderate variant discovery rate. The two study designs attempt to balance the probability of sampling a variant against the probability of discovering it, conditional on it being included in the analysis. Ionita-Laza et al[92, 93] showed

that, for realistic design parameters, the overall probability of discovering a variant is higher under the latter design.

Several methods have been proposed to improve the genotyping accuracy under the "large sample low coverage" design, e.g. borrowing information across samples and using linkage disequilibrium (LD) to improve genotype confidence. The information pooling methods perform well for common variants but exacerbate the genotype uncertainty for low frequency and rare variants. We analyze the performance of three variant calling strategies, viz., individual-based, population-based and LD-aware, with respect to estimating the SFS. All three calling strategies result in an underestimation of the rare end of the SFS. While combining information across samples and markers improves the SFS estimate for allele counts greater than 5, it further biases the estimate of the SFS for rarer variants.

Yi et al [94] presented Bayesian approach to estimate the SFS from short read sequence data. They applied their method to estimate the SFS using 50 sequenced Tibetan exomes. Comparing the SFS estimates between Tibetans and Han Chinese, they identified a variant under selection in the *EPAS1* gene, found to facilitate adaptation to hypoxic environments. Their method is not designed to estimate SFS from low coverage data.

In this work, we pose the SFS estimation problem in a parametric framework. We parameterize the SFS using scaling parameters to account for the change in the relative abundance of rare variants. We use the maximum likelihood estimates (MLE) of the scaling parameters to estimate the SFS. We use a simulation study to compare the performance of our algorithm to the counting estimate of the SFS, obtained by counting across the estimated genotypes. In the simulation study, our method provided a

more accurate estimate of the SFS compared to the counting estimates. Our method resulted in a overestimation in the absolute numbers of rare variants, especially for singletons. We also applied our algorithm to estimate the SFS for a subset of the samples sequenced as part of the quantitative proof of concept (QPOC) [95] study. In the real data example, our method underestimated the SFS in the QPOC data, but it outperformed the counting estimate of the SFS.

5.2 Methods

In this section, we describe the SFS in a mathematical framework and present its expectation under constant sized Wright Fisher population with neutral mutations. We introduce two estimates of the SFS from short read sequence data, viz. the counting estimate based on the estimated genotypes and maximum likelihood estimates (MLE) based on our parameterization of the SFS. Finally, we detail our simulation study to test the performance of our approach.

5.2.1 Site Frequency Spectrum

Consider a sample of n diploid individuals with S sites being analyzed. Let $X = \{\xi_1, \ldots, \xi_{2n-1}\}$ represent the $2n-1$ bins of the SFS. Here, ξ_i counts the number of variant sites that have exactly i derived alleles in the sample, i.e.

$$\xi_i = \sum_{s=1}^{S} \mathbb{I}(D_s = i) \tag{5.1}$$

Here D_s is the total number of derived alleles in the sample at site s. In the absence of information regarding the ancestral and derived allele at each site, we cannot compute the SFS as above. We can instead compute the folded version of the SFS which counts minor alleles instead of derived alleles. Let $X_f = \{\eta_1, \ldots, \eta_n\}$ represent the folded

SFS. If we denote the total number of minor alleles in the sample at site s by M_s, we have

$$\eta_i = \sum_{s=1}^{S} \mathbb{I}(M_s = i) \tag{5.2}$$

For the remainder of this paper, we focus on the folded SFS. We can use coalescent theory for a constant sized Wright-Fisher population with neutral sequence evolution to compute the expected SFS.

$$\mathbb{E}(\eta_i) = \theta \left(\frac{1}{i} + \frac{1}{2n - i} \right), \qquad i = 1, \ldots, n \tag{5.3}$$

where θ is the scaled mutation rate given by $4N\mu$. Here N is the coalescent effective population size and μ is the mutation rate per base per generation. The neutral expected SFS is parameterized by a single parameter θ which controls the overall amplitude of the SFS. The shape of the expected neutral SFS is fixed for a fixed sample size, n, and does not depend on θ. Violations of the assumptions under which the expectation was computed lead to changes in the shape of the SFS. Population growth leads to an excess of rare variants while population decline results in fewer rare variants than expected. In addition, selection pressures can also result in deviation from the expected neutral SFS.

5.2.2 Counting estimate of SFS

We can use a genotype calling algorithm to estimate the genotype for each sample at each site. We can obtain the number of minor alleles observed at each site using the estimated genotypes. In this framework, we assume each variant site to be bi-allelic. We can construct the folded SFS by counting the number of sites with each minor allele count.

$$\hat{\eta}_{i,count} = \sum_{s=1}^{S} \mathbb{I}(\hat{M}_{s,count} = i|\mathbf{C}_s), \qquad i = 1, \ldots, n \tag{5.4}$$

Here, $\hat{\eta}_{i,count}$ is the counting estimate of the i^{th} bin of the SFS, $\hat{M}_{s,count}$ is the counting estimate of the total number of minor alleles at site s. $\mathbf{C}_s$ is the matrix of counts of bases A, C, G, T at site s for each sample, obtained by counting the bases on the short reads aligned to the genomic location of site s.

5.2.3 Parameterizing the SFS

We tackle the SFS estimation in a parametric framework. The natural choice for a parameterized folded SFS is the expected folded SFS under a constant sized Wright Fisher population with neutral sequence evolution, given in eqn. (5.3). We propose an alternative parameterization for the SFS to accommodate the deviations from the neutral SFS caused by violations of assumptions such as constant population size and no selection on variants. We introduce an additional parameter for each bin, z_i to scale the number of variants with i minor alleles in the sample. Table 5.1 shows the expected bin counts under the neutral and our parameterization of the SFS.

Table 5.1: Expected SFS bin counts under neutral and our parameterization

Minor allele count	Neutral model	Our model
$i = 1$	$\theta\left(1 + \frac{1}{2n-1}\right)$	$z_1\theta\left(1 + \frac{1}{2n-1}\right)$
$i = 2$	$\theta\left(\frac{1}{2} + \frac{1}{2n-2}\right)$	$z_2\theta\left(\frac{1}{2} + \frac{1}{2n-2}\right)$
$i \geq 3$	$\theta\left(\frac{1}{i} + \frac{1}{2n-i}\right)$	$z_i\theta\left(\frac{1}{i} + \frac{1}{2n-i}\right)$

5.2.4 Estimating parameters of the SFS

We set up a likelihood framework for the estimation of the parameters of the SFS. Given the alignment of the set of short reads, we can write the likelihood of the parameters, $Z = z_i : 1 \leq i \leq 2n-1$, using the probability of observing $\mathbf{C}_s$, the matrix of counts of bases A, C, G, T at site s. Assuming independence between sites, we can write the likelihood as

$$L(Z, \theta) = \prod_{s=1}^{S} P(\mathbf{C}_s | Z, \theta) \tag{5.5}$$

For each site, we can compute the probability term $P(\mathbf{C}_s|Z, \theta)$ by further conditioning on M_s, the number of minor alleles observed at site s. Leveraging the fact that $\mathbf{C}_s$ depends on Z and θ only through M_s, we get

$$P(\mathbf{C}_s | Z, \theta) = \sum_{j=0}^{n} P(\mathbf{C}_s | M_s = j) P(M_s | Z, \theta) \tag{5.6}$$

In order to ease the computational burden, we make two simplifying assumptions. Firstly, we assume that the counting estimate of the folded SFS is accurate for sites with minor allele counts greater than m, an arbitrary minor allele threshold, i.e. $P(M_s = k | \hat{M}_{s,count} = k, k > m) = 1$. We also assume that m is small enough that we can ignore the probability of finding a homozygote for the minor allele for sites with less than m minor alleles. Using the first assumption we can reduce the summation in (5.6) to m terms by conditioning on $M_s \leq m$. Now, we can rewrite (5.6) and (5.5) as

$$P(\mathbf{C}_s | Z, \theta, M_s \leq m) = \sum_{j=0}^{m} P(\mathbf{C}_s | M_s = j) P(M_s = j | Z, \theta, M_s \leq m) \tag{5.7}$$

$$\tilde{L}(Z, \theta) = \prod_{s \in S_m} \sum_{j=0}^{m} P(\mathbf{C}_s | M_s = j) P(M_s = j | Z, \theta, M_s \leq m) \tag{5.8}$$

where S_m is the set of sites where the counting estimate of the number of minor alleles is $\leq m$.

We can write the probability $P(M_s|Z, \theta, M_s \leq m)$ using our parameterization of the folded SFS. We obtain these probabilities using table 5.1 as

$$P(M_s = j|Z, \theta, M_s \leq m) = \begin{cases} 1 - \sum_{k=1}^{m} P(M_s = k|Z, \theta, M_s \leq m) & \text{if } j = 0 \\ z_i \theta \left(\frac{1}{i} + \frac{1}{2n-i}\right) / |S_m| & \text{if } j \geq 1 \end{cases} \tag{5.9}$$

In order to calculate the probability of observing the count configuration $\mathbf{C}_s$ conditional on minor allele count, M_s, we need to identify the minor allele and the samples that carry the minor allele. Since we cannot ascertain the minor allele or the carrier status for the samples, we integrate them out by summing over all three possible minor alleles and all possible carrier statuses, i.e. all ways of splitting the M_s into the n samples. Here, we leverage the assumption of no homozygotes of the minor allele so that each sample can carry at most one copy of the minor allele. Let $k \in \{A, C, G, T\}$ be the minor allele and $K \in \{A, C, G, T\}$ represent the major allele. Let $T_j = (t_1, \ldots, t_n)$ be the vector of carrier status for each sample with j total minor alleles, where $t_i = 1$ if sample i carries the minor allele and $t_i = 0$ if not.

$$\begin{aligned} P(\mathbf{C}_s|M_s = j) &= \sum_{k \in \{A,C,G,T\}/K} P(k) \sum_{T_j} P(\mathbf{C}_s|T_j) P(T_j) \\ &= \sum_{k \in \{A,C,G,T\}/K} P(k) \sum_{T_j} \left(\prod_{i=1}^{n} P(C_s^i|t_i) \right) P(T_j) \end{aligned} \tag{5.10}$$

The probability of the minor allele, $P(k)$ in the first sum above, is calculated assuming a transition-transversion ratio of 2:1 [96]. If the major allele K and the chosen minor allele k are both purines or pyrimidines, i.e. the mutation that gave rise to the variation was a transition mutation, we set $P(k) = 2/3$. Similarly, if the mutation that gave rise to the variation was a transversion, we set $P(k) = 1/6$. The probability

of each carrier vector T_j is the same and is given by $P(T_j) = \binom{n}{j}^{-1}$, as there are $\binom{n}{j}$ ways of choosing j carriers from n samples. The carrier status and the minor allele completely determine the genotypes of all the samples. Given the genotype of sample i, we can compute the probability of the count vector for sample i, C_s^i, as

$$P(C_s^i|t_i) = \begin{cases} P(C_s^i|G_i = (k, K)) & \text{if } t_i = 1 \\ P(C_s^i|G_i = (K, K)) & \text{if } t_i = 0 \end{cases} \tag{5.11}$$

We have presented the likelihood as the function of genotype likelihoods of the samples at each site. While our method is based on the genotype likelihoods, it is independent of the method or model used to estimate the genotype likelihoods. The genotype likelihoods can be computed using variant calling methods such as *MAQ* [7] and *soapSNP* [97].

Under the complete model with parameters Z and θ, the parameters are unidentifiable. In order to overcome the identifiability issue, we fix θ to be the Watterson's estimator obtained using the counting estimate of the SFS. We further reduce the parameter space by assuming that scaling factors for all bins with greater than three minor alleles are the same, i.e. $z_3 = z_4 = \cdots = z_m$. We obtain the maximum likelihood estimates of Z using Newton-Raphson algorithm to find the maximum of $\tilde{L}(Z, \theta)$.

5.2.5 Simulation

We used a simulation study to quantify the bias in the counting estimate of SFS obtained using various genotype callers. We used *cosi*[98] to generate 1 Mb long haplotypes matched to CEU HapMap [16] on multiple summary statistics. We used a randomly selected 1 Mb region on chromosome 21 as the ancestral state. At the

positions of the derived alleles on the simulated haplotypes, we introduced variant positions with a transition-transversion ratio of 2:1. Subsequently, we selected a pair of haplotypes at random to simulate an individual. We generated short read data for the samples by randomly placing short reads on the haplotypes. We introduced errors with a per base error rate of 0.5%. We simulated a high depth dataset with 30-fold average coverage and a low depth dataset with 4-fold average coverage. For both coverage depths, we generated 10 replicates each for two different sample size settings with 200 and 400 samples each.

We used three types of genotype calling algorithms to obtain the counting estimate of the SFS. We used individual-based, population-based and Linkage Disequilibrium (LD) aware callers on the simulated data to estimate the genotypes at each site for all samples. We used *SOAPSnp* [97], *glfMultiples* [99] and *Thunder* [100] as representative individual, population and LD-aware genotype callers respectively. We filtered the genotype calls using a phred scaled quality threshold of 20. We assumed a major allele homozygote genotype when a genotype call did not pass the quality threshold. We computed the counting estimate of the SFS by counting minor alleles across the filtered genotypes. Using the counting estimates of the SFS, we qualitatively compared the biases in the estimates obtained by using the genotypes from the three different classes of genotype callers.

We also applied our method on the simulated datasets using a minor allele threshold of 10, i.e. we set $m = 10$. We estimated the parameters Z independently in each replicate. We restricted the comparison between the counting estimate and our estimate of the SFS to the first 10 bins, since our method estimates only the first m bins of the SFS.

5.2.6 Validation using QPOC data

We validated our method by using the Glaxo-SmithKline Quantitative Proof of Concept (QPOC)[95] study. We used 219 Iberian samples sequenced as part of the GSK QPOC study. The coding sequence of 208 genes were sequenced in these samples at 25-fold average depth across the target sequence. We thinned the data by sub-sampling the reads for each individual. We generated two sub-sampled datasets, with an average of 4-fold and 10-fold coverage depths. We applied our method to estimate the SFS for the sub-sampled datasets. We compared the estimated SFS to the counting estimate of the SFS obtained from the full (25-fold) dataset.

5.3 Results

First we present the count estimates of the SFS. For all subsequent results, we present only the first 10 bins of the SFS. Fig. 5.1 shows the first 10 bins of the estimated SFS, combined across the 10 replicates, using an individual based caller for a high coverage depth dataset with an average 30-fold coverage across 200 samples. The estimated SFS recapitulates the true SFS well for the rare and less common allele frequencies. Population based and LD aware callers perform similarly.

Figs. 5.2(a)-5.2(c) show the estimated SFS using the three different genotype callers on a low coverage dataset. The individual level caller underestimates the number of singletons and doubletons significantly. In addition, it suffers from underestimation problems at intermediate minor allele counts(between 3 and 10). In comparison, the population level callers perform better at the intermediate minor allele counts, but worse for singletons and doubletons. In particular, the LD aware callers estimate the SFS accurately for minor allele counts greater than 5. This observation validates our assumption about the accuracy of the counting estimate and lets us limit our

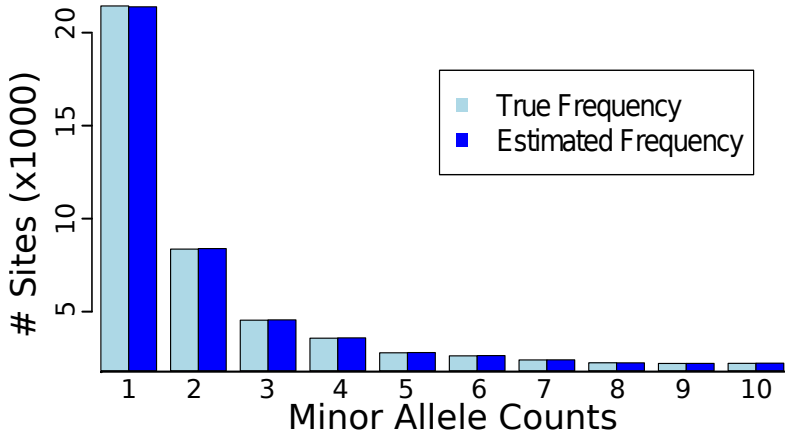


Figure 5.1: Estimated SFS using individual based genotype calls using 200 samples sequenced at 30-fold average depth

estimation procedure to site with minor allele counts ≤ 10.

Figs. 5.3(a) and 5.3(b) show the SFS estimated using our maximum likelihood approach, combined across 10 replicates, with 4-fold average coverage across the region.

The graphs of the true and estimated SFS show that our method is able to recover the site frequency spectrum well. Specifically, we are able to recover the number of singletons and doubletons present in the data. We observe a slight overestimation of the number of singletons, irrespective of the sample size. The mean and standard error of the ratio of true to estimated site frequency for each bin is shown in figure 5.4. For singletons, our estimate is upwardly biased; for minor allele counts greater than one, our estimate is not statistically different from the true SFS.

We applied our method to the Iberian samples present in the QPOC study. Fig.(5.5(a)) shows three different estimates of the SFS, viz., our estimate and the counting estimate of the SFS, both using the 4-fold average coverage data and the "true" SFS estimated by a counting estimate using the 25-fold coverage data. Fig.(5.5(b)) shows the same three estimates of the SFS for the 10-fold average coverage data. The MLE and counting estimate of the SFS underestimate the SFS. The MLE method is

significantly less biased than the counting estimate, for both coverage levels.

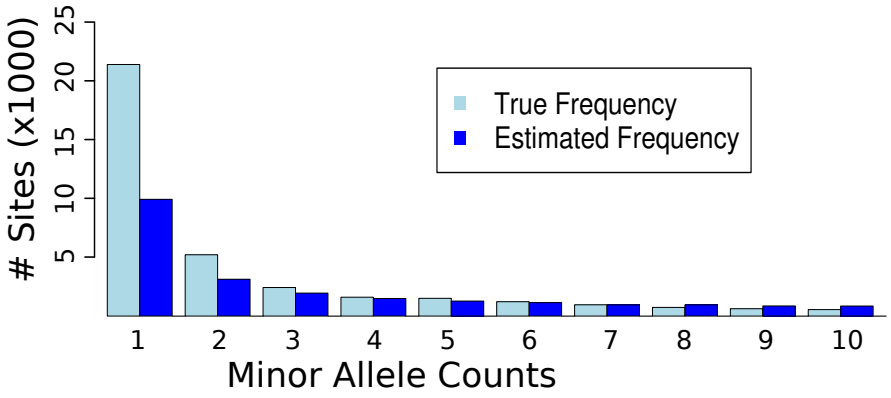
5.4 Discussion

The goal of this project was two-fold. The first aim was to test the estimation of the SFS using genotypes obtained from various genotype calling algorithms. All three genotype calling strategies resulted in an underestimation of the rare part of the SFS. The individual based caller resulted in an uniform underestimation over the entire range of the SFS. Population and LD-aware callers recovered the SFS accurate for low frequency variants, but performed significantly worse for the rarest variants. Since information for the rarest variants in the sample is limited, using population or LD based genotype calling results in rare variants being misclassified as errors. For example, singletons occur in exactly one individual in the sample; combining information across samples does not improve the genotyping accuracy for singletons. The lack of evidence from other individuals leads to further loss in quality of the genotype. LD-aware callers suffer from a similar problem. As singletons do not occur on a shared haplotype, the LD-aware callers are unable to incorporate information from neighboring markers leading to a reduction in genotype confidence.

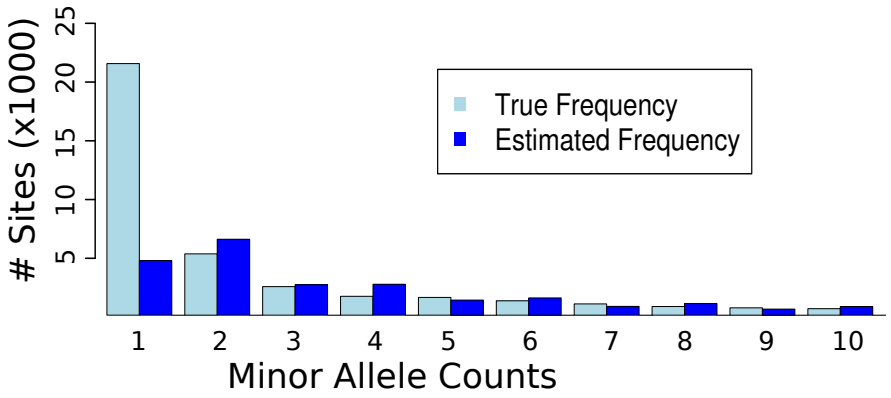
The second aim of the project was to devise a maximum likelihood based method to estimate the SFS from low coverage short read sequence data. In simulation studies, our method successfully recaptures the true SFS for low coverage short read datasets. Although our method overestimates the true number of singletons, it accurately estimates the rare part of the SFS for sites with minor allele counts greater than 1. In this study, we have limited the estimation to a one-dimensional frequency spectrum. The estimation procedure can be extended to multiple populations. While the estimation framework remains unchanged, much larger sample sizes would be required to overcome the instability in the estimation due to the increase in the number of parameters.

Our approach amalgamates the advantages of LD-aware calling with MLE estimation. LD aware callers borrow information from adjoining markers to refine genotype calling at each marker, thus accurately estimating the genotypes at markers with more than 5 minor allele copies in the sample. We leverage the accuracy of these genotype calls to limit our SFS estimation to sites with minor allele counts less than 10, vastly improving our computation time.

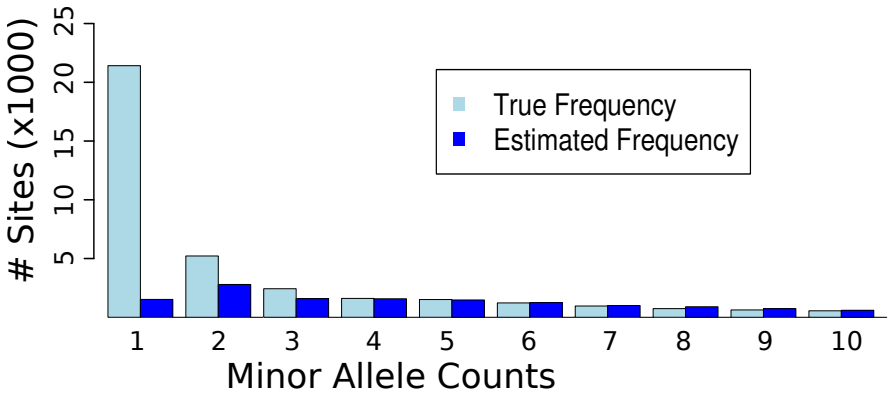
We have presented a maximum likelihood framework for SFS estimation which integrates out the inherent uncertainty present in low coverage sequencing datasets. Although we have used this framework only to estimate the SFS, it can be extended to estimate any function of the genotypes, such as heterozygosity, inbreeding coefficient etc. Our method provides a statistical framework to account for the uncertainty in low coverage sequence data. With an increasing number of low coverage medical sequencing studies underway, it can be used to augment the tools currently used to analyze rare variants in the sample.



(a) SOAPSnp: Individual-based genotype caller

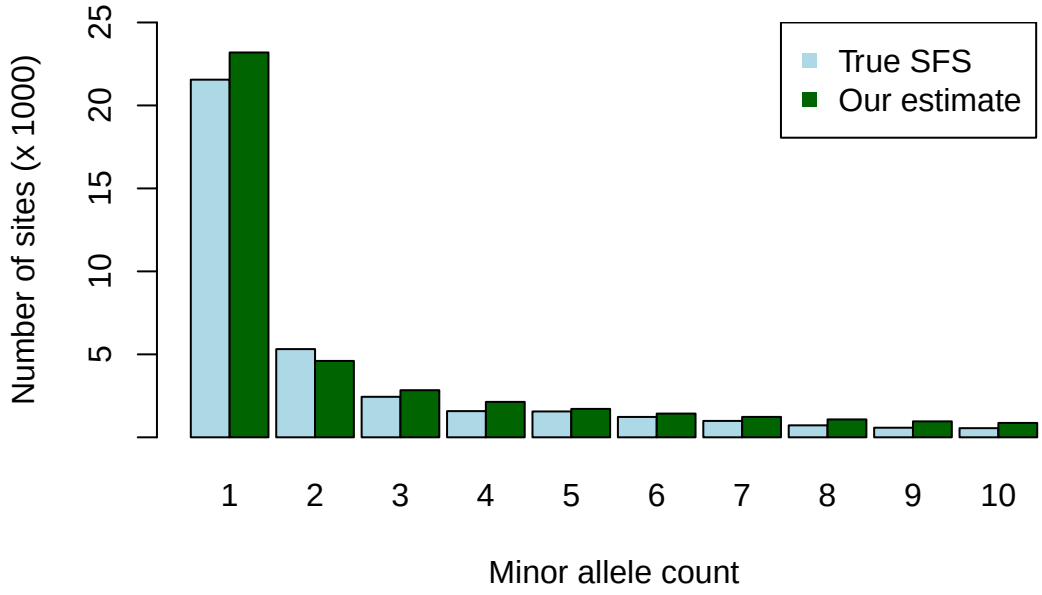


(b) glfMultiple: Population-based genotype caller

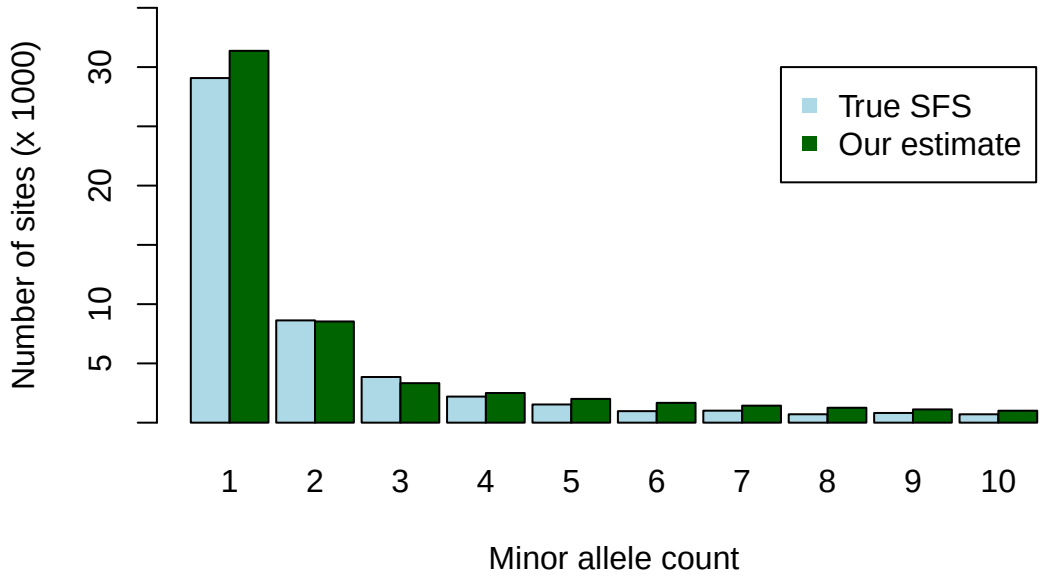


(c) Thunder: LD-aware genotype caller

Figure 5.2: Estimated SFS for low pass, 4-fold, short read sequencing data. The panels show the first 10 bins of the estimated SFS using genotypes from (a) an individual level caller, (b) population level caller and (c) population level LD aware caller.



(a) Our estimate of SFS with 200 samples



(b) Our estimate of SFS with 400 samples

Figure 5.3: MLE estimate of the SFS using simulated data

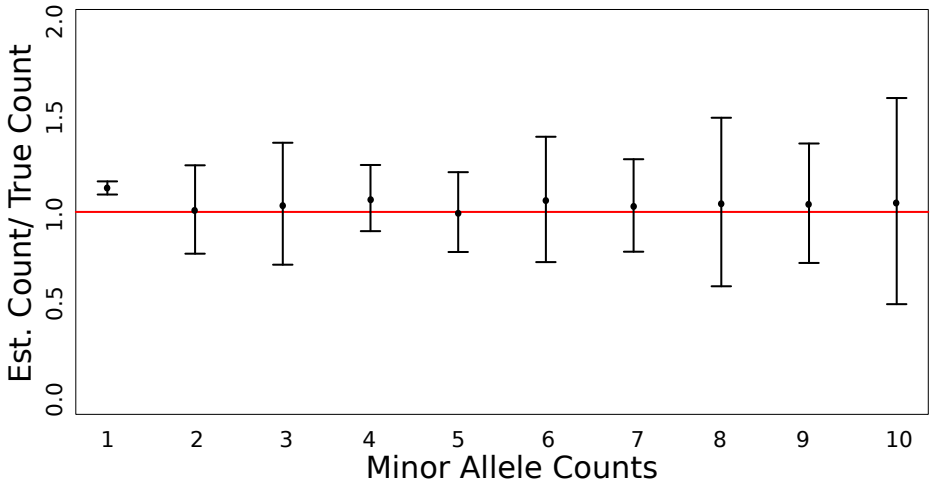
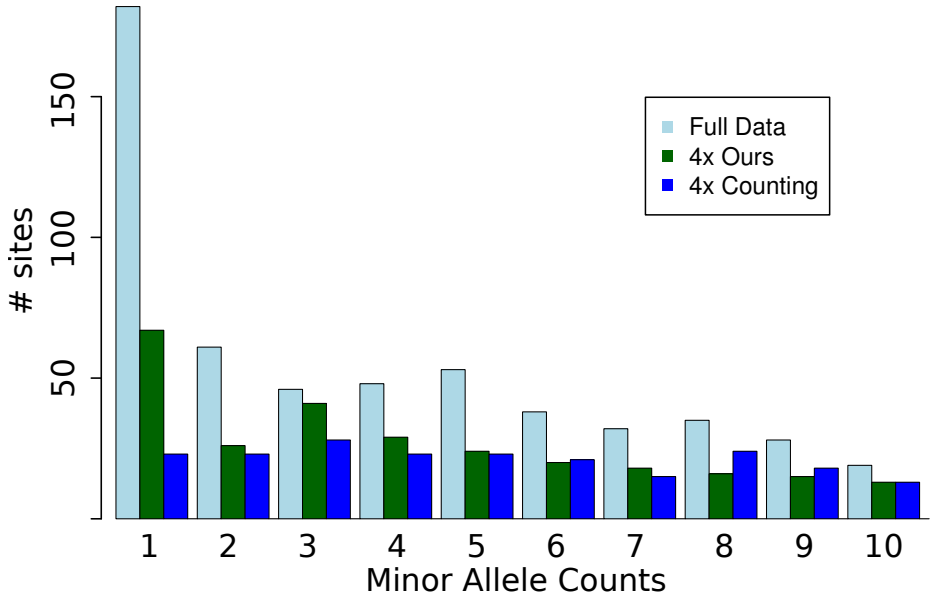
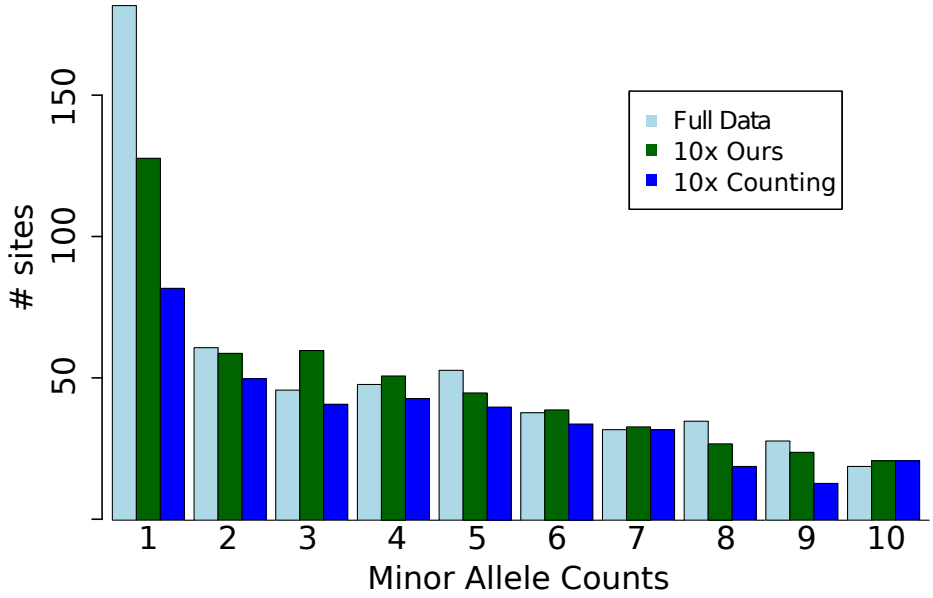


Figure 5.4: Mean and standard error of the ratio of true to estimated SFS for each bin



(a) SFS estimate on QPOC data: 4 fold coverage



(b) SFS estimate on QPOC data: 10 fold coverage

Figure 5.5: QPOC data: Comparison of MLE and counting estimate of SFS

CHAPTER VI

Conclusion

My dissertation focused on developing novel statistical methods for a wide array of problems in the fields of disease and population genetics.

In chapter 2, I developed a graph based algorithm to select the smallest set of tagging SNPs, based on the linkage disequilibrium measure r^2. We used a "divide and conquer" approach to break the tagSNP selection problem in the region to an easier problem of selecting tagSNPs in much smaller clusters. We applied our method to select tagSNP for chromosome 2 in HapMap. Compared to the greedy algorithm for tagSNP selection, our method required 5% fewer tagSNPs at an r^2 threshold of 0.5 and 2% fewer tagSNPs at an r^2 threshold of 0.8.

Since the LD measure r^2 is directly related to the power of indirect association,we have used r^2 to define the ability of one marker to act as a proxy for another. Our method can easily be applied to any other measure for connectedness, such as D'.

Our method identifies equivalent sets of tagSNPs for each precinct. A potential improvement would be to further optimize the tagSNP set based on an external cost function, such as genotyping error rate for each marker, based on the neighboring

sequence. From a computation standpoint, the search for tagSNPs in the disjoint clusters are independent processes and can be carried out concurrently. A parallel computing approach would further alleviate the computational burden.

We provided additional constraints, such as inclusion/exclusion of selected variants in the tagSNP set, minor allele frequency threshold for tagSNPs, minimum distance between tagSNPs and robust coverage using two tagSNPs to cover each variant. We have implemented our method in a publicly available software package.

In chapter 3, I developed a Gibbs sampling algorithm for identifying the true genomic location of multiply mapped reads. We used a simulation study to test the performance of our algorithm. In a chromosome wide simulation study, our method placed $\sim$87% of multiply mapped reads to their true location. Adding re-aligned multiply mapped reads led to an additional 3% variants being discovered.

The same algorithm that was used for identifying a unique alignment for each multiply mapped read can provide the posterior distribution of their alignments. Obtaining multiple samples from the posterior distribution would allow robust quantitative analysis.

In this project, we have focused on single end reads. An extension to our algorithm would be to incorporate paired end reads. The underlying framework would remain unchanged, since we model the counts of bases at location in the reference region and this is unaffected by the sequence study design. A simple solution to accommodate paired end reads would be to adjust the prior to reflect the distance between the read pair.

In chapter 4, we explore the feasibility of admixture mapping to identify rare disease susceptibility variants. We used a simulation study to test the power of admixture mapping to identify regions harboring rare susceptibility loci. We simulated African-American case-control data under settings that result in low power for single marker association tests. For cumulative risk allele frequencies greater than 1%, admixture mapping had considerable power to detect association of the region to disease.

In this work, we focused on Africans and Europeans as the founding populations for the admixed population due to the high divergence between the two populations. We would extend the study to include Native American populations. Recent studies suggest a strong bottleneck event in the Native American founding population before the peopling of the Americas. Combined with the serial founder effect, it might result in a large difference in the number of rare variants.

There are many admixture mapping studies underway in African-American and Latino populations. Most of these studies consist of relatively small sample sizes, or the order of 1000 cases and controls each. Our simulation studies suggest that such small sample sizes are inadequate to detect the presence of rare susceptibility loci.

The admixture and marker association signals provide orthogonal information. Tang et al [101] proposed a framework to combine the two signals into a single test of association. I would like to explore this idea further with respect to burden tests. Combining the admixture association signal with the burden test might improve the power to detect associated regions. Conversely, since both, burden tests and admixture mapping, collapse information across the region, they may not represent independent signals.

In chapter 5, we presented a maximum likelihood estimate of the site frequency spectrum(SFS) obtained from low coverage short read sequence data. The counting estimate of the SFS obtained from estimated genotypes from all three callers, viz., individual based, population based and LD-aware, significantly underestimate the number of singletons and doubletons in the sample. In simulation studies, our method overestimates the rare part of the SFS. The upward bias in our estimate of the SFS is much smaller than the downward bias observed in the counting estimates. We applied our method to estimate the SFS in Iberian samples from the QPOC study. Both our method and the counting estimate underestimated the SFS.

The upward bias in our estimate of the SFS can lead to biased estimates of various functions of the SFS such as tajima's D, Fay and Wu's H and estimates of the mutation rate, θ. In addition, it can lead to incorrect inferences about demographic history and selection parameters. As a follow up to this project, I would like to quantify the effects of the bias in SFS on these population genetics estimates. The SFS estimation framework can be extended to other functions of the genotypes, such as heterozygosity, inbreeding coefficient, two-population SFS etc.

In summary, in this dissertation, I have addressed several methodological issues in a wide array of problem in statistical and population genetics. It is my hope that the methods elaborated in this dissertation provide statistical tools in our efforts to gain a better understanding of genomic architecture.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 1996.

[2] L. J. Scott, K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li, W. L. Duren, M. R. Erdos, H. M. Stringham, P. S. Chines, A. U. Jackson, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 316(5829):1341–1345, 2007.

[3] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23):9362–7, 2009. ISSN 10916490. doi:10.1073/pnas.0903103106.

[4] B. Maher. The case of the missing heritability. *Nature*, 456(November):18–21, 2008. ISSN 14764687. doi:10.1057/palgrave.pcs.2100148.

[5] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. Mccarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009. ISSN 00280836. doi:10.1038/nature08494.Finding.

[6] J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, 69(1):1–14, 2001.

[7] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851–8, November 2008. ISSN 1088-9051. doi:10.1101/gr.078212.108.

[8] M. Zawistowski, S. Gopalakrishnan, J. Ding, Y. Li, S. Grimm, and S. Zöllner. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *American journal of human genetics*, 87(5):604–17, November 2010. ISSN 1537-6605. doi:10.1016/j.ajhg.2010.10.012.

[9] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–9, November 2008. ISSN 1476-4687. doi:10.1038/nature07517.

[10] R. M. Durbin, D. L. Altshuler, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, F. S. Collins, F. M. De La Vega, P. Donnelly, M. Egholm, et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010. ISSN 00280836. doi:10.1038/nature09534.

[11] J. C. Cohen, R. S. Kiss, A. Pertsemlidis, Y. L. Marcel, R. McPherson, and H. H. Hobbs. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science (New York, N.Y.)*, 305(5685):869–72, August 2004. ISSN 1095-9203. doi:10.1126/science.1099870.

[12] J. C. Cohen, E. Boerwinkle, T. H. Mosley, and H. H. Hobbs. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *The New England journal of medicine*, 354(12):1264–72, March 2006. ISSN 1533-4406. doi:10.1056/NEJMoa054013.

[13] S. Anders and W. Huber. Differential expression analysis for sequence count

data. *Genome biology*, 11(10):R106, January 2010. ISSN 1465-6914. doi: 10.1186/gb-2010-11-10-r106.

[14] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, 2007.

[15] X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science (New York, N.Y.)*, 329(5987):75–8, July 2010. ISSN 1095-9203. doi:10.1126/science.1190371.

[16] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, 2005. ISSN 14764687. doi: 10.1038/nature04226.

[17] F. S. Collins, M. S. Guyer, and A. Charkravarti. Variations on a theme: cataloging human DNA sequence variation. *Science*, 278(5343):1580–1581, 1997.

[18] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–933, 2001.

[19] G. C. L. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, et al. Haplotype tagging for the identification of common disease genes. *Nature Genetics*, 29(2):233–237, 2001.

[20] L. R. Cardon and G. R. Abecasis. Using haplotype blocks to map human complex trait loci. *Trends in Genetics*, 19(3):135–140, 2003.

[21] N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–1723, 2001.

[22] K. Zhang, M. Deng, T. Chen, M. S. Waterman, and F. Sun. A novel efficient dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences of the United States of America*, 267(2):7335–7339, 2002.

[23] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, et al. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229, 2002. ISSN 10959203.

[24] Z. Meng, D. V. Zaykin, C.-F. Xu, M. Wagner, and M. G. Ehm. Selection of Genetic Markers for Association Analyses, Using Linkage Disequilibrium and Haplotypes. *The American Journal of Human Genetics*, 73(1):115–130, 2003.

[25] P. Sebastiani, R. Lazarus, S. T. Weiss, L. M. Kunkel, I. S. Kohane, and M. F. Ramoni. Minimal haplotype tagging. *Proceedings of the National Academy of Sciences of the United States of America*, 100(17):9900–9905, 2003.

[26] H. I. Avi-Itzhak, X. Su, and F. M. De La Vega. Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity. *Pacific Symposium On Biocomputing*, pages 466–77, 2003. ISSN 17935091.

[27] X. Ke and L. R. Cardon. Efficient selective screening of haplotype tag SNPs. *Bioinformatics*, 19(2):287–288, 2003.

[28] D. B. Goldstein, K. R. Ahmadi, M. E. Weale, and N. W. Wood. Genome scans

and candidate gene approaches in the study of common diseases and variable drug responses. *Trends in Genetics*, 19(11):615–622, 2003.

[29] D. O. Stram, C. A. Haiman, J. N. Hirschhorn, D. Altshuler, L. N. Kolonel, B. E. Henderson, and M. C. Pike. Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Human Heredity*, 55(1):27–36, 2003.

[30] J. Hampe, S. Schreiber, and M. Krawczak. Entropy-based SNP selection for genetic association studies. *Human Genetics*, 114(1):36–43, 2003. ISSN 03406717. doi:10.1007/s00439-003-1017-2.

[31] J. M. Chapman, J. D. Cooper, J. A. Todd, and D. G. Clayton. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Human Heredity*, 56(1-3):18–31, 2003.

[32] Z. Lin and R. B. Altman. Finding Haplotype Tagging SNPs by Use of Principal Components Analysis. *The American Journal of Human Genetics*, 75(5):850–861, 2004.

[33] B. V. Halldórsson, V. Bafna, R. Lippert, R. Schwartz, F. M. De La Vega, A. G. Clark, and S. Istrail. Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. *Genome Research*, 14(8):1633–1640, 2004.

[34] A. Rinaldo, S.-A. Bacanu, B. Devlin, V. Sonpar, L. Wasserman, and K. Roeder. Characterization of multilocus linkage disequilibrium. *Genetic Epidemiology*, 28(3):193–206, 2005.

[35] K. Zhang and L. Jin. HaploBlockFinder: haplotype block analyses. *Bioinformatics*, 19(10):1300–1301, 2003.

[36] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *The American Journal of Human Genetics*, 74(1):106–120, 2004.

[37] W. G. Hill. Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, 33(2):229–239, 1974.

[38] B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2):311–22, 1995. ISSN 08887543. doi: 10.1006/geno.1995.9003.

[39] W. G. Hill and A. Robertson. The effects of inbreeding at loci with heterozygote advantage. *Genetics*, 60(3):615–628, 1968.

[40] D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, et al. Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204, 2001.

[41] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29(2):229–232, 2001.

[42] A. J. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, 29(2):217–222, 2001.

[43] E. Dawson, G. R. Abecasis, S. Bumpstead, Y. Chen, S. Hunt, D. M. Beare, J. Pabial, T. Dibling, E. Tinsley, S. Kirby, et al. A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, 418(6897):544–548, 2002.

[44] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms 2nd Edition*. MIT Press, 2001.

[45] T. Niu, Z. S. Qin, X. Xu, and J. S. Liu. Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms. *The American Journal of Human Genetics*, 70(1):157–169, 2002.

[46] D. O. Stram. Software for tag single nucleotide polymorphism selection. *Human Genomics*, 2(2):144–151, 2005.

[47] P. I. W. De Bakker, R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly, and D. Altshuler. Efficiency and power in genetic association studies. *Nature Genetics*, 37(11):1217–1223, 2005.

[48] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, 42(1):30–5, January 2010. ISSN 1546-1718. doi:10.1038/ng.499.

[49] J. K. Pickrell, J. C. Marioni, A. a. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–72, April 2010. ISSN 1476-4687. doi: 10.1038/nature08872.

[50] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics (Oxford, England)*, 26(4):493–500, February 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp692.

[51] J. Rainer. Identification of GC receptor binding sites in lung carcinoma cells employing ChIPseq technology. *Quest*, 2010.

[52] H. Li and R. Durbin. Fast and accurate short read alignment with BurrowsWheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[53] R. Li, Y. Li, K. Kristiansen, and J. Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics (Oxford, England)*, 24(5):713–4, March 2008. ISSN 1367-4811. doi:10.1093/bioinformatics/btn025.

[54] M. Taub, D. Lipson, and T. P. Speed. Methods for allocating ambiguous short-reads. 10(2):69–82, 2010.

[55] A. Mortazavi, B. A. Williams, K. Mccue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):1–8, 2008. doi:10.1038/NMETH.1226.

[56] G. J. Faulkner, A. R. R. Forrest, A. M. Chalk, K. Schroder, Y. Hayashizaki, P. Carninci, D. A. Hume, and S. M. Grimmond. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics*, 91(3):281–288, 2008.

[57] Y. Ji, Y. Xu, Q. Zhang, K.-W. Tsui, Y. Yuan, C. Norris Jr, S. Liang, and H. Liang. BM-Map: Bayesian Mapping of Multireads for Next-Generation Sequencing Data. *Biometrics*, 1(ii), April 2011. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2011.01605.x.

[58] R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[59] H. M. Kang and G. R. Abecasis. GlfMultiples, 2010.

[60] D. E. Reich and E. S. Lander. On the allelic spectrum of human disease. *Trends in genetics : TIG*, 17(9):502–10, September 2001. ISSN 0168-9525.

[61] E. N. Smith, W. Chen, M. Kähönen, J. Kettunen, T. Lehtimäki, L. Peltonen, O. T. Raitakari, R. M. Salem, N. J. Schork, M. Shaw, et al. Longitudinal Genome-Wide Association of Cardiovascular Disease Risk Factors in the Bogalusa Heart Study. *PLoS Genetics*, 6(9):11, 2010.

[62] W. Satake, Y. Nakabayashi, I. Mizuta, Y. Hirota, C. Ito, M. Kubo, T. Kawaguchi, T. Tsunoda, M. Watanabe, A. Takeda, et al. Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nature Genetics*, 41(12):1303–7, 2009. ISSN 15461718. doi:10.1038/ng.485.

[63] J. Simón-Sánchez, C. Schulte, J. M. Bras, M. Sharma, J. R. Gibbs, D. Berg, C. Paisan-Ruiz, P. Lichtner, S. W. Scholz, D. G. Hernandez, et al. Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nature Genetics*, 41(12):1308–1312, 2009.

[64] I. P. M. Tomlinson, E. Webb, L. Carvajal-Carmona, P. Broderick, K. Howarth, A. M. Pittman, S. Spain, S. Lubbe, A. Walther, K. Sullivan, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nature Genetics*, 40(5):623–630, 2008.

[65] A. Tenesa, S. M. Farrington, J. G. D. Prendergast, M. E. Porteous, M. Walker, N. Haq, R. A. Barnetson, E. Theodoratou, R. Cetnarskyj, N. Cartwright, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nature Genetics*, 40(5):631–637, 2008. ISSN 15461718. doi:10.1038/ng.133.

[66] R. Saxena, B. F. Voight, V. Lyssenko, N. P. Burtt, P. I. W. de Bakker, H. Chen, J. J. Roix, S. Kathiresan, J. N. Hirschhorn, M. J. Daly, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.

Science (New York, N.Y.), 316(5829):1331–6, June 2007. ISSN 1095-9203. doi: 10.1126/science.1142358.

[67] S. Kathiresan, O. Melander, C. Guiducci, A. Surti, N. P. Burtt, M. J. Rieder, G. M. Cooper, C. Roos, B. F. Voight, A. S. Havulinna, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature genetics*, 40(2):189–97, February 2008. ISSN 1546-1718. doi:10.1038/ng.75.

[68] E. K. Speliotes, C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson, and et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, 42(11):937–948, 2010.

[69] C. J. Willer, E. K. Speliotes, R. J. F. Loos, S. Li, C. M. Lindgren, I. M. Heid, S. I. Berndt, A. L. Elliott, A. U. Jackson, C. Lamina, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature genetics*, 41(1):25–34, January 2009. ISSN 1546-1718. doi: 10.1038/ng.287.

[70] J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? *American journal of human genetics*, 69(1):124–37, July 2001. ISSN 0002-9297. doi:10.1086/321272.

[71] B. E. Madsen and S. R. Browning. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genetics*, 5(2):11, 2009.

[72] 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*, 7(3):e1001322, 2011. ISSN 15537404. doi:10.1371/journal.pgen.1001322.

[73] A. L. Price, G. V. Kryukov, P. I. W. de Bakker, S. M. Purcell, J. Staples, L.-J. Wei, and S. R. Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *American journal of human genetics*, 86(6):832–8, June 2010. ISSN 1537-6605. doi:10.1016/j.ajhg.2010.04.005.

[74] 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. *American journal of human genetics*, July 2011. ISSN 1537-6605. doi: 10.1016/j.ajhg.2011.05.029.

[75] J. C. Cohen, A. Pertsemlidis, S. Fahmi, S. Esmail, G. L. Vega, S. M. Grundy, and H. H. Hobbs. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proceedings of the National Academy of Sciences of the United States of America*, 103(6):1810–5, February 2006. ISSN 0027-8424. doi:10.1073/pnas.0508483103.

[76] G. Montana and J. K. Pritchard. Statistical Tests for Admixture Mapping with Case-Control and Cases-Only Data. *The American Journal of Human Genetics*, 75(5):771–789, 2004.

[77] W. H. L. Kao, M. J. Klag, L. A. Meoni, D. Reich, Y. Berthier-Schaad, M. Li, J. Coresh, N. Patterson, A. Tandon, N. R. Powe, et al. MYH9 is associated with nondiabetic end-stage renal disease in African Americans. *Nature Genetics*, 40(10):1185–1192, 2008.

[78] C. A. Winkler, G. W. Nelson, and M. W. Smith. Admixture mapping comes of age. *Annual Review of Genomics and Human Genetics*, 11(June):65–89, 2010.

[79] S. Sankararaman, G. Kimmel, E. Halperin, and M. I. Jordan. On the inference of ancestries in admixed populations. *Genome Research*, 18(4):668–675, 2008.

[80] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000. ISSN 00166731.

[81] V. Plagnol and J. D. Wall. Possible Ancestral Structure in Human Populations. *PLoS Genetics*, 2(7):8, 2006.

[82] X. Zhu, H. Tang, and N. Risch. Admixture mapping and the role of population structure for localizing disease genes. *Advances in genetics*, 60(07):547–69, January 2008. ISSN 0065-2660. doi:10.1016/S0065-2660(07)00419-1.

[83] G. Kang, G. Gao, S. Shete, D. T. Redden, B.-L. Chang, T. R. Rebbeck, J. S. Barnholtz-Sloan, N. M. Pajewski, and D. B. Allison. Capitalizing on Admixture in Genome-Wide Association Studies: A Two-Stage Testing Procedure and Application to Height in African-Americans. *Frontiers in Genetics*, 2(March):1–16, 2011. ISSN 16648021. doi:10.3389/fgene.2011.00011.

[84] N. A. Rosenberg, L. M. Li, R. Ward, and J. K. Pritchard. Informativeness of genetic markers for inference of ancestry. *American journal of human genetics*, 73(6):1402–22, December 2003. ISSN 0002-9297. doi:10.1086/380416.

[85] G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276, 1975. ISSN 00405809.

[86] G. T. Marth, E. Czabarka, J. Murvai, and S. T. Sherry. The Allele Frequency Spectrum in Genome-Wide Human Variation Three Large World Populations. *Evaluation*, 372(January):351–372, 2004.

[87] A. M. Adams and R. R. Hudson. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide

polymorphisms. *Genetics*, 168(3):1699–712, November 2004. ISSN 0016-6731. doi:10.1534/genetics.104.030171.

[88] F. Tajima. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, 123(3):585–595, November 1989.

[89] J. C. Fay and C. I. Wu. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):1405–13, July 2000. ISSN 0016-6731.

[90] Y. Fu. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, 147(2):915, 1997.

[91] R. Nielsen, M. J. Hubisz, and A. G. Clark. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics*, 168(4):2373–82, December 2004. ISSN 0016-6731. doi: 10.1534/genetics.104.031039.

[92] I. Ionita-Laza, C. Lange, and N. M Laird. Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 106(13):5008–13, March 2009. ISSN 1091-6490. doi:10.1073/pnas.0807815106.

[93] I. Ionita-Laza and N. M. Laird. On the optimal design of genetic variant discovery studies. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article33, 2010.

[94] X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science (New York, N.Y.)*, 329(5987):75–8, July 2010. ISSN 1095-9203. doi:10.1126/science.1190371.

[95] M. R. Nelson, M. G. Ehm, D. Wegmann, P. St. Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. Bacanu, D. Fraser, L. Warren, et al. Sequencing 202 drug target genes in 14002 people finds excess of rare deleterious variants. *In Preparation*, 2011.

[96] D. W. Collins and T. H. Jukes. Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics*, 20(3):386–96, April 1994. ISSN 0888-7543. doi:10.1006/geno.1994.1192.

[97] R. Li, Y. Li, X. Fang, H. Yang, J. Wang, K. Kristiansen, and J. Wang. SNP detection for massively parallel whole-genome resequencing. *Genome Research*, 19(6):1124–1132, 2009.

[98] S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, pages 1576–1583, 2005. doi:10.1101/gr.3709305.

[99] S. J. Kang, C. W. K. Chiang, C. D. Palmer, B. O. Tayo, G. Lettre, J. L. Butler, R. Hackett, A. A. Adeyemo, C. Guiducci, I. Berzins, et al. Genome wide association of anthropometric traits in African and African derived populations. *Human molecular genetics*, 19(13):2725–38, April 2010. ISSN 1460-2083. doi: 10.1093/hmg/ddq154.

[100] Y. Li, C. Sidore, H. Kang, M. Boehnke, and G. Abecasis. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Research*, 6:940–51, June 2011.

[101] H. Tang, D. O. Siegmund, N. A. Johnson, I. Romieu, and S. J. London. Joint testing of genotype and ancestry association in admixed families. *Genetic Epidemiology*, 34(8):783–791, 2010.