Title

Methods for the analysis of human genetic variation in the search for the genetic basis of human disease

Permalink

Author

Zaitlen, Noah

Publication Date

2009

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Methods for the Analysis of Human Genetic Variation in the Search for the Genetic Basis of Human Disease

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics

by

Noah Zaitlen

Committee in charge:

University of California, San Diego

Professor Vineet Bafna, Chair
Professor Daniel O'Connor
Professor Pavel Pevzner
Professor Nicholas Schork

University of California, Los Angeles

Professor Eleazar Eskin

2009

The dissertation of Noah Zaitlen is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____

____

Chair

University of California, San Diego

2009

DEDICATION

To my friends and family for their enduring love and support.

EPIGRAPH

In the field of observation,
chance favors only the prepared mind.
—Louis Pasteur

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Thanks are owed to many fantastically awesome people. First and foremost, my advisor Eleazar Eskin for years of sharing insights, providing guidance, and giving encouragement when I was stuck. I have been extremely lucky to be surrounded by great friends and colleagues in my lab Hyun Min Kang, Sean O'Rourke, Buhm Han, Jimmie Ye, and the rest of Zarlab. We've had countless scientific and "civilian" conversations, travelled to far flung conferences, and been an all around great crew. I hope we get the chance to work together in the future. My friends from the bioinformatics program at UCSD Juan Lorenzo Rodriguez Flores, Mark Chaisson, and Ali Bashir with whom I've shared many good times surfing, climbing, bullshitting, drinking beers, and sharing ideas. Rachael Cleghorn for teaching me that everything will be ok, and for my sanity, occasionally at the expense of her own. Fantastic collaborators Eran Halperin, David Heckerman, and Nebojsa Jojic. If I am lucky we will get to work together again soon. My teachers and committee members Vineet Bafna, Pavel Pevzner, Nicholas Schork, and Dan O'Connor for creating a phenomenal program at UCSD and sharing their time and knowledge with the next generation. Sasha Tchir for her support, encouragement, nourishment, and much much more. Damien Clark for the use of his "office" where much of this was written, and also his leftovers. Miriam Udler for last minute freak-outs and other great conversations. Most importantly, my family Nan, Richard, Ben, and Zach for their love and faith.

I was supported by NSF Grant Number 0513612 and NIH Grant Number 1K25HL080079 and the Microsoft Research Fellowship. Thanks to NBCR and FWGRID for letting me hammer their cluster for many cpu hours. Special thanks to all the folks developing and maintaining all the software I use, Emacs, gcc, gdb, Perl, octave, R, latex, and many others.

Chapter 2, was published in The American Journal of Human Genetics, Vol 80, pp 683-91, 2007. Noah A. Zaitlen, Hyun Min Kang, Eleazar Eskin, and Eran Halperin, "Leveraging the HapMap correlation structure in association studies". The dissertation author was the primary investigator and author of this paper.

Chapter 3, was published In Proceedings of the 8th Workshop on Algo-

rithms in Bioinformatics, (WABI-2008), Karlsruhe, Germany, September 15-17, 2008. Arthur Choi, Noah Zaitlen, Buhm Han, Knot Pipatsrisawat, Adnan Darwiche, E. Eskin, “Efficient Genome Wide Tagging by Reduction to SAT”. The dissertation author and Arthur Choi were the primary investigators and authors of this paper.

Chapter 4, was published in Human Heredity, Vol 68, pp73-86, 2009. Noah A. Zaitlen, Hyun Min Kang, and Eleazar Eskin, “Linkage effects and analysis of finite sample errors in the HapMap”.

Chapter 5, in part is currently being prepared for submission for publication of the material. Noah Zaitlen, Eleazar Eskin. The dissertation author is the primary investigator and author of this material.

Chapter 6, was published in The Journal of Computational Biology, Vol 15, pp 927-942, 2008. Noah Zaitlen, Manuel Reyes-Gomez, David Heckerman, Nebojsa Jojic, “Shift Invariant Adaptive Double Threading: Learning MHC II Peptide Binding”. The dissertation author was the primary investigator and author of this paper.

Chapter 7, was published in Genome Research, Vol 15, pp 1594-600, 2005. Noah A. Zaitlen, Hyun Min Kang, Michael L. Feolo, Stephen T. Sherry, Eran Halperin, and Eleazar Eskin, “Inference and analysis of haplotypes from combined genotyping studies deposited in dbSNP”. The dissertation author was the primary investigator and author of this paper.

VITA

2001	Bachelor of Science in Computer Science, Mathematics, and Cognitive Science *cum laude*, University of California, Berkeley
2009	Doctor of Philosophy in Bioinformatics, University of California, San Diego

PUBLICATIONS

Hyun Min Kang, Noah Zaitlen, Buhm Han, Eleazar Eskin, "An adaptive and memory efficient algorithm for genotype imputation", *In Proceedings of the Thirteenth Annual Conference on Research in Computational Biology (RECOMB-2009)*, Tucson, AZ: May 18 - 21, 2009.

Noah A. Zaitlen, Hyun Min Kang, and Eleazar Eskin, "Linkage effects and analysis of finite sample errors in the HapMap", *Human Heredity*, 68(2):73-86, 2009.

Chiara Sabatti, Susan K Service, Anna-Liisa Hartikainen, Anneli Pouta, Samuli Ripatti, Jae Brodsky, Chris G Jones, Noah A Zaitlen, Teppo Varilo, Marika Kaakinen, Ulla Sovio, Aimo Ruokonen, Jaana Laitinen, Eveliina Jakkula, Lachlan Coin, Clive Hoggart, Andrew Collins, Hannu Turunen, Stacey Gabriel, Paul Elliot, Mark I McCarthy, Mark J Daly, Marjo-Riitta Jrvelin, Nelson B Freimer, & Leena Peltonen, "Genome wide association analysis of metabolic traits in a birth cohort from a founder population", *Nature Genetics*, 41:35-46, 2008.

Arthur Choi, Noah Zaitlen, Buhm Han, Knot Pipatsrisawat, Adnan Darwiche, E. Eskin, "Efficient Genome Wide Tagging by Reduction to SAT", *In Proceedings of the 8th Workshop on Algorithms in Bioinformatics*, (WABI-2008), Karlsruhe, Germany, September 15-17, 2008.

Buhm Han, Hyun Min Kang, Myeong Seong Seo, Noah A. Zaitlen, and Eleazar Eskin, "Efficient association study design via power-optimized tag SNP selection", *Annals of Human Genetics*, 72:834-47, 2008.

Noah Zaitlen, Manuel Reyes-Gomez, David Heckerman, Nebojsa Jojic, "Shift Invariant Adaptive Double Threading: Learning MHC II Peptide Binding", *Journal of Computational Biology*, 15(7): 927-942, 2008.

Hyun Min Kang, Noah A. Zaitlen, Claire M. Wade, Andrew Kirby, David Heckerman, Mark J. Daly, and Eleazar Eskin, "Efficient control of population structure in model organism association", *Genetics*, 178:1709-23, 2008.

Sean O'Rourke, Noah Zaitlen, Nebojsa Jojic, Eleazar Eskin, "Reconstructing the Phylogeny of Mobile Elements", *In Proceedings of the Eleventh Annual Conference on Research in Computational Biology*, (RECOMB-2007), Oakland, CA: April 21st-25th, 2007.

Noah Zaitlen, Manuel Reyes-Gomez, David Heckerman, Nebojsa Jojic, "Shift Invariant Adaptive Double Threading: Learning MHC II Peptide Binding", *In Proceedings of the Eleventh Annual Conference on Research in Computational Biology*, (RECOMB-2007). Oakland, CA: April 21st-25th, 2007.

Noah A. Zaitlen, Hyun Min Kang, Eleazar Eskin, and Eran Halperin, "Leveraging the HapMap correlation structure in association studies", *American Journal of Human Genetics*, 80:683-91, 2007.

Noah A. Zaitlen, Hyun Min Kang, Michael L. Feolo, Stephen T. Sherry, Eran Halperin, and Eleazar Eskin, "Inference and analysis of haplotypes from combined genotyping studies deposited in dbSNP", *Genome Research*, 15:1594-600, 2005.

FIELDS OF STUDY

Major Field: Bioinformatics
Eleazar Eskin

ABSTRACT OF THE DISSERTATION

Methods for the Analysis of Human Genetic Variation in the Search for the Genetic Basis of Human Disease

by

Noah Zaitlen
Doctor of Philosophy in Bioinformatics

University of California San Diego, 2009

Professor Vineet Bafna, Chair

Recent technological advances in the field of molecular biology have ushered in the genome wide association era of human genetics. Researchers can now simultaneously examine hundreds of thousands of single nucleotide polymorphisms (SNPs) in an individual at continually decreasing costs. In an effort to characterize distributions of SNPs in human populations a set of four million SNPs was collected in 269 individuals from four populations. This HapMap data set in combination with high throughput genotyping technology has caused a fundamental shift in the methodologies of scientists searching for the relationship between genotype and phenotype. The genome wide association study (GWAS) has become mainstream practice, leading to the discovery of a growing number of loci associated with the genetics basis of complex phenotypes including many human diseases.

This work describes novel methods, resources, tools, and techniques designed to improve our ability to interpret and utilize GWAS and HapMap data.

The Weighted Haplotype (WHAP) association method leverages the linkage structure information from the HapMap to improve GWAS power by providing accurate statistics for unobserved SNPs without the costs of additional genotyping. The SAT based tagging algorithm SATTagger identifies which SNPs to genotype as part of an association study, and provides the first optimal genome wide solution to this classic bioinformatics problem. The HapMap suffers from the fundamental limitation that at most 60 unrelated individuals are available per population. An analytical framework for analyzing the implications of a finite sample HapMap is presented. The results of the first round of GWAS studies showed that effect sizes of causal variants were small and that larger sample sizes were required for adequate power. Meta-analysis provides a mechanism for overcoming this problem with the cost of additional genotyping. A new statistic for imputation based meta analysis in a GWAS is given.

Additional research is presented on MHC Class II binding prediction, which is a useful tool in understanding auto-immune and pathogenic diseases. A physics based binding model is presented with an EM like solution to find the optimal binding conformation.

Chapter 1

Introduction to Human Genetic Variation Methods

Even a simple examination of living organisms reveals that children inherit traits from their parents. Humans have exploited this property in agriculture and animal husbandry for thousands of years. The ancient Greek philosophers theorized on its mechanisms, but it was not until the mid 19^{th} Century and the arrival of Gregor Mendel that we began a scientific study of heritability. It took almost another 100 years before Oswald Avery, Colin MacLeod, and Maclyn McCarty discovered that DNA is the molecule responsible for transferring genetic information. By that time, the visionaries Alfred Sturtevant and Sir Ronald Aylmer Fisher had laid down the framework for modern genetics, which contains as one of its fundamental goals, uncovering the relationship between variation in an organism's DNA and its phenotypes. The words genotype and phenotype themselves divulge the centrality of this problem, dichotomizing an individual into its genetic material and all other constituent parts.

The last five years have seen a dramatic shift in the way this problem is approached. This is largely due to the combination of four factors both technological and social. First, work on the human genome project identified millions of new genetic variants believed to be the core elements driving genetic heritability. Second, technological advances in high-throughput genotyping based on these new variants, provided a means of quickly and cheaply performing experiments that had previously taken years and cost millions of dollars. Third, the availability of the HapMap reference data set characterized the nature of genetic variation on a genome wide scale. Finally, the massive increase of funds, both public and private available to researchers working on human disease phenotypes paid for the new large scale experiments that are becoming commonplace. Together, these elements have brought about the genome wide association era of human genetics. This dissertation is primarily concerned with a new set of methods, resources, tools, and techniques designed to address the problems and improve the power of genome wide association studies.

Before outlining the specific novel contributions of this work, some background is given to lend them context and show their relevance to the field. The human genome consists of 23 chromosomes comprised of 2.3 billion base pairs of DNA. If we examine the DNA of two individuals, the differences in their genome will include individual nucleotides changes called *single nucleotide polymorphisms* (SNPs), changes in the number of copies of a segment of DNA called *copy number variations* (CNVs), and other structural changes such as inversions and translocations. It is believed that heritability is mostly due to changes such as these, with some growing evidence for epigenetic effects. If a particular genetic variant has a functional property that effects a phenotype, such as susceptibility to a disease, it called *causal* with respect to that phenotype. While the causal genetic variants of many phenotypes have been discovered, the search for variants related to human disease are particularly important and unfortunately remain amongst the most elusive to discover.

The fundamental property of genetic variation that allows the mapping of loci was first discovered by William Bateson and Reginald Punnett at the beginning of the 20^{th} Century. *Genetic Linkage* is the relationship between variants in proximal genomic regions. Figure 1.1 illustrates this property and why it exists. Only cross-over events during meiosis can break apart alleles on the same chromosome. The closer together two alleles are, the lower the probability that there will be a recombination event between them. After many generations, distant alleles will independent of one another. Proximal alleles however will only exist in certain combinations. Thus knowing the genotype of a particular SNP may give some information about the genotype of its neighbors. Association mapping studies are experiments used to identify causal variants, which rely on this property.

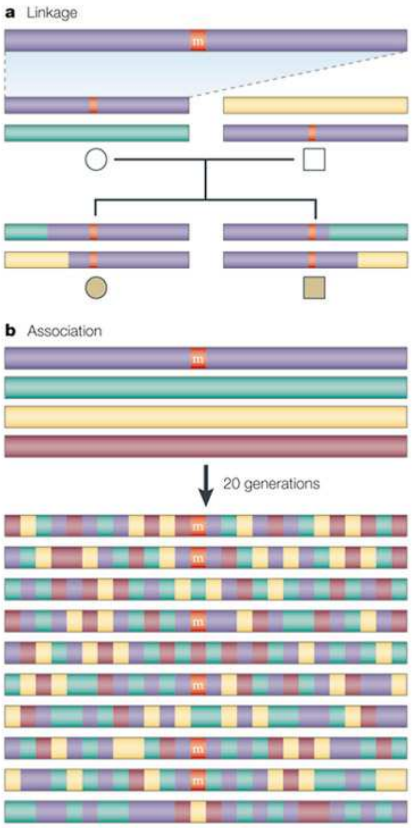

Figure 1.1: Proximal regions on a chromosome can only be separated by a recombination event. The orange segment marked m maybe a causal mutation for a disease. (a) shows how recombination breaks the haplotype backgrounds in one generation. The background around m is purple and extends very far. In (b) haplotype backgrounds are given for a large number of individuals after 20 generations have passed from four original chromosomes. Again the region around m is purple, but the regions lengths are much shorter. Association studies work on the principal that if a genotyped SNP falls into the purple region, the effect of m may be observed.

Association mapping has until recently served as a means to narrow the region identified as part of a linkage mapping study. In this technique, individuals are collected from two groups, the cases who have the disease of interest, and the controls that are members of the same populations but do not have the disease. The individuals are genotyped and differences in the allele frequencies of the SNPs between the cases and controls are searched for. If the causal variant is amongst the genotyped SNPs then its distribution may show a significant difference between the cases and controls. However, if it is not genotyped, there is still the possibility of recovering a signal. Due to the linkage properties of the genome described above, a proximal SNP may be linked to the causal variant and its distribution may therefore also exhibit a significant difference between cases and controls. Notice that the mutation m in Figure 1.1 b usually has a purple haplotype background. Without this local linkage structure individuals would need to be completely sequenced in order to identify causal variants via association studies. The nature of linkage is therefore a central aspect of association mapping and for this reason it has been extensively studied.

The HapMap project characterized the extent of linkage in four human populations and showed how local linkage structure exists throughout our genomes. Eric Lander and colleagues were strong proponents of developing and using this information to extend association mapping to the whole genome. Many researchers believed that haplotypes would exist in local "blocks" that could be completely identified with a small number of SNPs. Genome wide association studies (GWAS) could be conducted by using makers (i.e. SNPs) laid at intervals across the genome based on the block structure garnered the HapMap. The "block" theory died when the HapMap data was released, but the local linkage structure was strong enough for the GWAS principal to work anyway. Figure 1.2 shows the correlations between proximal SNPs in one region of the genome. In 2007, shortly after the HapMap release, the first GWAS was completed by Robert Sladek and colleagues. Since then hundreds of such studies have been funded with many producing novel loci associated with human disease and related phenotypes. Although there has been a lot of argument about the value of these experiments and their results, there is no doubt that they are an extremely active research area.

The following briefly outlines my contributions to this field. Each Chapter is based on a paper, already published or in submission, that addresses issues core

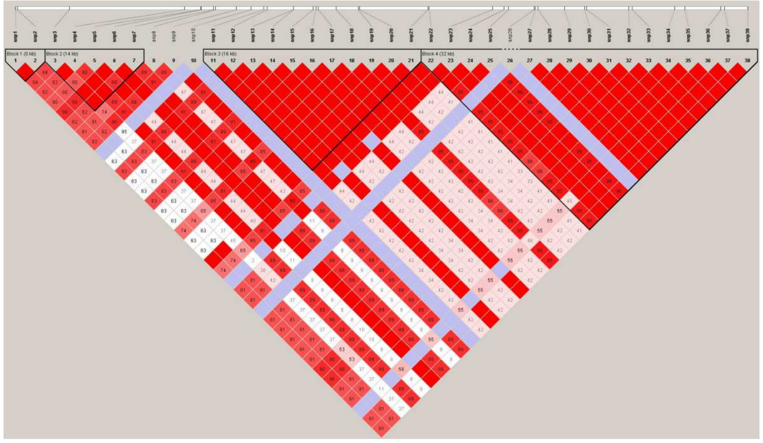

Figure 1.2: The linkage structure of the gene GCH1 in the HapMap CEU population. The SNPs are numbered from 1 to 38. For each pair a square is colored a shade of red depending on the strength of the correlation between them. Red represents complete LD and white represents independent SNPs. Potential "blocks" are outlined in black with very strong LD between all SNPs in the block. Although the "block" theory of haplotypes is no longer widely accepted, the local linkage structure in the genome is evident from the figure.

to GWAS and the current study of the genetic basis of human disease.

Chapter 2 describes a novel method to improve the power of GWAS studies by utilizing the HapMap data set. Recent high-throughput genotyping technologies, such as the Affymetrix 500k array and the Illumina HumanHap 550 beadchip, have driven down the costs of association studies and have enabled the measurement of single-nucleotide polymorphism (SNP) allele frequency differences between case and control populations on a genome wide scale. A key aspect in the efficiency of association studies is the notion of "indirect association", where only a subset of SNPs are collected to serve as proxies for the uncollected SNPs, taking advantage of the correlation structure between SNPs. Recently, a new class of methods for indirect association, multi-marker methods, has been proposed. Although the multi-marker methods are a considerable advancement, current methods do not fully take advantage of the correlation structure between SNPs and their multi-marker proxies. We propose a novel multi-marker indirect-association method, WHAP, that is based on a weighted sum of the haplotype frequency differences. In contrast to traditional indirect-association methods, we show analytically that there is a considerable gain in power achieved by our method compared with both single-marker and multi-marker tests, as well as traditional haplotype-based tests. Our results are supported by empirical evaluation across the HapMap reference panel data sets, and a software implementation for the Affymetrix 500k and Illumina HumanHap 550 chips is available for download.

Chapter 3 presents a new solution to the classical tag SNP selection problem. Whole genome association has recently demonstrated some remarkable successes in identifying loci involved in disease. Designing these studies involves selecting a subset of known single nucleotide polymorphisms (SNPs) or tag SNPs to be genotyped. Even though how to chose the tag SNPs is a well studied research problem, questions remain on how to choose the optimal set of tag SNPs. Since the standard formulations of the problem are NP-hard, most algorithms for selecting tag SNPs are either heuristics which do not guarantee selection of the optimal set of tag SNPs or are exhaustive algorithms which are computationally impractical. We present the first practical algorithm for optimal tag SNP selection. We reduce the tag SNP selection problem to a variant of the much studied satisfiability problem, encoding a given instance into conjunctive normal form (CNF). We take advantage of the local structure inherent to the problem, as well as progress in knowledge compilation, and convert our CNF encoding into a tractable representation known as DNNF, from which solutions to our original problem can be easily enumerated. We demonstrate our methods by constructing the optimal tag set for the whole genome and show that we significantly outperform previous exhaustive search-based methods. We also present optimal solutions for the harder problem of selecting multi-marker tags, a problem for which no optimal algorithms have been proposed. We also show how our methods can be adapted to discovering the tag set that maximizes statistical power given a budget of SNPs to collect. This problem is more challenging than the traditional tag SNP selection problem and we show how it can be reduced to weighted Max-SAT.

Chapter 4 presents a framework for analyzing finite sample issues in statistics computed over the HapMap data. The HapMap provides a valuable resource to help uncover genetic variants of important complex phenotypes such as disease risk and outcome. Using the HapMap we can infer the patterns of LD within different human populations. This is a critical step for determining which SNPs to genotype as part of a study, estimating study power, designing a follow-up study to identify the causal variants, imputing untyped SNPs, and estimating recombination rates along the genome. Despite its tremendous importance, the HapMap suffers from the fundamental limitation that at most 60 unrelated individuals are available per population. We present an analytical framework for analyzing the implications of a finite sample HapMap. We present and justify simple approximations for deriving analytical estimates of important statistics such as the square of the correlation coefficient r^2 between two SNPs. Finally, we use this framework to show that current HapMap based estimates of r^2 and power have significant errors, and that tag sets highly overestimate their coverage. We show that a reasonable increase in the number of individuals, such as that proposed by the 1000 genomes project, greatly reduces the errors due to finite sample size for a large proportion of SNPs.

Chapter 5 covers a new method we developed for performing meta-analysis over imputed data. Genome wide association studies have identified many new loci which may be involved in complex human diseases. The newly discovered variants often have weak effects requiring studies with large numbers of individuals to achieve the statistical power necessary to identify them. Likely, there exist even more associated variants which remain to be found if even larger association studies can be assembled. Meta-analysis provides a straightforward means of increasing study sample sizes without collecting new samples by combining existing data sets. A difficulty in combining studies is that they are collected on different platforms and collect different markers. Current studies combine results from different genotyping platforms by imputing genotypes missing from either study and then performing standard meta-analysis techniques. We show that this approach will result in a loss of power since errors in imputation are not accounted for. We present a new method for performing meta-analysis over imputed SNPs, show that it is optimal with respect to power, and discuss practical implementation issues. Through simulation experiments, we show that our imputation aware meta-analysis approach outperforms or matches standard meta-analysis approaches.

Chapter 6 describes the phasing and curation of haplotypes at NCBI. In addition it provides new results about haplotype block theory and genotyping error. In the attempt to understand human variation and the genetic basis of complex disease, a tremendous number of single nucleotide polymorphisms (SNPs) have been discovered and deposited into NCBI's dbSNP public database. More than 2.7 million SNPs in the database have genotype information. This data provides an invaluable resource for understanding the structure of human variation and the design of genetic association studies. The genotypes deposited to dbSNP are unphased, and thus, the haplotype information is unknown. We applied the phasing method HAP to obtain the haplotype information, block partitions, and tag SNPs for all publicly available genotype data and deposited this information into the dbSNP database. We also deposited the orthologous chimpanzee reference sequence for each predicted haplotype block computed using the UCSC BLASTZ alignments of human and chimpanzee. Using dbSNP, researchers can now easily perform analyses using multiple genotype data sets from the same genomic regions. Dense and sparse genotype data sets from the same region were combined to show that the number of common haplotypes is significantly underestimated in whole genome data sets, while the predicted haplotypes over the common SNPs are consistent between studies. To validate the accuracy of the predictions, we benchmarked HAP's running time and phasing accuracy against PHASE. Although HAP is slightly less accurate than PHASE, HAP is over 1000 times faster than PHASE, making it suitable for application to the entire set of genotypes in dbSNP.

In addition to methods to related to the design and analysis of GWAS, I include research from the area of computational immunology. Although it is a break from the above chapters it is still in a general sense related to the study of human genetic variation and disease. In this case the diseases are auto-immune or the result of viral or bacterial infections. The objective is not to identify the genes in related to the disease, but to characterize how genetic variation affects a core aspect of immune resoponse.

Chapter 7 presents work on modelling binding of protein's from the major histocompatibility complex (MHC). The MHC plays important roles in the workings of the human immune system. Specificity of MHC binding to peptide fragments from cellular and pathogens' proteins has been found to correlate with disease outcome and pathogen or cancer evolution. In this work we propose a novel approach to predicting binding configurations and energies for MHC class II molecules, whose epitopes are generally predicted less well than the MHC I epitopes due in part to larger variation in bound peptide length. We treat the relative position of the peptide as a hidden variable, and model the ensemble of different binding configurations, rather than use a separate alignment procedure to narrow it down to one. Thus, our predictor infers a distribution over peptide positions from the MHC II and peptide sequences, and computes the total binding affinity. The training procedure iterates the predictions with re-estimation of the parameters of the binding groove model. For a given relative peptide position, any MHC class I prediction model can be used. Here we choose the physics based model of Jojic et al. [66]. We show that the parameters of the binding model can be learned efficiently from the training data and then used to estimate binding energies for previously untested peptides. Our technique performs on par with previous approaches to MHC II epitope prediction. Furthermore, our model choice allows generalization to new MHC class II alleles, which were not a part of the training set.

Before beginning a full description of each method I point out that this dissertation focuses on the details of the methods listed above as opposed to their application. However, it is important to note that many of these methods have been and are currently being used as part of GWAS and other studies. It is beyond the scope of this document to describe in depth the application of the methods as each study would require its own chapter and I am not the primary researcher in any of those projects. However, I have put a lot of effort into establishing collaborations and building software tools. In fact, the software engineering and implementation of several of the methods consumed significantly more time and energy than the original research and paper writing. This is an important but less recognized component of methods development, as it prevents the work from remaining just interesting ideas, but puts them into practice.

Chapter 2

Weighted Haplotype Analysis

2.1 Introduction

Large scale case-control association studies are a potentially powerful tool for discovering the genetic basis of human disease [39, 98, 22]. Recent high-throughput genotyping technologies such as the Affymetrix 500k array and the Illumina HumanHap550 beadchip have driven down the costs of association studies and allow us to measure allele frequency differences between case and control populations on a genome wide scale [81, 50]. A key aspect in the efficiency of association studies is the notion of “indirect association”. By leveraging the linkage disequilibrium (LD) structure of the genome, frequency differences between case and control populations do not need to be measured in all SNPs, but only in a subset, or a set of “tag SNPs” which serve as proxies for the remaining uncollected SNPs (we refer to the uncollected SNPs as *hidden SNPs*) [64]. A chromosome carrying a particular allele of a tag SNP has a high probability of carrying a particular allele of a proximal hidden SNP. Thus an allele frequency difference in an uncollected hidden SNP will manifest itself as an allele frequency difference in a tag SNP. This correlation is often measured by the correlation coefficient r^2 between two SNPs. The r^2 measure is widely used in the design and analysis of association studies because the relation between the power of detecting an association at the hidden SNP while only observing the tag SNP has been well understood for some time (see, for example, [93, 103]).

Tag SNPs are chosen by examining the linkage disequilibrium structure of a *reference* panel (such as the HapMap[25]), which is a data set that contains a complete set of genotypes for 270 individual on over 3.9 million SNPs across the genome. Choosing a set of tag SNPs is a challenging problem since the linkage disequilibrium structure is quite complex and varies through the genome. To date, many tag SNP selection methods have been proposed (e.g., [38, 18]). These methods employ different statistical criteria, the most common being procurement of a set of tag SNPs, for which every hidden SNP is 'covered' by a tag SNP, such that the correlation coefficient r^2 between the two SNPs in the reference set is higher than a certain threshold (see, e.g. [18]). These methods vary greatly in the optimization methods used to obtain the tag SNPs.

Recently, a new class of methods, multi-marker methods, have been proposed [38, 19, 118, 110]. These methods take advantage of the fact that some pairs (or groups) of SNPs serve as better proxies for the hidden SNPs, than any single SNP. As multi-marker proxies have more than two possible alleles, the frequencies of a specific sequence of alleles in these SNPs (a *haplotype*) is compared between the cases and the controls. Thus, a specific haplotype, instead of a single SNP, is used as a proxy for a hidden SNP. It has been shown empirically that these methods can reduce the number of required tags in order to achieve equivalent power[38]. In addition, it has been empirically shown that even if the set of tag SNPs is fixed, such as in the case when a commercial high throughput genotype product is used, one can choose a set of multi-markers for each hidden SNP, and considerably increase the r^2 (and therefore the power) between that proxy haplotype and the hidden SNP [90].

While multi-marker methods are a considerable advance, current methods do not fully take advantage of the correlation structure between SNPs and their multi-marker proxies. For example, consider the scenario given in Figure 2.1. In this example, we assume that the first two SNPs are collected as tag SNPs for the association study which will be used as proxies for the three remaining SNPs. The third SNP is in perfect disequilibrium with the first SNP ($r^2 = 1$), and thus the first SNP serves as a perfect proxy for the third SNP. While the fourth SNP is

Haplotypes					Freq.
1	**2**	3	4	5	
A	A	A	A	A	.25
A	G	A	G	G	.15
A	G	A	G	A	.10
G	A	G	G	G	.25
G	G	G	G	G	.25

Figure 2.1: A sample haplotype distribution for 5 SNPs where the first 2 SNPs are collected as tag SNPs and the remaining 3 SNPs are uncollected.

not in perfect disequilibrium with either of the first two SNPs, the haplotype AA at the first two SNPs can serve as a perfect proxy for the fourth SNP. The most interesting case is the fifth SNP, for which no haplotype serves as perfect proxy. The best haplotype proxy for this SNP is the haplotype AA, for which $r^2 = 0.619$. However, by restricting ourselves to the haplotype AA, we ignore the additional information given by the other haplotypes. For example, the allele A in the fifth SNP occurs occasionally with haplotype AG but never with haplotypes GA or GG.

To take advantage of this additional information, we propose a new method, WHAP, and a new statistic, ρ-test, that is based on a Weighted sum of all the Haplotype frequency differences. We show both empirically and analytically that there is a considerable gain in power achieved by this statistic, as opposed to using a χ^2-statistic on a single SNP, or a group of haplotypes. We show that the ρ-test is χ^2 distributed with one degree of freedom, regardless of the weight assignments. We then develop an equivalent notion to r^2 defined by the haplotype weights, r_h^2, with values ranging from 0 to 1. Analogously to Pritchard and Preworzski [93], we show that if a multi-marker set has a correlation of r_h^2 with a causal SNP, then using the ρ-test with n/r_h^2 individuals for this set is equivalent to directly testing the causal SNP for association with n individuals. We show analytically that the r_h^2 for a set of tag SNPs is always at least as large as the best r^2 for any single haplotype or single SNP. Empirically, we observe that in many cases r_h^2 is in fact quite larger than r^2, leading to a significant increase in power. For instance, in the above example, the correlation coefficient between the weighted average of the haplotypes and the fifth SNP is 0.85 while it is only 0.619 for the best single haplotype. Finally, we show that the ρ-test is always more powerful than the standard χ^2-test over a set of haplotypes.

Previous approaches for tag SNPs such as single markers and multi-marker approaches involving one haplotype, fall into our framework, since these can be seen as specific assignments of weights to the haplotypes (i.e., letting the weight of the haplotype be 1 and all the other haplotypes have weight 0). We present a method to find the optimal set of weights which maximizes the power of the ρ-statistic and we show both analytically and empirically that our method always performs at equal or greater power to standard multi-marker methods. Furthermore, we show that asymptotically one can only gain power by using a larger number of SNPs as a proxy to the hidden SNP; in practice, as sample size is limited, "over-fitting" effects may reduce power, and we therefore empirically show that for haplotypes of moderate length there is an increase in power. To the best of our knowledge, this is the first analytical rigorous proof that demonstrates that haplotype and multi-marker indirect association is *asymptotically* more powerful than indirect association based on single SNPs.

Our methods and power analysis relies on accurate haplotype frequency estimates. Since the accuracy of haplotype frequency estimation depends on different factors, such as the number of SNPs used, their physical location, and the

LD structure, we evaluated our analytical results via simulation. We first demonstrate that r_h^2 is always greater than r^2 for both SNPs and multi-marker tags over the marker sets of the Affymetrix 500k and Illumina HumanHap550 chips. In particular, moving from multi-marker tags to our weighted haplotypes results in up to a 21.1% increase in the number of captured common SNPs (MAF ≥ 0.05 and r^2 or $r_h^2 \geq 0.8$). Second, we simulate case control panels under various disease models,and show that this increase in utility corresponds as expected to an increase in the power of our method over single SNPs and multi-marker tags.

We calculated the optimal weights for every HapMap Phase II SNP using the Affymetrix 500k and Illumina HumanHap 550 SNP sets. These data, as well as a software implementation for using our statistic over data from these chips is available upon request.

2.2 Material and Methods

The ρ-test is a statistic that is applied to a set of Weighted HAPlotype (WHAP) tag SNPs which are a proxy for the hidden SNP. It can be used in place of the standard χ^2 statistic applied to the tag SNPs. Informally, the ρ-test computes a weighted sum of all the tag SNP haplotype frequency differences between the case and control samples. A more formal description of the ρ-test is given below.

In traditional multi-marker methods, for a given hidden SNP, a set of SNPs is chosen as tag SNPs and a specific haplotype of the tag SNPs is used as the proxy. In contrast, in the ρ-test framework, once the tag SNPs are chosen, a weight for each of the haplotypes is determined. The specific values of the weights are estimated from the reference panel (e.g., the HapMap data set) and recorded for each hidden SNP.

The ρ-test is χ^2 distributed with one degree of freedom, and its power depends on the correlation coefficient r_h^2 between the statistic and the hidden SNP (see below). We show that r_h^2 is analogous to r^2 in standard association methods in the sense that it provides a direct linear relation to power.

We consider the setting in which an association study is performed on N

cases and N controls. We assume that the causal SNP s is not genotyped, but a set of SNPs $\S = \{s_1, \ldots, s_m\}$ in LD with s are genotyped. For simplicity of presentation, we assume that each of the SNPs is biallelic with allele values 0 and 1. In order to distinguish the allele notation of s from the other SNPs, we assume that the alleles of s are C and c. Let $h_1, \ldots, h_k \in \{0, 1\}^m$ be the set of haplotypes over the set of SNPs $\S$. We suggest a statistical test, which we call ρ-test, which is based on a convex combination of the haplotype frequencies. This combination depends on the joint distribution of the alleles c and C of s and the haplotypes in the HapMap data.

Formally, let $\vec{a} = \{a_1 \ldots, a_k\}$ be a set of haplotype weights. Let $\hat{p}_h^1$ and $\hat{p}_h^0$ be the observed frequencies of haplotype h in the case and control populations, and let $\hat{p}_h = \frac{\hat{p}_h^0 + \hat{p}_h^1}{2}$. We define the ρ-statistic as

$$\rho(\vec{a}) = \frac{N\left(\sum_{h=1}^{k} a_h(\hat{p}_h^1 - \hat{p}_h^0)\right)^2}{2\left(\sum_h a_h^2 \hat{p}_h - (\sum_h a_h \hat{p}_h)^2\right)}.$$

Under the null hypothesis, $\rho(\vec{a})$ is distributed as χ^2 with one degree of freedom, that is, the square of a standard normal distribution. Denoting by p_h^0 and p_h^1 the true frequency of haplotype h in the case and control populations respectively, under the alternate hypothesis, $\rho(\vec{a})$ is distributed as the square of a normal distribution with mean

$$\lambda_h = \frac{\sqrt{N}\sum_{h=1}^{k} a_h(p_h^1 - p_h^0)}{\sqrt{2}\sqrt{\sum_h a_h^2 p_h - (\sum_h a_h p_h)^2}},$$

and where the variance is approximately 1, assuming that $p_h^1 \approx p_h^0$, and that $p_h = \frac{p_h^0 + p_h^1}{2}$. Thus, the power of the $\rho(\vec{a})$-statistic depends on the frequencies p_h^0, p_h^1, and on the weight vector $\vec{a}$.

In order to evaluate the statistical power of the $\rho(\vec{a})$-statistic, we are interested in comparing its power to the power of detecting association directly with the causal SNP s by the χ^2 test. Let $\hat{p}_C^1$ and $\hat{p}_C^0$ be the observed frequencies of allele C at SNP s in the case and control populations assuming we directly genotype the SNP. The χ^2-statistic can be written as

$$X = \frac{N(\hat{p}_C^1 - \hat{p}_C^0)^2}{2p_C(1 - p_C)}.$$

Similarly to the $\rho(\vec{a})$-statistic, under the null hypothesis, X is distributed as the square of a standard normal distribution. Denoting the true SNP frequencies as p_C^0 and p_C^1, and $p_C = \frac{p_C^0+p_C^1}{2}$, under the alternative hypothesis, X is distributed as the square of a normal distribution with mean

$$\lambda_c = \frac{\sqrt{N}(p_C^1 - p_C^0)}{\sqrt{2}\sqrt{p_C(1-p_C)}},$$

and with a variance approximately 1, assuming $p_C^0 \approx p_C^1$. The relation between λ_h and λ_c determines the relation between the power of $\rho(\vec{a})$ and X.

The underlying assumption in any indirect association method is that the correlation structure of the cases and the controls is similar as long as the two groups are sampled from the same underlying population. For instance, the underlying correlation structure is assumed to be similar to the closest HapMap population, and therefore the set of tag SNPs and the expected power of these SNPs to detect association can be estimated from the HapMap data set. More formally, we assume that the conditional probability q_{hC} (or q_{hc}) of haplotype h given C (or c) are the same in the case and control populations. If the cases and controls are sampled from a population which is similar to one of the HapMap populations, these conditional probabilities can be estimated from the HapMap quite efficiently, as we show later.

Under these assumptions, we have

$$\begin{aligned}
\lambda_h &= \frac{\sqrt{N}\sum_h a_h(p_h^1 - p_h^0)}{\sqrt{2}\sqrt{\sum_h a_h^2 p_h - (\sum_h a_h p_h)^2}} \\
&= \frac{\sqrt{N}(p_C^1 - p_C^0)\sum_h a_h(q_{hC} - q_{hc})}{\sqrt{2}\sqrt{\sum_h a_h^2 p_h - (\sum_h a_h p_h)^2}} \\
&= \frac{\sqrt{N}(p_C^1 - p_C^0)\sum_h a_h(q_{hC} - q_{hc})}{\sqrt{2}\sqrt{\sum_h a_h^2 p_h - (\sum_h a_h p_h)^2}} \frac{\sqrt{p_C(1-p_C)}}{\sqrt{p_C(1-p_C)}} \\
&= \frac{\sqrt{N}(p_C^1 - p_C^0)}{\sqrt{2}\sqrt{p_C(1-p_C)}} \frac{\sum_h a_h(q_{hC} - q_{hc})\sqrt{p_C(1-p_C)}}{\sqrt{\sum_h a_h^2 p_h - (\sum_h a_h p_h)^2}} = \lambda_c r_{\vec{a}},
\end{aligned}$$

where $r_{\vec{a}} = \frac{\sum_h a_h(q_{hC} - q_{hc})\sqrt{p_C(1-p_C)}}{\sqrt{\sum_h a_h^2 p_h - (\sum_h a_h p_h)^2}}$. Thus, the power of detecting the causal SNP with a sample size of N individuals (using the χ^2 statistic) is the same as the power of detecting the causal SNP using the $\rho(\vec{a})$-statistic with $N' = N/r_{\vec{a}}^2$

individuals. When the indirect association is performed on one SNP (i.e., $m = 1$), $r_{\vec{a}}$ is $\sqrt{r^2}$ regardless of the weight vector $\vec{a}$. Thus, $r^2_{\vec{a}}$ can be seen as a natural generalization to the standard notion of r^2 measure of linkage disequilibrium.

2.2.1 Finding the best weight vector.

Clearly, it is desirable to perform the $\rho(\vec{a})$-test with a weight vector $\vec{a}$ that maximizes $r^2_{\vec{a}}$. We now show that $r_{\vec{a}}$ is maximized when a_h is the conditional probability of C given h (denoted by q_{Ch}). That is, we show the following theorem:

Theorem 1. *The power of the $\rho(\vec{a})$ statistic is maximized when for each haplotype h, $a_h = q_{Ch}$.*

Proof. As shown above, the power of the $\rho(\vec{a})$-test is directly determined by the value of $r^2_{\vec{a}}$. We set

$$\begin{aligned} \alpha_C &= \sum_h a_h q_{hC} \\ \alpha_c &= \sum_h a_h q_{hc}. \end{aligned}$$

With these notations, the numerator can be written as $(\alpha_C - \alpha_c)\sqrt{p_C(1-p_C)}$. Assuming that for the optimal solution $\alpha_C \neq \alpha_c$ (otherwise the optimum is zero, and then any vector $\vec{a}$ will satisfy this), it can be easily verified that without loss of generality, we can arbitrarily choose the values of α_C and α_c, as long as they are non-negative numbers. The latter follows from the fact that if $\vec{a}$ maximizes $r^2_{\vec{a}}$, then so does $\beta\vec{a}$ and $\vec{a}+\beta$ for every constant β. We thus set these values to satisfy $\alpha_C = \sum_h q_{Ch}q_{hC}$ and $\alpha_c = \sum_h q_{Ch}q_{hc}$.

The second term of the denominator can be written as

$$\begin{aligned} \sum a_h p_h &= \sum_h a_h(q_{hC}p_C + q_{hc}p_c) \\ &= p_C\alpha_C + p_c\alpha_c. \end{aligned}$$

At the same time, by the Cauchy-Schwartz inequality,

$$\sum_h a_h^2 p_h \cdot \sum_h \frac{q_{hC}^2}{p_h} \geq \left(\sum_h a_h q_{hC}\right)^2 = \alpha_C^2,$$

where equality holds if there is a constant β such that $a_h = \beta \dfrac{q_{hC}}{p_h} = \beta \dfrac{q_{Ch}}{p_C}$ for every haplotype h. By adding the definition of α_C and α_c, we can satisfy this equality by setting $\beta = p_C$. Put differently, the denominator is minimized when $a_h = q_{Ch}$ for every h. Since the numerator is now constant, the vector $\vec{a_h} = \vec{q_{Ch}}$ maximizes the value of $r_{\vec{a}}$.

□

Note that for the optimal selection of $\vec{a}$, i.e., when $a_h = q_{Ch}$ we observe that

$$\begin{aligned}
r^2_{\vec{a}} &= \frac{\left(\sum_h q_{Ch}(q_{hC} - q_{hc})\right)^2 p_C(1-p_C)}{\sum_h q^2_{Ch} p_h - (\sum_h q_{Ch} p_h)^2} \\
&= \frac{\left(\sum_h \frac{p^2_{Ch}}{p_h} - p_C \sum_h p_{Ch}\right)^2}{p_C(1-p_C)\left(\sum_h \frac{p^2_{Ch}}{p_h} - (\sum_h p_{Ch})^2\right)} = \frac{\left(\sum_h \frac{p^2_{Ch}}{p_h} - p^2_C\right)^2}{p_C(1-p_C)\left(\sum_h \frac{p^2_{Ch}}{p_h} - p^2_C\right)} \\
&= \frac{\sum_h \frac{p^2_{Ch}}{p_h} - p^2_C}{p_C(1-p_C)} = \frac{\sum_h q_{Ch}(p_{Ch} - p_C p_h)}{p_C(1-p_C)}.
\end{aligned}$$

We denote by $r^2_h = \frac{\sum_h q_{Ch}(p_{Ch} - p_C p_h)}{p_C(1-p_C)}$ the correlation coefficient between the haplotype distribution of $\{h_1, \ldots, h_k\}$ and the causal SNP. It is easy to see that $0 \leq r^2_h \leq 1$, and that r^2_h is always larger than the r^2 coefficient between any group of haplotypes and the causal SNP, and in particular, it is larger than the r^2 coefficient between any single tag SNP and the causal SNP. Furthermore, when the number of SNPs used for the ρ-test increases (i.e., m increases), the power of the association increases. To see this, consider the original haplotypes $\{h_1, \ldots, h_k\}$, and consider the haplotypes $\{h'_1, h''_1, h'_2, h''_2, \ldots, h'_k, h''_k\}$ that are formed by adding one more SNP. By definition, $p_{Ch_i} = p_{Ch'_i} + p_{Ch''_i}$, and $p_{h_i} = p_{h'_i} + p_{h''_i}$. Therefore, the r^2_h increases by

$$\frac{\sum_h((p^2_{Ch'}/p'_h + p^2_{Ch''}/p''^2_h) - p^2_{Ch}/p_h)}{p_C(1-p_C)} \geq 0,$$

where the latter is true since $\frac{(a+b)^2}{c+d} \leq \frac{a^2}{c} + \frac{b^2}{d}$ for every four numbers $a, b, c, d > 0$. Thus, increasing the number of SNPs can only amplify the power of detecting association with a hidden SNP. In practice, this is not exactly true, as the errors in the haplotype frequency estimates increases when the number of SNPs increases, and so does the effect of over-fitting.

2.2.2 The ρ-test compared to the χ^2-test.

As r_h^2 is larger than the maximal r^2 over all groups of haplotypes, we observe that the ρ-test has more power than the χ^2-test with one degree of freedom applied to any single haplotype. A natural question is whether the ρ-test is more powerful than the χ^2-test with $k-1$ degrees of freedom, applied to the set of haplotypes. This statistic can be written as

$$X_k = \frac{n}{2}\sum_h \frac{(p_h^0 - p_h^1)^2}{p_h}.$$

It is well known that for the null distribution X_k is distributed as χ^2 with $k-1$ degrees of freedom. Now, we can write

$$p_h^0 = p_C^0 q_{hC} + (1-p_C^0)q_{hc} = p_C^0\frac{p_{hC}}{p_C} + (1-p_C^0)\frac{p_{hc}}{1-p_C} = p_C^0\frac{p_{hC}-p_h p_C}{p_C(1-p_C)} + \frac{p_{hc}}{1-p_C}.$$

Therefore, $(p_h^0 - p_h^1)^2 = (p_C^0 - p_C^1)^2\frac{(p_{hC}-p_h p_C)^2}{p_C^2(1-p_C)^2}$. Thus, we observe that:

$$\begin{aligned} X_k &= \frac{n}{2}\sum_h \frac{(p_h^0 - p_h^1)^2}{p_h} \\ &= \frac{n}{2}\frac{(p_C^0 - p_C^1)^2}{p_C(1-p_C)}\cdot\frac{1}{p_C(1-p_C)}\sum_h\frac{(p_{Ch}-p_C p_h)^2}{p_h} = Xr_h^2. \end{aligned}$$

The last equality holds, as $r_h^2 = \frac{\sum_h q_{Ch}(p_{Ch}-p_C p_h)}{p_C(1-p_C)} = \frac{1}{P_C(1-p_C)}\cdot\left(\sum_h \frac{p_{Ch}^2}{p_h} - p_C^2\right)$, and on the other hand, $\sum_h \frac{(p_{Ch}-p_C p_h)^2}{p_h} = \sum_h \frac{p_{Ch}^2}{p_h} - p_C^2$. Under the alternative hypothesis, X_k is χ_{k-1}^2 distributed with mean λ_h, while the ρ-test is χ_1^2 distributed with mean λ_h. Therefore, one gains more power by using the ρ-test. We note that this conclusion is valid under the assumptions made in this analysis, and in particular under the assumption that in the studied region the disease is affected by one causal SNP. However, there are scenarios in which the statistic X_k has more power than the ρ-test; for instance, one such case would be when each of the different haplotypes affects the disease independently.

2.2.3 Estimating the values $q_{\vec{Ch}}$.

As Theorem 1 shows that the vector $\vec{a}$ that maximizes the power of the ρ-test is $q_{\vec{Ch}}$, we are interested in estimating the values q_{Ch} from the HapMap population closest to the cases and controls populations.

In order to do so, we first estimate the haplotype frequencies over the set of SNPs s, $s_1, \ldots, s_m$. The haplotype frequencies in a population can potentially be estimated by different methods such as EM [43] or PHASE [106]. For our needs, we use HaploFreq [55], which is based on a similar likelihood model to the one used by the EM algorithm, but it is provably more efficient and empirically more accurate than the EM algorithm. In particular, when performing whole genome association studies, the efficiency of these algorithms is crucial, as every hidden SNP s requires a new calculation of the haplotype frequencies in the HapMap population.

Given the haplotype distribution over the entire set of SNPs, it is easy to calculate the values q_{Ch} by setting $q_{Ch} = \dfrac{p_{Ch}}{p_{Ch} + p_{ch}}$. Since the frequencies p_{Ch} and p_{ch} are given by HaploFreq, we are able to calculate q_{Ch}.

2.3 Results

2.3.1 Benchmarks over HapMap ENCODE Regions.

In order to evaluate the relative utility of our ρ-test in comparison with single SNP and multi-marker methods we performed several benchmarks using the HapMap reference samples over the ENCODE regions. These data are made up of polymorphisms from 270 individuals from four populations over ten genomic regions spanning a total of 5Mb of sequence. These regions have been carefully studied and are believed to have complete ascertainment for SNPs with frequency greater than 5%. They are commonly used to estimate the performance of association statistics since there are still many ungenotyped and unknown common SNPs in the rest of the genome.

In a typical association study there is a set of marker SNPs (tag SNPs), which are genotyped, and a set of SNPs that are not observed (hidden SNPs). In order to replicate this scenario, we used the intersection of SNPs from current genotyping platforms and SNPs from each of the ENCODE regions as our marker sets. Following the example of others we measured the correlation between each SNP in the ENCODE regions with the best marker for the SNP from single tag

SNPs, multi-marker tags (HAPs), and our weighted haplotypes (WHAPs). We used the correlation coefficient r^2, and r_h^2 where appropriate, as measures of utility of the various methods. Sets with a higher correlation have a greater potential power as they are stronger proxies for the uncollected SNPs in the region.

The HAP and WHAP tags were selected by finding the strongest proxy via enumeration over all possible sets of two, three and four tag SNPs within 100Kb of each SNP in every ENCODE region. We limited the tag length to four in order to prevent over-fitting (see below for a further examination of the issue of over-fitting). We used two sets of tag SNPs for each ENCODE region: the SNPs contained in the Affymetrix 500k set and the SNPs contained in the Illumina HumanHap 550 set.

We compared the correlation coefficient of the weighted haplotypes used for the ρ-test (denoted by WHAPs) to the correlation coefficient with a single SNP (denoted SNPs) , and a single haplotype (denoted HAPs). Since the effective sample size is linearly related to the correlation coefficient, we measured the fraction of common SNPs (minimum allele frequency $\geq 5\%$) captured with correlation coefficient larger than a given threshold, for a range of thresholds. Figures 2.2 and 2.3 demonstrate this performance evaluation over the sets of tag SNPs, and the four HapMap populations. The figure demonstrates that the ρ-test outperforms each of the other methods in terms of correlation. Indeed, the ρ-test has significantly higher correlation for every population on every platform at all thresholds. This is especially pronounced in populations with complex LD structure (YRI). Although the improvement shown by our simulations is only a modest one, we expect this improvement to be more noticeable when haplotypes of more than four SNPs are used. As discussed below, this is currently prohibited due to effects of over-fitting, but larger reference data sets may allow such improvements in the future.

We explore the difference between HAPs and WHAPs by examining their relative increase in performance over using single SNPs. We observe that both WHAPs and HAPs are significantly stronger proxies than SNPs. In order to elucidate their differences, Tables 2.1 and 2.2 present the fraction of common SNPs captured with correlation coefficient ≥ 0.8 and the average correlation coefficient.

Evidently, the weighted haplotypes are a much better proxy to the hidden SNPs than the best haplotype (HAPs) or tag the best SNP. In fact, we observe that the ρ-test increases the correlation relative to the best haplotype or SNP for 50.4% of the SNPs. In Figure 2.4, we outline the distribution of weights for tags of these 50.4% of the SNPs. Unfortunately, even though in the majority of the cases the weighted haplotypes serve as a better proxy than the best haplotype or SNP, the average increase in r^2 is modest, since the increase is greater than 0.1 for 18.1% of the SNPs.

Table 2.1: Number of SNPs captured by each of the methods. Each row contains the fraction of common SNPs (MAF $\geq$ 0.05) captured with $r^2 \geq 0.8$ for each genotyping platform and population used in this study with tags up to length 4 SNPs. The Tag Set column specifies the genotyping platform as the Affymetrix 500K or Illumina HumanHap 550 set. For each hidden SNP, the four tag SNPs where chosen among all possible quartets of SNPs in 100kb distance from the SNP. The %*Inc.* column shows the % increase in the fraction of captured SNPs when moving from the HAPs to WHAPs. For example, the first row shows that in the CEPH population over the Affymetrix 500k chip, multi-marker tags capture 77% of SNPs while weighted haplotypes capture 84% of the SNPs. This is an 8.52% increase in the number of captured SNPs. We prove that WHAPs always perform at least as well as HAPs in the Methods section.

Tag Set	Pop	SNP	HAP	WHAP	%*Inc.*
Affymetrix 500k	CEU	0.61	0.77	**0.84**	8.52
Affymetrix 500k	CHB	0.62	0.76	**0.83**	8.95
Affymetrix 500k	JPT	0.59	0.73	**0.81**	11.67
Affymetrix 500k	YRI	0.37	0.61	**0.74**	21.06
Illumina HumanHap 550	CEU	0.88	0.97	**0.98**	1.60
Illumina HumanHap 550	CHB	0.80	0.91	**0.94**	3.49
Illumina HumanHap 550	JPT	0.78	0.90	**0.95**	4.48
Illumina HumanHap 550	YRI	0.52	0.83	**0.92**	10.63

2.3.2 Power Evaluation.

Although correlation is important in determining the power of a method, other factors such as the frequency of a causal SNP, number of individuals, disease model, prevalence, relative risk, and multiple hypothesis correction contribute to the overall power. In order to measure the increase in power in practice, we used

Table 2.2: Average r^2 obtained by the different methods. Each row contains average correlation coefficient for each genotyping platform and population used in this study with tags up to length 4 SNPs. The Tag Set column specifies the genotyping platform as the Affymetrix 500K or Illumina HumanHap 550 set. The $\%Inc.$ column shows the % increase in the average correlation coefficient when moving from the HAPs to WHAPs.

Tag Set	Pop	SNP	HAP	WHAP	$\%Inc.$
Affymetrix 500k	CEU	0.77	0.87	**0.91**	4.37
Affymetrix 500k	CHB	0.75	0.86	**0.91**	4.96
Affymetrix 500k	JPT	0.74	0.85	**0.90**	5.88
Affymetrix 500	YRI	0.59	0.79	**0.87**	9.17
Illumina HumanHap 550	CEU	0.92	0.97	**0.99**	1.26
Illumina HumanHap 550	CHB	0.86	0.95	**0.97**	2.42
Illumina HumanHap 550	JPT	0.86	0.94	**0.97**	2.77
Illumina HumanHap 550	YRI	0.71	0.91	**0.96**	4.84

the complete phased data for the ENCODE regions from NCBI [122] to simulate panels of 1000 cases and 1000 controls with a disease prevalence of 0.01, and relative risk of 1.5 For each SNP with MAF $\geq$ 0.05, we generated a panel in which the SNP is assumed to be the causal SNP. The total number of such panels was 32017, corresponding to the number of SNPs with MAF $\geq$ 0.05. We evaluated each statistic for these panels using the tag SNPs from the Affymetrix 500k and Illumina HumanHap 550 SNP sets in each region. For the HAP and WHAP tests, for every hidden SNP in the region, we found the tags with maximum correlation to that SNP by enumerating over all possible subsets of SNPs within a window of 100kb. We estimated p-values using a permutation test with 10,000 permutations in order to correct for multiple hypotheses. We consider a causal SNP as identified if its p-value adjusted for multiple hypothesis is less than 0.01. Table 2.3 presents the results of these power simulations. In order to illustrate the difference between multi-marker and our weighted haplotypes method, the table presents the average relative power taken over all ten ENCODE regions when compared to the ideal baseline situation in which we genotype every SNP. Comparing the power to genotyping every SNP helps remove bias caused by factors such as differing minor allele frequencies which are independent of correlation coefficient. As expected from the results of the correlation coefficient experiment, we observe that our method outperforms the multi-marker method.

2.3.3 Robustness to over-fitting.

Our method is based on the assumption that the linkage disequilibrium structure is consistent between the reference and case control panels. There are several reasons why this may not be the case and have the potential of limiting the power of of our method. First, it is not clear *a priori* whether the weights estimated from one population apply to another. In order to simulate discrepancies between the HapMap population and the case/control populations, we used the Han Chinese (CHB) genotype data to choose the best tags and estimate the weights of haplotypes while measuring the power (using the ρ-test) over simulations generated using the Japanese (JPT) population. For every hidden SNP in the region, we found the tags with maximum correlation to that SNP by enumerating over all possible subsets of SNPs within a window of 40kb in the CHB population. Using the Affymetrix 500k tags, the power of simulations using the JPT population was 74%, 76%, and 78% for the best SNPs, haplotypes and weighted haplotypes respectively obtained from the CHB population. Using the Illumina HumanHap 550 tags, the power of simulations using the JPT population was 83%, 88%, and 89% for the best SNPs, haplotypes, and weighted haplotypes respectively. Evidently, our method is not affected considerably by the difference in the population structure between the reference data set and the case-control populations.

Another complication may be the limited data size of the HapMap populations. Since the HapMap population is limited in size, there is the risk that the weights do not represent the true population haplotype frequencies but might be an artifact of over-fitting. In order to measure the effect of over-fitting on our results, we re-estimated the haplotype frequencies using only half of the individuals in the HapMap panels, and then measured the power on the rest of the individuals with weights derived from first half. As seen in Table 2.3, these two error source do not seem to affect our method considerably. If there was significant over-fitting, we expect power to drop significantly.

Table 2.3: Power simulations. Each row shows the power of HAP and WHAP tests relative to genotyping all SNPs averaging over all 10 ENCODE regions in simulated case control studies of 1000 cases and 1000 controls assuming a relative risk of 1.5. The Pop is the population used to generate the case and controls and find tags. The HAP and WHAP columns show the relative power of each method respectively. For any population marked with an "h", haplotype weights were estimated using only half of the individuals from the HapMap reference panel data and power was measured using simulations over the other half.

Tag Set	Pop	SNP	HAP	WHAP
Affymetrix 500k	CEU	0.92	0.94	0.96
Affymetrix 500k	CHB	0.90	0.94	0.95
Affymetrix 500k	JPT	0.90	0.93	0.95
Affymetrix 500k	YRI	0.77	0.88	0.92
Illumina HumanHap550	CEU	0.98	0.98	0.99
Illumina HumanHap550	CHB	0.95	0.97	0.98
Illumina HumanHap550	JPT	0.96	0.97	0.99
Illumina HumanHap550	YRI	0.86	0.95	0.96
Affymetrix 500k	CEUh	0.92	0.93	0.94
Affymetrix 500k	CHBh	0.90	0.91	0.91
Affymetrix 500k	JPTh	0.89	0.91	0.92
Affymetrix 500k	YRIh	0.77	0.87	0.90
Illumina HumanHap550	CEUh	0.96	0.97	0.98
Illumina HumanHap550	JPTh	0.96	0.96	0.96
Illumina HumanHap550	CHBh	0.95	0.96	0.96
Illumina HumanHap550	YRIh	0.87	0.95	0.95

In addition, if there was significant over-fitting, we would expect spurious correlation (high r_h^2 values) between weighted haplotypes and hidden SNPs due to the limited size of the HapMap populations. We measure the amount of spurious correlation by considering tag SNPs from all ENCODE regions as proxies for a random set of hidden SNPs from an ENCODE region on another chromosome. For each of the hidden SNPs, we found the best pair, triplet, and quartet of tag SNPs from other ENCODE regions, and the corresponding haplotype weights. In all cases no set of tag SNPs achieved a $r_h^2 > 0.5$ and the vast majority had very low r_h^2 which is evidence that our results are not due to over-fitting.

2.4 Discussion

r_h^2 and the ρ-test can be used as a natural criterion for tag SNP selection, according to a similar argument for which r^2 is currently used for tag SNP selection methods. Here, in contrast to previous methods, we suggest not to use the the LD between a specific haplotype and the causal SNP, but between a weighted combination of the haplotype and the SNP.

Intuitively, our method increases power over traditional multi-marker methods because the weighted haplotypes are effectively an unbiased estimate for the allele frequency of the hidden SNP. Traditional multi-marker methods are a biased estimate of this allele frequency and we believe that this discrepancy is the root cause of the difference in power. From this point of view, our approach is related to methods that attempt to predict the allele frequency of hidden SNPs[19, 118, 18]. In particular, our method has some similarities with the method proposed in Stram 2004[109] where the expectation of the hidden SNP is obtained from the haplotype frequencies with a block. However, our approach differs from the methods presented in Stram 2004[109, 117] because we do not rely on haplotype blocks and instead use the multi-marker tags that maximize the power of the indirect association (according to our analytic predictions), regardless of their location.

In this work we focused on the optimization of haplotype-based tests for association studies, when the set of genotyped SNPs (tag SNPs) is fixed. In cases where the tag SNPs are not fixed, it is also of interest to find a set of tag SNPs that will maximize the power of the study, when the genotyping is followed by the haplotype analysis suggested here. The design of such tag SNP selection algorithm is beyond the scope of this work, although it is likely that a greedy method such as the one used for TAGGER [38] would be a reasonable strategy to find such a set of SNPs.

The complete set of WHAP tags for the Affymetrix 500k and Illumina HumanHap 550 SNP sets as well as software for performing association tests are available upon request.

2.5 Preliminaries

Proof. For every haplotype $h \in \{0,1\}^{k-1}$ on the SNPs $s_1, \ldots, s_{k-1}$, we have that $w_{0,h} = \frac{w_{0,(h,0)}p_{(h,0)}+w_{0,(h,1)}p_{(h,1)}}{p_{(h,0)}+p_{(h,1)}}$, and $p_h = p_{(h,0)} + p_{(h,1)}$. Thus,

$$w_{0,h}^2 p_h \quad = \quad \frac{(w_{0,(h,0)}p_{(h,0)} + w_{0,(h,1)}p_{(h,1)})^2}{p_{(h,0)} + p_{(h,1)}} < w_{0,(h,0)}^2 p_{(h,0)} + w_{1,(h,1)}^2 p_{(h,1)},$$

and therefore,

$$R_{s_0,\ldots,s_{k-1}} = \frac{\sum_{h\in\{0,1\}^{k-1}} w_{0,h}^2 p_h - q_0^2}{q_0q_1} <$$
$$\frac{\sum_{h\in\{0,1\}^{k-1}}(w_{0,(h,0)}^2 p_{(h,0)} + w_{0,(h,1)}^2 p_{(h,1)}) - q_0^2}{q_0q_1} = R_{s_0,\ldots,s_k}$$

□

Note that in order to interpret the theorem correctly, one has to make sure that λ_2 can be approximated as a χ^2 distribution. This holds as long as the haplotype frequencies can be estimated with low error rate from the data. Thus, the theorem states that when k gets larger, we gain power by increasing R, but we actually may lose some power due to inaccuracies in haplotype frequency estimations. The latter is a caveat in any haplotype-based test. Since the accuracy of the haplotype frequency estimations has been shown empirically to be quite accurate over short regions [80], we expect that in practice the power of these test is maximized for haplotypes of short region (20kb-100kb).

2.5.1 Redefinition of R.

It is easy to see that the value of R can be written as

$$R_{s_0,\ldots,s_k} = \frac{\sum(x_{(0,h)} - q_0p_h)^2}{q_0q_1},$$

where p_h is the probability for the haplotype h (where h is taken over all SNPs except for s_0), $x_{(0,h)}$ is the frequency of $(0,h)$, and q_0, q_1 are the allele frequencies of s_0. This definition resembles the definition of r^2.

2.6 The Continuous Case

In order to improve the power of a study we want to use the HapMap linkage structure to develop statistical tests for SNPs not found on the genotyping platform. The above version of the WHAP test was designed to do this in the case control study setting using a χ^2 like test. In this continuous setting genotype dosages for every HapMap SNP are estimated for each individual and an additive linear model is used to generate a statistical test. In addition to providing a means of studying continuous data, the estimated dosages can be used in a logistic regression framework for case control data. This has the advantage of allowing correction for covariates such as population substructure, age, sex, and other confounding variables. This method was developed as part of a study on metabolic phenotypes in a Finnish Population [100].

Weighted Haplotype Association (WHAP) is a statistical method to test association between a trait of interest and untyped HapMap SNPs by relying on haplotypes of genotyped markers. We extended WHAP to deal with continuous phenotypes and we introduced new quality control procedures that we briefly describe as follows:

Let M be an untyped SNP with major allele A and minor allele a. Let S_M be a set of genotyped SNPs $s_1, s_2, \ldots, s_k$ which are shared between HapMap and the SNPs in the study and that are informative about M (we will discuss later how this set is selected). Let H=h be the set of haplotypes observed in HapMap for the SNPs in S_M. Each of these haplotypes can be extended to include M, leading to two possible haplotypes:h_a carrying the minor allele a and h_A carrying the major allele A. Let ph_a and ph_A be the respective frequencies of these haplotypes in HapMap. For each haplotype $h \in H$ over the set of genotyped SNPs S_M we can then calculate the conditional probability of a minor allele a at the untyped SNP M:

$$P(a \mid h) = \frac{ph_a}{ph_a + ph_A} \tag{2.1}$$

Assuming that the study population is similar to the HapMap population, we can evaluate the expected minor allele count for SNP M, in each genotyped individual, using the conditional probabilities above. Specifically, let h_{1i} and h_{2i} be the two haplotypes for individual i over the SNPs in S_M. Then the expected minor allele count at SNP M for individual i is

$$\hat{C}_i = P(a \mid h_{1i}) + P(a \mid h_{2i}) \tag{2.2}$$

To test for association between the expected minor allele count $\hat{C}$ and the continuous phenotype Y we use a standard linear regression with the phenotypes as the response variables:

$$Y_i = \hat{C}_i \beta + \varepsilon_i \tag{2.3}$$

We evaluate the significance of the association of SNP M to the phenotype using the p-values from the F-test of this regression.

For each untyped SNP M the set SM is selected by searching, within a window of 40kb around M, the collection S, of at most 3 SNPs, which maximizes the following measure:

$$r^2_{M,S} = \frac{\sum_{h \in H} P(a \mid h)(ph_a - p_a p_h)}{p_a(1 - p_a)} \tag{2.4}$$

where H is the haplotype collection for the SNP in S, p_h is the frequency of haplotype h, and p_a is the minor allele frequency for M. This measure takes on values between 0 and 1 and was introduced by Nicolae [85] and Zaitlen et al. [121].

The value of r^2_{M,S_M} (correlation of SNP M with its best tag set S_M) provides a measure of how well the SNP can potentially be imputed. In the present study, we have imputed only SNPs for which the value of r^2_{M,S_M} was larger then 0.7.

Making sure that $r^2_{M,S_M} > 0.7$, however, only guarantees that SNP M can be reasonably well imputed using a population identical to HapMap. The specific population under study, however, can have a haplotype distribution substantially different from those observed in HapMap. To avoid spurious results under this circumstance, we adopt the following criteria.

1. If the haplotype h observed for one individual across the SNPs in S is not observed in HapMap, we do not estimate the allele count for that individual.

2. If we observe more then 10% of individuals with haplotypes not in HapMap, we do not estimate allele counts for the SNP M .

Chapter 2, was published in The American Journal of Human Genetics, Vol 80, pp 683-91, 2007. Noah A. Zaitlen, Hyun Min Kang, Eleazar Eskin, and Eran Halperin, “Leveraging the HapMap correlation structure in association studies”. The dissertation author was the primary investigator and author of this paper.

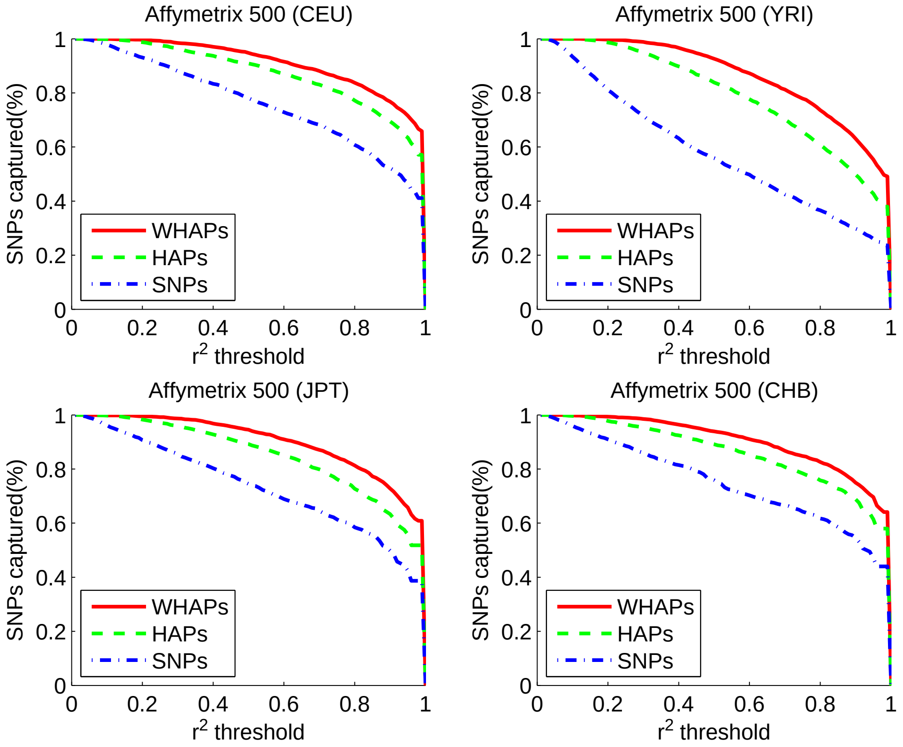

Figure 2.2: Fraction of SNPs captured by each of the methods, when tested on the marker set of the Affymetrix 500k array. This figure shows the fraction of SNPs with minimum allele frequency (MAF) $\geq 5\%$ that are captured by a marker SNP, haplotype (HAP), or weighted haplotype (WHAP). The notion of a hidden SNP being captured depends on the r^2 between the proxy and the SNP. For each of the figures, the x-axis represents the r^2 threshold, and the y-axis represents the fraction of common SNPs with r^2 greater than the threshold. The three lines correspond to single SNPs, HAPs, and WHAPs. The populations are the four ENCODE panels consisting of European Ancestry (CEPH), Yoruba people of Ibadan, Nigeria (YRI) , Han Chinese (CHB), and Japanese (JPT). Evidently, WHAPs significantly outperform both SNPs and HAPs over any platform and population, but do especially well in populations with more complex LD structure such as YRI.

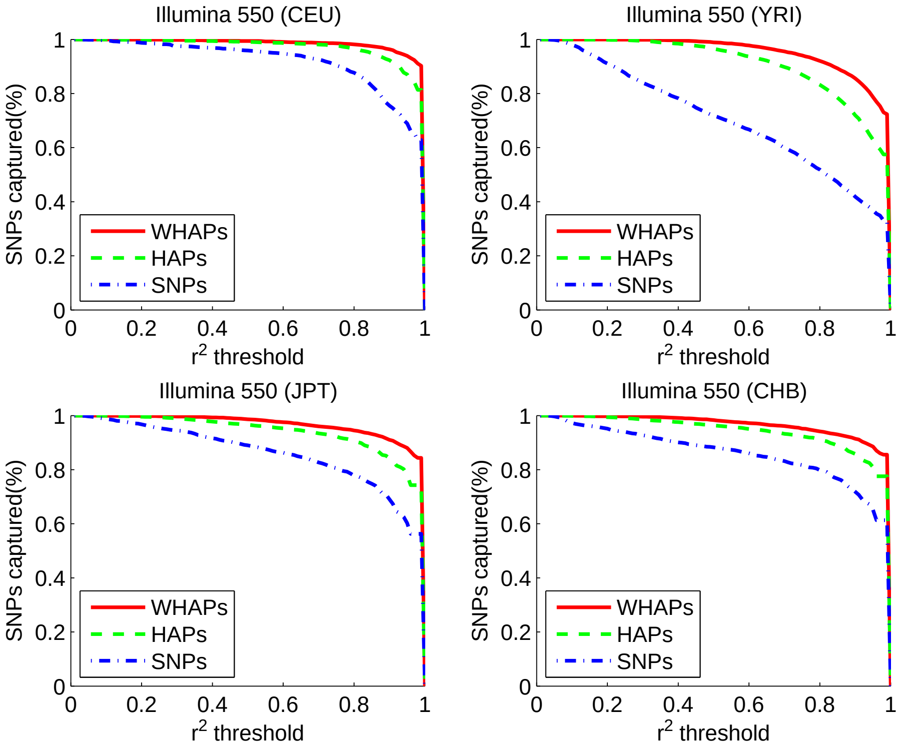

Figure 2.3: Fraction of SNPs captured by each of the methods, when tested the the marker set of Illumina HumanHap 550 BeadArray. This figure shows the fraction of SNPs with minimum allele frequency (MAF) $\geq$ 5% that are captured by a marker SNP, haplotype (HAP), or weighted haplotype (WHAP). The notion of a hidden SNP being captured depends on the r^2 between the proxy and the SNP. For each of the figures, the x-axis represents the r^2 threshold, and the y-axis represents the fraction of common SNPs with r^2 greater than the threshold. The three lines correspond to single SNPs, HAPs, and WHAPs. The populations are the four ENCODE panels consisting of European Ancestry (CEU), Yoruba people of Ibadan, Nigeria (YRI) , Han Chinese (CHB), and Japanese (JPT).

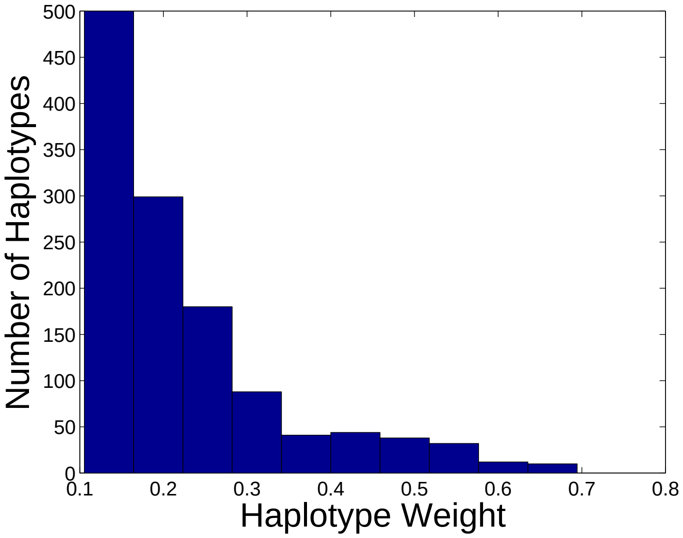

Figure 2.4: Histogram of the distribution of haplotype weights for SNPs in which weighted haplotypes provide a better proxy than a single haplotype or a single SNP. The weight distribution was generated from the CEPH population over ENCODE region ENm010.

Chapter 3

Single Nucleotide Polymorphism Tag Selection by Reduction to SAT

3.1 Introduction

Whole genome association is a powerful method for discovering the genetic basis of human diseases. Recently, it has been successfully employed to reveal novel loci correlated with risks for diseases including coronary artery disease, bipolar disorder, type 1 and type 2 diabetes, amongst many others[26]. In brief, a subset of all single nucleotide polymorphism (SNP) markers is genotyped in case and control populations. The distribution of each SNP's genotypes is compared between the populations via a statistical test in order to identify loci associated with the altered risk for the disease.

Even with the tremendous technological advances that have driven down the cost of collecting SNP genotypes, collecting all known SNPs is prohibitively expensive. Genetic association studies take advantage of the fact that genotypes at neighboring SNPs are often in linkage disequilibrium (LD) or are correlated with each other. This correlation allows for "indirect association" where a causal SNP is detected not by genotyping the SNP directly, but instead by genotyping a "tag" SNP or SNP that is correlated with the causal SNP. The availability of reference data sets such as those provided by the HapMap project[25] allow for us to measure the linkage disequilibrium patterns between SNPs. Naturally, using this information to determine which SNPs to select as tags is a central problem in designing association studies and has been extensively studied [52]. It is commonly referred to as the *tag SNP selection* problem.

Research on the tag SNP selection problem can be roughly split into two categories: the statistical criteria used for selecting tag SNPs and the algorithms for choosing a tag set given this statistical criteria. Many statistical criteria have been proposed for tag SNP selection [52]. The most popular criterion considers the square correlation coefficient r^2 between SNPs. Under this formulation of the tag SNP selection problem or the *Single SNP r^2 tag SNP selection* problem, the goal is to choose a subset of the SNPs as tags such that each SNP not selected in the tag set has an r^2 value with a tag SNP above a minimal threshold. Relatively few algorithmic approaches have been proposed for this problem with the greedy approach being the most widely used approach[18]. This form of the problem is NP-complete [4] and it was widely believed that an efficient optimal solution capable of whole genome scale data would not exist [52]. Several existing methods address this issue by various heuristics and/or alteration of the problem definition. Halldorson et al. [53] restrict LD patterns to a window and FESTA [94] solves the problem by partitioning the SNPs into precincts which do not have any linked SNPs in them and then exhaustively enumerating the solutions within the precinct.

In this work we present a solution to the tag SNP selection problem which can discover all optimal solutions efficiently and can scale to the whole genome. Our method encodes the tag SNP selection problem as an instance of the satisfiability (SAT) problem. Here, our SAT instances are clauses in conjunctive normal form (CNF) where a variable assigned to `true` corresponds to the inclusion of a SNP into the tag set.

A satisfying assignment of variables to truth values in the SAT instance yields a valid solution to the tagging problem (and vice versa). As we shall clarify, a "minimal" satisfying assignment yields an optimal solution to the tagging problem.

We compare the results of our method on the single SNP r^2 tag SNP selection problem to FESTA [94] and Halldorson's method [53] over the Encode [25] data set. We also demonstrate that our methods scales to the whole genome HapMap data. Consistent with previous studies, optimal solutions for the tag SNP selection method are only slightly more efficient than greedy solutions[94]. One advantage of our framework is that we can enumerate the entire set of optimal solutions in polynomial time allowing for flexible designs. Another advantage is that it extends to more challenging variants of the tag SNP selection problem. We provide the first optimal solution for selecting multi-marker tags. Multi-marker tags have been shown to significantly increase the power of association studies[90, 121]. We also demonstrate an approach for choosing the best-N tags to optimize power. This is a more challenging problem because when the budget of tag SNPs is fixed, increasing the number of SNPs in one region, requires removing a SNP in another region which creates long range dependencies in the optimization problem. We show how this problem can be reduced to Max-SAT from which we can obtain optimal solutions.

3.2 Methods

We present methods for optimally solving several variations of the *tagging* problem of selecting a subset of tag SNPs to be genotyped as part of an association study from a larger set of known SNPs.

First we show how to optimally solve the local single SNP r^2 tagging problem in which we search for a minimal set of tag SNPs which cover the remaining SNPs in a region of the genome with an r^2 above some minimum threshold. Second, we present a method for combining our optimal solutions in local regions to an optimal solution for the entire genome. Third, we extend our solution to multi-marker tags or tags which combine two or more SNPs. The use of multi-marker tags can significantly reduce the number of tags which need to be collected in order to cover a region, but the optimization procedure is much more difficult. Finally, we show how to optimally solve the best-N tagging problem variant where the number of tag SNPs is fixed (in this case N) and a function such as power, is maximized over the set of SNPs. This formulation of the problem can significantly increase the statistical power of a tag set, but raises additional computational challenges. Since the number of SNPs is fixed, adding an additional SNP in one region of the genome requires removing a SNP in another region of the genome which introduces long range dependencies in the problem.

3.2.1 Local Single SNP r^2 Tagging

Let $S = \{s_i\}_{i=1}^n$ be a set of SNPs. We say SNP s_i "covers" SNP s_j if their correlation coefficient r^2, exceeds some threshold r^2_{min}. If $T' \subseteq S$ and $\forall s_j \in S \exists s_i \in T$ such that $r^2_{ij} \geq r^2_{min}$ we call T' a *valid* cover of S. Our goal is to select the smallest set $T' \subseteq S$ that is a *valid* cover of S.

Consider the example in Figure 1, where we have 6 SNPs $s_1, \ldots, s_6$, and the pairwise r^2 values described in the table in Figure 1(a). Suppose that we have the threshold $r^2_{min} = 0.8$. We can represent the SNPs as the graph shown in Figure 1(b) where an edge denotes an r^2 above the minimum threshold. The standard greedy algorithm [18, 38] picks tag SNPs by repeatedly selecting the SNP with the largest number of uncovered neighbors. We can easily see that there are two optimal solutions, $T = \{s_4, s_2\}$ and $T = \{s_4, s_1\}$. Note that one greedy solution will select SNP s_3 in the first step resulting in a non-optimal solution $T = \{s_3, s_1, s_6\}$. Our solution to the tag SNP selection problem will characterize all optimal solutions in a compact DAG, which happens to be a tree in the example of Figure 1(c).

We shall reduce the problem of identifying a valid selection of SNPs to the problem of identifying a satisfying assignment to a propositional sentence in conjunctive normal form (CNF). In particular, we want a sentence in CNF where satisfying assignments correspond to a valid selection of SNPs. We create a literal for every SNP and a clause for every SNP consisting of literals that can cover that SNP.

Given a threshold r^2_{min}, consider a sentence in CNF: $\Phi = \phi_1 \wedge \cdots \wedge \phi_n$ with

(a)

r_{ij}^2	s_1	s_2	s_3	s_4	s_5	s_6
s_1	1.0	0.9	0.5	0.4	0.2	0.4
s_2	0.9	1.0	0.9	0.5	0.3	0.2
s_3	0.5	0.9	1.0	0.9	0.8	0.1
s_4	0.4	0.5	0.9	1.0	0.9	0.8
s_5	0.2	0.3	0.8	0.9	1.0	0.5
s_6	0.4	0.2	0.1	0.8	0.5	1.0

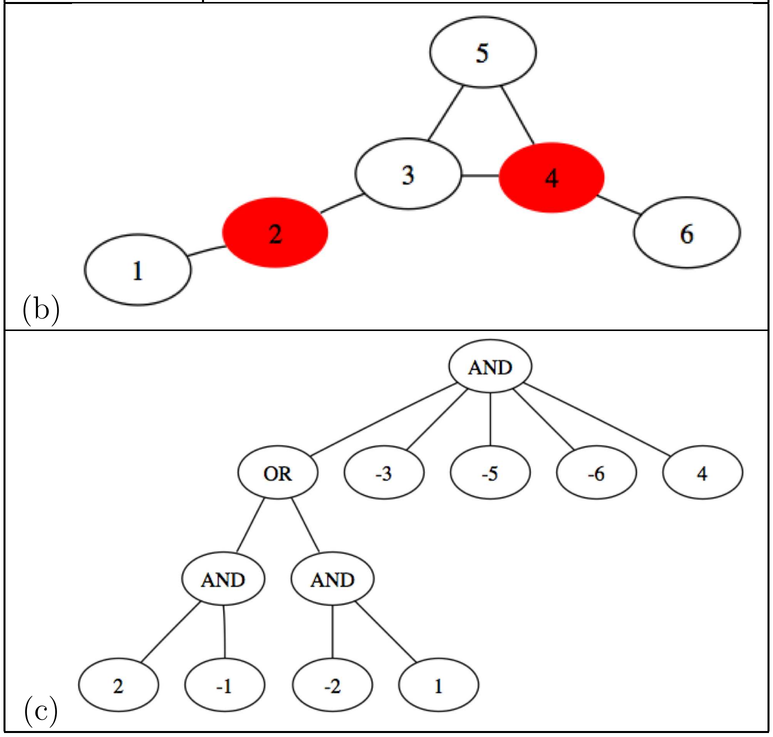

Figure 3.1:]

(a) Single SNP r^2 table (b) Graph of cover problem (c) NNF equivalent to CNF as many clauses ϕ_i as there are SNPs s_i, where each clause is of the form:

$$\phi_i = \bigvee_{r_{ij}^2 \geq r_{min}^2} s_j$$

Each SNP $s_j \in S$ is a positive literal in the CNF sentence Φ, and appears in clause ϕ_i if and only if SNP s_j can cover SNP s_i. A valid selection T' of SNPs then corresponds precisely to a satisfying assignment of Φ.

In order to find a *minimally* valid selection T of SNPs, we seek a minimum cardinality model of our propositional sentence, where a minimum cardinality model is a satisfying assignment with a minimal number of positive literals.

Consider the example in Figure 1 with six SNPs $s_1, \ldots, s_6$. Given the threshold $r_{min}^2 = 0.8$ we have the following CNF formula:

$$\begin{gathered}(s_1 \vee s_2) \wedge (s_1 \vee s_2 \vee s_3) \wedge (s_2 \vee s_3 \vee s_4 \vee s_5) \\ \wedge (s_3 \vee s_4 \vee s_5 \vee s_6) \wedge (s_3 \vee s_4 \vee s_5) \wedge (s_4 \vee s_6)\end{gathered}$$

We have two minimum cardinality models, $(\neg s_1, s_2, \neg s_3, s_4, \neg s_5, \neg s_6)$ and $(s_1, \neg s_2, \neg s_3, s_4, \neg s_5, \neg s_6)$, corresponding to our two minimally valid selection of SNPs.

Not surprisingly, identifying a minimum cardinality model for a given sentence in CNF is also an NP–hard problem. Our approach is based on converting our sentence Φ in CNF into a logically equivalent sentence Δ in *decomposable negation normal form* (DNNF) [29, 30, 32, 33]. DNNF is a logical representation that allows queries, that are in general intractable, to be computed in time polynomial in the size of the DNNF sentence. For example, if a conversion from CNF to DNNF does indeed result in a sentence of manageable size, we can efficiently test whether the original sentence is satisfiable, enumerate its models, and identify another sentence in DNNF that characterizes all its minimum cardinality models. By enumerating the models of the resulting sentence, we can enumerate all of the minimally valid selections of SNPs. In general, there are no guarantees that a CNF can be converted to a DNNF of reasonable size, but we demonstrate that for the tag SNP selection problem, due to the inherent local structure of the problem, our approach is tractable.

This conversion is performed here by the C2D compiler, which compiles CNF instances into DNNF [16].C2D further enforces the *determinism* property, and more specifically, compiles CNF instances into d-DNNF. The C2D compiler has already been successfully employed in a number of other applications, serving as a backbone reasoning system in support of higher level tasks. For example, C2D was used as the backbone for planning systems [12, 61, 51], for diagnostic systems [41, 8, 104, 62, 7], for probabilistic reasoning [115, 21, 101, 20], and for query rewrites in databases [119]. In each one of these applications, high level reasoning problems were encoded into CNF, which was compiled into DNNF by C2D. The resulting compilation was then used to solve the original problem by posing polytime queries to it.

Decomposable negation normal form

A negation normal form (NNF) is a rooted directed acyclic graph in which each leaf node is labeled either by a literal (say i for a positive literal, and $-i$ for a negative literal), or simply by true or false. Each internal node is labeled with a conjunction ($\wedge$ or AND) or a disjunction ($\vee$ or OR); Figure 3.1(c) depicts an example. A negation normal form is decomposable (DNNF) if it satisfies the *Decomposability* property: for each conjunction in the NNF, the conjuncts do not share variables. The NNF in Figure 3.1(c) is also in DNNF.

If we are able to efficiently compile a CNF instance into DNNF, many queries are straightforward to compute [29]. For example, we can test if a DNNF sentence is satisfiable by simply traversing the graph bottom-up, while visiting children before parents. If a leaf node is labeled by a literal (i or $-i$), or true, then it is satisfiable; otherwise, it is unsatisfiable (labeled with false). An OR node is satisfiable iff any of its children are satisfiable, while an AND node is satisfiable iff all of its children are satisfiable. We can compute the minimum cardinality of a sentence and enumerate its models in a similar way [29, 34].

Before we proceed to describe how to compile a sentence in CNF into DNNF, consider the *conditioning* of a sentence Δ on an instantiation α, denoted $\Delta \mid \alpha$. This operation yields a sentence that can be obtained by replacing every literal in

Δ with true (respectively, false) if it is consistent (inconsistent) with instantiation α. For example, conditioning the DNNF $(\neg a \wedge \neg b) \vee (b \wedge c)$ on instantiation $b \wedge d$ gives $(\neg a \wedge \mathsf{false}) \vee (\mathsf{true} \wedge c)$. Note that DNNF is closed under conditioning, i.e., conditioning a DNNF Δ on an instantiation α results in another DNNF. Moreover, the resulting sentence $\Delta \mid \alpha$ does not mention variables assigned by α.

Consider now the following theorem, proved in [29], which motivates the compilation procedure underlying C2D.

Theorem 2 (Case Analysis)**.** *Let Δ_1 and Δ_2 be two sentences in* DNNF*, and let Δ be the sentence $\bigvee_\alpha (\Delta_1 \mid \alpha) \wedge (\Delta_2 \mid \alpha) \wedge \alpha$, where α are instantiations of variables mentioned in both Δ_1 and Δ_2. Then Δ is in* DNNF*, and is equivalent to $\Delta_1 \wedge \Delta_2$.*

This theorem suggests a recursive algorithm $\mathsf{DNNF1}(\Phi)$ that converts a sentence Φ in CNF into a sentence Δ in DNNF:

1. If Φ contains a single clause ϕ, return $\mathsf{DNNF1}(\Phi) \leftarrow \phi$. Note that a clause is vacuously decomposable.

2. Otherwise, return

$$\mathsf{DNNF1}(\Phi) \leftarrow \bigvee_\alpha \mathsf{DNNF1}(\Phi_1 \mid \alpha) \wedge \mathsf{DNNF1}(\Phi_2 \mid \alpha) \wedge \alpha,$$

   where Φ_1 and Φ_2 is a partitioning of clauses in Φ, and α is an instantiation of the variables mentioned in both Φ_1 and Φ_2.

We can see that this procedure gives us the decomposability property, but at the expense of increasing the size of the original sentence. This increase is incurred primarily due to the case analysis performed, and the extent of this increase is sensitive to the way we decide to partition the clauses of the input sentence Φ. In particular, we would want to minimize the number of common variables between Φ_1 and Φ_2, as the complexity of case analysis is exponential in this number.

Partitioning can be guided by decomposition trees, or simply d-trees [29]. A d-tree $\mathcal{T}$ for a CNF Φ is a binary tree whose leaves correspond to the clauses in Φ. An example d-tree for the CNF given used for the example in Figure 3.1 is shown in Figure 3.2. Intuitively, the above compilation procedure traverses the

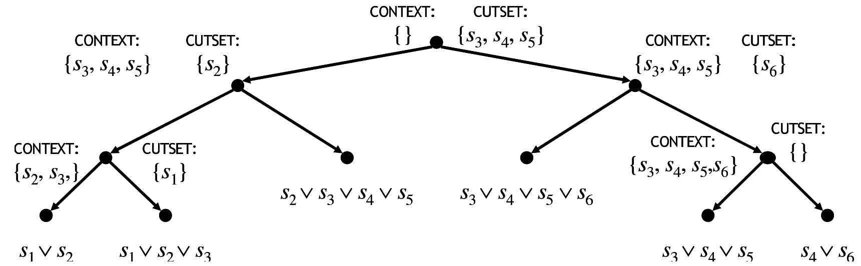

Figure 3.2: A d-tree for the CNF used for the example in Figure 3.1. Internal nodes are labeled with their contexts and cutsets.

d-tree, starting from the root, where case analysis is performed based on how the d-tree partitions the clauses of the CNF Φ. In particular, each interior node t is associated with the set of clauses that appear below it, and the partition is determined by the clauses of t's left and right children.

As we can see in Figure 3.2, each internal node is labeled with two variable sets: the *cutset* and the *context*. At a given node t, the cutset tells our compilation algorithm which variables to perform case analysis on. The context tells us those variables that appear in both of t's children, but have already been instantiated for case analysis by an ancestor. An instantiation α of the context variables can then be used as a key for a cache that stores the results of compiling the subset of clauses $\Phi \mid \alpha$. When a node is revisited with the same context, then the algorithm can simply return the DNNF sentence $\Delta \mid \alpha$ already computed. For example, at the root of the tree, the cutset contains $\{s_3, s_4, s_5\}$ since those variables appear in both children. If we follow the left branch twice, the context is now $\{s_2, s_3\}$, which was instantiated by the root and its left children. Note that this node will be visited multiple times for different instantiations of $\{s_4, s_5\}$, but only different instantiations of the context yield different subproblems. Thus, when this node is revisited with the same context instantiation, we simply fetch the result from the cache. It is this subformula re-use that allows compilation to moderate the exponential growth of the formula caused by case analysis.In particular, the complexity of compilation can be bounded in terms of the size of the context and the cutset [29]. The C2D compiler, while based on this approach, employs more advanced techniques to further improve on the efficiency of compilation [31].

Scaling to Whole Genome Tagging

The c2d compiler is capable of computing minimal tag sets, for several thousands of SNPs. Unfortunately, memory becomes an issue when we try to compile even larger regions of the genome. To encode the entire genome as a CNF, however, we must use 3.8 million literals and 3.8 million clauses. Clearly, we need new techniques to scale to this problem size. One idea is to compile sufficiently small regions of the genome into DNNF, which are then "stitched" together to construct a DNNF for the entire genome. We will then need to specify two things:

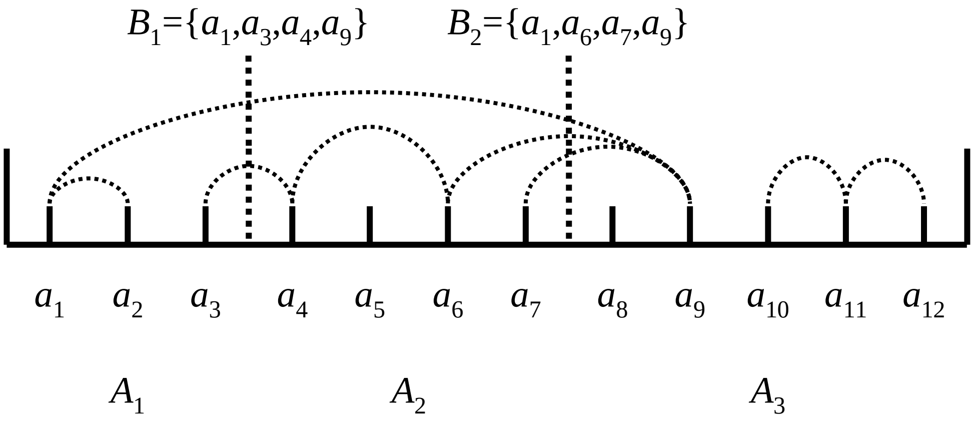

Figure 3.3: Partitioning the SNPs. We draw an edge between SNP (a_i) and (a_j) if (a_i) and (a_j) can cover a common SNP.

Consider Figure 3.3 where we have 17 SNPs in a region, and where we have drawn an edge between two SNPs if their r^2 exceeds a minimum r^2 threshold. We have further partitioned the region into 5 Siberians, $A_1, \dots, A_5$, which represent also a partitioning of the CNF Φ into corresponding sets of clauses $\Phi_1, \dots, \Phi_5$. Moreover, we have boundary sets $B_1, \dots, B_4$ that record the SNPs that interact across two halves of Φ. For example, B_2 contains all variables mentioned by both Φ_1, Φ_2 and Φ_3, Φ_4, Φ_5.

We could compile into DNNF each Φ_i in isolation, but the interaction among them prevents us from simply conjoining the result: it will neither be in DNNF, nor

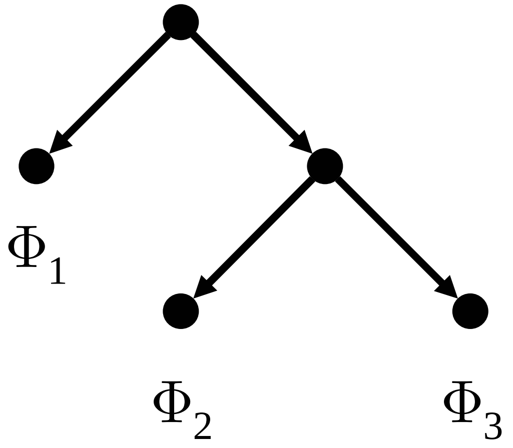

Figure 3.4: A d-tree with DNNF sentences for leaves.

will it represent the minimal models. We shall instead run a compilation algorithm, similar to the one given previously, except that it will operate on a d-tree whose leaves are assigned DNNF sentences (rather than just clauses); see Figure 3.4. If Δ_i are the respective DNNF sentences compiled by C2D, we can perform a similar compilation process, external to C2D. At an interior node with a child containing Δ_i, the boundaries guide us in case analysis: B_{i-1} is the context and B_i is the cutset. Going further, we can prime the compiler caches by precomputing the sentences $(\Delta_i \mid B_{i-1} \cup B_i)$. Intuitively, we are using these sentences as pieces of the desired DNNF, that are stitched together in a process guided by the compilation algorithm. Note that to perform this stitching, we need not keep all pieces in memory, as compilation (and C2D) normally would; this allows us to reconstruct the full DNNF. It is also possible to keep the sentence implicit, where queries are implemented to handle individual pieces independently.

When we want to compile into DNNF an instance where we have many partitions (as we would have for the entire genome), the d-tree extends linearly; see again Figure 3.4. By [29], we know further that the size of the resulting compilation is bounded. Suppose we have partitioned the instance into n regions. Let $w_i = |B_{i-1} \cup B_i|$ denote the size of the boundaries surrounding region A_i, and let $w = \max_i w_i$. A region A_i needs to construct 2^{w_i} pieces, so if every piece

$(\Delta_i \mid B_{i-1} \cup B_i)$ has size $O(S)$, then we can bound the size of the resulting DNNF instance by size $O(n2^w S)$.

Clearly, when we decide how to partition the instance, we need to identify boundaries with as few variables as possible. Upon searching for such boundaries in the genome for the single SNP r^2 tagging problem, we discovered that surprisingly there were enough boundaries where no interactions occur, and where C2D was sufficient to compile the resulting regions. In this case, each region is indeed independent, and we can simply conjoin each region without the need for the above "stitching." The maximum number of linked SNPs was 1012 in the CEU population which took under 1 minute to compile with C2D. However for more complicated variants of the tagging problem such as those described below, there are more interactions and independent regions are unlikely to exist.

3.2.2 Multi-Marker SNP Tagging

Recent work has shown that using statistical tests based on haplotypes over multiple SNPs improves the power of whole genome association studies[90, 121]. In the context of tagging, this permits combinations of tag SNPs (multi-marker tags) to cover a SNP, allowing for a smaller set of tags to cover the SNPs.

In this situation, an even smaller set T' may be a valid cover of SNPs.

Again, we reduce the problem of identifying a valid set of SNPs to satisfiability. Given a threshold r^2_{min}, we now have two classes of clauses Φ and Ψ. Clauses Φ as above enforces constraints that require each SNP in S is covered: $\Phi = \phi_1 \wedge \cdots \wedge \phi_n$, where there are as many clauses ϕ_i as there are SNPs s_i, but where each clause is now of the form:

$$\phi_i = \Big(\bigvee_{r^2_{j \to i} \geq r^2_{min}} s_j \Big) \vee \Big(\bigvee_{r^2_{j,k \to i} \geq r^2_{min}} p_{jk} \Big).$$

In this case, either a positive literal s_j representing SNP s_j or a positive literal p_{jk} representing a SNP pair (s_j, s_k) can also satisfy clause ϕ_i and cover SNP s_i.

Clauses Ψ enforce the constraints that if a pair literal p_{jk} is `true`, then both

s_j and s_k are true (i.e. in the tag set):

$$\Psi = \bigwedge_{r^2_{j,k\to i} \geq r^2_{min}} (p_{jk} \equiv s_j \wedge s_k)$$

where $p_{jk} \equiv s_j \wedge s_k$ are equivalence constraints that ensure that if p_{jk} is selected, the corresponding pair (s_j, s_k) is also selected (and vice versa). In clausal form, this equivalence constraint is given by three clauses: $\neg p_{jk} \vee s_j, \neg p_{jk} \vee s_k$ and $p_{jk} \vee \neg s_j \vee \neg s_k$.

Consider the example from the previous section with six SNPs $s_1, \ldots, s_6$. Suppose that the pair (s_1, s_3) can cover s_6, and that $r^2_{1,3\to 6} = 0.9$. Given the threshold $r^2_{min} = 0.8$, we gain a third optimal solution s_1, s_3, to go with the two solutions s_4, s_2 and s_4, s_1 from before. Encoding the problem, we have the following formula:

$$\begin{aligned} (s_1 \vee s_2) \wedge (s_1 \vee s_2 \vee s_3) \wedge (s_2 \vee s_3 \vee s_4 \vee s_5) \\ \wedge (s_3 \vee s_4 \vee s_5 \vee s_6) \wedge (s_3 \vee s_4 \vee s_5) \wedge \\ (s_4 \vee s_6 \vee p_{13}) \wedge (p_{13} \equiv s_1 \wedge s_3) \end{aligned}$$

We again want a minimum cardinality assignment, but minimizing only the number of positive s_i literals. We can introduce constraints $p_{jk} \equiv \neg q_{jk}$ to cancel out the contribution of the p_{jk}'s to the cardinality with the q_{jk}'s; we can then convert to DNNF as before. We can also existentially quantify out the p_{jk} literals; this operation is also supported by C2D [16].

3.2.3 Best-N Tag SNP Selection using a Penalty Term

While r^2 is a reasonable measure of correlation between a tag set and the full set of SNPs, a tag that achieves the min r^2 is not necessarily the set that maximizes the overall statistical power. Another variant of the tag SNP selection problem is where we have a fixed budget of tag SNPs and we want to obtain the best tag set for a given objective function defined over the tag set such as statistical power. The solution we present below for this problem can be applied to a large class of objective functions including statistical power. In this problem, each pair of SNPs is assigned a weight, and the goal is to maximize the sum of the maximum weight to any tag SNP for each SNP in the region. These weights can be set so that the problem is maximizing statistical power. However, due to space limitations, we instead describe our method in the context of maximizing the average r^2 between each SNP and the the best tag in our tag set given a fixed budget of tag SNPs.

For a given tag set T, we say that the r^2 for a given SNP in S is the maximum r^2 of any SNPs in T and we compute the average. This problem does not have the local structure property of single SNP r^2 tag SNP selection problem.

We must now keep track of which tag SNP is used to cover each SNP in the region. Let s_i and s_j be literals for two SNPs, and let c_{ji} be the literal for the event that SNP s_j covers SNP s_i. Note that c_{ij} is a different event from c_{ji}. Each c_{ji} is associated with a weight w_{ji} (e.g. the r^2 between s_i and s_j), and we want to maximize the average weight of the c_{ji}'s that are selected (equivalently, the sum $\sum_{i,j} w_{ij}c_{ij}$) given that we are selecting a maximum of N s_i's and each SNP can be covered by only one other SNP. For a consistent instantiation α of the literals s_i, c_{ij} where there are x positive literals s_i, we can write this objective function as $W_{\phi(x)}(\alpha)$, the weight of assignment α with respect to formula ϕ. $W_{\phi(x)}(\alpha) = \sum_{i,j} w_{ij}c_{ij}$ where the number of tags (positive literals s_i) is x. Our goal is to find the consistent instantiation α such that $W_{\phi(x)}(\alpha)$ is maximized.

We can reduce this problem to an instance of weighted Max-SAT. The maximum satisfiability problem (Max-SAT) is one of the optimization counterparts of the Boolean satisfiability problem (SAT). In the weighted version of this problem, every clause in the CNF formula is associated with a positive weight. The weight of any complete assignment is defined as the sum of the weights of the clauses that it satisfies. The weighted Max-SAT problem asks for a complete assignment that has a maximal weight.

The Max-SAT community is a fast growing research community with two international solver evaluations in the last two years [2, 3]. The Max-SAT problem has also been used as a model for solving many problems, in diverse areas such databases, FPGA routing, and automatic scheduling. As a result, many Max-SAT solvers have been developed in recent years (e.g., [113, 111, 75, 92, 1, 57, 73]).

Virtually all state-of-the-art complete Max-SAT solvers are based on depth-first branch-and-bound search in the space of all possible complete assignments. At every search node, which corresponds to a partial assignment, the solver compares the best seen weight against a bound computed for every completion of the current partial assignment. The solver prunes the current branch as soon as the bound becomes worse than the best seen weight. The method used for computing bounds varies among solvers. In this work, we used Clone [92], which computes its bounds using a relaxed version of the input problem, after having compiled it into a d-DNNF formula (by C2D only once at the beginning). At each search node, Clone computes a bound by operating on the compiled d-DNNF in time linear in its size.

Our weighted Max-SAT model $\Psi(p)$ has a number of mandatory constraints (clauses with infinite weight) which enforce the consistency of our solutions. First, each SNP may be only covered by either itself or only one other SNP. For each pair of SNPs s_j and s_k that can cover s_i, we add the constraint $c_{ji} \Rightarrow \neg c_{ki}$, or equivalently $\neg c_{ji} \vee \neg c_{ki}$. Second, if an SNP s_j covers s_i c_{ji} is TRUE, we must select SNP s_j, i.e., $c_{ji} \Rightarrow s_j$, or equivalently $\neg c_{ji} \vee s_j$. We also include weighted clauses consisting of individual literals. Each c_{ji} appears as a clause given a weight w_{ji}. Each s_i appears as a clause with a weight p corresponding to a fixed penalty per SNP. Each valid instantiation of the literals α will have weight $W_{\psi(p)}(\alpha)$ the weight of assignment α with respect to formula ψ, and $W_{\psi(p)}(\alpha) = \sum_{i,j} w_{ij} c_{ij} - p \sum_i s_i$.

Theorem 3. *For any maximal solution α for $\Psi(p)$ that contains x positive literals s_i, α is also a maximal solution for $\Phi(x)$.*

By contradiction, assume that there is an instantiation α' with x positive literals s_i such that $W_{\phi(p)}(\alpha') > W_{\phi(p)}(\alpha)$. Since $W_{\psi(p)}(\alpha) = W_{\phi(p)}(\alpha) - px$ and $W_{\psi(p)}(\alpha') = W_{\phi(p)}(\alpha') - px$, then $W_{\psi(p)}(\alpha') > W_{\psi(p)}(\alpha)$ which contradicts that α is the maximal solution.

We can use this result to find the maximum of the first term (the objective function) given a fixed number of tags. The penalty term lets us adjust the trade off between the objective function and the number of tags selected. For each value of the penalty term, the solution will be an optimal solution given that number SNPs in the solution. We can repeatedly solve the optimization problem with different penalty terms until we obtain a solution with the correct number of SNPs. Figure 5 illustrates the basis of the optimization procedure. In Figure 5(a), the curved line is the objective function and the straight lines are the penalty terms. Each maximum point of the difference between the objective function and penalty terms (Figure 5(b)) correspond to a optimal solution for a different number of SNPs. By varying the penalty weight, we can obtain solutions for any number of SNPs.

3.2.4 Best-N Tag SNP Selection using an Adder Circuit

Instead of using weights to enforce that fixed budget of N SNPs must be selected, we can model this explicitly in the CNF using an adder circuit. This is a set of clauses that adds all the SNP literals set to TRUE (i.e. SNPs in the tag set) together. This circuit will produce an n-bit number output, where n is the log of the number of SNPs. We then need to set this n-bit output to the desired number of SNPs, thereby enforcing the size of the tag set.

Consider a set of three SNPs s_1, s_2, s_3. We wish to express their sum in two new literals b_0, b_1 where b_0 represents the 0-bit and b_1 represents the 1-bit. To do this, we use a full adder circuit which can be represented in propositional logic as:

$$b_0 = (s_1 \oplus s_2) \oplus s_3$$

$$b_1 = (s_1 \vee s_2) \wedge (s_3 \vee (s_1 \oplus s_2))$$

Two fix the number of SNPs, we can force b_0, b_1 to be true or false. For example, if we add the clauses $(-b_0) \wedge (b_1)$, then exactly two of s_1, s_2, s_3 must be true if the entire CNF is satisfied.

In the full problem we construct a hierarchy of adders that will sum all the SNP variables $(s_1, s_2, \ldots, s_n)$ together and put the result if $(b_0, b_1, \ldots, b_l)$ where $l = log(n)$. Fixing k is then achieved by setting the individual b_i to true or false (see [45] for a description of adder circuits). The number of adders required will be less than 5N and the number of additional clauses/variables will therefore be in O(N) [45].

The adder circuit is fixed and independent of the target number of SNPs. Generating maxsat problems with different numbers of allowed SNPs, only requires

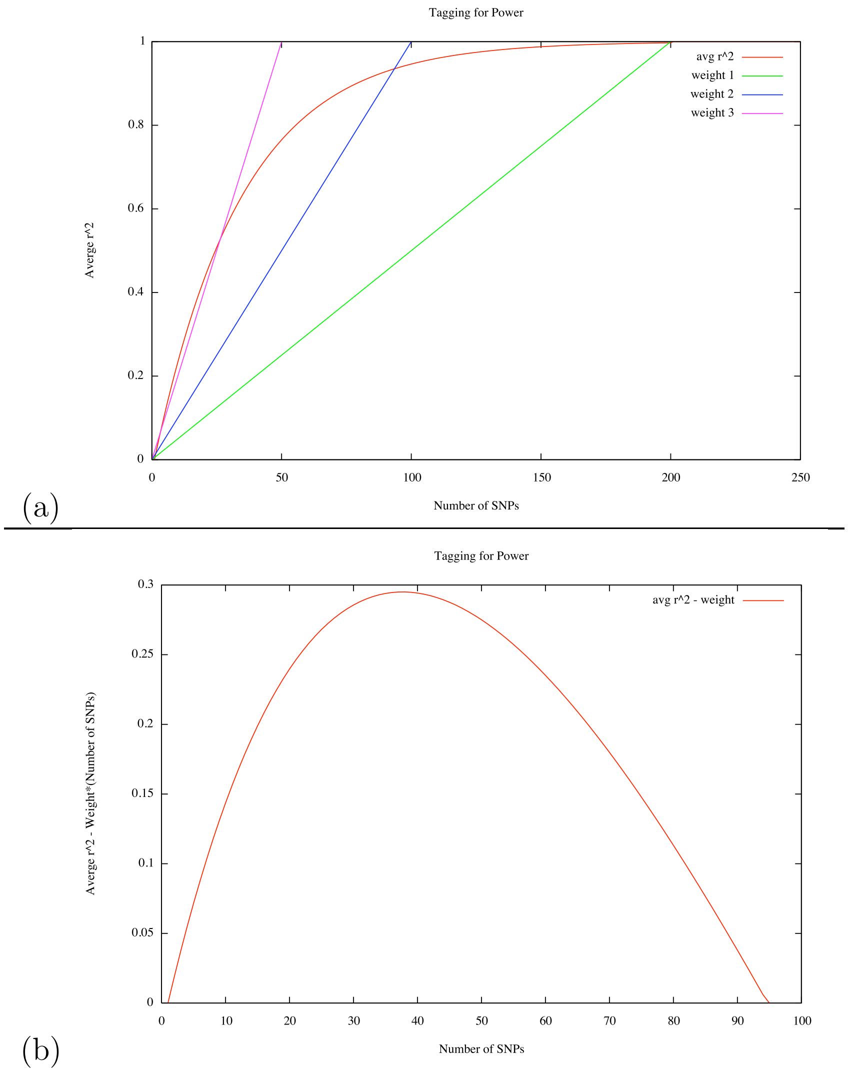

Figure 3.5: (a) The maximum average r^2 for different numbers of tags. The straight lines correspond to different penalty weights for the optimization procedure. (b) Each curve is the difference between the the maximum avg r^2 and the penalties for each different penalties. Max-SAT finds the tag set which achieves the maximum of the curves in (b). Each of these maximums correspond to a point on the maximum average r^2 curve in figure (a). By varying the penalty weight (corresponding to different slopes for the lines in (a)), we can recover optimal solutions for different number of tag SNPs.

modification of O(log(N)) clauses, a small portion of the problem. The CNF is the exactly as described in section 2.3 above with the exception that the adder circuit now replaces the negatively weighted SNPs. The optimal solution of maximum weight will then be the set of k SNPs (k specified by the adder circuit) that maximize the average r^2.

3.3 Results

We downloaded the complete HapMap build 22 data including all ENCODE regions. These data are genotypes on 270 individuals in 4 populations and over 3.8 million SNPs. They represent the most complete survey of genotype data currently available and are used as our test data sets. The 10 ENCODE regions span 5 MB and are believed to have complete ascertainment for SNPs with frequency greater than 5%. They are commonly used to estimate the performance of association study design methods and tag SNP selection methods since there are still many unknown common SNPs in the rest of the genome.

3.3.1 ENCODE Single SNP r^2 Comparison

We compared the performance of our method to the two other optimal methods as well as the non-optimal greedy algorithm over each of the ENCODE regions in each of the populations. Haldorsson *et al.*[52] restricts the maximum length of correlations and uses a dynamic programming procedure which guarantees to find an optimal solution. Given a window size W, Haldorsson *et al.*[52] examines all 2^W possible choices of tag SNPs in the window and then uses dynamic programming to extend this to a longer region. FESTA [94] extends the standard greedy algorithm in a natural way. Given an r^2 threshold, FESTA partitions the SNPs into precincts where SNPs are correlated only within the precinct. For a small precinct FESTA enumerates all possible tag sets in search of the minimal tag set. For a larger precinct, FESTA applies a hybrid exhaustive enumeration and greedy algorithm by first selecting some SNPs using exhaustive enumeration and then applying greedy algorithm. The user defines a threshold L so that the hybrid

Region	n_s	Greedy	Halldorson(W=15)	FESTA(L=10^7)	Optimal
ENm010	567	159	243(12m)	157(0m)	157
ENm013	755	93	309(23m)	90(0m)	90
ENm014	914	164	393(33m)	157(0m)	157
ENr112	927	180	340(34m)	173(34m)	173
ENr113	1072	179	395(46m)	176(274m)	176
ENr123	937	174	463(35m)	172(0m)	172
ENr131	1041	228	414(44m)	221(0m)	221
ENr213	659	122	248(17m)	122(0m)	122
ENr232	533	142	181(11m)	140(36m)	140
ENr321	599	132	202(14m)	131(0m)	131

Figure 3.6: Comparison of several *tagging* algorithms over the Encode regions in the CEU population. n_s is the number of SNPs in the region. The table shows the tag set size for each of the methods various methods (smaller is better). Running times are given in parentheses in minutes. All running times for Optimal are less than one minute (0m).

method is applied when $\binom{n}{k} > L$ where n is the number of SNPs in a precinct and k is the number of tags that need to be selected from the precinct We use $L = 10^7$ in our experiments.

The results are presented in table 3.6. Surprisingly, the Halldorson*et al.*[52] method, at the maximum limit of what is computationally feasible ($W = 15$) performs worse than the simple greedy algorithm and is much slower than than both FESTA and our approach. FESTA and our approach both recover optimal solutions for the ENCODE regions, however, FESTA ends up spending a very large amount of computational time in very large precincts, taking several hours to complete some of the data sets while our approach requires less than a minute.

3.3.2 Whole Genome Single SNP r^2

We ran our approach and the greedy approach over the entire genome wide HapMap data for the CEU population in order to find the minimal tag set to cover all SNPs with MAF ≥ 0.05 and $r^2 \geq 0.8$. Greedy resulted in 472729 tag SNPs while our approach needed only 468967 over the entire 1692323 SNP data set. This modest decrease shows that in the single SNP r^2 tag SNP selection approach, greedy search performs almost as well as optimal search. Our program required less than one day on a single CPU to run over the whole genome. Since the DNNF needs to be compiled only once, we quickly can list all possible optimal solutions, allowing for flexible design or optimization on a secondary criteria. For example, we can efficiently obtain the set of optimal solutions that contain some SNPs and do not contain other SNPs.

3.3.3 Multi-Marker Tag SNP Selection ENCODE Results

Although for the single SNP r^2 tag SNP selection problem, the greedy algorithm achieves a solution close to the optimal solution, this is not the case for multi-marker tag SNP selection. We compare our method to the popular Tagger [38] method over the Encode region ENm010 in the CEU population. This regions contains 567 SNPs with minor allele frequency (MAF) greater than 0.05. Tagger first computes a single SNP tag set using greedy search resulting in 159 tag SNPs and then uses a "roll back" procedure in which a SNP s_i is removed from the tag set if another pair of SNPs in the tag set cover s_i with $r^2 = 1.0$. That is, redundant SNPs are removed from the tag set. Tagger's multi-marker approach does reduce the number of SNPs required to cover an ENCODE region to 141 SNPs compared to 157 optimal single SNP tags, but the reduction is far from optimal. Our method requires only 72 SNPs to cover the ENCODE region a 54% and 40% reduction over single SNP tags and the Tagger's multi-marker tags.

3.3.4 Power

As a proof of concept, we ran the our approach over the first 100 SNPs Encode region ENm010 while maximizing average r^2 with a 4 SNP limit and compared this to the naive greedy approach of choosing the best 4 SNPs in terms of single SNP r^2 in the region. Using the naive approach, the average r^2 is 0.90. Using our method, the average r^2 is 0.98.

3.4 Discussion

We have presented efficient optimal methods for solving a variety of tag SNP selection problems. Although the general form of these problems is NP-complete, we showed that natural constraints of the genome's LD structure bound the problem complexity in practice. Tag SNP selection problem instances are first reduced into a SAT instance in CNF. The CNF representation represents SNPs as boolean literals and solutions as settings of these variables that satisfy the clauses. The CNFs are compiled into DNNFs which allow quick enumeration of all optimal solutions. This compilation is the crucial step in which the genome's LD structure naturally partitions SNP literals into distinct regions of the DNNF.

Improvements over the classic single SNP r^2 tagging problem are modest compared to greedy search. The FESTA [94] algorithm also achieves these results over the ENCODE regions, but is not guaranteed to be optimal in the general case. We outperform FESTA in terms of running time, and also our ability to characterize *all* optimal solutions as opposed to just those containing perfectly linked SNPs. This permits flexible tag set choice that can be further optimized over secondary criteria. This method is also extensible to other measures of SNP coverage besides r^2.

Using a "stitching" method we showed how to combine local solutions into a globally optimal genome wide tag set solution. Although the method is exponential in the number of SNPs that are shared between local solutions, we find that choosing regions to minimize this overlap allows efficient whole genome tag set optimization.

Recent work has shown the multi-marker methods are more power than single SNP techniques in the context of association studies. While a variety of multi-marker statistical tests exist, the current optimal tagging methods such as FESTA are not able to tag for multiple markers efficiently. Our SAT based method is able to find optimal multi-marker tags for pairs of SNPs over the dense ENCODE regions. The gain for optimal tagging over greedy search in this context is significantly better than for the single SNP with improvement over the popular Tagger [38] method reaching 40%. Since the cost of custom genotype arrays remains high, this tool is valuable for follow up association studies. That is, once genome wide results have been found, further genotyping must be done to localize the region containing the causal variant. Intelligent choice of tag sets for follow up studies can greatly improve their power and until now, multi-marker tagging for follow up has been non-optimal.

Finally, we showed how to extend these techniques to maximize average r^2 given a fixed SNP budget. Power, or any other convex function can be used in place of r^2.

Our method and whole genome optimal data sets are available for use via web server at http://whap.cs.ucla.edu.

Chapter 3, was published In Proceedings of the 8th Workshop on Algorithms in Bioinformatics, (WABI-2008), Karlsruhe, Germany, September 15-17, 2008. Arthur Choi, Noah Zaitlen, Buhm Han, Knot Pipatsrisawat, Adnan Darwiche, E. Eskin, “Efficient Genome Wide Tagging by Reduction to SAT”. The dissertation author and Arthur Choi were the primary investigators and authors of this paper.

Chapter 4

Finite Sample Effects of the HapMap

4.1 Introduction

The development of the HapMap[25] has ushered in the genome wide association era of human genetics and a tremendous number of studies have reported associations to novel genes for a variety of complex diseases[26, 78, 77]. These genome wide association studies have been made possible by two recent developments. The first is the development of high throughput genotyping technology enabling the simultaneous genotyping of hundreds of thousands of single nucleotide polymorphisms (SNPs) from an individual at a reasonable cost. The second is the development of the HapMap which provides genotype information for the majority of common SNPs in panels from four populations. Although it is still prohibitively expensive to genotype all common polymorphisms in an association study, linkage disequilibrium (LD) or correlation between SNPs allows association studies to genotype a subset of the SNPs referred to as "tag" SNPs[108]. Association to a phenotype at an ungenotyped SNP can be detected if a nearby correlated SNP is one of the tag SNPs[93]. The HapMap allows us to infer the patterns of LD between SNPs.

The most relevant measure of LD between two SNPs to the statistical power of an association study is r^2, the square of the correlation coefficient between the two SNPs. As shown in Pritchard and Przeworski[93], to achieve the equivalent power of detecting an association at a SNP with N individuals requires N/r^2 individuals at a neighboring marker. Using this insight, the standard approach to choosing tag SNPs is to select a subset of the SNPs which have an $r^2 > 0.8$ with the remaining SNPs[38]. A SNP that is correlated with a tag SNPs with an r^2 greater than 0.8 is referred to as being "captured". The development of the current generation of commercial genotyping products have been strongly motivated by this notion and tag SNPs are chosen to capture as many of the common HapMap SNPs as possible.

In addition to aiding in the design of genome wide association studies, the HapMap genotypes allow us to estimate study power for a given disease model. Recent imputation methods [121, 79] have demonstrated how to further improve study power with the HapMap data by estimating allele frequencies of untyped variants. The HapMap genotypes have also been used to estimate recombination rates across the genome, search for regions under natural selection, and locate structural variations such as deletions and inversions.

Despite its tremendous importance as a resource, the HapMap suffers from the fundamental limitation that it is based on only 60 unrelated individuals or 120 chromosomes per population. Current methods which use the HapMap effectively assume that the HapMap has infinitely many individuals and that the observed correlation patterns are the true correlation patterns. In reality, the HapMap is not a large enough sample to accurately measure the LD patterns between SNPs. This limitation has significant implications for association studies. First, many of the SNPs which are believed to be captured by tag sets developed using the HapMap are in fact not well captured, but only appear to be captured in the HapMap due to sampling bias. For these SNPs, even very large association studies will not detect associations. Second, the estimates of r^2 are very inaccurate which leads to inaccurate estimates of the power of association studies. Several groups have previously pointed out this limitation[112] and have performed empirical studies exploring this and the related issue of the transferability of tag SNPs to different populations[83, 36, 99]. Finally, the linkage structure is used by imputation methods to estimate frequencies of untyped SNPs in association studies. Their accuracy is necessarily tied to that of the HapMap.

In this work, we present an analytical framework for analyzing the implications of a finite sample HapMap. We observe that most of the error in the estimates of correlation patterns stems from the difficulty in estimating conditional probabilities from small samples. We present and justify simple approximations for obtaining confidence intervals for r^2 estimates from the HapMap and verify our confidence intervals through simulations using both the HapMap and Welcome Trust Case Control Consortium data[26] (WTCCC). We show how the current HapMap may perform very poorly for estimating the power of an association study at certain SNPs. Consider a case control study in which 5,000 cases and 5,000 controls are genotyped on 500,000 independent SNPs, and the causal variant has a relative risk 1.5 and minor allele frequency of 0.05. If the causal variant is genotyped directly the study has a power of 93.1% to detect an association at that variant. If the causal variant is not genotyped, but has an r^2 of 0.8 with a nearby tag SNP, then the study has a power of 55% to detect an association due to the causal variant. Using our framework, we show that approximately 8.2% of the SNPs with a minor allele frequency of 0.05 and an observed r^2 of 0.8 in the HapMap have a true r^2 below 0.5. The power to detect an association is only 2.7% for these SNPs, yet using the HapMap we believe the power is 55%. In order to better ground our results in an actual association study context, we extend this analysis using the WTCCC [26] data and show that many of the SNPs in these studies may be affected by finite sample errors in the HapMap. In addition, we show that procedure of selecting tag SNPs is upwardly biased and results in an overestimation of power.

While larger reference sets such as the 1000 genomes project are being proposed to catalogue rare human variation, it is not appreciated how these reference sets will profoundly affect our ability to detect common variation involved in human disease. In addition to providing a highly valuable fine grained picture of low frequency SNPs and a more extensive list of SNPs human populations, these additions to the HapMap reference panels will help resolve many of the issues illustrated in the above examples. To estimate how large of a sample we need to collect in order to avoid these problems we examine confidence intervals around "captured" SNPs with an r^2 of 0.8. As the sample size or MAF of a SNP increases, the confidence interval for r^2 will tighten. If a SNP's 95% confidence interval is greater than 0.1 or equivalently that the probability that $r^2 < 0.7$ is less than 2.5% we are confident that the estimated r^2 is close to the true r^2. The current HapMap cannot provide a bound this tight even for SNPs with MAF 0.5. Increasing the HapMap to 238 individuals would achieve this confidence interval for SNPs with MAF as low as 0.2, and increasing it to 502 individuals would provide for SNPs with MAF as low as 0.1. In order to have a tight confidence interval for low frequency SNPs with MAF of 0.05, 1042 individuals are needed.

4.2 Material and Methods

4.2.1 Case Control Studies.

In a typical association study, individuals are collected from two populations, a case population consisting of individuals with a disease and a control population consisting of individuals without the disease. The populations differ along the phenotype of interest but individuals are carefully selected so that they are otherwise members of the same population. Each individual in the study is genotyped on a set of tag SNPs such as those on the Affymetrix and Illumina high throughput genotyping platforms. SNPs with alleles that cause an alteration in risk for the phenotype potentially occur in different frequencies in the cases and controls. These causal SNPs may not be included in the set of tag SNPs. It is the primary objective of a case control study to identify these causal polymorphisms.

Suppose that there is a tag SNP A and a causal SNP B in a case control study with $N/2$ cases and $N/2$ controls. We denote the frequency of the minor allele of SNP A in the cases, controls, and entire population as p_A^+, p_A^-, and p_A respectively. We denote $\mathrm{p}_\mathrm{a} = (1 - \mathrm{p}_\mathrm{A})$ as the frequency for the major allele of SNP A and use p_B, p_b for the equivalent frequencies over SNP B. $\hat{\mathrm{p}}_\mathrm{B}$ and $\hat{\mathrm{p}}_\mathrm{A}$ denote the observed frequencies of SNP B and SNP A in the collected samples respectively.

Consider a case control study in which we genotype the causal SNP B directly. We compute the following statistic S_B

$$S_B = \frac{\hat{\mathrm{p}}_{\mathrm{B}}^{+} - \hat{\mathrm{p}}_{\mathrm{B}}^{-}}{\sqrt{2/N}\sqrt{\hat{\mathrm{p}}_{\mathrm{B}}(1-\hat{\mathrm{p}}_{\mathrm{B}})}} \sim \mathcal{N}\left(\frac{(\mathrm{p}_{\mathrm{B}}^{+} - \mathrm{p}_{\mathrm{B}}^{-})\sqrt{N}}{\sqrt{2\mathrm{p}_{\mathrm{B}}(1-\mathrm{p}_{\mathrm{B}})}}, 1\right) = \mathcal{N}\left(\lambda_B\sqrt{N}, 1\right) \tag{4.1}$$

S_B measures the difference in the frequency of SNP B in the cases ($\hat{\mathrm{p}}_{\mathrm{B}}^{+}$) and the controls ($\hat{\mathrm{p}}_{\mathrm{B}}^{-}$) in the collected sample normalized such that the variance is 1. We refer to $\lambda_B\sqrt{N}$ as the non-centrality parameter (NCP) for SNP B. In the null hypothesis $\mathrm{p}_{\mathrm{B}}^{+} = \mathrm{p}_{\mathrm{B}}^{-}$ and $\lambda_B\sqrt{N}$ is 0. In the alternative hypothesis $\mathrm{p}_{\mathrm{B}}^{+} \neq \mathrm{p}_{\mathrm{B}}^{-}$ and $\lambda_B\sqrt{N}$ is the mean of the distribution of S_B. The NCP $\lambda\sqrt{N}$ is a function of study size, disease model, SNP minor allele frequency (MAF), and is the fundamental measure of study power. Power is calculated from the NCP and significance threshold (α) as

$$\mathcal{P}(\alpha, \lambda\sqrt{N}) = \Phi\left(\Phi^{-1}(\alpha/2) + \lambda\sqrt{N}\right) + 1 - \Phi\left(\Phi^{-1}(1-\alpha/2) + \lambda\sqrt{N}\right) \tag{4.2}$$

where $\Phi(x)$ is the normal cumulative distribution function and $\Phi^{-1}(x)$ is the normal quantile function. Fixing α (e.g. 0.05), the power is solely a function of the NCP.

4.2.2 Indirect Association.

In general we do not expect the causal variant SNP B to be amongst the set of genotyped tag SNPs, but instead rely on the correlation or LD between proximal tag SNPs and the causal variant to discover the association. Consider a case control study in which the causal variant SNP B is not genotyped but is near a tag SNP A. If SNP A is in strong enough LD with SNP B, and the study is sufficiently powered, it may be possible to detect a significant difference in the frequencies of SNP A between the cases and controls due to its correlation with the causal variant SNP B.

In this section we derive the NCP and power for SNP A given that SNP B is causal. This relies on the conditional probability of observing the minor allele at SNP A given that the minor allele at SNP B is observed. This is denoted $\mathrm{p_{A|B}} = p_{AB}/p_B$, where p_{AB} is the frequency of the haplotype made from minor alleles of both SNPs A and B. The conditional probability of observing the minor allele of SNP A given an observation of the major allele of SNP B is similarly denoted $\mathrm{p_{A|b}}$. It is a standard assumption of association studies that, if SNP B is causal then the conditional probability $\mathrm{p_{A|B}}$ is equal in the cases and controls. Formally, $\mathrm{p_{A|B}} = \mathrm{p^+_{A|B}} = \mathrm{p^-_{A|B}}$ and $\mathrm{p_{A|b}} = \mathrm{p^+_{A|b}} = \mathrm{p^-_{A|b}}$. Note that $\mathrm{p^+_{B|A}} \neq \mathrm{p^-_{B|A}}$ if $\mathrm{p^+_B} \neq \mathrm{p^-_B}$ under this assumption.

The frequencies of SNP A can be written in terms of the conditional probabilities and the frequency of SNP B. The frequency in the cases is $\mathrm{p^+_A} = \mathrm{p^+_{A|B}p^+_B} + \mathrm{p^+_{A|b}}(1 - \mathrm{p^+_B})$ and the frequency in the controls is $\mathrm{p^-_A} = \mathrm{p^-_{A|B}p^-_B} + \mathrm{p^-_{A|b}}(1 - \mathrm{p^-_B})$. Combining these two equations, the difference in frequency of the genotyped SNP A between the cases and controls is expressed in terms of the conditional probabilities and the frequency of the causal SNP B as $\mathrm{p^+_A} - \mathrm{p^-_A} = \mathrm{p_{A|B}}(\mathrm{p^+_B} - \mathrm{p^-_B}) + \mathrm{p_{A|b}}(1 - \mathrm{p^+_B} - 1 + \mathrm{p^-_B}) = (\mathrm{p_{A|B}} - \mathrm{p_{A|b}})(\mathrm{p^+_B} - \mathrm{p^-_B})$.

We can now derive the NCP for SNP A which will in turn give the power for SNP A in terms of the NCP of SNP B and the conditional probabilities following the derivation of Pritchard and Przeworski, 2001[93].

$$
\begin{aligned}
S_A &= \frac{\mathrm{\hat{p}^+_A} - \mathrm{\hat{p}^-_A}}{\sqrt{2/N}\sqrt{\mathrm{\hat{p}_A}(1-\mathrm{\hat{p}_A})}} \sim \mathcal{N}(\lambda_A\sqrt{N}, 1) \\
\lambda_A\sqrt{N} = \frac{\mathrm{p^+_A} - \mathrm{p^-_A}}{\sqrt{2/N}\sqrt{\mathrm{p_A}(1-\mathrm{p_A})}} &= (\mathrm{p_{A|B}} - \mathrm{p_{A|b}})\frac{(\mathrm{p^+_B} - \mathrm{p^-_B})}{\sqrt{2/N}\sqrt{\mathrm{p_A}(1-\mathrm{p_A})}} \\
&= (\mathrm{p_{A|B}} - \mathrm{p_{A|b}})\frac{(\mathrm{p^+_B} - \mathrm{p^-_B})}{\sqrt{2/N}\sqrt{\mathrm{p_A}(1-\mathrm{p_A})}}\frac{\sqrt{\mathrm{p_B}(1-\mathrm{p_B})}}{\sqrt{\mathrm{p_B}(1-\mathrm{p_B})}} \\
&= (\mathrm{p_{A|B}} - \mathrm{p_{A|b}})\frac{\sqrt{\mathrm{p_B}(1-\mathrm{p_B})}}{\sqrt{\mathrm{p_A}(1-\mathrm{p_A})}}\lambda_B\sqrt{N}
\end{aligned}
\tag{4.3}
$$

The correlation coefficient $r_{AB} = (\mathrm{p_{A|B}} - \mathrm{p_{A|b}})\sqrt{\frac{\mathrm{p_B}(1-\mathrm{p_B})}{\mathrm{p_A}(1-\mathrm{p_A})}}$ between SNP A and SNP B relates λ_A and λ_B and is algebraically equivalent to the standard form of $r_{AB} = \frac{P_{AB} - \mathrm{p_A p_B}}{\sqrt{\mathrm{p_A}(1-\mathrm{p_A})\mathrm{p_B}(1-\mathrm{p_B})}}$.

Equation (4.3) above shows that the NCP at SNP A is a function of r_{AB} and the NCP at SNP B, $\lambda_A\sqrt{N} = r_{AB}\lambda_B\sqrt{N}$. Finally, we can express the power at SNP A if SNP B is causal as:

$$\mathcal{P}(\alpha, \lambda_A\sqrt{N}) = \mathcal{P}(\alpha, r_{AB}\lambda_B\sqrt{N}) \tag{4.4}$$

As expected, the higher the correlation between the SNPs, the greater the power of using the tag SNP as a proxy for the causal variant.

4.2.3 Estimating Correlation Variance from the HapMap.

The HapMap data do not provide the exact frequencies or conditional probabilities of the SNPs in a population, but is commonly used to estimate these quantities as well as the correlation coefficient between SNPs, the NCP, and study power under a given disease model. The finite sample size of each of the HapMap populations introduces a source of error into each of these estimates. In the previous section we derived the NCP for the causal SNP given a tag SNP. Here we extend this derivation to calculate the mean and variance of the NCP and the correlation coefficient assuming a finite reference sample. The variance of the correlation coefficient found in this section uses a simplifying assumption and is therefore denoted σ_S. A more complex estimate is derived in the next section.

We derive an approximation for the correlation coefficient between SNPs A and B.

$$\hat{r}^H_{AB} \sim \mathcal{N}(r_{AB}, \sigma_S^2) \tag{4.5}$$

$$\sigma_S \equiv var(\hat{r}^H_{AB}) = \left(\frac{\mathrm{p}^{\mathrm{H}}_{\mathrm{A|B}}(1-\mathrm{p}^{\mathrm{H}}_{\mathrm{A|B}})}{N^H_B} + \frac{\mathrm{p}^{\mathrm{H}}_{\mathrm{A|b}}(1-\mathrm{p}^{\mathrm{H}}_{\mathrm{A|b}})}{N^H_b} \right) \frac{\mathrm{p}^{\mathrm{H}}_{\mathrm{B}}(1-\mathrm{p}^{\mathrm{H}}_{\mathrm{B}})}{\mathrm{p}^{\mathrm{H}}_{\mathrm{A}}(1-\mathrm{p}^{\mathrm{H}}_{\mathrm{A}})} \tag{4.6}$$

where superscript H denotes values over the HapMap data N^H and denotes the number of chromosomes in the HapMap. $\mathrm{p}^{\mathrm{H}}_{\mathrm{B}}$ denotes the true frequency of the minor allele SNP B in the HapMap population. The observed frequency of the minor allele of SNP B in the HapMap samples is denoted by $\hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{B}}$. The true and observed conditional probabilities are $\mathrm{p}^{\mathrm{H}}_{\mathrm{A|B}}$ and $\hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{A|B}}$ respectively, and $N^H_B = \hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{B}} N^H$ is the number of chromosomes with the minor allele.

To derive an approximation for the variance of r_{AB}^H in equation 4.5 we begin with the derivation of the distribution for the estimate of the NCP at SNP A. Assuming normal approximations:

$$\begin{aligned}
\hat{\mathrm{p}}_{\mathrm{A|B}}^{\mathrm{H}} &\sim \mathcal{N}(\mathrm{p}_{\mathrm{A|B}}^{\mathrm{H}}, \frac{\mathrm{p}_{\mathrm{A|B}}^{\mathrm{H}}(1-\mathrm{p}_{\mathrm{A|B}}^{\mathrm{H}})}{N_B^H}) \\
\hat{\mathrm{p}}_{\mathrm{A|b}}^{\mathrm{H}} &\sim \mathcal{N}(\mathrm{p}_{\mathrm{A|b}}^{\mathrm{H}}, \frac{\mathrm{p}_{\mathrm{A|b}}^{\mathrm{H}}(1-\mathrm{p}_{\mathrm{A|b}}^{\mathrm{H}})}{N_b^H}) \\
\hat{\mathrm{p}}_{\mathrm{B}}^{\mathrm{H}} &\sim \mathcal{N}(\mathrm{p}_{\mathrm{B}}^{\mathrm{H}}, \frac{\mathrm{p}_{\mathrm{B}}^{\mathrm{H}}(1-\mathrm{p}_{\mathrm{B}}^{\mathrm{H}})}{N^H}) \\
\hat{\mathrm{p}}_{\mathrm{b}}^{\mathrm{H}} &\sim \mathcal{N}(\mathrm{p}_{\mathrm{b}}^{\mathrm{H}}, \frac{\mathrm{p}_{\mathrm{b}}^{\mathrm{H}}(1-\mathrm{p}_{\mathrm{b}}^{\mathrm{H}})}{N^H})
\end{aligned} \tag{4.7}$$

The estimates of the conditional probability are based on far fewer observations than the estimates of the frequency, thus the variance of the estimates of the conditional probabilities are much larger than the variance of the estimates of the allele frequencies.

In order to use the HapMap data for power estimation we make several assumptions about the relation between our case, control, and HapMap populations. The fundamental assumption of the HapMap is that the SNP frequencies conditional on a causal SNP in case and control samples are equal to the conditional frequencies in the HapMap. That is, $\mathrm{p}_{\mathrm{A|b}} = \mathrm{p}_{\mathrm{A|b}}^{+} = \mathrm{p}_{\mathrm{A|b}}^{-} = \mathrm{p}_{\mathrm{A|b}}^{\mathrm{H}} = \frac{P_{Ab}^H}{\mathrm{p}_{\mathrm{b}}^{\mathrm{H}}}$.

Under these assumptions, we can estimate the NCPs using the estimates of the conditional probabilities and allele frequencies directly from the HapMap. Using these terms we define the NCP $\hat{\lambda}_A^H$ in terms of the empirical conditional probabilities and frequencies in the HapMap, and the true NCP λ_B:

$$\hat{\lambda}_A^H \sqrt{N} \equiv (\hat{\mathrm{p}}_{\mathrm{A|B}}^{\mathrm{H}} - \hat{\mathrm{p}}_{\mathrm{A|b}}^{\mathrm{H}}) \sqrt{\frac{\hat{\mathrm{p}}_{\mathrm{B}}^{\mathrm{H}}(1-\hat{\mathrm{p}}_{\mathrm{B}}^{\mathrm{H}})}{\hat{\mathrm{p}}_{\mathrm{A}}^{\mathrm{H}}(1-\hat{\mathrm{p}}_{\mathrm{A}}^{\mathrm{H}})}} \lambda_B \sqrt{N} \tag{4.8}$$

In our estimate, the term $\lambda_B\sqrt{N}$ is considered a constant because we are interested in the relative strength of the association at the tag SNP compared to the strength at the causal SNP. Under this assumption, the only observed values appearing in the equation for the NCP are HapMap SNP frequencies and conditional probabilities.

We make the simplifying assumption that the empirical frequency $\hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{A}}$ is close to the true frequency $\mathrm{p}^{\mathrm{H}}_{\mathrm{A}}$ relative to the much larger variance in the estimates of conditional probability. This allows us to derive a simple estimate for the error due to the finite sample.

$$\hat{\lambda}^H_A\sqrt{N} \sim \mathcal{N}(\lambda_A\sqrt{N}, \sigma^{2H}_{\lambda_A}) \tag{4.9}$$

$$\sigma^{2H}_{\lambda_A} = \left(\frac{\mathrm{p}^{\mathrm{H}}_{\mathrm{A|B}}(1-\mathrm{p}^{\mathrm{H}}_{\mathrm{A|B}})}{\hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{B}}N^H} + \frac{\mathrm{p}^{\mathrm{H}}_{\mathrm{A|b}}(1-\mathrm{p}^{\mathrm{H}}_{\mathrm{A|b}})}{(1-\hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{B}})N^H}\right)\frac{\hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{B}}(1-\hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{B}})}{\hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{A}}(1-\hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{A}})}\lambda^2_B N \tag{4.10}$$

Although there is additional variance due to the assumption that empirical frequency $\hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{A}}$ is close to the true frequency $\mathrm{p}^{\mathrm{H}}_{\mathrm{A}}$, the results section shows experimentally that simulation of NCPs from linked SNPs are surprisingly close to the derived distribution. We will derive a more sophisticated estimate by dropping this assumption below.

The correlation coefficient is commonly used to measure the strength of SNP A to serve as a proxy for SNP B. We showed above that the decrease in power due to using SNP A as a proxy for SNP B is directly proportional to the correlation coefficient. The NCP and the correlation coefficient can be used together to measure study power. Using the fact that $\lambda_A\sqrt{N} = r_{AB}\lambda_B\sqrt{N}$ and equation (4.9) above we derive mean and variance of $\hat{r}^H_{AB}$

$$var(\hat{r}^H_{AB}) \approx (var(\hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{A|B}} + \hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{A|b}}))\frac{\mathrm{p}^{\mathrm{H}}_{\mathrm{B}}(1-\mathrm{p}^{\mathrm{H}}_{\mathrm{B}})}{\mathrm{p}^{\mathrm{H}}_{\mathrm{A}}(1-\mathrm{p}^{\mathrm{H}}_{\mathrm{A}})} \tag{4.11}$$

$\hat{r}^H_{AB}$ is an unbiased estimate of r^H_{AB}, and our experiments validate the assumptions of this approximation with experimental simulation.

4.2.4 Estimating Variance with the Delta Method.

The above "simple" estimate for the variance of the correlation coefficient σ_S assumes that the minor allele frequency is accurately estimated from the data. However, when the frequency of one of the SNPs is very low, this assumption no longer holds. In order to accurately estimate the distribution of $\hat{r}^H_{AB}$ we employ the delta method [88] and derive variance σ_Δ. We let $x = \hat{\mathrm{p}}_{\mathrm{AB}}$, $y = \hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{B}}$, and $z = \hat{\mathrm{p}}^{\mathrm{H}}_{\mathrm{A}}$ and rewrite the formula for the correlation coefficient in terms of x,y,z.

$$f = \hat{r}^H_{AB} = \left(\frac{x}{y} - \frac{z-x}{1-y}\right)\left(\frac{\sqrt{y-y^2}}{\sqrt{z-z^2}}\right) \tag{4.12}$$

We compute the variance covariance matrix Σ for x,y,z with $\sigma_{xx} = \mathrm{p_{AB}}(1-\mathrm{p_{AB}})$,$\sigma_{yy} = \mathrm{p_B}(1-\mathrm{p_B})$,$\sigma_{zz} = \mathrm{p_A}(1-\mathrm{p_A})$,$\sigma_{xy} = \mathrm{p_{AB}}(1-\mathrm{p_B})$,$\sigma_{xz} = \mathrm{p_{AB}}(1-\mathrm{p_A})$,$\sigma_{yz} = \mathrm{p_{AB}} - \mathrm{p_A p_B}$.

$$\Sigma = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{pmatrix} \tag{4.13}$$

We compute the gradient of f:

$$\Delta^T = \begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} & \frac{\partial f}{\partial z} \end{pmatrix} \tag{4.14}$$

Finally, the variance is estimated as:

$$\sigma_\Delta \equiv var(\hat{r}^H_{AB}) = var(f) = \Delta^T_\mu \Sigma \Delta_\mu \tag{4.15}$$

where Δ_μ is the gradient evaluated at the means of x,y,z. The results section demonstrates that the "delta method" estimate is accurate over low frequency SNPs with experimental simulation.

4.2.5 Overestimation of r^2 in Tagging.

We showed above that the finite sample size of the HapMap results in error for the estimation of r_{AB} between SNP A and SNP B but this estimate is unbiased. However, when selecting tags to genotype, since the goal is to choose the smallest subset of SNP which cover as many of the remaining SNPs as possible, the HapMap estimates of the correlation are significantly biased. This bias is compounded by the overestimation described by Bhangale et. al. [10] that is observed when SNPs are examined in addition to those contained in the HapMap.

Consider the 3 SNPS, A, B, C, where we are choosing one of A or B to serve as a proxy for SNP C. We will select the SNP which has a stronger correlation with SNP C. For this example, suppose that the correlations coefficients r_{AC} and r_{BC} are positive since exchanging major and minor alleles will flip the sign. Using the HapMap will result in estimates of these correlations with variances σ_{AC}, σ_{BC} and means r^H_{AC} and r^H_{BC} with expected values close to the true correlation. The estimated coverage of SNP C is the max of the empirical measurements of the correlation coefficients in the HapMap max($\hat{r}^H_{AC}$,$\hat{r}^H_{BC}$) $= \frac{\hat{r}^H_{AC}+\hat{r}^H_{BC}}{2} + \frac{|\hat{r}^H_{AC}-\hat{r}^H_{BC}|}{2}$.

We show that this maximum is

$$max(\hat{r}^H_{AC}, \hat{r}^H_{BC}) = max(r^H_{AC}, r^H_{BC}) + bias(r^H_{AC}, r^H_{BC}) \tag{4.16}$$

and we prove that the term $bias(r^H_{AC}, r^H_{BC})$ is always positive.

To calculate the expectation of this maximum we note that $\hat{r}^H_{AC}$ and $\hat{r}^H_{BC}$ are normally distributed random variables. The term $\frac{\hat{r}^H_{AC}+\hat{r}^H_{BC}}{2}$ will have expected value $\frac{r^H_{AC}+r^H_{BC}}{2}$. The expected value of the other term is more complicated due to the absolute value. We let $\sigma^2 = \frac{\sigma^2_{AC}+\sigma^2_{BC}}{4}$ and $\mu = \frac{r^H_{AC}-r^H_{BC}}{2}$. The expected value of $\frac{|\hat{r}^H_{AC}-\hat{r}^H_{BC}|}{2}$ is:

$$\frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}}\int_{-\infty}^{\infty}|x|e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx \tag{4.17}$$

$$\begin{aligned} &= \frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}}\int_{-\infty}^{\infty}|x|e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx \\ &+\frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}}\int_{-\infty}^{\infty}\mu e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx + \frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}}\int_{-\infty}^{\infty}-\mu e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx \end{aligned} \tag{4.18}$$

Let u=$\frac{(x-\mu)^2}{2\sigma^2}$ and rewrite the integrals as:

$$\begin{aligned} &\frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}}\int_{0}^{\infty}(x-\mu)e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx + \frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}}\int_{-\infty}^{0}-(x-\mu)e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx \\ &+\frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}}\int_{0}^{\infty}\mu e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx + \frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}}\int_{-\infty}^{0}-\mu e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx \end{aligned} \tag{4.19}$$

$$= \frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}} \int_{\frac{\mu^2}{2\sigma^2}}^{\infty} \sigma^2 e^{-u} du + \frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}} \int_{\infty}^{\frac{\mu^2}{2\sigma^2}} -\sigma^2 e^{-u} du \\ +\mu(1 - \Phi(-\frac{\mu}{\sigma})) - \mu\Phi(-\frac{\mu}{\sigma}) \tag{4.20}$$

The expectation of the entire max is:

$$\frac{r_{AC}^H + r_{BC}^H}{2} + \frac{2\sigma^2}{\sqrt{\sigma^2}\sqrt{2\pi}} e^{-\frac{\mu^2}{2\sigma^2}} + \mu(1 - 2\Phi(-\frac{\mu}{\sigma})) \tag{4.21}$$

The function $\mu(1 - 2\Phi(-\frac{\mu}{\sigma})) = -\mu(1 - 2\Phi(-\frac{-\mu}{\sigma}))$ because it is symmetric about 0 with respect to μ, and we substitute $|\mu|$ for μ and recall that $\mu = \frac{r_{AC}^H - r_{BC}^H}{2}$:

$$\frac{r_{AC}^H + r_{BC}^H}{2} + \frac{2\sigma}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2\sigma^2}} + |\mu|(1 - 2\Phi(-\frac{|\mu|}{\sigma})) \tag{4.22}$$

$$max(r_{AC}^H, r_{BC}^H) + 2\sigma(\Phi'(\frac{\mu}{\sigma}) - \frac{|\mu|}{\sigma}\Phi(-\frac{|\mu|}{\sigma})) \tag{4.23}$$

The expected maximum is

$$max(\hat{r}_{AC}^H, \hat{r}_{BC}^H) = max(r_{AC}^H, r_{BC}^H) + 2\sigma(\Phi'(\frac{\mu}{\sigma}) - \frac{|\mu|}{\sigma}\Phi(-\frac{|\mu|}{\sigma})) \tag{4.24}$$

and the bias in the expectation is therefore:

$$2\sigma(\Phi'(\frac{\mu}{\sigma}) - \frac{|\mu|}{\sigma}\Phi(-\frac{|\mu|}{\sigma})) \tag{4.25}$$

We show that the bias is positive by proving the following lemma:

For $x \geq 0$, $x\Phi(-x) \leq \Phi'(x) = \Phi'(-x)$:

$$x\Phi(-x) = x \int_{-\infty}^{-x} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt \tag{4.26}$$

$$\leq \int_{-\infty}^{-x} -t \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt \tag{4.27}$$

$$= \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} = \Phi'(x) \tag{4.28}$$

$$x\Phi(-x) \leq \Phi'(x) = \Phi'(-x) \tag{4.29}$$

Letting $x = \frac{\mu}{\sigma}$

$$\frac{\mu}{\sigma}\Phi(-\frac{\mu}{\sigma}) \leq \Phi'(\frac{\mu}{\sigma}) = \Phi'(-\frac{\mu}{\sigma}) \tag{4.30}$$

$$2\sigma(\Phi'(\frac{\mu}{\sigma}) - \frac{|\mu|}{\sigma}\Phi(-\frac{|\mu|}{\sigma})) \geq 0 \tag{4.31}$$

Thus the bias is positive and the expected estimate of the maximum will always be greater than the actual maximum.

4.3 Results

4.3.1 Increasing Sample Size.

Increasing the size of the HapMap samples will reduce the variance of statistics calculated over the data. Currently a SNP is considered covered by a tag SNP if the r^2 between the SNP and the tag is greater than 0.8. However, this fails to take into account the minor allele frequency of the SNPs in question, and hence the uncertainty of the HapMap's estimate of the correlation. We derive two analytical distributions for the correlation coefficient and non-centrality parameter in the context of an association study using a finite data set such as the HapMap. The first approximation for correlation coefficient is named σ_S and is a simple approximation which assumes estimates of conditional probability have higher variance than estimates of MAF. The second approximation is named σ_Δ and uses the Delta method to avoid this assumption. Our derivations show that SNPs with low minor allele frequencies have a high variance in the estimated correlation coefficient compared to SNPs with higher minor allele frequencies. For example A SNP with MAF 0.05 has a 10.2% chance of having a true r^2 less than 0.8 if its estimated r^2 is 0.9, while a SNP with MAF 0.15 has only a 1.1% chance.

We use our analytical estimates to produce confidence intervals for the correlation coefficient in order to address this issue and ensure that SNPs we estimate to be captured by a tag set are not missed due to errors from finite sample size. We examine the 95% confidence interval of 0.1 to measure the coverage of tag SNPs.

For SNPs with an empirical r^2 of 0.8 this is the probability that $r^2 < 0.7$ is less than 2.5%. For correlation coefficients that fall in this confidence interval, we are confident that the estimated r^2 is close to the true r^2. In a sample the size of the HapMap, a tag SNP with MAF of 0.15 requires an estimated r^2 of 0.95 to lie in this confidence interval, while a tag SNP with MAF 0.3 requires only an estimated r^2 of 0.91. SNPs with lower minor allele frequencies require higher values of estimated r^2 before they can be considered good proxies.

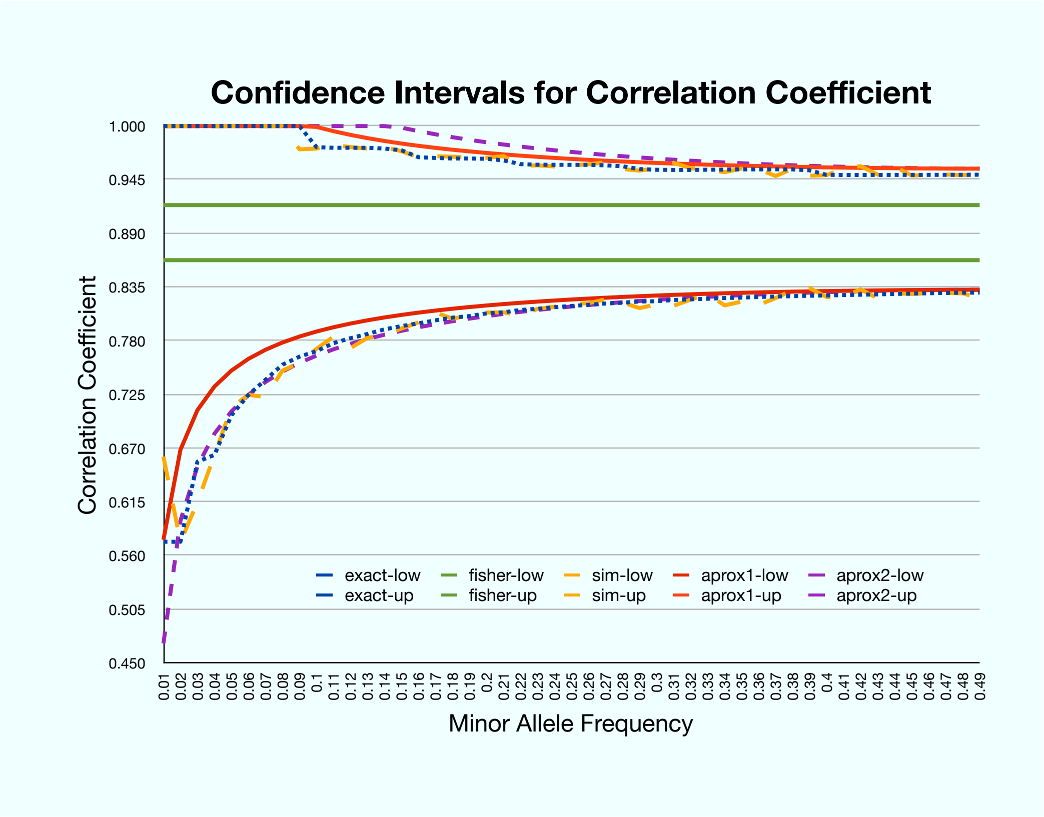

Figure 4.1: The 95% confidence intervals for SNPs with $r^2 = 0.8$ and N=200 chromosomes over a range of minor allele frequencies. Confidence intervals are reported for the exact distribution (exact), simulated empirical estimated (sim), Fishers estimate for correlation coefficients (fisher), our delta methods based estimate (approx1), and our simple estimate (approx2). Our approximations are accurate especially at higher minor allele frequencies.

For SNPs in the current HapMap, the sample size is so low that even for SNPs of very high minor allele frequency, we cannot be sure if their true r^2 falls within a 95% confidence interval of 0.1. We computed the number of chromosomes needed for accurate calculation of r^2 values near 0.8 over a range of minor allele frequencies using our simple approximation for the variance of the correlation coefficient. Figure 4.1 shows this approximation is accurate for estimation of the lower bound of a confidence interval. For minor allele frequencies between 0.05 and 0.5 we calculated the number of chromosomes needed to get a variance of 0.0008689 for SNPs with an r^2 of 0.8. This is the variance required such that the empirical r^2 has an 95% confidence interval of 0.1. As seen in Figure 4.2, we would need to extend that HapMap to 1003 chromosomes if we wanted accurate estimates of r^2 for SNPs with minor allele frequency of 0.1. The 1000 genomes project will provide almost twice as many haplotypes as this greatly reducing the error due to finite sample size for a large proportion of SNPs.

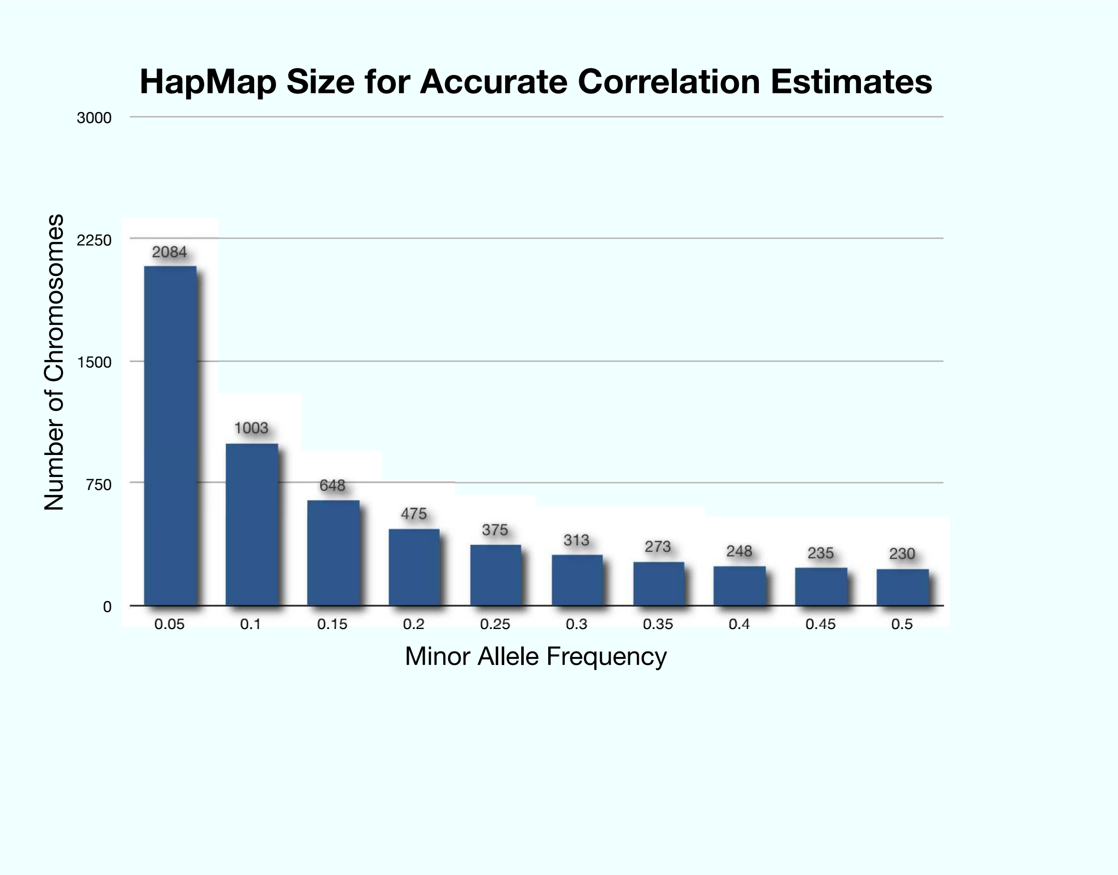

Figure 4.2: Number of chromosomes required for estimates of correlation coefficient to fall within a 95% confidence interval of 0.1. As the minor allele frequency decreases the number of chromosomes required for accurate estimation increases. Note that the current HapMap is not able to accurately estimate correlation coefficients for any minor allele frequency at its current size.

4.3.2 Effects of Finite Sample Size on r^2 and Power.

Figure 4.3 shows the error in estimation of the correlation coefficient due to the finite size of the HapMap. For a range of MAFs 120 correlated pairs of genotypes are generated 10,000 times and their empirical r^2 is used to compute the average value of r^2 in the simulation. The simulations are run with a true value of $r^2 = 0.8$. SNPs with low minor allele frequency are much less accurate in determining r^2 compared to high frequency SNPs. In the context of choosing tag SNPs or follow up SNPs for a case control association study, large numbers of strong tags will have poorly estimated correlation, and many SNPs estimated to have r^2 greater than 0.8 will have much weaker true LD.

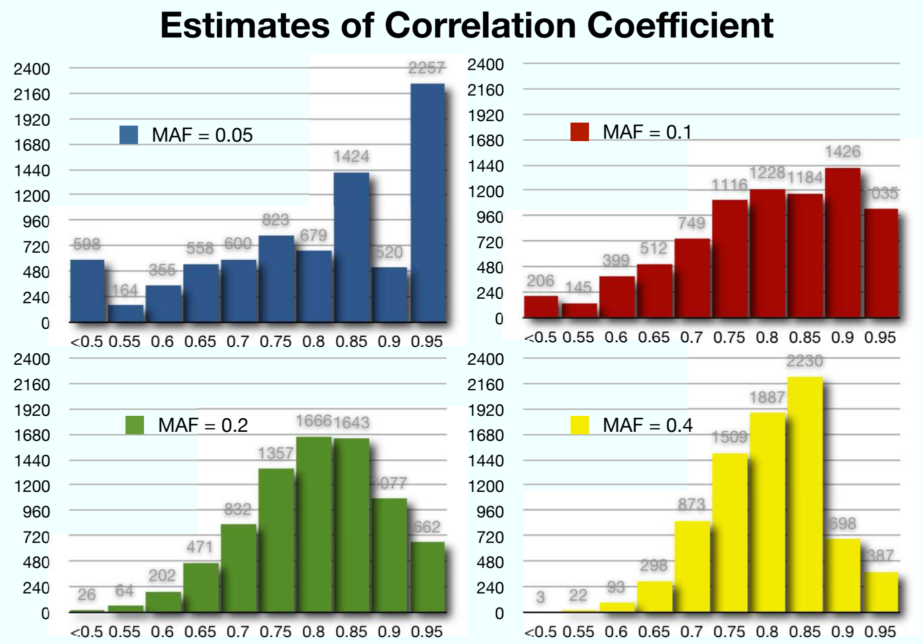

Figure 4.3: Histogram of empirical estimates of the correlation coefficient from a finite sample of 120 pairs of correlated data. The true value of the correlation coefficient is 0.8. The data are simulated such that both SNPs have minor allele frequency 0.05, 0.1, 0.2, or 0.4. The distributions for SNPs with higher minor allele frequency, green and yellow, are more tightly clustered and symmetric around 0.8. The distribution for SNPs with low minor allele frequency, blue and red, are very wide due to the high variance of the correlation coefficient for such SNPs.

The effect of finite sample size on power estimation is measured by comparing power estimates at genotyped SNPs and untyped SNPs based on simulation over a finite data set. This technique for estimating power is common practice as in the methods of [90, 121]. Case control panels of 1000 cases and 1000 controls are generated from 120 chromosomes with a causal SNP minor allele frequency of 0.1 and relative risk chosen such that the power is exactly 50%. The process is repeated 1000 times and the power computed. This entire power estimation process is then repeated 1000 times and the power of each simulation is recorded as shown in Figure 4.4. This estimate of power at a typed SNP is compared to estimating power at an untyped SNP by repeating the experiment above with a correlated SNP with $r^2 = 0.8$ and minor allele frequency of 0.1. Power is measured at the correlated marker and case control status is generated by the original marker. Figure 4.4 shows the power is more accurately estimated at the typed than untyped SNP. The loss in accuracy is due in large part to the finite sample size of the data in the simulation. We calculated the expected number of individuals required to achieve a range of powers between 44% and 56% in the untyped case and included this histogram in Figure 4.4 for reference. Almost 20% of the untyped SNPs have power overestimated by 6%, which is equivalent to having another 160 individuals in the study.

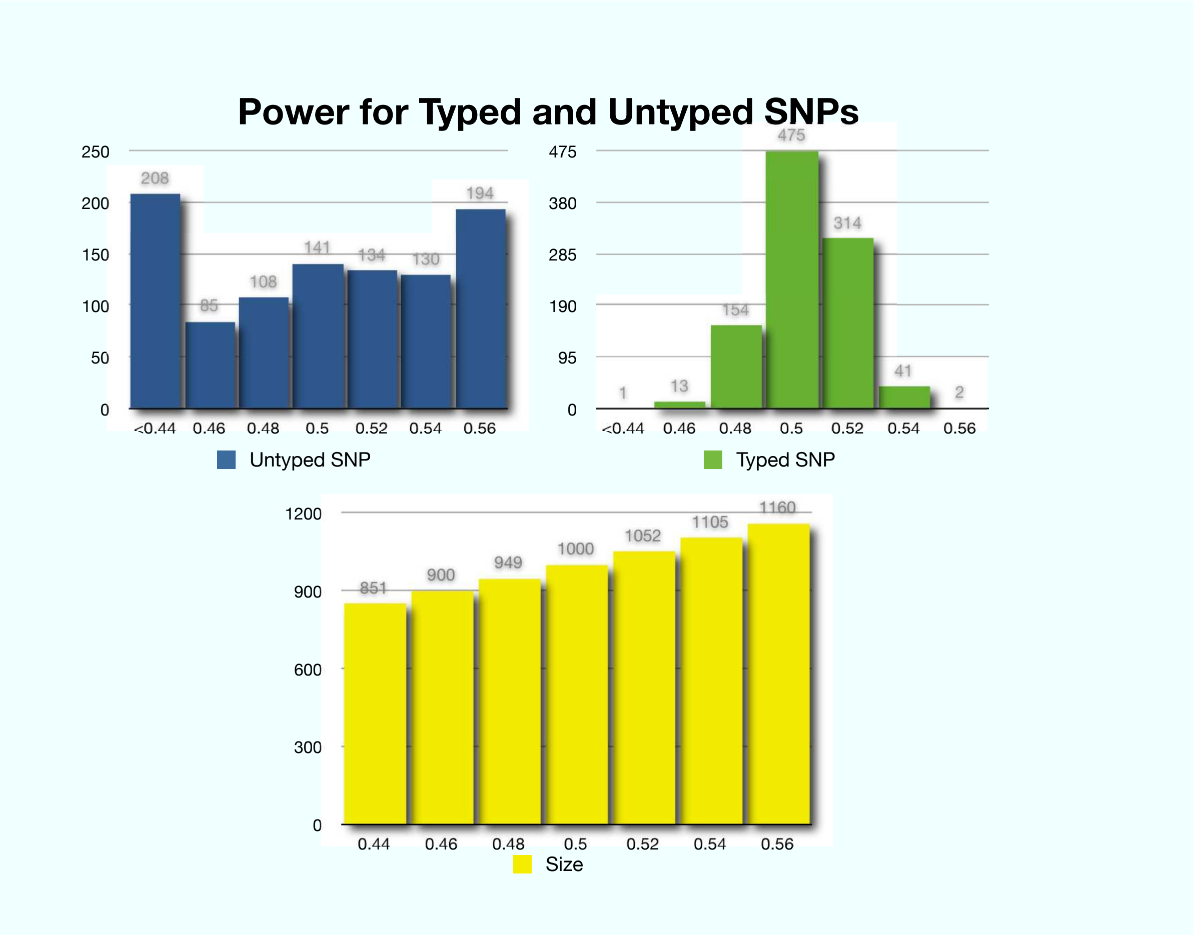

Figure 4.4: Histograms of power for 1000 simulations of case control studies where the causal SNP is typed (green) or untyped (blue). Simulated case control studies were generated by sampling from 120 chromosomes to achieve 1000 cases and 1000 controls. The estimated minor allele frequency of the SNPs were 0.1 and the estimated r^2 between the typed and untyped SNPs was 0.8. Each of the 1000 power estimates is calculated from 1000 simulated case control studies. The relative risk is chosen such that the simulations at typed and untyped SNPs both have an expected 50% power. The error in power estimation is much higher for the untyped case and leads to severe over and under estimation problems. A histogram of the number of individuals to needed achieve power between 44% and 56% in the untyped simulation is included for reference (yellow).

We further examine the issue of power estimation by using the WTCCC data [26] and our analytical estimate of power. First we created a set of tag SNPs by selecting all WTCCC SNPs that passed quality control and are found in the HapMap. Then, for each SNP in the HapMap, we found the best tag SNP in our tag set in terms of r^2. This best tag approach is commonly used to estimate study power. Given a disease model and study size, the power at a HapMap SNP is estimated using its best tag. Given a study size of 2400 cases and 2400 controls, similar to the WTCCC study size, a relative risk of 1.5, and Bonferroni correction factor based on 400,000 SNPs, we measured the probability that the true power of detecting an association at each SNPs was at least 10% less than the estimated power under our simple analytical estimate of the distribution of correlation. We found that over 10% of SNPs have a power at least 10% lower than that estimated under best tag. For example, if the study estimated power for detecting a SNP was 90%, there is a 10% chance that the power is actually $\leq 81\%$ of detecting that SNP.

4.3.3 Effects of Sample Size on Coverage.

Suppose that two SNPs had true correlation coefficients 0.75, but due to the finite size of the HapMap had a variance of 0.01. Then the expected value of the estimated coverage of the third SNP is 0.8. As the number of SNPs in the max increases this problem gets worse, and so current estimates of coverage are highly inflated. There exist many different algorithms for selecting tag SNPs which will each potentially result in different levels of estimated coverage. We examine the two SNP case because it is a subproblem used in many tagging algorithms that try and optimize on an r^2 criterion. In practice the maximization is over many more SNPs and the inflation is therefore even worse. We compared our analytical estimates of this inflation to empirical estimates of the maximum correlation coefficient of 3 SNPs over a range of minor allele frequencies and values of r^2. The minor allele frequency ranges from 0.1 to 0.5 in increments of 0.1 and the r^2 ranges from 0.5 to 0.9 in increments of of 0.1. The number of individuals is fixed at 60. The average error did not exceed 0.006 and is therefore accurate.

We use the WTCCC data[26] to examine the inflation of r^2 in real genotype data. First, we selected 60 individuals from the control population of the WTCCC data set at random. For each SNP in the data set, we found the best r^2 to that SNP amongst the the 300 most proximal SNPs (i.e. the best tag). We then repeated this procedure with 1000 individuals from the WTCCC control population to measure the overestimation in r^2 in this population. The results are shown in Figure 4.5. Although the WTCCC SNPs are much less dense than the HapMap there was still significant inflation in the r^2 values when only 60 individuals are used to estimate r^2. This is due to the higher variance of the correlation when the sample size is smaller. Amongst SNPs with minor allele frequency between 0.2 and 0.3 over 15% of SNPs have an estimated r^2 0.3 greater in the smaller sample size.

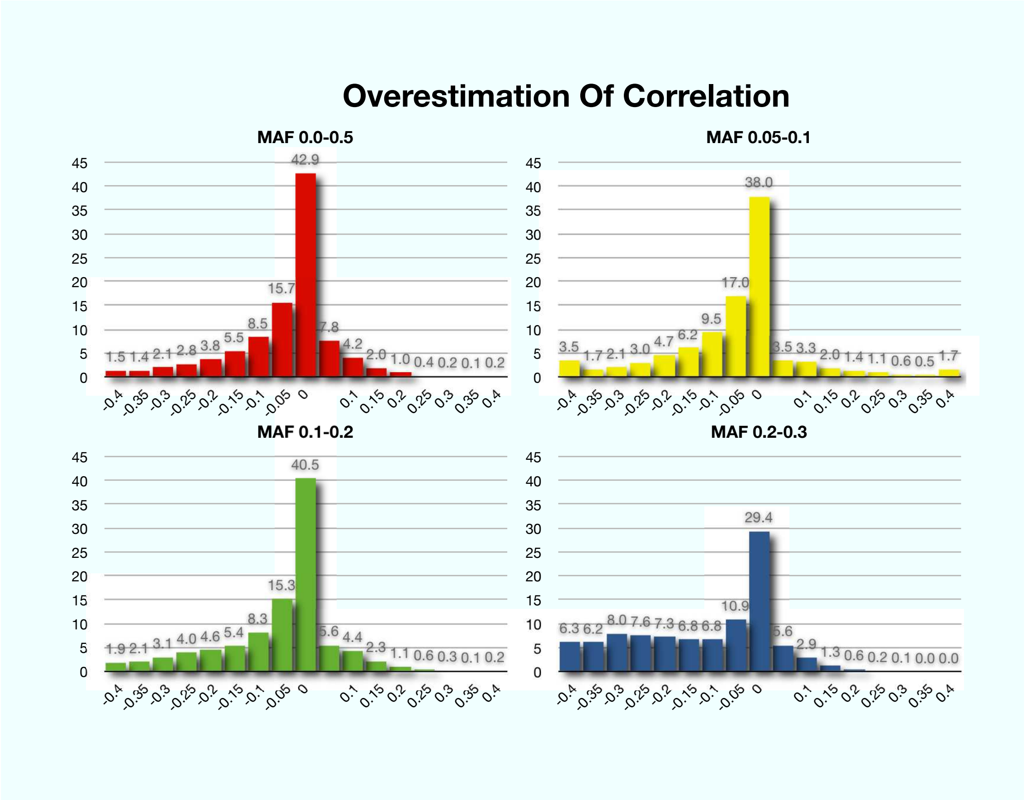

Figure 4.5: Overestimation of r^2 due to finite sample size. We measured r^2 for the best tag of each SNP in the WTCCC data using 60 then 1000 control individuals. We then measured the difference between these to values. The images show the a histogram over the percentage of SNPs with a given difference in r^2. In all ranges of minor allele frequency the values of r^2 over 60 individuals were significantly higher than with 1000 individuals. The difference is greatest amongst the SNPs with minor allele frequency 0.2 - 0.3. This is most likely due to the greater number of SNPs with minor allele frequency in the range, and hence appearing in the maximization term.

4.3.4 Validation of Analytical Results.

We examine the error in empirical estimates of minor allele frequency and conditional probability due to the finite sample size of the HapMap. Sets of correlated binary random variables were simulated to represent SNP genotypes. In the case of the CEU and YRI HapMap populations, there are 60 unrelated individuals or equivalently 120 chromosomes drawn independently from the population. For a range of MAFs we sample 120 binomial random variables and compute the empirical minor allele frequency $\hat{\mathrm{p}}_{\mathrm{A}}$ and the % error $\frac{|\hat{\mathrm{p}}_{\mathrm{A}}-\mathrm{p}_{\mathrm{A}}|}{\mathrm{p}_{\mathrm{A}}}$. This process is repeated 1000 times to get the average estimated minor allele frequency and the average % error. Figure 4.6 shows the results of this simulation.

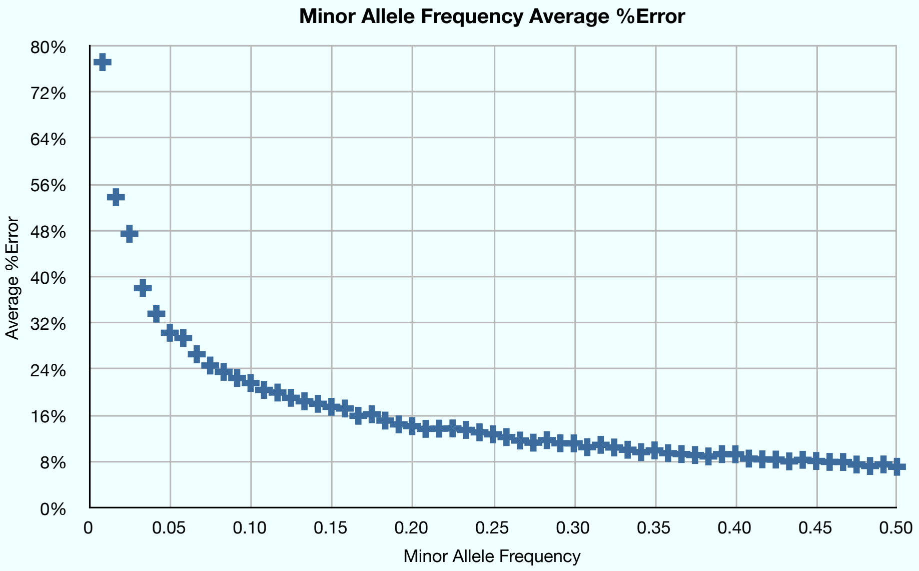

Figure 4.6: The average % error in empirical estimates of minor allele frequency when the sample size is the size of the HapMap. Minor allele frequency error is measured for a given value of MAF by sampling 120 binomial random variables with that MAF and and calculating the % error. This process is repeated 1000 times to get the average error.

We apply a similar procedure to measure the average error in conditional probability estimates due to the finite sample size of the HapMap. We sample 120 pairs of SNPs with a given conditional probability, compute empirical estimates of the frequency and conditional probability from the simulated data, and measure the error. To compute the average error, this process is repeated 1000 times. Figure 4.7 shows the average error of the conditional probability when the first sampled SNP has a minor allele frequency of $0.05, 0.15, 0.25$.

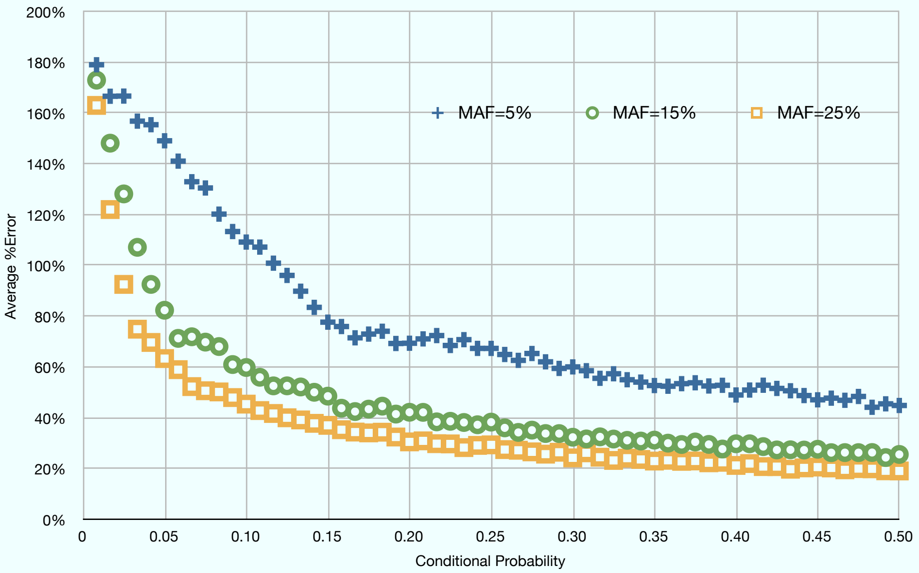

Figure 4.7: The average % error in empirical estimates of conditional probability when the sample size is equivalent to the size of the HapMap. The x-axis shows the true conditional probability, and the y-axis shows the average % error in samples drawn from that distribution 1000 times. Values are given for simulations where the minor allele frequency is 0.05, 0.15, and 0.25. The errors for conditional probability are significantly higher than those for minor allele frequency as shown in Figure 4.6 and contribute greater to the error in estimates of more complicated statistics such as the correlation coefficient and the NCP.

Although these distributions are well known, they are included to demonstrate the substantially larger error of the conditional probability compared to the minor allele frequency. This is due to the lower number of expected observations of the haplotype made from two minor alleles. Consider the case of two SNPs with minor allele frequency 0.1 and conditional probability 0.5. In the HapMap we expect 12 chromosomes with the minor allele for each SNP, but only 6 where both SNPs have the minor allele. Small changes in the sample will therefore have a much larger effect on conditional probability than on minor allele frequency. Our analytical derivations of the distribution of r and NCP given in the Materials and Methods section rely on this relative accuracy of MAF compared to conditional probability. We show empirically that for most values of MAF, our assumption is valid and our estimates are accurate.

We examine the effect of this assumption on our analytical estimates of the mean and variance of the correlation coefficient by sampling correlated binomial random variables and comparing their distribution to the analytical ones we derived. For a range of correlation coefficients and MAFs, 120 pairs of variables were sampled. This was repeated 1000 times to get a mean and variance for the correlation coefficient. Table 4.1 shows the results. Our analytical predictions of the mean and variance are very close to the results of empirical simulation demonstrating that our approximations are valid. The simple analytical estimate that assumes a correct minor allele frequency estimate is not as accurate for low minor allele frequencies. However, the analytical method based on the delta method accurately estimates mean and variance even for low minor allele frequencies.

We used the genotypes from the Welcome Trust Case Control Consortium [26] in order to examine possible deviations in real genotype data as opposed to the sampled binomial random variables. We computed the correlation coefficient for randomly selected pairs of SNPs using 3008 available individuals. Then, we subsampled random collections of 120 genotypes 10000 times and computed the mean and variance of these subsets correlation coefficients. Similarly to the case for simulations using binomial random variables, the analytical methods derived in the Materials and Methods section is highly accurate, with an average error

Table 4.1: Pairs of correlated genotypes were sampled from a distribution with a given correlation coefficient and minor allele frequency as noted in the r^2 and MAF columns respectively. 120 pairs are generated and their empirical correlation coefficients $\hat{\sigma}$ are measured. This process is repeated 1000 times to get the mean and standard deviations of the distribution of the correlation coefficient. We compute the same values using the analytical methods described in the Materials and Methods section to estimate the error in the analytical methods. σ_Δ and $|\hat{\sigma} - \sigma_\Delta|$ are the estimates of standard deviation and the error in the estimate for the delta-method based estimate respectively. σ_S and $|\hat{\sigma} - \sigma_S|$ are the equivalent measurements for the simple analytical estimate. The error is higher for SNPs with low minor allele frequency due to the assumption of a correct minor allele frequency.

r^2	MAF	$\hat{\sigma}$	σ_Δ	σ_S	$\lvert\hat{\sigma} - \sigma_\Delta\rvert$	$\lvert\hat{\sigma} - \sigma_S\rvert$
0.1	0.1	0.138	0.135	0.143	0.003	0.004
0.1	0.3	0.094	0.093	0.094	0.001	0.001
0.1	0.5	0.087	0.087	0.087	0.001	0.001
0.2	0.1	0.136	0.134	0.146	0.002	0.010
0.2	0.3	0.088	0.089	0.091	0.001	0.003
0.2	0.5	0.080	0.082	0.082	0.001	0.001
0.4	0.1	0.126	0.120	0.137	0.006	0.011
0.4	0.5	0.072	0.071	0.071	0.001	0.001
0.8	0.1	0.071	0.069	0.085	0.002	0.014
0.8	0.3	0.046	0.045	0.048	0.002	0.001
0.8	0.5	0.042	0.041	0.041	0.001	0.001
0.9	0.1	0.050	0.048	0.061	0.001	0.011
0.9	0.3	0.033	0.032	0.034	0.002	0.000
0.9	0.5	0.029	0.029	0.029	0.001	0.001

Table 4.2: The standard deviation of the correlation coefficient computed over SNPs rs2381104 and rs4819534 in the control population of the welcome trust case control consortium study [26] computed over a range of sample sizes (N). The empirical standard deviation is denoted $\hat{\sigma}$. σ_Δ and $|\hat{\sigma} - \sigma_\Delta|$ are the estimates of standard deviation and the error in the estimate for the delta-method based estimate respectively. σ_S and $|\hat{\sigma} - \sigma_S|$ are the equivalent measurements for the simple analytical estimate. The HapMap sample size of 120 has a much higher high variance for this SNP than the larger sample sizes. The errors are similar to those for simulated binomial random variables, and demonstrate that the assumption of correct minor allele frequency does not affect the accuracy of the analytical values for SNPs when the minor allele frequency is not low.

N	$\hat{\sigma}$	σ_Δ	σ_S	$\lvert\hat{\sigma} - \sigma_\Delta\rvert$	$\lvert\hat{\sigma} - \sigma_S\rvert$
60	0.142	0.134	0.153	0.009	0.011
120	0.097	0.095	0.108	0.003	0.011
180	0.079	0.077	0.088	0.002	0.010
240	0.068	0.067	0.077	0.001	0.009
300	0.060	0.060	0.069	0.001	0.008
360	0.055	0.055	0.063	0.000	0.008
420	0.052	0.051	0.058	0.001	0.006
480	0.047	0.047	0.054	0.000	0.007
540	0.045	0.045	0.051	0.000	0.006
600	0.042	0.042	0.048	0.000	0.006

of less than 0.001 in estimating the variance. We chose two SNPs rs2381104 and rs4819534 and repeated the experiment with a variety of sample sizes. The results are shown in Table 4.2.

The analytical estimates derived in the Materials and Methods section assume that the distribution of the correlation coefficient is normal. However genotype data are discrete and the correlation coefficient is discrete. The distribution is therefore not normal and moves further from a normal distribution as r^2 approaches 1 or the minor allele frequency approaches 0. We measure the utility of our analytical estimates by measuring confidence intervals for the distribution of the correlation coefficient. We estimated the confidence interval for a range of minor allele frequencies when the true $r^2 = 0.8$ and N=200 chromosomes. First we generated all possible contingency tables for pairs of SNPs and measured their probability with Fisher's exact test. This gave us the exact distribution of correlation and exact confidence intervals. Second, we simulated pairs of correlated binomial random variables and recorded their correlation. The 95% confidence intervals were estimated from 10,000 rounds of simulation. Third, we used Fisher's estimate for confidence intervals of the correlation coefficient. The Fisher estimate was not designed for this setting, and does not depend on the minor allele frequency. It is included for reference. Finally, we used our analytical estimates to generate the 95% confidence intervals. Figure 4.1 shows the results. Surprisingly, the simpler estimate is a more accurate estimate of the lower bound for lower minor allele frequencies. The two estimates converge as the minor allele frequency increases. This is due to the deviation from the normality assumption when minor allele frequency is low. The simpler estimate overestimates the variance (Table 4.1) for low minor allele frequencies in such a way that it more accurately estimates the true confidence intervals.

4.4 Discussion

In summary, we derived analytical distributions for the correlation coefficient and non-centrality parameter in the context of an association study design using a finite data set such as the HapMap. We showed via extensive simulation over real and generated data that our distributions very closely followed the true distributions of these statistics in the same context. This permits quick and accurate examination of the effect of sample size on these commonly used measures, and gives the first exploration on the central data set in genome wide association studies (i.e. the HapMap).

Throughout the work we used a 95% confidence interval of 0.1 and an r^2 threshold of 0.8. Although somewhat arbitrary, they served as a means to ground the results in a familiar setting. The analytical estimates we derived in this study allow quick and accurate examination of alternative thresholds.

We used our analytical distributions to examine the error of current estimates of LD and power based on the current HapMap. We found that the HapMap at its current size does not provide enough data for accurate estimation of these

Table 4.3: The overestimation of r^2 in SNP tagging when taking the maximum of two tag SNPs to cover a third. We generated correlated triples of binomial data such that two tag SNPs had the same minor allele frequency and were associated with a third SNP at the same level of r^2. 1000 samples of 120 triples were used to estimate the empirical maximum of two SNPs ($\hat{max}$), and compared to our analytical formula for the maximum of two SNPs (max). The error (err) is the absolute difference between the empirical and analytical estimates. The inflation of coverage is very significant and suggests that estimates of power based on the best tag for each SNP in the HapMap are overly generous.

MAF	r^2	$\hat{max}$	max	err
0.1	0.5	0.595	0.588	0.006
0.1	0.7	0.780	0.781	0.001
0.1	0.9	0.951	0.954	0.003
0.2	0.5	0.566	0.564	0.002
0.2	0.7	0.761	0.760	0.002
0.2	0.9	0.940	0.938	0.001
0.3	0.5	0.561	0.557	0.005
0.3	0.7	0.751	0.751	0.001
0.3	0.9	0.933	0.935	0.002
0.4	0.5	0.549	0.552	0.003
0.4	0.7	0.751	0.747	0.004
0.4	0.9	0.930	0.931	0.001
0.5	0.5	0.547	0.550	0.003
0.5	0.7	0.749	0.746	0.003
0.5	0.9	0.931	0.931	0.000

central statistics. The variance is especially high for SNPs with low minor allele frequency. This error has impact in the field of human disease genetics, especially for researchers conducting genome wide association studies.

In the design phase of an association study, the appropriate number of individuals to achieve desired power cannot be accurately estimated with the current sample for certain SNPs. Although the effect is small for well covered SNPs recent evidence has shown that genome coverage is much worse than currently estimated [10]. This is further hindered by the overestimate of LD in the Affymetrix and Illumina genotyping platforms. We showed that selecting SNPs with maximal r^2 to find a tag set is heavily upwardly biased. That is, the expected empirical r^2 under such a procedure is significantly higher than the true r^2. The current high throughput genotyping platforms utilized hundreds of thousands of such maximizations in selecting their tag SNPs, and therefore, the true average correlation coefficient of these platforms is likely much lower than found by measurement over the HapMap data. This also produces overestimates in power, since the NCPs are linked to r^2 as described in the Materials and Methods section.

During the analysis phase of an association study, one may select SNPs for follow up based on LD to the tag SNPs found to be significant. The strength of this LD is commonly measured from the HapMap data. As show in Figure 4.3, these estimates have very high variance, and it would not be surprising to have strongly linked SNPs with reported r^2 of less than 0.5, or weakly linked SNPs with reported r^2 greater than 0.8. Thus, one may incorrectly choose to genotype SNPs without strong LD, and miss SNPs that do have strong LD. Our analytical estimates provide a simple way of estimating errors due to finite sample size so that future association studies may avoid these type of errors.

The HapMap has also been used to estimate global significance levels for genome wide association studies. The finite size of the HapMap as evidenced by the high variance of r^2 will lead to observed long range LD even though it does not exist. This reduces the effective number of hypotheses being tested, and therefore alters the global significance level for association. Long range LD has also been examined in the HapMap data [120]. Our findings suggest that at least some of the long range LD is expected due to the size of the HapMap.

Increasing the size of the HapMap will improve its utility to researchers working on discovering the genetics basis of human disease. A larger HapMap, such as that proposed by the 1000 genomes project, will address all of the issues described above and provide a foundation for the new and growing high throughput sequencing technology. While genome wide association studies are well powered for common diseases with causes due to common variants, sequencing can examine rare variants. We demonstrated clearly that statistics for SNPs with low minor allele frequency have the greatest variance. If the community decides to make a similar investment for sequence based studies as they did for genotype based studies, a significantly larger number of individuals must be collected.

Chapter 4, was published in Human Heredity, Vol 68, pp73-86, 2009. Noah A. Zaitlen, Hyun Min Kang, and Eleazar Eskin, “Linkage effects and analysis of finite sample errors in the HapMap”.

Chapter 5

Meta-Analysis of Genome Wide Association Studies

Introduction

The genome wide association study (GWAS) has proven to be a successful method for identifying loci contributing to the genetic basis of complex human diseases. While the list of SNPs and genes correlated with phenotypes continues to grow, many of the discovered variants exhibit only a weak to moderate effect and account for just a small fraction of the total phenotypic variance. Over 75% of loci from completed case control GWAS reporting significant results had SNPs with relative risks (RR) less than 1.4 with 39% less than 1.2. In order to achieve 90% power to capture a SNP with RR=1.2, minor allele frequency (MAF) of 0.2, and genome-wide cutoff of 10^{-6} under a multiplicative model, 15248 individuals must be collected in a balanced study. Over 82% of discovered loci from completed case control GWAS are from studies with significantly fewer individuals and are underpowered to reliably discover these associations [59].

Given this observation, GWAS must be designed with larger numbers of individuals to have sufficient power to identify weaker variants. This requires a large scale effort to collect potentially tens of thousands of individuals, who are then genotyped at hundreds of thousands of single nucleotide polymorphisms (SNPs).

Although the cost of genotyping is dropping, it remains difficult to find, screen, and approve individuals suited for a study. Despite these difficulties, multiple groups are performing association studies on the same disease, each collecting is own case and control cohorts. A natural approach to addressing the lack of power problems of each of these individual studies is to combine the cohorts using meta-analysis.

Meta-analysis is a well studied problem and is currently widely used in the genetics community in the planning and analysis of GWAS. For a review of meta-analysis techniques and pitfalls see Kavvoura et al[69]. Traditional approaches to meta-analysis combine the statistics at each marker from both studies. This approach requires individuals to be genotyped on the same set of SNPs. Since studies often employ different genotyping platforms and different SNPs pass quality control filters in each study, many markers are not shared between studies and are unable to be combined using traditional meta-analysis methods.

Recently, several methods have been proposed which use a reference set such as the HapMap[25] to "impute" ungenotyped SNPs in a study[74, 79, 49]. Provided that the study population is similar to one of the HapMap populations, these *imputation* methods are highly accurate for many of the HapMap SNPs. A straightforward approach to combining studies with different marker sets is to impute the ungenotyped SNPs in each study so that all HapMap SNPs are either genotyped or imputed in both studies. A traditional meta-analysis method may then be applied to the genotyped and imputed SNPs. Indeed, a recent meta-analysis of several GWAS for type 2 diabetes adopts this approach[123]. Unfortunately, not all SNPs are imputed with perfect accuracy. In fact, this accuracy may vary greatly from SNP to SNP. Traditional meta-analysis methods do not take this into account, leading to a reduction in the power of the combined study.

In this work we develop a new method which corrects for potential inaccuracies of imputation by "weighting" each association study depending on the accuracy of the imputation at each marker. We analytically derive an optimal set of weights for combining results from each study and show that it can result in significant increase in power compared to the standard approach. Unfortunately, the optimal weights cannot be computed directly from the data since we do not know the true accuracy of the imputation. However, several methods were recently proposed for estimating these weights. We empirically examine each of these methods to determine which should be used in conducting a meta-analysis of imputed data. Recently, de Bakker et al [37] have analyzed issues relating to conducting meta-analysis in the context of GWAS. In particular they suggested incorporating estimates of imputation accuracy into the meta-analysis statistic by scaling the number of individuals by the SNP information measure. In this work, we demonstrate that such scaling by the (unknown) correlation between true and imputed genotypes maximizes the statistical power of the study. Our method for estimating the weights are equivalent to estimating this correlation.

We conduct several experiments to show that our new method for handling imputed genotypes from distinct SNP sets improves the power of meta-analysis. We simulate case control studies using the HapMap and Welcome Trust Case Control Consortium (WTCCC) data sets with distinct SNP sets. For each pair of studies we show that our meta-analysis method improves the power of the overall study compared to the traditional method of combining Z-scores based on study size.

Material and Methods

Case Control Studies. In a case control study individuals are collected from two groups, the cases and the controls. The individuals in each group differ along a phenotype of interest, such as disease state, but are otherwise members of the same population. The individuals are genotyped on a set of single nucleotide polymorphisms (SNPs), and the allele frequency of each SNP s_i is measured in the cases $\hat{p}_i^+$ and in the controls $\hat{p}_i^-$. Assuming a study with $N/2$ cases and $N/2$ controls where the true SNP frequencies in the population, cases, and controls are p_i, p_i^+, and p_i^- respectively, the Z-score statistic Z_i in equation (5.2) is computed for each SNP. It is normally distributed with mean equal to the non-centrality parameter (NCP) $\lambda_i\sqrt{N}$ and variance 1. Those SNPs with statistic $|Z_i| > \phi^{-1}(\alpha/2)$ where $\phi^{-1}(x)$ is the quantile function of the standard normal distribution and α is the significance threshold, are considered significant and maybe linked to a causal variant for the phenotype.

$$Z_i = \frac{\hat{\mathrm{p}}_\mathrm{i}^+ - \hat{\mathrm{p}}_\mathrm{i}^-}{\sqrt{2/N}\sqrt{\hat{\mathrm{p}}_\mathrm{i}(1-\hat{\mathrm{p}}_\mathrm{i})}} \sim \mathcal{N}\left(\lambda_i\sqrt{N}, 1\right) \tag{5.1}$$

$$\lambda_i\sqrt{N} = \frac{(\mathrm{p}_\mathrm{i}^+ - \mathrm{p}_\mathrm{i}^-)\sqrt{N}}{\sqrt{2\mathrm{p}_\mathrm{i}(1-\mathrm{p}_\mathrm{i})}} \tag{5.2}$$

Traditional Meta-Analysis. A standard approach in meta-analysis is combining the Z-scores of several association studies. This approach has recently been applied in the meta-analysis of several type 2 diabetes GWAS studies [123]. The data required from each study are the statistics $\mathrm{Z}_\mathrm{i}^\mathrm{j}$ for each SNP i in each study j, and the number of individuals N^j in each study j, and assume an equal number of cases and controls (although this is easily changed), and that the case and control frequencies at SNP i are the same across all studies, that is, $p_i^{j+} = p_i^+$ and $p_i^{j-} = p_i^-$ for all j, which implies that $\lambda_i^j = \lambda_i$. This is equivalent to assuming that the relative risk of the causal SNP is the same for all studies. These assumptions maybe unrealistic in the case that the studies are performed over different populations and that the causal variant is different or acts differently in different populations.

For each SNP s_i in the studies we compute the meta-analysis statistic $\mathbf{M}_i$ which is a weighted sum of Z-scores defined in equation 5.3.

$$\mathbf{M}_i = \sum_j \frac{w^j \mathrm{Z}_\mathrm{i}^\mathrm{j}}{\sqrt{\sum_j (w^j)^2}} \sim \mathcal{N}\left(\sum_j \frac{w^j \lambda_\mathrm{i}^\mathrm{j}\sqrt{N^j}}{\sqrt{\sum_j (w^j)^2}}, 1\right) \tag{5.3}$$

$\mathbf{M}_i$ is defined for any weights w_i^j which are positive and with at least one w_i^j greater than zero. The statistical power of the statistic $\mathbf{M}_i$ to detect associations depends on the weights and is maximized when the weights $w_i^j = \sqrt{N^j}$. Intuitively we are assigning larger weights to studies with more individuals and therefore with more power to detect an association. The optimality of these weights is shown with a direct application of the Cauchy Schwartz inequality $\sqrt{\sum_j (w^j)^2}\sqrt{\sum_j (\lambda_\mathrm{i}^\mathrm{j}\sqrt{N^j})^2} \geq \sum (w^j \lambda_\mathrm{i}^\mathrm{j}\sqrt{N^j})$. Since $\lambda_\mathrm{i}^\mathrm{j} = \lambda_i$ for all j there is equality when $w^j = \sqrt{N^j}$.

Imputation. Unfortunately, the set of SNPs genotyped in a GWAS, or "tag" SNPs, are not identical between studies, so the $\mathrm{Z}_\mathrm{i}^\mathrm{j}$ required for meta-analysis are not immediately available. Furthermore, the set of tag SNPs is much smaller than the total number of SNPs in the population and it is likely that the causal variants are not contained in the tag SNP set. Recently, several methods have been developed to leverage existing data sets with millions of genotyped SNPs, such as the HapMap, to improve the power of association studies. If the study population is closely matched to a HapMap population it is possible to measure statistics over SNPs not included in the set of tag SNPs. In addition to improving the power of association studies, imputation methods can be used to aid meta-analysis of association studies that used different sets of tag SNPs by computing statistics at SNPs missing from either study but contained in the HapMap. Meta-analysis is performed by imputing the missing SNPs in each study and computing a statistic $\mathrm{Z}_\mathrm{i}^\mathrm{j}$ for each SNP i in the HapMap and each study j. Provided that all of the tag SNPs in each study are contained in the HapMap, this procedure will provide the required statistics to perform meta-analysis at all SNPs in both studies as well as all HapMap SNPs not contained in either study.

Due to linkage disequilibrium (LD) between proximal SNPs in the genome, a difference in frequency between cases and controls at a causal SNP may cause a similar difference at a nearby tag or imputed SNP. The NCP at a tag SNP is a function of relative risk, disease model, MAF, study size, and correlation coefficient to the causal variant. Let $\lambda_i\sqrt{N}$ be the NCP of tag SNP s_i in a case control study. Imputing s_i instead of genotyping directly will alter its NCP. We define $r_{i,j}^2$ as the square of the correlation coefficient between the imputed genotypes and the true genotypes of SNP s_i in study j. Intuitively, if $r_{i,j}^2$ is close to 1 then the SNP is imputed well and the NCP will be close to $\lambda_i\sqrt{N}$, and if $r_{i,j}^2$ is close 0 then little information is known about the true genotypes of s_i and the NCP will be close to 0. The NCP of an imputed SNP is equal to $r_{i,j}\lambda_i\sqrt{N}$, a function of the NCP of the SNP it is imputing as well as the correlation coefficient between the imputed and true genotypes.

5.0.1 Imputation Aware Meta-Analysis.

The statistic $\mathrm{Z}_\mathrm{i}^\mathrm{j}$ computed for an imputed SNP does not necessarily share non-centrality parameters across studies. The assumption that $\lambda_\mathrm{i}^\mathrm{j} = \lambda_i$ from the simple meta-analysis described above is still valid. However, the correlation between the imputed and true genotypes my vary from study to study affecting the NCP. Consider the situation presented in Table 5.1. Two different tag sets are used to impute a HapMap SNP s_H. The linkage patterns between s_H and the two different tag sets result in poorer imputation for Tag Set 1 then Tag Set 2. Suppose that two studies of N/2 individuals, study 1 and study 2, use these tag sets and that $r_{H,1} = 0.7$ and $r_{H,2} = 0.95$. Since both studies have N/2 individuals the NCPs will be $0.7\lambda_i\sqrt{N}$ in study 1 and $0.95\lambda_i\sqrt{N}$ in study 2. Given this result, the derivation for M_i in the simple case above no longer holds. Treating the statistics $\mathrm{Z}_\mathrm{i}^\mathrm{j}$ as the equivalent of directly genotyped SNPs may weaken the meta-analysis power. Our objective is to develop a new meta-analysis statistic which accounts for the imputation error.

Table 5.1: Two tag sets with different markers s_1, s_2, s_3 and s_2, s_4, s_6 will have different accuracies in imputing the HapMap SNP s_H. Four example individuals are shown for each tag set with genotype dosages 0, 1, and 2 representing homozygous minor, heterozygous and homozygous major alleles. In this case Tag Set 2 is more accurate than Tag Set 1 with imputed genotypes $\hat{s_H}$ much closer to the true genotypes s_H.

Tag Set 1					Tag Set 2				
s_1	s_3	s_5	s_H	$\hat{s_H}$	s_2	s_4	s_6	s_H	$\hat{s_H}$
0	1	1	1	0.8	1	0	0	1	0.95
2	1	2	2	1.6	1	2	2	2	1.95
2	2	0	2	1.5	2	0	1	2	1.99
0	0	2	0	0.5	1	2	2	0	0.04

Adopting the same framework as the traditional method we wish to find a set of weights w_i^j such that a weighted combination of the $\mathrm{Z}_\mathrm{i}^\mathrm{j}$ from each study will maximize $\mathbf{M}_i$. The w_i^j we propose is $\lambda_\mathrm{i}^\mathrm{j}\sqrt{N^j} = r_i^j\lambda_i\sqrt{N^j}$. Since the λ_i is fixed across studies this is equivalent to $w_i^j = r_i^j\sqrt{N^j}$. In this case we consider not only study size, but also the quality of the imputed genotypes. Provided that the imputed genotypes are accurate estimates of the probability of the true genotype given the observed tag SNP genotypes, poorly imputed SNPs will have low non-centrality parameters because their $r_{i,j}$ will be close to zero. A large study with poorly imputed genotypes for a SNP will not alter the meta-analysis statistic significantly if there exists a smaller study that genotypes the SNP directly. The proof of optimality once again follows from a direct application of the Cauchy Schwartz inequality.

To understand the effect of this new statistic consider a SNP s_i in a two study meta-analysis where each study has $N/2$ cases and $N/2$ controls. Suppose study 1 genotypes the SNP directly and that in study 2 the SNP is imputed, that is, $r_{i,1} = 1$ and $r_{i,2} = r$. Then in order to maximize power we must maximize the NCP of the meta-analysis statistic $\mathbf{M}_i$. We set $w_i^1 = 1$ and $w_i^2 = r$ and get NCP of $\mathbf{M}_i = \sqrt{1+r^2}\lambda_i\sqrt{N}$. If instead we choose to follow the standard method for meta-analysis and set $w_i^j = 1$ for all j, then we get NCP of $\mathbf{M}_i = \frac{1+r}{\sqrt{2}}\lambda_i\sqrt{N}$. In this case if r $\leq \sqrt{2} - 1.0$ then the meta-analysis will have even less power than than either study alone. If both studies impute the SNP then the potential for loss of power compared to our method is even greater.

5.0.2 Estimating Imputation Correlation.

We showed that the correlation between the true and imputed genotypes $r_{i,j}$ are the weights which maximize the power of the meta-analysis. These weights can not be computed directly since the true genotypes are unknown. There exist several methods for estimating $r_{i,j}$ which we describe here.

First we describe two methods using the HapMap haplotypes. They both assume the linkage patterns in the HapMap resemble those in the study population. The conditional probability of observing a genotype at the imputed SNP given the tag SNPs can be computed in the HapMap data. This in turn can be used to estimate the correlation coefficient. The derivation is presented as part of the WHAP method described in Zaitlen et. al[121]. We call denote this estimate of correlation $E_W(r_{i,j})$. Equivalent empirical measurements can be made (at higher computational expense) for other imputation methods such as MACH, IMPUTE, and BIMBAM [74, 79, 49]. This second estimate of $r_{i,j}$ for an imputation method is computed by restricting the HapMap to the set of SNPs used in the study. Then, leaving out each HapMap individual in turn, the remaining individuals are used to impute the genotypes. The empirical estimate for $r_{i,j}$ is the correlation coefficient between the imputed and true HapMap genotypes. It is denoted $E_C(r_{i,j})$ and is called the cross validation estimate since we are using leave one out procedure to compute it.

In addition to these HapMap based approaches several estimates of imputation quality relying solely on the imputed genotypes have been proposed. One such estimate of $r_{i,j}$ proposed by MACH [74] is called $r^2 - hat$, which we write $E_H(r_{i,j})$. It is the ratio of the empirical variance of the imputed genotypes to the expected variance given the imputation estimate of the minor allele frequency. Provided that the imputed genotypes are the expected dosages given the observed genotypes then this will be the expected correlation coefficient.

Differences between the study population and the HapMap, the genotyping density, and the finite size of the HapMap can effect both the empirical and analytical estimates of correlation. We examine the relation between the true $r_{i,j}$ and the estimates $E_C(r_{i,j}), E_H(r_{i,j})$ of imputation quality over several data sets. We show that the correlation is estimated closely enough in most cases to warrant the use of our new meta-analysis statistic over the traditional method when combining imputed genotypes.

5.1 Results

5.1.1 Power Simulations.

The difference in power between traditional meta-analysis and our imputation aware meta-analysis method is explored by simulating pairs of case control studies. For every pair we record the power of each study as well as the power of each type of meta-analysis. Figure 5.1 shows the results of three such simulations. In each of these simulations both studies contain 2000 individuals with equal numbers of cases and controls. The disease model is multiplicative with an odds ratio of 1.203 and a causal SNP minor allele frequency of 0.05, giving an expected power of 50%. The genotypes in each study are generated as conditional binomial random variables with some correlation coefficient r to the causal variant. An r of 1.0 means that the causal variant and the generated genotypes are identical. For each study we compute the Z-score and if the corresponding p-value is less than 0.05 we consider it successful. We also compute the weighted combination of the Z-scores from both studies according to the traditional method and our imputation aware method. This process is repeated 1000 times and the power of the four methods is computed as the fraction of times a successful test occurred. In each simulation our imputation aware meta-analysis statistic matched or beat the power of the traditional method. The difference between the methods is especially large when the quality of imputation is poor. In some circumstances traditional meta-analysis power can be even lower than the power of an individual study, but this is never the case for the imputation aware statistic. Filtering poorly imputed SNPs has been suggested as means for addressing this issue [123]. This may prevent power loss beyond each of the individual studies if the threshold is high enough, but it will not prevent a power loss compared to the imputation aware statistic.

5.1.2 Correlation Coefficient Estimates.

Unfortunately the optimal weighting of the Z-scores from individual studies cannot be computed from the data. Instead, the correlation between the true and imputed genotypes must be estimated. We can estimate this correlation using the HapMap as described in the Methods section. Since this procedure is computational expensive for most imputation methods we also examine the estimate of r^2 called $\hat{r^2}$ defined by MACH [74].

We examine the quality of these approaches over real genotype data in order to asses the feasibility of using our imputation aware meta-analysis method without access to the true value of r_i^j. Using the controls from the WTCCC we randomly removed one quarter, one half, and one third of the genotyped SNPs producing three new data sets for chromosomes 1,2, and 22. For each data set we imputed the removed SNPs and computed the true value of r_i^j for each SNP. We then estimated

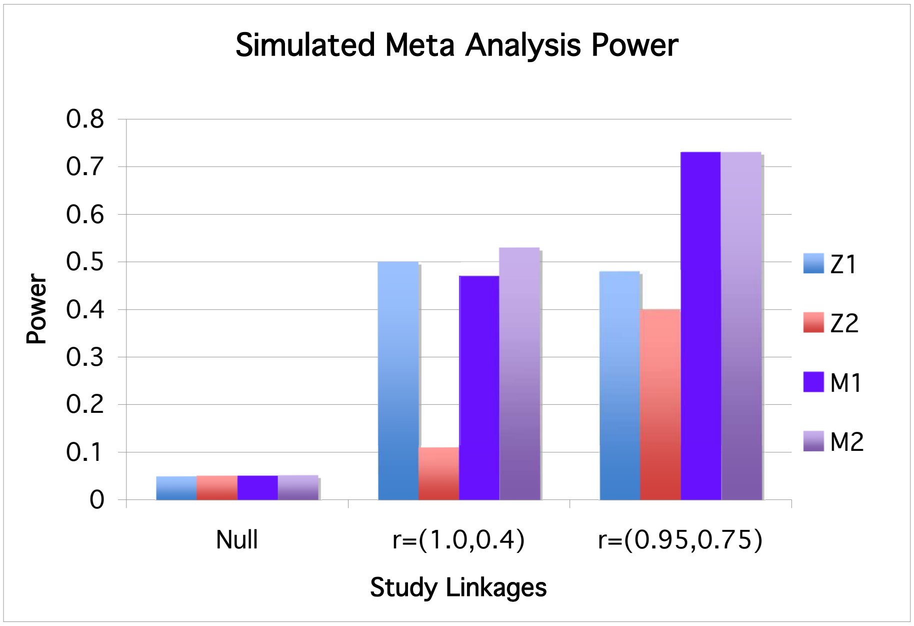

Figure 5.1: Power of simulated studies. Z1 is the power of study 1, Z2 is the power of study 2, M1 is the power of the traditional meta-analysis method, and M2 is the power of the imputation aware meta-analysis method. In the Null example the genotypes are completely unlinked to the causal variants in both study 1 and study 2. In the second example, study one genotypes the causal variant directly and study 2 imputes it with r = 0.4. In the third example study one and study two both impute the SNP with r = 0.95 and r = 0.75 respectively. Notice that the imputation aware meta-analysis method matches or beats the power of the traditional method in each case, and that in the second example the power actually drops in the traditional method due to poor imputation quality that is not accounted for in the second study.

this correlation coefficient using each of the methods described above. For all but the most sparsely genotyped data set and SNPs with low MAF the value of $\hat{r^2}$ very closely approximates the true r_i^j. In the densest data set both the $\hat{r^2}$ and the cross validation method exceeded 0.7. The performance is noticeable poorer on the sparsest data set, but most current studies and genotyping platforms exceed this density considerably. The imputation method EMINIM used in this analysis permits a quick estimate of this estimated correlation accuracy by automatically performing a leave one out imputation estimate of all genotyped SNPs.

5.1.3 Power Simulation with Error.

We repeated the experiments show in Figure 5.1 with values of r sampled from the error observed in the experiments above. Although there was a slight drop in power, this method is still more power than traditional meta-analysis for most error levels, and we conclude that the error in estimating r is not large enough to warrant abandoning the imputation aware meta-analysis statistic.

5.1.4 Power Simulation with Real Data.

We used the WTCCC data set for T2D to generate two studies of equal size with different marker sets. We randomly partitioned the cases and controls into into two studies. For each of these new studies we removed every other, every third, or every fourth SNP. In the first study we started at the first SNP, and in the second study we started at the second SNP so that different SNPs were removed from each group. We then imputed in each of the studies and combined the results using the traditional and imputation aware meta-analysis methods. We compared these results to those of the original WTCCC study.

In both cases all the significant SNPs remained significant. However, the order of the top 100 SNPs changed between the original, traditional, and imputation aware methods. Our method was the most concordant with the original study. The traditional method output the same significant results but the rank of SNPs was further from the original study. We took this as evidence that our imputation aware meta-analysis was more robust to errors from imputation.

5.2 Discussion

Currently, meta-analysis of genome wide association studies are commonly performed using a weighted sum of Z-scores approach. This well established method linearly combines the results of each study weighting them by their size. In this way, larger studies are up-weighted relative to smaller ones and their results have greater influence in the final meta-analysis statistic. GWAS do not necessarily contain the same set of genotyped SNPs and so additional work must be done before meta-analysis can be conducted. Specifically, an imputation method is used to estimate the genotypes of SNPs absent from either study. Typically, Z-scores over these imputed SNPs are then combined between studies using the traditional method.

Although this method is optimal under certain reasonable assumptions it does not take into account errors from imputation of genotypes. Thus a large study that poorly imputes a genotype will be given more weight than a smaller study that imputes it perfectly. In this work we introduce a novel meta-analysis statistic to deal with this issue of imputed genotypes in meta-analysis. Specifically, we adjust the weighting scheme of the traditional method to take into account the accuracy of the imputed genotypes. The new weights are function of both sample size and the correlation coefficient between the imputed and true genotypes. We show that our method is optimal under the same set of assumptions as the traditional approach. In addition, we show that for many cases our new statistic not only improves the meta-analysis power, but prevents a loss in power compared to each individual study that can occur when SNPs are poorly imputed.

Unfortunately, the optimal weights in our statistic are not computable from the results of GWAS and imputation. However, there exist several techniques for estimating them either directly from the imputed data or with a secondary data set such as the HapMap. We performed several experiments to examine the accuracy of these approaches and found that although there are slight differences in accuracy depending on minor allele frequency and tag set density, for most current studies either approach is accurate enough to estimate the weights effectively. That is, the power of the meta-analysis will still be more power using our new method with estimated correlation coefficients than using the previous method, which ignores imputation issues altogether.

Finally, we simulated case control studies using real genotype data from the Welcome Trust Case Control Consortium studies. In each study we imputed genotypes using EMINIM and performed a meta-analysis using traditional and our new imputation aware statistics. We showed that our method matches our outperforms the traditional approach in all scenarios examined. Thus we suggest the adoption of our statistic for future meta-analysis of GWAS studies that use imputed genotypes.

Chapter 5, in part is currently being prepared for submission for publication of the material. Noah Zaitlen, Eleazar Eskin. The dissertation author is the primary investigator and author of this material.

Chapter 6

NCBI Phasing

6.1 Introduction

Many risk factors for human disease are accounted for by variation in DNA sequence[18]. The most common type of human sequence variation consists of differences in individual base pairs termed single nucleotide polymorphisms (SNPs) [116, 17, 56]. It has been estimated that there are about 7.1 million common biallelic SNPs with a minimum minor allele frequency of 5%. These SNPs appear on average once every 450 base pairs [72]. In recent years, a tremendous number of single nucleotide polymorphisms (SNPs) have been discovered and deposited into NCBI's dbSNP public database. Today, dbSNP contains information for over 10 million human SNPs with over 5 million of them validated. More recently, a significant amount of genotype data has been deposited as well. More than 2.7 million human SNPs in the database have genotype information. This data resource consists of over 286,757,371 genotypes over 3,285 individuals split into 417 data sets. This data resource provides an invaluable resource for understanding the haplotype structure of human variation and discovering the genetic basis of human disease. Analysis of these data sets will allow for the design of effective whole genome association studies designed to identify the genetic contribution to the manifestation of complex diseases.

The database contains two whole genome human variation maps, one deposited by the HapMap project[24] covering four populations, and the other deposited by Perlegen Sciences[60]. The database also contains a significant amount of sequenced gene data from the Environmental Genome Project and the SeattleSNPs project in addition to many other smaller data sets. Each data set has different properties such as the number of individuals genotyped, average SNP density, genome coverage and types of genomic regions genotyped. These inherent properties may bias inferences drawn from the analysis of any one of these data sets. Analysis of multiple data sets with different properties genotyped over the same region in the human genome provides opportunity for a more complete analysis of human variation.

Alleles of SNPs which are physically located in close proximity to each other on a chromosome are often correlated (i.e. in "linkage disequilibrium") with each other. Thus, within most short regions, there is limited genetic variability, and only a small number of allele sequences (haplotypes) exist in a population. Empirical studies investigating different regions of the genome show that haplotype structure varies considerably. In a typical region or "block of limited diversity", three or four common haplotypes often account for at least 80% of the sequence variation in a population [89, 28, 46]. Studies show that some blocks can extend over 100 kb while others only extend less than 10 kb. The haplotype structure of a given region depends on evolutionary and population genetic factors such as mutation and recombination rates, selection, and population history.

Obtaining the haplotypes and partitioning the region into blocks of limited diversity are the first steps for many types of analysis of human variation. However, since humans are diploid, phase (or haplotype) information is not immediately available. Therefore, the construction of haplotypes from the diploid genotype information (i.e., phasing the genotypes) requires statistical inference or the financially prohibitive collection of extended pedigrees. Consider for example two SNPs lying on the same chromosome, both with alleles A and G. If both SNPs are observed as heterozygous, it is unclear whether one chromosome contains allele A at both loci and the other chromosome contains allele G in both loci, or whether one chromosome contains allele A at the first locus and allele G at the second locus and the other chromosome contains alleles G and A, respectively (Fig. 6.1).

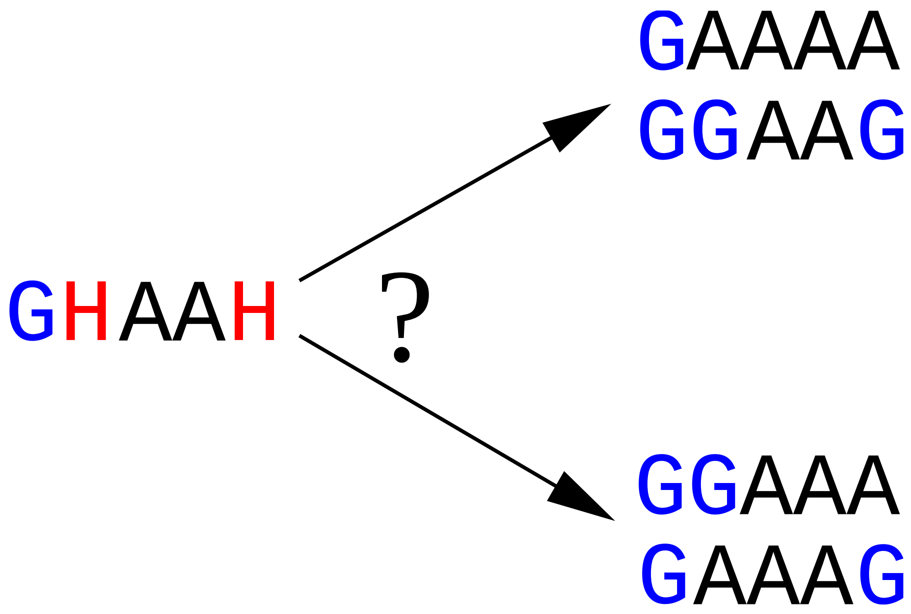

Figure 6.1: A genotype for 5 SNPs (left) and two possible phasings of the genotype into pairs of haplotypes (right) demonstrating the inherent ambiguity of haplotype phasing. Each SNP has possible bases of "A" and "G". "A" and "G" positions in the genotype represent homozygous genotypes at a particular SNP and an "H" position represents a heterozygous genotype at a particular SNP. From only the observed data, it is impossible to determine which haplotype phasing is correct.

In order to overcome this problem many computer programs have been designed to estimate and assign phase from diploid genotype data [106, 87, 54]. In order to compute the full set of haplotypes for dbSNP, we used HAP [54], a phasing program which determines haplotypes by exploiting the correlation between SNPs in physical proximity due to linkage disequilibrium using a genealogy based model (perfect phylogeny [63]). The perfect phylogeny model assumes that in short regions there has been no recombination nor recurrent mutation throughout human history (Fig. 6.2). HAP assumes that over a short genomic region the haplotype structure is close to a perfect phylogeny.

Computing the haplotypes for the complete set of more than 286 million genotypes over 417 data sets of dbSNP is a tremendous computational task. In this work, we describe how we modified the phasing program HAP[54] to scale up to this task and generate a set of haplotypes for all of the genotype data sets in dbSNP. The main advantage of HAP is its speed since the algorithms running time scales linearly with the number of SNPs. HAP is able to process up to 40,000 SNPs at a time which allows for phasing and partitioning into blocks the entire dbSNP genotype database in under 24 hours on a 30 node cluster. HAP has also been extended to incorporate pedigree information into the phasing which is available for many of the data sets. The predicted haplotypes are deposited in dbSNP and will be made publicly available to the community. The accuracy of HAP has previously been tested on various regions of the genome [42, 54] and it has proven to phase correctly 97% of the heterozygous SNPs, which is comparable in accuracy to other established methods. In addition, we measured the accuracy of HAP on unrelated individuals by considering genotypes collected from mother, father, child pedigrees in the HapMap data. We use HAP to phase the parents and compared the predicted haplotypes to the haplotypes inferred by using the child genotypes. The error rate of HAP is 1.5% which is comparable to the error rate of PHASE[106].

Other phasing methods such as PHASE [106] could potentially have also been used for the phasing task, only that the computation time for PHASE is too long to make the phasing feasible over the whole database. We measured the running time of PHASE over randomly selected regions of the HapMap data obtained from dbSNP. From these experiments it is not clear how long it would take for PHASE to predict the haplotypes for the database because of the high variance in running time and the fact that it does not appear that PHASE scales linearly with the number of SNPs. Nevertheless, it appears that applying PHASE to the entire database is computationally infeasible.

One of the main contributions of this work is the organization of the data sets in a way that corrects for errors in the strand and physical location annotations of the SNPs submitted to dbSNP. Since all of the data sets are submitted to dbSNP and mapped to the current human genome build, using dbSNP we can easily extract multiple genotype data sets for the same genomic region. For example, researchers interested in the ABO gene, can easily obtain haplotype and genotype data from multiple data sets including the HapMap, Perlegen, and SeattleSNPs.

Since many of the data sets were mapped to different human genome builds, reconciling the original data sets and mapping them to a common genome build is a very time consuming task. In addition, if there are changes to the physical location of SNPs or strand errors in the genotypes which are corrected, the data sets in dbSNP will reflect these changes.

We perform preliminary analysis on the haplotypes focusing on measuring the consistency of the haplotypes in the same region from different data sets. Since some of the data sets are computed from the same individuals, we can observe how the SNP density significantly affects the inferred haplotype and block structure in a region. By combining high density data from Seattle SNPs and the Perlegen whole genome analysis, we show how the number of haplotypes in the blocks defined by the Perlegen data sets are underestimated by a factor of 3.6. These differences illustrate the advantage of examining multiple data sets when inferring human variation structure.

We find the chimpanzee haplotypes corresponding to each human haplotype block by mapping all the SNPs typed to the UCSC BLASTZ alignment of the human and chimp genomes. These haplotypes will also be made available for download at dbSNP.

6.2 Results

6.2.1 Data Description

The human portion of the dbSNP database contains 286,757,371 total genotypes from 4,284 individuals over 2.7 million SNPs partitioned into 417 data sets. 835 of the individuals have genotypes from two or more data sets. The CEPH families for example were used in several different genotyping studies.

Two whole genome data sets compose 94.2% of the genotypes. The HapMap data set which contains 159,862,776 genotypes taken from four populations consisting of a total of 270 individuals over 954,302 SNPs and the Perlegen data sets which consists of 110,385,051 genotypes taken from three populations consisting of a total of 71 individuals over 1,576,578 SNPs. In addition to these data sets there are an additional 16,509,544 genotypes from other data sets. dbSNP contains a significant amount of genotypes derived from sequenced data which includes the SeattleSNPs (PGA/UW) data which consists of 573,194 genotypes taken from two populations consisting of a total of 48 individuals over 15,981 SNPs in a total of 177 sequenced genes and the Environmental Genome Project (EGP) sequenced genes which contains 3,184,170 genotypes over 37,737 SNPs in a total of 304 sequenced genes in 90 individuals. The 48 individuals in SeattleSNPs are the same individuals in the Perlegen data. Some of these data sets contain a much larger number of individuals such as the SNP Consortium (TSC) Celera CEPH data set which contains 691 individuals and a data set from Perlegen containing 655 individuals from Mexico City. Others data sets contain many populations (such as the TSC data set which contains 17 populations). Table 6.1 summarizes the contents of the top 10 data sets contained in dbSNP.

Table 6.1: Summary of genotype data contained in dbSNP. The NIHPDR data contains a single mixed population. Pops is the number of populations, Inds is the number of individuals, Density is the average SNP density, and Ref is the reference to the publication about the data set.

Data set	Genos	SNPs	Pops	Inds	Density	Ref
HapMap	159,862,776	954,302	4	270	3,149	[24]
PERLEGEN WG	110,385,051	1,576,578	3	71	1,938	[60]
Affymetrix	6,189,466	125,778	6	116	24,029	[70]
TSC	4,932,382	19,048	17	1963	312,754	[48]
EGP	3,184,170	37,737	1	90	72,443	[76]
PGA/UW	573,194	1,5981	2	47	153,861	[27]
IIPGA	176,162	3,801	3	47	430,361	[91]
NIHPDR	159,549	1,982	1*	448	1,419,125	[23]
WICVAR	33,240	1,462	1	130	2,011,277	
HG_BONN_CNS	24,522	320	1	143	5,284,550	[44]

In dbSNP, each genotype is mapped to the human genome consistent with the latest available build. Since many of the original data sets were released at different times, the data sets were mapped to different human genome builds. Since the position of SNPs change slightly from build to build, the genome positions listed for the SNPs in each build are not necessarily consistent between the different data sets. The same SNP genotypes in two different data sets may appear to be at a different position due to mapping to different builds.

Each build of dbSNP maps each SNP to the correct position of the human genome. Each genotype data set in dbSNP contains references to the dbSNP identifier for each genotyped SNP. Any strand or mapping errors corrected for SNPs are propagated to all genotype data sets which contain that SNP.

The HapMap, Perlegen, and PGA/UW groups each maintain interfaces for viewing their data available on their project websites. However, combining information from these three groups can be difficult because HapMap is using dbSNP build 122, Perlegen is using build 123, PGA/UW does not list dbSNP identifiers. The current dbSNP build is 124, and it maps all three groups on to the most recent genome build.

Within dbSNP, the complete set of genotypes mapped to the correct positions in the genome are available for download, and the haplotypes resulting from this study will be made available. The data is available in multiple formats including XML which allows the data in dbSNP to be easily integrated into other databases. Since multiple data sets can be mapped to the same locations, dbSNP provides a resource for comparing and combining genotype data between different studies with ease. As shown below, combining data from multiple sources and performing a joint analysis of the data can significantly alter the picture of a region. In addition, mapping data sets to the same location is useful for providing quality control.

6.2.2 Phasing the genotypes

We applied HAP to phase all of the genotypes in dbSNP. We phased each of the data sets separately. Where mother-father-child pedigree information is available in dbSNP, we used that information in our phasing. The haplotypes were partitioned into blocks of limited diversity so that 5 haplotypes covered at least 80% of the total number of haplotypes. A set of tag SNPs was chosen to minimize the number of SNPs needed to distinguish between the common haplotypes of each block [124]. The full phasing of dbSNP, partitioning all of the haplotypes in blocks of limited diversity, and determining a set of tag SNPs took under 24 hours. Table

Table 6.2: Summary of block partitions and tag SNPs for the largest 6 data sets in dbSNP.

Data set	Pop	Genotypes	SNPs	Inds	Blocks	Tag SNPs
HapMap	CEU	84,727,965	954,302	90	73,986	179,351
HapMap	CHB	18,443,054	411,568	45	41,381	94,583
HapMap	JPT	18,030,239	411,627	44	20,671	31,466
HapMap	YRI	38,661,518	431,505	90	67,111	157,287
PERLEGEN	Afr	35,568,060	1,569,392	23	235,139	569,182
PERLEGEN	Asi	37,417,872	1,572,384	24	86,636	211,972
PERLEGEN	Eur	37,399,120	1,570,560	24	109,212	274,153
Affymetrix	AfAm	885,135	125,776	20	24,526	40,050
Affymetrix	Cau	1,534,726	125,778	20	27,561	47,957
Affymetrix	Asian	884,091	125,772	20	20,671	31,466
Affymetrix	CEPH	50	30	3	18,453	26,018
Affymetrix	PD	2869641	125,776	24	35,048	67,154
Affymetrix	APE	15,823	9,027	2	6,253	6,262
TSC	ALL	4,932,382	19,048	1,963	31,886	46,789
EGP	ALL	3,184,170	37,737	90	3,847	6,643
PGA/UW	Afr	363,643	15,981	24	2,833	5,375
PGA/UW	Eur	209,551	9,525	23	1,086	2,378

6.2 summarizes the block partitions and the number of tag SNPs for each data set.

6.2.3 Haplotype Coverage

The combined set of haplotypes in dbSNP provide a significant amount of coverage of the genome. We measure coverage by two criteria: *density* and *depth*. Density defines the minimum gap between genotyped SNPs. The depth of a data set is defined as the number of individuals for whom haplotypes are available in the region. The coverage is the percentage of the genome covered by haplotypes with the minimum number of individuals and with a minimum gap between SNPs.

The coverage of the HapMap and Perlegen data as well as the combined two data sets is shown in Table 6.3. As can be seen from the table, the HapMap and Perlegen data sets provide excellent coverage for 10kb and more, but they give poor coverage for 1kb density, and for 5kb density they cover about 50% of the genome. On the other hand, when the two data sets are combined with the remaining data sets of dbSNP, the coverage significantly increases the coverage

Table 6.3: Coverage of Whole Genome Data Sets in dbSNP

Data Set	Density				
	1kb	5kb	10kb	20kb	50kb
HapMap	3.56%	54.50%	85.13%	89.52%	90.46%
PERLEGEN	10.79%	48.69%	63.06%	78.07%	88.24%
Combined	15.12%	72.70%	87.51%	90.02%	90.84%

Table 6.4: Coverage of Combined Data Sets in dbSNP

Depth	Density				
	1kb	5kb	10kb	20kb	50kb
1	15.62%	73.02%	87.60%	90.09%	90.89%
10	15.61%	73.01%	87.60%	90.08%	90.88%
50	15.48%	72.68%	87.22%	89.70%	90.48%
100	4.73%	28.84%	36.49%	37.49%	37.73%
200	3.23%	20.75%	26.67%	27.37%	27.51%
300	1.36%	8.51%	10.53%	10.72%	10.75%
350	0.62%	4.11%	5.14%	5.23%	5.26%

at 1kb density, and at 5kb. In addition, the remaining data in dbSNP provides higher coverage of the genome at higher depths since the Perlegen data set has 71 individuals and the HapMap data has 270 individuals. The coverage of the haplotypes in dbSNP is summarized in Table 6.4.

6.2.4 Haplotype Accuracy and Consistency Analysis

Since the haplotypes are obtained by statistical inference, a natural concern is that the results of analysis of this data may be biased due to errors in the inference. We benchmarked HAP over data collected in the HapMap project to obtain an estimate of the error rate for phasing unrelated individuals. The error rate for phasing related individuals has been shown to be very low in a recent benchmarking study performed by the HapMap analysis group[80]. We use mother, father, child pedigree information to measure the inference of haplotypes over the parents treating them as unrelated and then compare these predictions to what can be inferred from the pedigrees. The error rate of HAP for unrelated individuals is only 1.5% which is on the order of the amount of missing genotypes in the region. In addition, the haplotypes inferred from the whole genome variation data sets are consistent with the haplotypes inferred from the high density data sets obtained from resequencing studies. The accuracy and consistency of the haplotypes appear to minimize this concern.

HAP Error Estimation

In order to benchmark the accuracy of the predicted phase, we considered 5000 SNPs in chromosome 19 obtained from the HapMap CEPH data. The data set contains 30 mother, father, child trios from families in Utah with European ancestry. We used the trios to resolve haplotypes for heterozygous SNPs, whenever Mendelian genetics determines the phase. We then phased only the 60 parents, excluding the children from each of the trios, thus resulting in a set of 60 unrelated individuals. Among 300,000 genotypes in the parents, 73,333(24.4%) are heterozygous and 57,913(19.3%) can be resolved into haplotypes using the trio information. The predictions for the parents genotypes treating them as unrelated are then compared to the haplotypes resolved using trios.

We evaluated the benchmark on both HAP and the widely used phasing algorithm PHASE[106]. We also measured the discrepancies between the predictions of PHASE and HAP. Since the running time of PHASE increases rapidly as the number of SNPs to be phased increases, the 5000 SNPs were split into 50 regions of 100 SNPs each. We used PHASE 2.1.0 with its default option, and the default parameters of HAP.

Our results show that PHASE and HAP give identical results in 97.6% of the genotypes and 90.1% of heterozygous SNPs. We measured the accuracy of the results using the switch error rate. The switch error rate measures the proportion of heterozygous positions for which the phase is erroneously inferred relative to the previous heterozygous position. In terms of switch error rate, PHASE and HAP show 5.44% and 8.20% of switch error rates, respectively. When compared to the total number of genotypes, these switch errors occurs only 1.05% and 1.58% of genotypes respectively, and these are comparable to the rate of missing SNPs in these region, which is 1.17%.

As opposed to the accuracy of the phase prediction, the running time of HAP and PHASE, differs considerably. In Table 6.5 we provide the summary of the running times of HAP and PHASE on 10 different regions in chromosome 19 with different number of SNPs. As can be seen from the table, the running time of HAP is several order of magnitude faster than PHASE in most cases. Extrapolating from these results, by assuming that the PHASE algorithm is run with 100 SNPs sequentially on a single CPU, it would take PHASE at least 75,000 hours to phase the whole dbSNP database

In the benchmark performed by the HapMap analysis group, HAP was able to phase unrelated individuals over 1000 times faster than PHASE[80].

Table 6.5: Comparison of running time in seconds between HAP and PHASE. The running time is measured by running both methods from ten different positions in chromosome 19, with different length of genotypes. Intel Xeon 3.20GHz CPU is used in the measurement.

	Mean		Stdev.		Max.	
SNPs	Hap	Phase	Hap	Phase	Hap	Phase
10	0.06	19.12	0.03	8.88	0.10	37.74
20	0.56	109.71	0.30	68.78	1.08	237.78
30	1.10	327.55	0.55	257.59	2.24	887.82
40	1.53	833.99	0.61	831.93	2.58	2906.84
50	1.99	1643.49	0.75	1454.08	3.32	5013.80
60	2.45	3719.40	0.83	4352.68	3.74	14554.47
70	3.02	5931.03	0.91	5680.70	4.48	18593.30
80	3.43	8071.75	1.00	7495.58	5.12	26016.98
90	3.82	10585.10	1.10	9307.13	5.72	32363.89
100	4.42	13409.43	1.25	12113.96	6.53	40183.36
110	4.86	16082.93	1.21	12598.09	6.83	44603.33
120	5.25	20283.20	1.20	14935.60	7.14	54431.03
130	5.70	25249.62	1.35	18740.87	8.09	63775.21
140	6.16	30643.41	1.39	18292.52	8.53	69463.15
150	6.63	35768.83	1.46	20482.31	9.05	74459.95
160	7.05	42161.60	1.49	23714.27	9.73	91346.27
170	7.53	51597.25	1.59	30670.41	10.4	113281.51
180	8.09	63743.02	1.72	37621.29	11.08	138096.67

Haplotype Consistency Analysis

We performed a joint analysis of haplotypes from three data sets over the same genomic regions to measure the consistency of inferred haplotypes and block partitions. We considered regions where resequenced genes are available from the SeattleSNPs[27] and compared the haplotypes and blocks inferred from these data sets with the haplotypes and blocks inferred from the HapMap and Perlegen data. The European population was used for comparison because there is a corresponding population in each data set and there are overlapping individuals in the data sets.

We observe that the number of blocks and tag SNPs in the high density sequence data is much higher than in the corresponding HapMap or Perlegen data sets. This shows that there is a considerable amount of information loss when the data is sampled every 5kb such as in the HapMap data set. We examined 41 blocks in the Perlegen data set that overlapped with SNPs typed in the Seattle data set. There are 91 common haplotypes over the Seattle individuals on these SNPs. We then added in the additional Seattle SNPs typed on the blocks and reexamined the haplotypes for each individual. From the 91 original common haplotypes 369 haplotypes were found with 72 common ones. On average, 1.2 common haplotypes were created for every original common haplotype, and 30 of the original haplotypes were split into only rare haplotypes. One may hypothesize that this is due to the rare SNPs in the Seattle data. However, we performed the same analysis using only Seattle SNPs with a minor allele frequency of 10% or greater. The 91 original haplotypes were split into 330 haplotypes with 73 common ones. On average, each original common haplotype was split into 1.16 new common haplotypes, and 28 common haplotypes were split into only rare haplotypes when the Seattle SNPs were added. This demonstrates the utility of high density genotype data. The LD blocks and common haplotypes found by examining only the Perlegen data are significantly different than those found over the same individuals in the Seattle data. The common haplotypes are smaller than predicted by sparse maps, and suggests that more SNPs are necessary to type in whole genome association projects than suggested by the blocks found in the HapMap and Perlegen data sets.

We also measured the consistency of the haplotypes between data sets by comparing the phased data of individuals that exist in more than one study. The haplotypes inferred from the sequenced based genotypes of SeattleSNPs data sets were compared to the haplotypes from the HapMap and Perlegen studies. The HapMap and Perlegen studies contain genotypes for 1545 and 2426 common positions respectively. The predicted haplotypes over the HapMap data and the SeattleSNPs (PGA) data differ by 679 switches and the Perlegen data and PGA data differ by 11,071 switches. There differences correspond to switch differences of 0.4% and 2.4% respectively. The whole genome studies and the PGA data are not completely consistent in terms of genotypes. Between the HapMap and the PGA data, there are 17,424 genotype differences (3.5%) that occur in 602 different SNPs. Between the Perlegen and PGA data, there are 179,906 differences (3.8%) that occur in 6,758 SNPs. These switch distance between the inferred haplotypes are comparable to the amount of differences in the genotypes between the data sets.

Chimpanzee Haplotypes

We used the blastz alignments of the human and chimp genomes from the UCSC Genome Browser to determine the chimp haplotypes corresponding to the human ones. For each human haplotype block found, we examined the chimp allele corresponding to each SNP in the human haplotype block. These chimp haplotypes will also be publicly available for download at NCBI. These chimp haplotypes can serve as out groups or starting points in determining human haplotype phylogeny. They may also serve a purpose in comparative genomics methods when searching for functional haplotypes.

6.3 Discussion

Understanding the structure of common variation is an important step which will give insights into designing effective strategies for whole genome association analysis. Analysis of a single data set may bias any drawn inferences to properties of the data set. Performing joint analysis over multiple data sets may provide more robust analyses.

Our analyses show that the use of a combination of the various data sets of dbSNP increases the coverage of the genome considerably for high density markers. Furthermore, we show that when the density of the sampled SNPs increases, the block partition and the set of tag SNPs changes considerably. This can be interpreted as a quantitative measure for the amount of information provided by the data set. Therefore, when the sampled SNP density increases, the amount of information increases considerably. Furthermore, using the combination of all data sets in dbSNP, we increase the density of the sampled SNPs, and therefore increase the amount of information.

A challenge in analyzing multiple data sets is the time consuming pre processing that is required to map the data sets to the same build of the human genome. By using the haplotypes that we have submitted to dbSNP, researchers can more easily perform these joint analyses. We hope that the haplotypes, block partitions and tag SNPs will be useful for researchers in designing association studies.

6.4 Methods

6.4.1 HAP Phasing of Genome Wide Data

We used the HAP algorithm in order to phase the dbSNP data sets. HAP was run on a 30 CPU cluster consisting of 15 2GB RAN Nodes dual Intel Xeon 3.96 GHz processors.

The HAP algorithm assumes that the ancestral history of the haplotypes can be described by a perfect phylogeny tree. A perfect phylogeny tree is a genealogy tree with no recombinations, and no recurrent mutations. HAP considers all phases that result in a set of haplotypes that are almost consistent with a perfect phylogeny. HAP then efficiently enumerates over all such phases, and gives a score to each phase according to the likelihood of the solution under the assumption that the haplotypes were randomly picked from the population. HAP then chooses the phase with the highest score. In order to phase a long region, HAP applies the perfect phylogeny model in a sliding window to short overlapping regions. These overlapping predictions are then combined using a dynamic programming based tiling algorithm that chooses the optimal phase for the long region that is most consistent with the overlapping predictions of phase in the short regions. We considered all tiles of length 10 to 12 when constructing the haplotypes.

HAP is capable of phasing data sets up to 40,000 SNPs. The computational bottleneck is the size of the data structure necessary to perform the tiling. Since we only phased one chromosome at a time, the vast majority of the data in dbSNP was smaller than this limit. For some of the chromosomes in the HapMap and Perlegen data, we had to split the data set into 2-4 regions in order to perform phasing. Whenever the data sets were partitions, we picked a gap at least 50 kilobases between SNPs. Similarly, when computing block partitions, we only considered blocks which do not span a gap in SNPs greater than 50kb.

6.4.2 Partition into Blocks of Limited Diversity

We applied the dynamic programming based algorithm as described in Zhang et al.[124] to partition the inferred haplotypes into blocks of limited diversity. Their algorithm is based on the minimization of the number if tag SNPs so that the common haplotypes of each block could be distinguished by the tag SNPs. We consider regions where the common haplotype ($> 5\%$ frequency) account for more than 80% of the population a candidate block. We only consider SNPs with a minor allele frequency greater than 5%. We partitioned the haplotypes into candidate blocks where the partition minimizes the total number of SNPs that are necessary to distinguish between the common haplotypes in the blocks. HAP implements the Zhang et al. approach in a very efficient manner that can allow for partitioning of whole genome data sets. In order to compute the number of representative SNPs in a block, we apply a branch and bound algorithm which significantly reduces the computational time compared to the traditional exhaustive approach.

6.4.3 Extension of HAP to Trios

We extended the phasing algorithm HAP [54] in order to allow it to cope with genotypes typed from mother, father and child trios. Within a short region, the extension of HAP to trios must take into account the fact that the haplotypes of the children are copies of the haplotypes of the parents. We assume there are no recombinations or mutations between the parents and the children in the trios. This allows us to first unambiguously resolve the phase of the trios in many of the positions. For the remaining positions, we use HAP in order to enumerate over all possible phases. This results in a set of haplotypes that are almost consistent with a perfect phylogeny. In that enumeration, we exclude the solutions which contradict Mendelian heredity within a trio. For each such solution we give the likelihood score, which is the probability to observe the parents' haplotypes in our sample. We pick the solution with maximum likelihood as a candidate solution. In order to further improve the solution, we use a local search algorithm. The local search algorithm starts from the solution given by HAP, and it repeatedly changes the phase of one of the trios to a different possible phase, and checks whether the likelihood function has increased. If it has increased, we use the new solution as the candidate solution and repeat this procedure. If no local change can be applied in order to increase the likelihood, we stop and use the solution as a putative solution for this region. The resulting algorithm is very efficient and running times are comparable to the running time of HAP over unrelated individuals[80].

Chapter 6, was published in Genome Research, Vol 15, pp 1594-600, 2005. Noah A. Zaitlen, Hyun Min Kang, Michael L. Feolo, Stephen T. Sherry, Eran Halperin, and Eleazar Eskin, “Inference and analysis of haplotypes from combined genotyping studies deposited in dbSNP”. The dissertation author was the primary investigator and author of this paper.

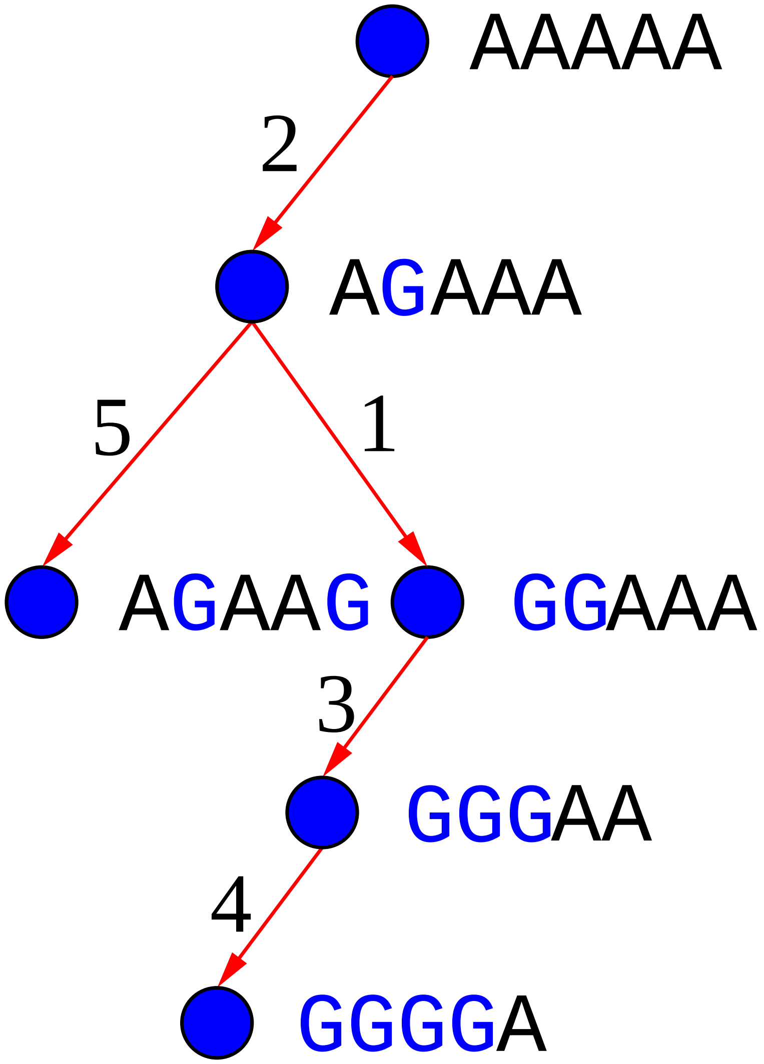

Figure 6.2: A perfect phylogeny model consists of a tree where each vertex corresponds to a haplotype, and each edge corresponds to a mutation in one of the positions of the haplotype. An edge is labeled with the position of the mutation. The tree fits the perfect phylogeny model if there are no recurrent mutations and no obligate recombination events. A set of haplotypes fits the perfect phylogeny model if it satisfies the four gamete test, that is, at most three allele combinations are observed for any pair of marker positions.

Chapter 7

MHC Class II Epitope Binding Prediction

7.1 Introduction

The open binding pocket of the MHC class II molecules allow for a much greater variation in peptide length relative to the closed pocket of the MHC class I molecules. This difference combined with the relative lack of sequence similarity across binding peptides makes MHC binding prediction significantly more challenging for the class II molecules. Two widely held beliefs about the physics of class II binding allow for some simplifying assumptions that have been used to make the problem more tractable. The first is that a majority of the binding is due to a consecutive group of nine amino acids along the peptide. The second is that there is overlap between the peptides that bind to different alleles. Working under the first or both of these assumptions, recent efforts for MHC class II binding have been focused on methods to identify the nine amino acid binding core of the peptide. This is then combined with one of the various methods for predicting MHC class I binding over the derived nonamers.

Several approaches to binding core identification have been explored. Many of these search for an optimal alignment of nonamers across the binding peptides. [97] and [58] use MEME [5] to identify and align the over represented nonamers.

Gibbs sampling is used by [86], evolutionary algorithms by [14], and an ant colony search strategy by [68]. The Linear Programming method of [84] effectively produces an alignment or choice of nonamers during training.

The alignment can be a pre-processing step as in [97] and [58] who use the set of nonamers in the alignment as direct input into their MHC class I predictors. The Gibbs sampling method of [86] uses the PSSM of the alignment as input into the binding prediction method. The closest method to our work in the Linear Programming model proposed by [84]. They use a sliding window over each peptide and a set of LP constraint which attempts to identify the window for each peptide which will maximize their ability to separate binders from nonbinders with a PSSM model.

The set of MHC class I tools that have been applied to the nonamer alignments are motif based, machine learning based, and structure based. Motif based methods such as RANKPEP [97], attempt to identify amino acids at particular positions that are characteristic of binding for a given allele. A variety of machine learning base methods such as neural networks, support vector machines, and hidden Markov Models have been applied. Structure based methods such as ours and [35] attempt to model the physics of MHC binding using the growing number of MHC class I and II molecules that have been solved by X-ray crystallography.

In this work, we develop a new method for predicting binding to arbitrary MHC class II alleles. Our model is based on the structure of the class I and class II molecules and treats the possible peptide alignments as an ensemble of possible configurations. Rather than assuming simply that any peptide alignment is equally possible, or turning to separate methodology to provide best alignment, we infer the distribution over possible states for each peptide-MHC combination based on the predicted state energy. This distribution is not treated as a distribution over a variable with mutually exclusive and exhaustive states, but rather as population frequencies in the thermodynamics sense, and the equivalent total binding energy is estimated accordingly. This is the key difference between our approach and previous approaches to MHC class II binding prediction, which enabled us to outperform, to the best of our knowledge, all previously published techniques.

7.2 MHC II - peptide binding model, inference and learning

In this section, we present a physics-based model of MHC class II molecules bound to peptides of variable length. The model uses the binding groove area of available crystal structures of MHC II molecules as exemplars, and treats the peptide alignment with the groove as a hidden variable. Using an energy model, we can perform inference of both the optimal structure and the optimal peptide-groove alignment. We also derive a learning algorithm that estimates the parameters of the energy model to fit the available binding energy data.

7.2.1 Modeling variable peptide position in the MHC II groove

As discussed in the introduction, the main source of prediction errors for MHC II binding is the fact that, as opposed to MHC I molecules that tend to have a fixed binding configuration, largely robust to changes in the peptide's amino acids, the MHC II molecules may exhibit a variety of binding configurations. Thus it is important to model an ensemble of configurations with hidden variables describing them, where model for each configuration (given the hidden variables), is simpler to model directly (e.g., using the model in [66])

Since a longer peptide (15-30 amino acids, for example) has only a part of it in the groove of the MHC class II molecule, we introduce a hidden random integer variable ℓ that represent the unknown alignment of the peptide with the groove. We use only a single "shift" variable to represent this main component of variation in binding configuration because the peptides are unfolded and the parts that interact with the groove have largely similar configurations. The largest difference in the binding of different peptides to the same MHC II allele is in where the bound part of the peptides start. We represent the starting index of this segment with the variable $\ell \in [1, N-8]$, where N is the length of the peptide, and we assume the segment that is inside the groove is at least 9 amino acids long (Fig. 7.1). There are, of course, other hidden variables that describe the binding configuration, such as a selection of particular geometric configuration m of the amino acids in the groove of the MHC molecule from the available crystal structures. In this section, we will denote all such hidden variables with $\mathbf{h}$, and in the next section, we will define $\mathbf{h} = (\ell, m)$ as the hidden variables in our shifted adaptive double threading model.

For now, we simply assume the existence of a model $E(\mathbf{s}, \mathbf{e}, \mathbf{h})$, where $\mathbf{s}$ denotes a particular MHC allele, and $\mathbf{e}$ denotes a particular peptide (potential epitope), such that if the setting for the hidden variables $\mathbf{h}$ are provided, the model can produce a good estimate of the binding energy for the pair $\mathbf{s}, \mathbf{e}$. In many past approaches to MHC binding, the settings of $\mathbf{h}$ (in particular the alignment analogous to our variable ℓ), where provided by a separate routine, unrelated to the energy model E.

We treat states indexed by $\mathbf{h}$ as different energy states (with energies $E(\mathbf{s}, \mathbf{e}, \mathbf{h})$) that an MHC-peptide complex can assume, with the partition function

$$Z = \sum_{h} e^{-E(\mathbf{s},\mathbf{e},\mathbf{h})}, \tag{7.1}$$

and the free energy per particle of the system of such particles is $F = -\log Z$, where the kT factors are omitted, as the reported measured binding energies are dimensionless $\log IC50$ values. Thus, we can model the measured binding energy $\log IC50$ as

$$E(\mathbf{s}, \mathbf{e}) = -\log \sum_{\mathbf{h}} e^{-E(\mathbf{s},\mathbf{e},\mathbf{h})}. \tag{7.2}$$

In particular, for the case of a shift as the hidden variable $\mathbf{h} = \ell$, the energy of a particular configuration $E(\mathbf{s}, \mathbf{e}, \ell)$, can be derived from a model $E_{mod}(\mathbf{s}, \mathbf{e})$ that does not deal with peptide shifts, and requires $\mathbf{e}$ to be a known k-mer sitting in the MHC groove, if the assumption is that k amino acids are in the pocket. In this case, $E(\mathbf{s}, \mathbf{e}, \ell) = E_{mod}(\mathbf{s}, \mathbf{e}_{\ell:\ell+k-1})$. Choices for models of binding given a known alignment include most previous MHC I and MHC II binding models (e.g. pssm [97], logistic regression,support vector machine[40], motif search), although they may have to be retrained in this new context. In what follows, we describe how this retraining may be done, on the example of the adaptive double -threading model [66], into which we add hidden variables according to the above recipe, and then derive an EM-like learning algorithm that can fit the parameters of the model in the presence of multiple possible conformations.

An important feature of our treatment of variable peptide alignment with the groove is that the distribution over possible alignments is effectively determined by the model's energy predictions alone, rather than by the fit of these predictions to the energy data. This means that the proper alignment can be inferred not only in training, but also in testing on new peptides for which the true (measured) binding energy is not provided to the predictor. Since the energy in (7.7) is dominated by the minimum energy state $E(\mathbf{s}, \mathbf{e}, \mathbf{h})$, the preferred alignments will have lower energy, rather than better fit to the data.

Next, we review the adaptive double threading model [66], and then introduce hidden variables into this model according to the above recipe.

7.2.2 Shift-invariant double threading

Our basic binding energy model is based on the geometry of MHC-peptide complexes, and is motivated by the *threading* approach [67]. As in [66], its implementation in [102] is here augmented by including learnable parameters. The parameters are estimated from the experimental data.

In general, threading aims at evaluating the compatibility of a certain protein sequence with a certain protein structure: The sequence is threaded onto the structure, and a list of contacting amino acid pairs is extracted, based on contacting residue positions (defined as residues in close proximity, e.g. that have at least one pair of atoms less than 4.5A apart). In order to allow estimation of the binding energy of any peptide with an MHC molecule whose structure in complex with some other peptide is known, we assume that the proximity pattern to the peptide in the groove does not change dramatically with the peptide's sequence.

Assuming that energy is additive, and that the pairwise potentials depend only on the amino acids themselves - and not on their context in the molecule - the energy becomes a sum of pairwise potentials taken from a symmetric 20×20 matrix of pairwise potentials between amino acids. These parameters are computed based on the amino acid binding physics, or from statistical analyses of amino acid pair contact preferences in large sets of available protein structures. Several sets of pairwise potentials have been described in the literature, each derived in a different way (for review see [82]. Obviously, the choice of pairwise potential matrix can dramatically alter performance of the energy predictor [102]. Previously, we have shown that estimating these parameters from training data leads to the better performance on the test set, than using a previously published, rationally derived potential matrix. Part of the reason for this is the possible specialization to the class of molecules under consideration, but the model still preserves its physics basis, and the learned weights tend to reveal contact aminoacids [66].

In the adaptive double threading model of MHC I - peptide binding, the binding energy is estimated as

$$E(m, \mathbf{s}, \mathbf{e}) \approx \sum_i \sum_j w_{i,j}^m \boldsymbol{\phi}_{\mathbf{s}_i, \mathbf{e}_j} h(d_{i,j}^m), \tag{7.3}$$

where MHC-specific weights $w_{i,j}^m$ and a trainable soft threshold function h provide added parameters whose role is to correct for the drastic approximations in the original threading approach. (The predictions of the original threading approach correspond to using the above equation with all weights w set to one, and the threshold function h set to a hard step with a threshold decided upon in advance.) In the above equation, m denotes a particular binding configuration (inferred through crystallography), $\mathbf{s}$ denotes the molecule's amino acid sequence, indexed by i, and $\mathbf{e}$ denotes the peptide (epitope), whose aminoacids are indexed by j. The distances $d_{i,j}^m$ are computed in the crystal structure. Only a small fraction of indices i correspond to molecules amino acids in contact with the peptide. [102] used $d_{i,j}^m < 4.5A$ to determine such aminoacids, which form a binding groove of the molecule. In the above equation, this would correspond to setting

$$h(d) = \begin{cases} 1, & d \leq d_{thr} \\ 0, & d > d_{thr} \end{cases}. \tag{7.4}$$

In our MHC I predictors [66], we used a soft h function [66],

$$h(d) = \frac{1}{1 + e^{-a(d - d_{thr})}} \tag{7.5}$$

whose threshold parameter (d_{thr}) and the step softness a are estimated together with the contact potentials and weights. (Still, only a small number of aminoacids in the molecule are close enough to the peptide to lead to nonzero values of h, and our “soft” groove is also small.) The adaptive soft step function and the addition of the weights w are meant to absorb the errors of the model assumptions [66].

The basic idea behind threading approaches is that, even though the structure information d is inferred from a known binding configuration of a *particular* peptide-MHC I combination, substituting *a different* peptide of the same length (or even another MHC molecule, as in our previous work) in the above equations would still lead to a reasonable estimate of the binding energy for the new MHC-peptide combination. This is due to the fact that relative positions and the basic chemistry of the amino acid-amino acid interactions are fixed. Even the light changes over different geometries of peptide-groove configurations (indexed by m have a small (though measurable) effect on the accuracy of the model. The success of the previous work on MHC I binding energy prediction attests that this main assumption holds well for MHC I molecule.

In [66], we have shown that the parameters of the above model can be estimated so that the error of approximation is minimized on the training set, and then the model's predictive power can be tested on a separate test set. When the training data sets are too small, sparsity priors on w and cross-validation are used to avoid overtraining. However, we also showed that multiple different MHC types can be trained together as they can share some or all parameters. Parameter sharing leads to a negligible drop in performance, while the main benefit is not merely avoidance of overtraining, but the ability to generalize the predictions to *new* MHC alleles, for which little or no binding or epitope data is available.

Essentially the same basic modeling strategy can be used for modeling MHC class II except for one very important difference. While the fixed chemistry of the amino acid interactions and the fixed overall geometry of the MHC molecule are still relatively mild assumptions, the fixed relative position of the peptide is a gross over-approximation. The binders to MHC class II molecules are often much longer than 8-11 aminoacids, while the binding groove is roughly the same size.

This means that only a part of the peptide is sitting snugly in the groove, while the tails on either side have much smaller influence on the binding affinity (Fig. 7.1). Thus, the model needs to be extended to account for *variable position* of the peptide, as discussed above.

To estimate the energy of the binding configuration for a particular shift ℓ, we update our model in the following way:

$$E(m, \mathbf{s}, \mathbf{e}, \ell) \approx \sum_{i} \sum_{j=1+\ell}^{N+\ell} w^{m}_{i,j-\ell} \boldsymbol{\phi}_{\mathbf{s}_i, \mathbf{e}_{j-\ell}} h(d^{m}_{i,j-\ell}), \tag{7.6}$$

In practice, the proper shift is not known, unless the structure has been solved, but the shifts that yield lower energy values should be considered more likely. In order to fit the model to the binding essays, we need to express the total affinity of the peptide by summing over all the binding configurations. The experimentally measured binding energy is usually reported in terms of an IC50 value, which approximates the dissociation constant. The energy is assumed to be proportional to the negative log of this value, and so energy estimators are typically trained on the $E = -\log n_{IC50}$ values. When many copies of the same longer peptide are mixed with many copies of the same MHC class II molecule, binding configurations with all different shifts ℓ may form. Therefore, according to 7.2, we sum over the two unknown variables that affect meaningfully the binding energy used in (7.6):

$$E(\mathbf{s}, \mathbf{e}) = -\log \sum_{m,\ell} e^{-E(m, \mathbf{s}, \mathbf{e}, \ell)}. \tag{7.7}$$

Variable m, as in the case of the MHC class I molecule (7.3), represents the geometry of the configuration of the MHC molecule and the peptide's segment that is in the groove. In case of MHC class I molecules, this is all the geometry variability we need to consider. The variable m influences the energy estimate through the distance matrix $d^{m}_{i,j}$. As the variability in the binding configurations of the groove is low, the influence of variable m is existent, but mild. In case of MHC class II molecule, this variability has a much smaller effect on the energy estimate than the shift variable ℓ – upon 3D alignment of different MHC structures, the relative positions of molecules close to the binding grooves change very little. While the slight geometry changes in the groove have an effect on the prediction, the shift variable ℓ influences the prediction much more dramatically as it alters the predicted amino acid composition of the peptide's segment sitting in the groove.

Short inspection (or simulation) of (7.7) reveals that the energy estimate is indeed dominated by the state (m, ℓ) with the smallest energy. However, as we will discuss later, it is typically dangerous to assume that the observed energies are equal to the minimum among the estimated energies for different states (m, ℓ). The reason for this is that the predictors are inherently noisy, and the more states we consider, and the more predicted variability across the states we find, the more likely it becomes that the wrong minimum energy state will be picked with a dramatically wrong predicted energy value. Taking more states into account in the estimate, on the other hand will lead to more robust estimates.

7.2.3 Parameter estimation and binding configuration inference

In our training and testing procedures, we assume that the data is given in a form of a list of triples, each consisting of an MHC class II sequence $\mathbf{s}$, a peptide $\mathbf{s}$ and the measured binding energy $E(\mathbf{s}, \mathbf{e})$. During training, we wish to determine the model parameters w, phi, d_{thr}, a which minimize the error of approximation in (7.7). Any number of optimization or search algorithms can be used for this. Since the error of approximation in (7.7) depends on the parameters in a highly nonlinear way, in our implementation, we introduce new auxiliary variables for each training case, in order to simplify the optimization criterion into a simple quadratic form. The price to pay is the EM-style iteration the parameter optimizations step with re-estimation of the case-specific auxiliary variables. To derive the algorithm, we first introduce an auxiliary probability distribution over states $q(m, \ell)$, so that, of course, $0 \leq q(m, \ell) \leq 1$, for all states, and $\sum_{m,\ell} q(m, \ell) = 1$. Next, we observe that, as log is a concave function,

$$
\begin{aligned}
E(\mathbf{s},\mathbf{e}) &= -\log \sum_m \sum_{\ell=1}^{N-8} e^{-E(m,\mathbf{s},\mathbf{e},\ell)} \\
E(\mathbf{s},\mathbf{e}) &= -\log \sum_m \sum_{\ell=1}^{N-8} q(m,\ell) \frac{e^{-E(m,\mathbf{s},\mathbf{e},\ell)}}{q(m,\ell)} \\
&\geq -\sum_m \sum_\ell q(m,\ell) \log \frac{e^{-E(m,\mathbf{s},\mathbf{e},\ell)}}{q(m,\ell)} \\
&= \sum_m \sum_\ell q(m,\ell) E(m,\mathbf{s},\mathbf{e},\ell) + \sum_m \sum_\ell q(m,\ell) \log q(m,\ell).
\end{aligned}
$$

Since for a given state m, ℓ, the energy depends on each subset of model parameters w and ϕ linearly, this bound on the energy is also bilinear in model parameters, and the same iterative linear regression reported in our previous work can be used to minimize the approximation error. The above bound is true for any auxiliary probability distribution q, but it becomes tight (exact equality is accomplished) when

$$
q(m,\ell) = \frac{e^{-E(m,\mathbf{s},\mathbf{e},\ell)}}{\sum_{m,\ell} e^{-E(m,\mathbf{s},\mathbf{e},\ell)}}, \tag{7.8}
$$

i.e., the distribution q is the exact distribution over states according to the energy model. This distribution depends on the sequence content of both the MHC molecule **s** and the peptide **e**, and so it has to be recomputed for each training or test case. It is important to note that this distribution is not treated as a distribution over a variable with mutually exclusive and exhaustive states, but rather as population frequencies in the thermodynamics sense. In the former case, the hidden shift variable could only be inferred from a given binding energy, and in prediction, energies of different possible shifts would have to be averaged. In the latter case, the distribution over shifts depends on the predicted energies for individual shifts, and not on the observed energies, and so it can be equally used in training and testing. In Section 7.3.1 we experimentally test the accuracy of the inference of the binding configuration using this approach. To learn the model parameters, the configuration inference step has to be iterated with re-estimation of model parameters. Such an iterative learning algorithm consists of the following steps:

- Initialize model parameters (e.g., setting all weight w to one, d_thr and a so that the step function h is smooth and has a larger threshold, e.g. 6 or 7, and the ϕ matrix to either uniform or the one previously estimated for MHC I in our previous work or to amino acid contact potential published by others Some care has to be taken regarding normalizing the parameters. If the potentials are initialized to be too large, for example, than weights w may absorb the problem, but some other order of updates may lead to local minima.

- Initialize $q^t(m,\ell)$ to uniform for each training sample $(\mathbf{e}^t, \mathbf{s}^t, E^t)$.

- Re-estimate the model parameters w, ϕ, d_{thr}, a so that $\sum_t (E(\mathbf{e}^t, \mathbf{s}^t) - E^t)^2$ is minimized, where

$$E(\mathbf{e}^t, \mathbf{s}^t) = \sum_{m,\ell} q^t(m,\ell) E(m, \mathbf{s}^t, \mathbf{e}^t, \ell) + \sum_{m,\ell} q^t(m,\ell) \log q^t(m,\ell). \qquad (7.9)$$

  Since the model is linear in w and linear in ϕ, iterative linear regression to solve for one set of parameters at a time is efficient. Step function parameters d_{thr}, a are updated every few steps by gradient descent.

- Using the new parameters, re-estimate the distribution

$$q^t(m,\ell) = \frac{e^{-E(m,\mathbf{s},\mathbf{e},\ell)}}{\sum_{m,\ell} e^{-E(m,\mathbf{s},\mathbf{e},\ell)}}. \qquad (7.10)$$

- Iterate the last two steps until convergence.

This procedure has some similarity with transformation-invariant generative models developed primarily for vision applications in [65]. However, the important difference is that the possible shifts are not considered as equally likely a priori. In fact, they depend on the peptide and MHC sequences. Consequently, the distribution over states m, ℓ can be determined both for training and testing peptides, and in prediction, the state energies are not averaged. Rather, the possible binding configurations are considered as an ensemble with population frequencies defined by q. It is also different form the LP approach discussed in the introduction, which tries to infer a single best alignment for each peptide in training.

7.2.4 Using temperature to account for modeling errors during learning

The update of the position distribution in (7.10) and the estimate of the energy in (7.9) are highly sensitive to the errors in prediction due to the nonlinearity of estimating the equivalent energy by summing over all configurations (7.7). This can cause local minima problems for the EM-like procedure described in the previous section, as the parameters, and therefore the predictions, are less reliable in the early iterations of learning.

To illustrate how the prediction errors may be propagated through (7.7), we present the following simple experiment. Assuming the total number of different shifts ℓ is 10, and that the true binding energy for fake MHC-peptide configurations $E_\ell$are drawn randomly form a uniform distribution on the interval $[0, 10]$, we computed total binding energies according to $E_{true} = -\log \sum_\ell e^{-E_\ell}$ for 100 such configurations.In this synthetic experiment, we ignore variability in configurations m, as they have a smaller effect on the variability of the energy prediction. Then, we computed

$$E_{estimate} = -T \log \sum_\ell e^{-\frac{\tilde{E}_\ell}{T}}, \tag{7.11}$$

where $\tilde{E}_\ell = E_\ell + v_\ell$, and v_ℓ, a random variable drawn from a zero mean Gaussian distribution with some variance σ^2, simulates a modeling error. A choice of the auxiliary temperature parameter $T > 1$ leads to smoothing of the energy estimate in the following sense: By reducing the differences between the energies of different states, it becomes possible for more states to significantly influence the estimate. This is potentially useful as the wrong state may have the lowest energy due to the prediction errors, and the state with the lowest energy dominates the estimate at $T = 1$. For larger parameter T, on the other hand, the lowest energy state would contribute more to the estimate of the energy, but the other states would contribute, as well.

We assume for the moment that the measurement procedure which would in practice provide a direct measurement of E_{true} is perfect, and that a potential inability of a predictor to match it is only due to predictor's errors in predicting the binding energy of the groove-peptide segment configurations for different shifts. In Fig. 2 we show how well the tempered prediction $E_{estimate}$ using the noisy predictions E_ℓ correlate with the true energies E_{true}. In particular, for different levels of error variance σ^2, we show how the Spearman correlation factor between E_{true} and $E_{estimate}$ varies with the temperature T. The graph shows that a rise in modeling error σ^2 can, to some extent, be absorbed by raising temperature factor T.

Adding the temperature factor into (7.7) leads to the following change in (7.9) and (7.10) in the algorithm of the previous section:

$$E(\mathbf{e}^t, \mathbf{s}^t) = \sum_{m,\ell} q^t(m,\ell) E(m, \mathbf{s}^t, \mathbf{e}^t, \ell) + T \sum_{m,\ell} q^t(m,\ell) \log q^t(m,\ell). \tag{7.12}$$

$$q^t(m,\ell) = \frac{e^{-\frac{E(m,\mathbf{s},\mathbf{e},\ell)}{T}}}{\sum_{m,\ell} e^{-\frac{E(m,\mathbf{s},\mathbf{e},\ell)}{T}}}. \tag{7.13}$$

In training, rather than annealing the temperature according to some fixed training schedule, we *search* for the optimal temperature parameter after every few updates of the model parameters. Upon convergence of all model and auxiliary parameters, the temperature typically settles to a value close to 1, which might indicate that the physical measurement errors are higher than the modeling errors.

7.3 Experiments

We downloaded the complete set of MHC class II structures that contain an epitope of at least seven amino acids from the PDB [9]. The resulting set consisted of 12 HLA-DR, 3 HLA-DQ, 3 H2-K, and 18 H2-D alleles. Although the MHC class II allele HLA-DP are missing from this set, they share relatively high sequence similarity with HLA-DR alleles. As discussed above, precise structure is less important than sequence and shift variability, suggesting that it is possible to predict for these alleles using the structures of their closest matching alleles in the PDB. These structures are used as exemplars m of the groove structures in the experiments. To evaluate the prediction accuracy, we used our method both as an epitope predictor and a binding energy predictor and tested it on the available epitope and energy data. In addition to comparisons with existing techniques for epitope prediction, we analyze the ability of our model to predict the binding configuration (in terms of the simplified state m, ℓ), predict for new alleles for which training data is not given, and assist in association studies in immunology.

MHCPEP Data set

The MHCPEP data set has recently been used to evaluate the performance of the MHC class II binding predictors *DistBoost* and RANKPEP. Following the procedure of [58] and [97], we downloaded the contents of the MHCPEP database [13] in order to compare the relative performance of our method. The data are peptide sequences paired with MHC alleles and binding affinities. As in [58] and [97], we removed all peptides classified as low binders or with unknown residues at some position. We removed peptides from all non human MHC alleles (although our method can be applied to these as well), leaving 1265 peptides from 17 MHC class II alleles. We verified via email correspondence that our data set matched the corresponding subset of [58]. Unlike [58] and [97] our method does not require an alignment step and was therefore omitted.

We compared our method to *DistBoost* and RANKPEP [97] by replicating the exact same experimental setup. The MHCPEP data set described above was used as the set of positive binders. Non-binders were taken from random protein sequence from the SwissProt database [6], so that there were twice as many non-binders as binders per allele. Training was performed using half of the binders for each allele with twice as many non-binders. Testing was performed on the remaining set. We used 5-fold cross validation over the training set to find an optimal set of parameters, and then evaluated the method on the test set. This setup was repeated 10 times to measure average performance and standard deviation.

We plotted ROC curves for our model and compared the AUC of our method with the published results of RANKPEP and *DistBoost*. Our method outperformed both *DistBoost* and RANKPEP on 15 out of the 17 data sets (p-value $<$ .00014 binomial) see Table 1. The average AUC for our method was .87 compared to .78 for *DistBoost* and .71 for RANKPEP. In addition, our average standard deviation was lower than either method, 0.04 compared to 0.044 and 0.05, showing our method is as robust or better. Like *DistBoost* our method is able to take advantage of peptides of other alleles when training for a particular allele, giving improvement in alleles with a small amount of training data. [58] also compared *DistBoost* to the SVMHC web server [40] and the NetMHC web server [15], and outperformed them on an MHCBN [11] data set.

MHCBench Data set

The MHCBench data set was constructed for the purpose of evaluating MHC class II binding predictors. Recently, [84] and [86] have evaluated their methods over this data set after training on similar training data. In order to evaluate the relative performance of our method, we followed their training and testing procedures. We downloaded the set of HLA-DRB1*0401 binding peptides from the SYFPEITHI [96] database that were added before 1999. [86] does not require negative training examples for his method, so we followed the example of [84] and added the HLA-DRB1*0401 non-binders from the MHCBN database [11]. Although we do not align our peptides, and therefore do not have an initial putative position of the peptide in the MHC molecule, we followed their example and removed peptides that have a hydrophobic residue in the first position according to their model. Peptides that were more than 75% alanine were also removed. This left a data set of 462 binding and 177 non-binding peptides and is the training data set. Our method also has the capability to incorporate information from other alleles in training. We therefore created another training data set which consists of that described above in addition to the set of non HLA-DRB1*0401 peptides contained at MHCBN. All peptides overlapping the test data (see below) with alignment over 90% were removed, leaving a set of 2997 peptides.

The test data sets used by [84] and [86] consist of the 8 data sets described in [95], the data set from [105], and the data set from [47]. In the [95] data set, any peptide with a non-zero value is considered a binder and is a non-binder otherwise. For the other data sets, any peptide with affinity of less than 1000nM was considered a binder, and a non-binder otherwise. Since there is a significant overlap between the peptides in the training and test data sets, we removed any peptide with $> 90\%$ sequence identity to a peptide in the training set. We verified via email correspondence that our training and test data sets matched those of [84] and [86].

We used 5 fold cross validation over the training set to estimate the optimal set of parameters for our model. ROC curves were generated for each test set and the AUC was computed for comparison with the published results of LP, Gibbs, and Tepitope. In addition, we trained on another training data set which contained peptides from other alleles to show how our method can incorporate other data to improve performance. The results are shown in Table 2. Our method has a higher average ROC than any other method, and it is further improved by adding non DRB1*0401 alleles to the training set. We beat the other methods on 8 out of 10 data sets (p-value < 0.017 binomial). In training our model we assume a different cutoff for good versus bad binders than the 1000 nM cutoff used for the Southwood and Geluk data sets in the test data. Using our cutoff of $e^{6.2}$ improves our performance on these data sets, but can not be compared with the above methods since the training set would be different.

IEDB Data set

In order to allow others to easily compare their methods against ours, we created a new training and test set from the IEDB data set (described below). We selected 1175 peptides from 13 HLA-DRB alleles for the training set that each have at least 100 training examples. Thus, no transfer to unlearned alleles is required to compare with our method. The full makeup of the training and test set are described at We used 5-fold cross validation over the training set to learn the parameters for our model. Table 3 shows our performance over the test data set.

The size of all data sets used in this study are shown in Table 4.

7.3.1 Binding configuration inference

Our results compare favorably with previously published approaches, and we have argued above that the novelty of our approach is in proper inference of the binding configuration, consisting of the groove geometry m and the peptide shift ℓ. In this section we illustrate that a trained model indeed predicts well the binding configuration. We downloaded the set of 12 protein structures for human MHC class II allele from the PDB [9] with a bound peptide of length 9 or greater. For each of these peptides we compared the groove structure choice and the shift choice of our method with the ground truth. For each available structure, we threaded the MHC allele corresponding the peptide onto all available structures and estimated the binding of the peptide to each of the structures under our model. We then ranked the energies of each of the structures from lowest (strongest binder) to highest (worst binder) and found the rank of the 'true' structure peptide pair in the list. We took the average of this value over all twelve peptides. In order to estimate the significance of this result, we randomly generated 10000 lists of ranks from 1 to twelve, computed their average, and counted the number of times the average beat the average rank of our experimental results. This gave us a p-value of 0.021 and shows that the correct structure has relatively lower energy.

To verify that our technique chooses the correct shifts we measured the binding energy of each nonamer of the peptide to its corresponding MHC allele and compared the shift of minimum energy to the true nonamer in the binding pocket of the structure (minimum energy state has the highest probability in (7.8). Out of the 12 structures, we predicted the nonamer in the binding groove exactly in 8 structures, while the rest of the predictions were off by a single amino acid ($p - value < 0.0001$). In all cases, the chosen shift resulted in the energy estimate above the cutoff threshold for a good binder. These experiments suggest that correct identification of shift outweighs the importance of the slight variations in structure of the various MHC alleles. This in turn lends support to the idea of our double threading approach. Threaded alleles will have slight inaccuracies in structural position, but the correct shift can still be recovered.

7.3.2 Generalizing to new alleles

One of the important features of our approach as opposed to most others is that after training, any MHC sequence may be threaded onto a structure

Table 7.1: Comparison of RANKPEP,*DistBoost*, and our Shift Invariant Double Threading (SIDT) method over the MHCPEP data set. Best values shown in bold font. Columns A and B for *DistBoost* refer to training without and with negative constraints. Column RP B for RANKPEP refers to PSSMs constructed using BLK2PSSM. Those using PROFILEWEIGHT performed worse on average and were never the top across all studies.

Allele RP B	std	*DB* A	std	*DB* B	std	SIDT	std
QA 0501 0.88	0.06	**0.93**	0.03	**0.93**	0.04	0.87	0.08
QA 0301 0.7	0.06	0.75	0.04	0.77	0.05	**0.87**	0.02
PA 0201 **0.88**	0.1	0.75	0.12	0.74	0.09	**0.88**	0.03
RB 0101 0.75	0.04	0.81	0.02	0.8	0.02	0.**87**	0.02
RB 0102 0.72	0.04	0.9	0.07	0.83	0.05	**0.91**	0.06
RB 0401 0.6	0.04	0.71	0.01	0.73	0.02	**0.87**	0.01
RB 0402 0.72	0.04	0.74	0.06	0.69	0.04	**0.88**	0.01
RB 0405 0.82	0.04	0.86	0.03	0.86	0.04	**0.89**	0.06
RB 0404 0.61	0.05	0.74	0.05	0.7	0.05	**0.84**	0.04
RB 0701 0.72	0.04	0.79	0.05	0.76	0.04	**0.89**	0.04
RB 0901 0.78	0.06	0.89	0.03	0.91	0.04	**0.97**	0.07
RB 1101 0.54	0.04	0.76	0.03	0.73	0.02	**0.85**	0.02
RB 1501 0.6	0.07	0.73	0.07	0.75	0.06	**0.87**	0.05
RB 0101 0.81	0.03	0.83	0.07	0.8	0.05	**0.87**	0.04
RB 0801 0.52	0.06	0.67	0.09	0.65	0.05	**0.84**	0.06
RB 1104 **0.92**	0.02	0.87	0.04	0.88	0.03	0.83	0.04
RB 0301 0.52	0.09	0.54	0.44	0.62	0.06	**0.83**	0.03
AVG 0.71	0.05	0.78	0.07	0.77	0.04	**0.87**	0.04

Table 7.2: Performance of our shift invariant double threading method (SIDT), the Gibbs sampler, TEPTITOPE, and the Linear Programming method over 10 homology reduced data sets. *This is our method trained with additional data for different alleles. It demonstrates the ability of our method take advantage of information across alleles.

Method	Set1	Set2	Set3a	Set3b	Set4a	Set4b	Set5a	Set5b	Geluk
SIDT	**0.76**	0.71	**0.73**	**0.79**	**0.77**	0.72	0.71	0.79	**0.78**
SIDT*	0.75	**0.73**	0.72	0.74	**0.77**	**0.73**	**0.83**	**0.85**	**0.78**
Gibbs	0.68	0.66	0.6	0.69	0.67	0.68	0.59	0.59	0.69
Tepi	0.6	0.65	0.6	0.7	0.59	0.66	0.66	0.68	0.66
LP2	0.67	0.7	0.67	0.76	0.65	0.7	0.73	0.76	0.66

Table 7.3: Performance of our method over the IEDB test set.

Allele	num	AUC
DRB1*0101	100	0.89
DRB1*0301	100	0.73
DRB1*0401	100	0.82
DRB1*0404	100	0.87
DRB1*0405	100	0.80
DRB1*0701	100	0.80
DRB1*0802	73	0.81
DRB1*0901	39	0.93
DRB1*1101	100	0.87
DRB1*1302	100	0.76
DRB1*1501	100	0.66
DRB4*0101	63	0.80
DRB5*0101	100	0.81
Summary	1175	0.81

Table 7.4: Description of data sets used in this work. Train is the training set used for the MHCBench test set. Train2 is the same training set with the addition of peptides belonging to different alleles.

Data Set	Total	Binders	Non Binders
Set 1	531	248	283
Set 2	416	161	255
Set 3a	355	151	204
Set 3b	325	128	197
Set 4a	403	120	283
Set 4b	375	120	255
Set 5a	110	65	45
Set 5b	84	47	37
Southwood	99	19	80
Geluk 1	21	15	6
Train	639	462	177
Train2	2997	1782	121
IEDB	6272	3136	3136
MHCPEP	3111	1037	2074

and used for binding prediction Some techniques attempt something similar. For example, TEPITOPE learns individual binding pockets, allowing it some level of generalization. This allows us to predict peptide binding for alleles with little or no experimental data. For MHC class I molecules there are hundreds of alleles. MHC class II molecules are polymers of two different molecules called the alpha and beta chains. HLA-DQ has several hundred alpha and beta chains, with thousands of possible combinations, each of which binds different peptides. Since peptide binding experiments are currently costly and time consuming,the ability to predict binding for unseen alleles is an extremely useful feature of our method.

In performing the comparison experiments above we discovered that there are significant differences in the data available from different sources. Instead of using one of these data sets for transfer experiments, we searched for a meticulously curated data set of peptide binding data. The data set we found comes from the IEDB [114] database. This resource maintains a hand curated list of the epitopes, and carries continuous IC50 values instead of just marking peptides as binding or non-binding. We felt that this represented one of the best online resources for MHC binding data, and incorporates data from the comparison data sets above as well as others. We downloaded the complete IEDB MHC and TCell binding data from IEDB, removing peptides from before 1993, and any peptide marked as a good binder with an IC50 of greater than 3000 and any peptide marked as a non-binder with IC50 less than 500. In order to guarantee an equal number of binding and non-binding peptides in each allele set, we added random human peptides from SwissProt [6] until each allele was balanced. This data is described at

Using the IEDB database described above, we created transfer data sets by removing all epitopes of each allele in turn. For each of these data sets, we trained the model using 5 fold cross validation to estimate the optimal parameters. We then threaded the MHC sequence of the allele that was left out onto the structure of the allele that had the closest sequence alignment. We then ran the model using this sequence structure combination over all of the alleles from the data set. Since there is significant overlap between peptides that bind to different alleles, we compared our transfer results to two different voting based methods for predicting binding of unseen alleles. We ran our standard trained model for all observed alleles in the training data over the set of peptides of the unobserved allele. We called a peptide a binder if the majority of the alleles called it a binder. In another voting setup, we called a peptide a good binder if a majority of the alleles in the supertype of the left out allele called it an good binder. We plotted a ROC curves for the performance of each method and calculated their average AUC. The results are show in Fig. 3. As can be seen in the figures, our threading method significantly (p-value ¡ .00001 binomial) outperforms either voting mechanism. We are able to predict peptide binding for MHC class II alleles having learned over both alpha and beta chains, a single alpha or beta chain, or without any previous exposure to either chain of the allele.

7.3.3 Myelin binding

There are several auto-immune diseases in which the nerve insulating material called myelin is degraded. This degradation disrupts signal passage through the nervous system and can cause severe health problems. Myelin Basic Protein (MBP) has been shown to bind to the MHC class II allele HLA-DRB1*1501, and is a candidate autoantigen for multiple sclerosis (MS), an auto-immune disease of the central nervous system. We demonstrate how our MHC class II binding predictor can be used in autoimmune research by replicating several MS experimental results in silico. The HLA-DR2 supertype has been repeatedly shown to positively associate with MS [71]. We ran our method over the MBP using the HLA-DR2 allele HLA-DRB1*1501 (Fig. 4) and found four potential binders. Of these, the strongest signal was located at amino acid 91 of the MBP. The peptide consisting of residues 85-99 which contains our predicted binding site has been shown experimentally to be an immunodominant epitope for HLA-DRB1*1501 [107]. Furthermore, there is an approved drug to treat certain forms of MS that works by disrupting this binding, and there is active research to find new candidate peptides that will displace MBP 85-99 by competitively binding to the HLA-DRB1*1501 allele. These drugs have been shown to suppress relapse rates of certain forms of MS by 30% [107].

The drug and two other competitive binding peptides take the form of coplymers 1 poly(Y,E,A,K)n, 2 poly(F,Y,A,K)n, and 3 poly(V,W,A,K)n. These are peptide sequences of random combinations of each the amino acids inside the in the poly groups. We measured the number of predicted binders to HLA-DRB1*1501 over 20 random peptides of each of these polymers and found that in 20 polymers of length 50, there were 60, 80, and 155 predicted binders with a binding strength greater than that predicted for MBP 85-99, for polymers 1, 2, and 3 respectively. When 20 random SwissProt proteins of equivalent length were used, there were only 10 predicted stronger binders. This shows our method predicts the potential therapeutic uses of these coplymers. Recently, [107] examined the properties of the copolymers and synthesized non-random peptides of length 15. Three of these J2, J3, and J5 were experimentally found to suppress MBP 85-99 binding with the relative strength of suppression $J5 > J3 > J2$. We ran our method over each of these 15 amino acid long peptides and found that all three had predicted binding energies lower than MBP 85-99 (they form stronger bonds). Furthermore, the order of binding strength matched that of the relative levels of suppression. That is, J5 was the strongest binder followed by J3 and then J2. Thus our model may be used as a testbed for screening other potential auto-immune drugs that work on the same principle.

7.4 Conclusions

We have developed a novel MHC class II binding model which can be trained on examples of measured binding affinities for a number of allele-peptide combinations, as well as lists of good and bad binders for various alleles. In the latter case, the good binders are given low and bad binders high nominal energy. To the best of our knowledge, our method outperforms significantly all previously published class II epitope prediction techniques, due to its unique treatment of the variable position of the peptide with respect to the binding groove. Our method is physics-based, and treats the binding configurations with different possible peptide positions as a statistical ensemble in a thermodynamic sense. However, as opposed to other structure-based techniques [35], our approach is both accurate in binding energy prediction *and* computationally efficient. For instance, due to the computational cost, [35] reports results for only six peptides. Our model, while guided by the known MHC II structures, is simplified and enriched with trainable parameters, which allows us to refine it using published binding data. Testing a new peptide takes a fraction of a second. One of the most appealing properties of our technique is that it naturally generalizes well to previously unseen MHC II alleles (or unseen combinations of alpha and beta chains). We illustrated the accuracy of our technique on a biological problem: identifying targets and drugs for an autoimmune disorder. We are also investigating the uses of the model to explain certain evolutionary trends in pathogens.

Chapter 7, was published in The Journal of Computational Biology, Vol 15, pp 927-942, 2008. Noah Zaitlen, Manuel Reyes-Gomez, David Heckerman, Nebojsa Jojic, “Shift Invariant Adaptive Double Threading: Learning MHC II Peptide Binding”. The dissertation author was the primary investigator and author of this paper.

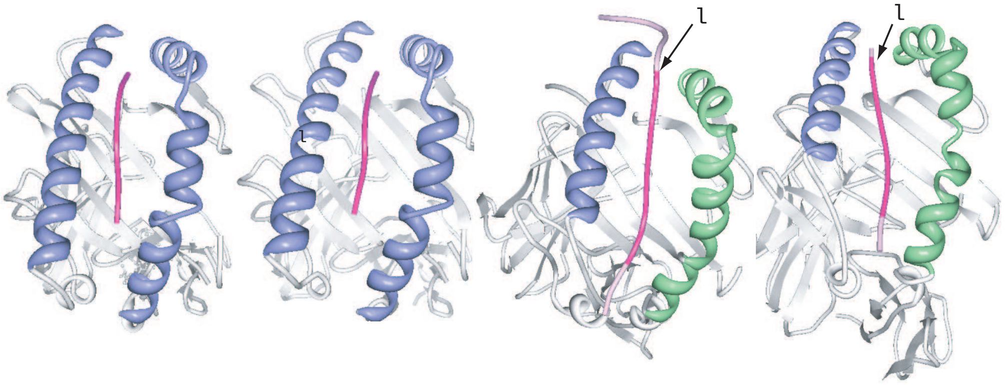

Figure 7.1: Examples of binding configurations of MHC class I molecules (first two renderings) and MHC class II (last two renderings) bound two different peptides. The class I molecules are rendered in gray, except for the alpha helices forming the groove, which are shown in blue to accentuate the peptide (pink) sitting snugly inside it. The class II molecules consist of two separate chains. Their alpha helices (rendered in blue and green) form a similar groove to that of class I molecules. While a class I molecule binds to different short peptides in a relatively constant configuration, a class II molecule can bind to peptides of much more variable length, with only a short segment (dark pink) of the bound peptide captured in its groove, and the peptide tails (light pink) sticking out and having a smaller effect on the configuration strength. The start of the segment that fits the groove is modeled by the random variable ℓ in Section 7.2.1, while the much smaller variability in the relative configuration of the peptide segment in the groove is modeled by the variable m denoting different available crystal structures. We used 90 available MHC II structures as exemplars for the groove-segment geometries. Note that the configurations of different MHC-peptide configurations in the figure are shown from slightly different viewing angles to help understand the 3D structure, but the segments that sit in the grooves of all four molecules are highly constrained, and they vary modestly in their configuration with respect to the MHC molecule. (This is what made it possible to train an energy predictor for the whole family of MHC I molecules in [66].) On the other hand, for MHC class II molecules, the usually unknown segment start ℓ can have dramatic consequences on the energy predictions as it determines the amino acids that sit in the groove.

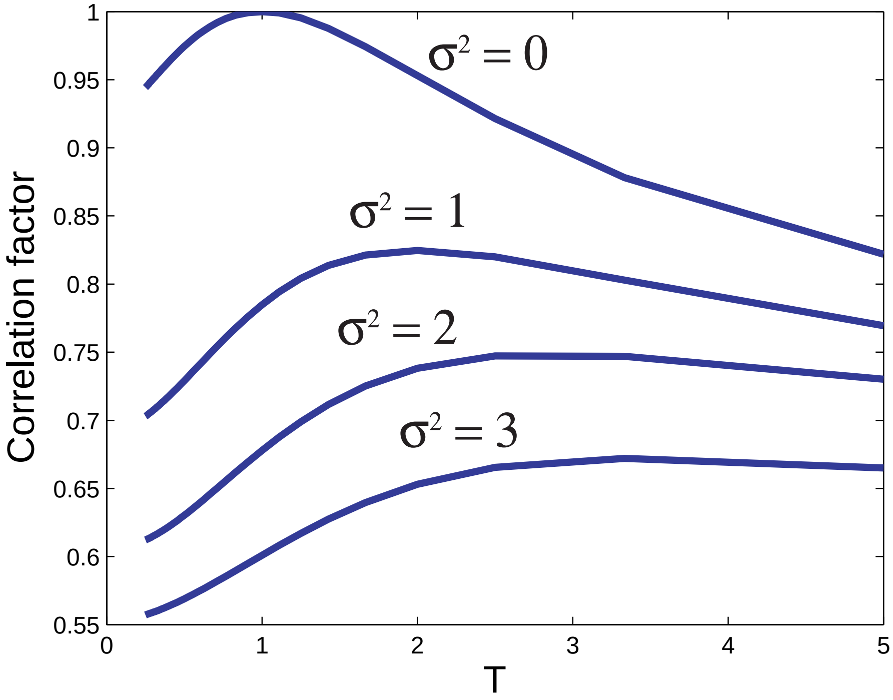

Figure 7.2: The effect of the temperature T used in energy estimate (7.11) on the prediction accuracy, here measure in terms of the correlation between the estimate and the "true" energy in the synthetic experiment described in the text. The curves correspond to the variance of the modeling error σ^2 of 0,1,2, and 3. Higher error variance leads to lower Spearman correlation factors, and the best correlation is achieved at optimal temperatures which increase with the error variance, as expected. The optimal estimated temperature for integrating states of the model is thus a symptom of modeling errors. In our experiments, this temperature converges to values close to one, indicating the possibility that most of the prediction errors are due to causes of variability not represented by the data, e.g., measurement noise, or higher level effects such as epitope competition.

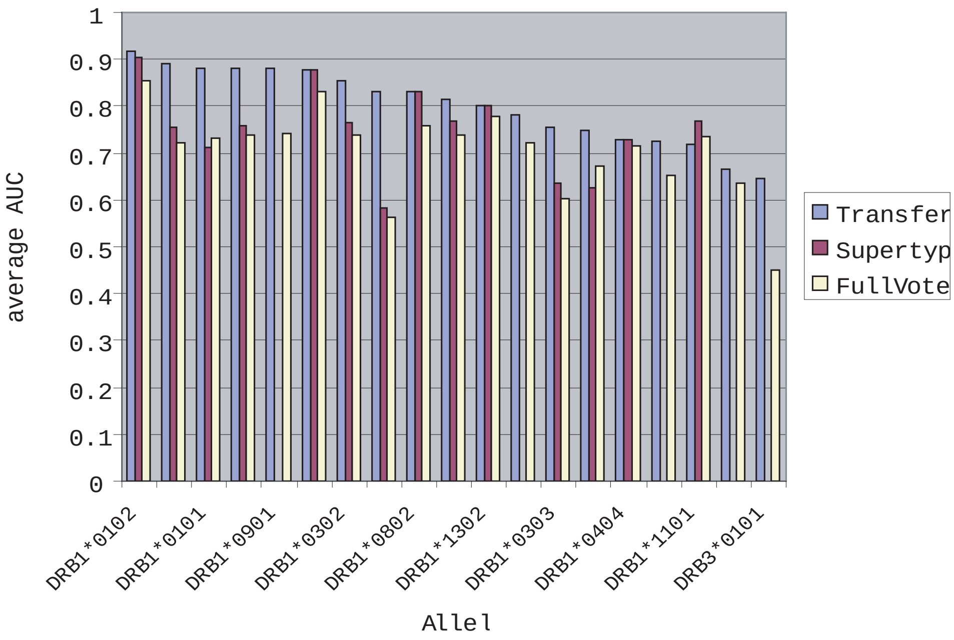

Figure 7.3: The capability of generalizing epitope prediction to alleles not found in the training set allows our method to be applied to a much larger set of MHC molecules. This figure shows the significantly greater predictive power of our method over two voting based mechanisms for binding across alleles. Note that the supertype method has zero values if there were no other members of the supertype in the training data.

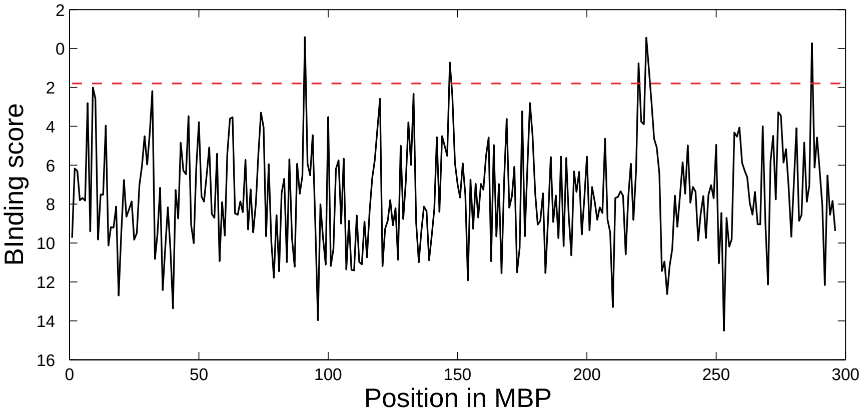

Figure 7.4: Binding score from our double threading model over the myelin basic protein (MBP). Higher scores are better binders. The red dashed line denotes our threshold for a positive binder. There are four clear peaks above the threshold, the largest of which is at position 91. This falls directly inside of the immunodominant epitope MBP 85-99.

Chapter 8

Conclusion

We have developed several new methods for improving our ability to conduct genome wide association studies. We have also worked out the details of some statistical issues relevant to GWAS and the use of the HapMap in general. In this very fast paced field some of our methods have been replaced, improved upon by ideas from other groups, or outworn their usefulness as technology has changed. This is the nature of working in a field where technology is so intertwined with methodology. It is exciting and there is always a demand for something new. I give a brief summary of the methods described in this text, some open problems in the field, and some discussion of where I think the field maybe going.

The weighted haplotype association method WHAP utilizes the linkage structure information learned from the HapMap project to improve the power of GWAS by computing statistics for untyped SNPs. It can also be used as part of a meta-analysis project in which SNPs are genotyped on different platforms, solving the issues of how to combine the genotype data for the same phenotypes. It has been used as part of several genome wide association studies, most notably an examination of several important disease related metabolic phenotypes from a Finnish population. We extended the method to handle continuous data in order to participate in this collaboration. At this point many of the newer methods outperform WHAP in both the case control and continuous setting. However, it was one of the first to solve this problem, and demonstrates some fundamental proprieties of haplotype structure and SNP distributions in and between human populations.

The SATTagger software offers the first complete solution to the SNP tagging problem. Although the classic version of this problem is no longer central to the field of genetics, it has been worked on by dozens of groups, and is considered by some to be a classic bioinformatics problem. The reduction to SAT takes advantage of the local linkage structure of the genome to reduce complexity and run time. Since the development of imputation methods such as WHAP, the number of markers required for tagging has been significantly reduced and may serve some utility in the development of cheap genotyping platforms.

The imputation aware meta-analysis method is likely the most relevant at the time of writing. There are many ongoing large meta-analysis projects of GWAS data. These will almost certainly use imputation to solve the problem of non-overlapping marker sets, and would benefit from the use of our method. Indeed, one such study on Bipolar disorder is nearing completion and we have recently received the data necessary to test our method. The technique is simple enough that no software package is necessary to aid researchers in its implementation.

All of the above methods have in common the use of the HapMap data set. Indeed, it is the quintessential data set of GWAS and is partly responsible for our ability to conduct these studies. It has been used repeatedly to calculate numerous statistics, most often r^2, to help us understand the nature of SNPs in human populations. We developed a new framework for understanding the finite sample issues of these statistics. We showed that they can cause significant errors in some cases, we bounded the size of these errors, and showed how many more individuals would be needed to correct them. The ongoing 1000 genomes project will help address this problem.

The phasing and analysis of the NCBI genotypes was published just at the release of the first version of the HapMap. Unlike the HapMap groups phasing of the data, we did not require one of the largest supercomputers in the world, and managed to have results on par in terms of accuracy. The NCBI haplotype resource was not extensively used, but the XML developed to manage the data solved some interesting issues, which are increasingly relevant in this field. Namely, how to properly store and index the volumes of information that are being produced by new technologies. Additionally, the paper has been used to study genotyping error rates, and argue against haplotype blocks.

The last method given comes from the domain of computational immunology. It does loosely fall into the domain of disease and human genetic variation as a goal of the method is to characterize how viral proteins interact with different immune systems. The MHC has been one of the most frequent and strong loci in GWAS results. Using our method in combination with GWAS results of auto-immune disorders may help uncover the proteins illiciting the unwanted immune response. This was partially demonstrated in the MS case given at the end of the chapter. In addition to auto-immune disease there potential applications in infectious disease and cancer.

8.0.1 Open Problems

We have recently seen the completion of the first rounds of genome wide association studies and they were met with some controversy. The single largest problem in the field right now is understanding where all the missing signal is. Years of twin and family studies have provided estimates of heritability for many of the same diseases in GWAS. Many novel loci have been discovered but only a small fraction of the total heritability has been explained by these results. The odds ratios are commonly between 1.2 and 1.4. It maybe the case that we just need larger studies to reveal hundreds of SNPs of small effect, but several alternatives has also been suggested and methods need to be developed to test for them. Gene environment interactions, errors from population substructure, admixture mapping in populations with complex ancestral populations, inferring population histories, and epigenetics have been and continue to be areas of active development.

Rare variants are SNPs with very low minor allele frequency and are nearly impossible to discover under current GWAS unless they have extreme effect sizes. Even if they are identified in all individuals in a study conducted with next generation sequencing, there is not yet a standard procedure or statistical test to apply. In fact, there are multiple hypotheses within this single problem. It could be that genes are disrupted under a mutational load or one single SNP is acting but is rare. Methods to understand the effects of rare variants are much needed and currently underexplored.

Next generation sequencing has already provided a wealth of new problems. At this point they are mostly technical in the sense that answers are needed before the technology can be fully utilized. Mapping, phasing, and assembly have already seen dozens of publications, and with the quality of the data changing so rapidly likely many more will appear soon. Several of the groups working on imputation have seen opportunity here and are working on reducing the number of individuals required for sequencing, or the number of reads per individual by imputing across the sample. It still remains an open problem as the running times and quality of results is not yet adequate.

Using multiple sources of data instead of genotypes in isolation could provide new insights into each data set as well as the phenotypes. Genotyping as well as measuring expression, protein (via mass spec), methylation patterns, and intermediate phenotypes gives a much richer set of data then genotyping on its own. In addition to helping identify the loci associated with disease, it maybe possible to learn about aspects of the mechanism of disruption or protection. Many people working previously strictly in statistical genetics are beginning to write and work on problems in “systems biology”, although they may use alternative terminology.

Within the field of computational immunology next generation sequencing offers some very interesting new problems. The viruses are so small that it is possible to cheaply sequence samples from many individuals at multiple time points. It may be possible to observe in much shorter time periods viral evolution and connect it with host/virus/environment interactions. This could aid not only in the understanding of the pressures on viral populations, but also in the development of new classes of vaccines and treatments.

Bibliography

[1] J. Argelich and F. Manya. Partial max-sat solvers with clause learning. In *Proceedings of SAT'07*, pages 28–40, 2007.

[2] Josep Argelich, Chu Min Li, Felip Manya, and Jordi Planes. First evaluation of max-sat solvers, 2006. http://www.iiia.csic.es/∼maxsat06/.

[3] Josep Argelich, Chu Min Li, Felip Manya, and Jordi Planes. Second evaluation of max-sat solvers, 2007. http://www.maxsat07.udl.es/.

[4] V. Bafna, B. V. Halldorsson, R. Schwartz, A. G. Clark, and S. Istrail. Haplotypes and informative snp selection algorithms: don't block out information. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, pages 19–27. ACM New York, NY, USA, 2003.

[5] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic acids research*, 34(Web Server issue):W369, 2006.

[6] A. Bairoch and R. Apweiler. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic Acids Research*, 28(1):45, 2000.

[7] Anthony Barrett. From hybrid systems to universal plans via domain compilation. In *Proceedings of the 14th International Conference on Planning and Scheduling (ICAPS)*, pages 44–51, 2004.

[8] Anthony Barrett. Model compilation for real-time planning and diagnosis with feedback. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1195–1200, 2005.

[9] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, and others others. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58(6):899–907, 2002.

[10] T. R. Bhangale, M. J. Rieder, and D. A. Nickerson. Estimating coverage and power for genetic association studies using near-complete variation data. *Nature Genetics*, 2008.

[11] M. Bhasin, H. Singh, and G. P. S. Raghava. Bioinformatics, 2003.

[12] Blai Bonet and Hector Geffner. Heuristics for planning with penalties and rewards using compiled knowledge. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pages 452–462, 2006.

[13] V. Brusic, G. Rudy, and L. C. Harrison. Mhcpep, a database of mhc-binding peptides: update 1997. *Nucleic acids research*, 26(1):368, 1998.

[14] V. Brusic, G. Rudy, G. Honeyman, J. Hammer, and L. Harrison. Prediction of mhc class ii-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics*, 14(2):121–130, 1998.

[15] S. Buus, S. L. Lauemoller, P. Worning, C. Kesmir, T. Frimurer, S. Corbet, A. Fomsgaard, J. Hilden, A. Holm, and S. Brunak. Sensitive quantitative predictions of peptide-mhc binding by a'query by committee'artificial neural network approach. *Tissue Antigens*, 62(5):378, 2003.

[16] The c2d compiler. http://reasoning.cs.ucla.edu/c2d/.

[17] M Cargill, D Altshuler, J Ireland, P Sklar, K Ardlie, N Patil, N Shaw, CR Lane, EP Lim, N Kalyanaraman, J Nemesh, L Ziaugra, L Friedland, A Rolfe, J Warrington, R Lipshutz, GQ Daley, and ES Lander. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.*, 22:231–8, 1999.

[18] Christopher S. Carlson, Michael A. Eberle, Mark J. Rieder, Qian Yi, Leonid Kruglyak, and Deborah A. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet*, 74(1):106–20, 1 2004.

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

[20] Mark Chavira and Adnan Darwiche. Compiling Bayesian networks with local structure. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1306–1312, 2005.

[21] Mark Chavira, Adnan Darwiche, and Manfred Jaeger. Compiling relational Bayesian networks for exact inference. *International Journal of Approximate Reasoning*, 42(1–2):4–20, May 2006.

[22] F. S. Collins, L. D. Brooks, and A. Chakravarti. A dna polymorphism discovery resource for research on human genetic variation. *Genome Res*, 8(12):1229–31, 12 1998.

[23] FS Collins, LD Brooks, and A. Chakravarti. A dna polymorphism discovery resource for research on human genetic variation. *Genome Research*, 8:1229–1231, 1998.

[24] International HapMap Consortium. The international hapmap project. *Nature*, 426:789–796, 2003.

[25] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, 10 2005.

[26] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, 6 2007.

[27] DC Crawford, CS Carlson, MJ Rieder, DP Carrington, Q Yi, JD Smith, MA Eberle, L Kruglyak, and DA Nickerson. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am J Hum Genet.*, 74:610–22, 2004.

[28] 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 Genet.*, 29(2):229–32, 2001.

[29] Adnan Darwiche. Decomposable negation normal form. *Journal of the ACM*, 48(4):608–647, 2001.

[30] Adnan Darwiche. On the tractability of counting theory models and its application to belief revision and truth maintenance. *Journal of Applied Non-Classical Logics*, 11(1-2):11–34, 2001.

[31] Adnan Darwiche. A compiler for deterministic, decomposable negation normal form. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI)*, pages 627–634, Menlo Park, California, 2002. AAAI Press.

[32] Adnan Darwiche. New advances in compiling CNF to decomposable negational normal form. In *Proceedings of European Conference on Artificial Intelligence*, pages 328–332, 2004.

[33] Adnan Darwiche and Pierre Marquis. A knowledge compilation map. *Journal of Artificial Intelligence Research*, 17:229–264, 2002.

[34] Adnan Darwiche and Pierre Marquis. Compiling propositional weighted bases. *Artificial Intelligence*, 157(1-2):81–113, 2004.

[35] M. N. Davies, C. E. Sansom, C. Beazley, and D. S. Moss. A novel predictive technique for the mhc class ii peptide–binding interaction. *Molecular Medicine*, 9(9-12):220, 2003.

[36] P. I. De Bakker, N. P. Burtt, R. R. Graham, C. Guiducci, R. Yelensky, J. A. Drake, T. Bersaglieri, K. L. Penney, J. Butler, S. Young, and others others. Transferability of tag snps in genetic association studies in multiple populations. *Nature genetics*, 38(11):1298–1303, 2006.

[37] Paul I. W. de Bakker, Manuel A. R. Ferreira, Xiaoming Jia, Benjamin M. Neale, Soumya Raychaudhuri, and Benjamin F. Voight. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet*, 17(R2):R122–8, 10 2008.

[38] Paul I. W. de Bakker, Roman Yelensky, Itsik Pe'er, Stacey B. Gabriel, Mark J. Daly, and David Altshuler. Efficiency and power in genetic association studies. *Nat Genet*, 37(11):1217–23, 11 2005.

[39] B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2):311–22, 9 1995.

[40] P. Dnnes and A. Elofsson. Prediction of mhc class i binding peptides, using svmhc. *BMC bioinformatics*, 3(1):25, 2002.

[41] Paul Elliott and Brian Williams. Dnnf-based belief state estimation. In *Proceedings of the* 21*st National Conference on Artificial Intelligence (AAAI-06)*, 2006.

[42] E. Eskin, E. Halperin, and R.M. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. *J. Bioinform. Comput. Biol.*, 1(1):1–20, 2003.

[43] L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 12(5):921–7, 9 1995.

[44] Y Freudenberg-Hua, J Freudenberg, N Kluck, S Cichon, P Propping, and MM Nothen. Single nucleotide variation analysis in 65 candidate genes for cns disorders in a representative sample of the european population. *Genome Res*, 13:2271–6, 2003.

[45] Zhaohui Fu and Sharad Malik. On solving the partial max-sat problem. *Proceedings of Theory and Applications of Satisfiability Testing*, pages 252–265, 2006.

[46] GB. Gabriel, SF. Schaffner, H. Nguyen, JM. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, SN. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, ES. Lander, MJ. Daly, and D. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, 2002.

[47] A. Geluk, K. E. Van Meijgaarden, N. C. Schloot, J. W. Drijfhout, T. H. Ottenhoff, and B. O. Roep. Hla-dr binding analysis of peptides from islet antigens in iddm. *Diabetes*, 47(10):1594–1601, 1998.

[48] The International SNP Map Working Group. A map of human genome sequence variation containing 1.4 million snps. *Nature*, 409:928–933, 2001.

[49] Yongtao Guan and Matthew Stephens. Practical issues in imputation-based association mapping. *PLoS Genet*, 4(12):e1000279, 12 2008.

[50] Kevin L. Gunderson, Frank J. Steemers, Grace Lee, Leo G. Mendoza, and Mark S. Chee. A genome-wide scalable snp genotyping assay using microarray technology. *Nat Genet*, 37(5):549–54, 5 2005.

[51] A. Darwiche H. Palacios, B. Bonet and H. Geffner. Pruning conformant plans by counting models on compiled d-dnnf representations. In *Proceedings of the 15th International Conference on Planning and Scheduling (ICAPS)*, pages 141–150. AAAI Press, 2005.

[52] B. V. Halldorsson, S. Istrail, and F. M. De La Vega. Optimal selection of snp markers for disease association studies. *Hum Hered*, 58(3-4):190–202, 2004.

[53] B. V. Halldrsson, V. Bafna, R. Lippert, R. Schwartz, F. M. De La Vega, A. G. Clark, and S. Istrail. Genome research, 2004.

[54] E. Halperin and E. Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20(12):1842–9, 2004.

[55] Eran Halperin and Elad Hazan. Haplofreq–estimating haplotype frequencies efficiently. *J Comput Biol*, 13(2):481–500, 3 2006.

[56] MK Halushka, JB Fan, K Bentley, L Hsie, N Shen, A Weder, R Cooper, R Lipshutz, and A. Chakravarti. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.*, 22:239–47, 1999.

[57] F. Heras, J. Larrosa, S. de Givry, and T. Schiex. Toolbar max-sat solver homepage. http://mulcyber.toulouse.inra.fr/projects/toolbar/.

[58] T. Hertz and C. Yanover. Pepdist: a new framework for protein-peptide binding prediction based on learning peptide distance functions. *BMC bioinformatics*, 7(Suppl 1):S3, 2006.

[59] LA Hindorff, HA Junkins, JP Mehta, and TA Manolio. A catalog of published genome-wide association studies. *www.genome.gov/26525384*, Accessed [April 24, 2009].

[60] D. A. Hinds, L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin, D. G. Ballinger, K. A. Frazer, and D. R. Cox. Whole-genome patterns of common dna variation in three human populations. *Science*, 307(5712):1072–1079, 2005.

[61] Jinbo Huang. Complan: A conformant probabilistic planner. In *Proceedings of the 16th International Conference on Planning and Scheduling (ICAPS)*, 2006.

[62] Jinbo Huang and Adnan Darwiche. On compiling system models for faster and more scalable diagnosis. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI)*, pages 300–306, 2005.

[63] RR Hudson. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, 7:1–44, 1991.

[64] G. C. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, R. C. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S. C. Gough, D. G. Clayton, and J. A. Todd. Haplotype tagging for the identification of common disease genes. *Nat Genet*, 29(2):233–7, 10 2001.

[65] N. Jojic and B. J. Frey. Topographie transformation as a discrete latent variable. In *Advances in Neural Information Processing Systems 12: Proceedings of the 1999 Conference*, page 477. MIT Press, 2000.

[66] N. Jojic, M. Reyes-Gomez, D. Heckerman, C. Kadie, and O. Schueler-Furman. Learning mhc i-peptide binding. *Bioinformatics*, 22(14), 2006.

[67] D. T. Jones, W. R. Taylort, and J. M. Thornton. A new approach to protein fold recognition. *Nature*, 358(6381):86–89, 1992.

[68] O. Karpenko, J. Shi, and Y. Dai. Prediction of mhc class ii binders using the ant colony search strategy. *Artificial Intelligence in Medicine*, 35(1-2):147–156, 2005.

[69] Fotini K. Kavvoura and John P. A. Ioannidis. Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Hum Genet*, 123(1):1–14, 2 2008.

[70] GC Kennedy, H Matsuzaki, S Dong, WM Liu, J Huang, G Liu, X Su, M Cao, W Chen, J Zhang, W Liu, G Yang, X Di, T Ryder, Z He, U Surti, MS Phillips, MT Boyce-Jacino, SP Fodor, and KW Jones. Large-scale genotyping of complex dna. *Nat Biotechnol.*, 10:1233–7, 2003.

[71] M. Krogsgaard, K. W. Wucherpfennig, B. Cannella, B. E. Hansen, A. Svejgaard, J. Pyrdol, H. Ditzel, C. Raine, J. Engberg, and L. Fugger. Visualization of myelin basic protein (mbp) t cell epitopes in multiple sclerosis lesions using a monoclonal antibody specific for the human histocompatibility leukocyte antigen (hla)-dr2-mbp 85-99 complex. *Journal of Experimental Medicine*, 191(8):1395–1412, 2000.

[72] L Kruglyak and DA Nickerson. Variation is the spice of life. *Nature Genet.*, 27(234):234–236, 2001.

[73] Chu Min Li, Felip Manya, and Jordi Planes. New inference rules for max-sat. *JAIR*, 2007.

[74] Y Li and GR Abecasis. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet*, S79(2290), 2006.

[75] Han Lin and Kaile Su. Exploiting inference rules to compute lower bounds for max-sat solving. In *IJCAI*, pages 2334–2339, 2007.

[76] RJ Livingston, A von Niederhausern, AG Jegga, DC Crawford, CS Carlson, MJ Rieder, S Gowrisankar, BJ Aronow, and DA Nickerson. Patterns of sequence variation across 213 environmental response genes. *Genome Research*, 14:1821–31, 2004.

[77] J. B. Maller, J. A. Fagerness, R. C. Reynolds, B. M. Neale, M. J. Daly, and J. M. Seddon. Variation in complement factor 3 is associated with risk of age-related macular degeneration. *Nature genetics*, 39(10):1200–1201, 2007.

[78] D. M. Maraganore, M. de Andrade, T. G. Lesnick, K. J. Strain, M. J. Farrer, W. A. Rocca, P. V. K. Pant, K. A. Frazer, D. R. Cox, and D. G. Ballinger. High-resolution whole-genome association study of parkinson disease. *The American Journal of Human Genetics*, 77(5):685–693, 2005.

[79] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–913, 2007.

[80] Jonathan Marchini, David Cutler, Nick Patterson, Matthew Stephens, Eleazar Eskin, Eran Halperin, Shin Lin, Zhaohui S. Qin, Heather M. Munro, Goncalo R. Abecasis, Peter Donnelly, and International HapMap Consortium. A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet*, 78(3):437–50, 3 2006.

[81] Hajime Matsuzaki, Shoulian Dong, Halina Loi, Xiaojun Di, Guoying Liu, Earl Hubbell, Jane Law, Tam Berntsen, Monica Chadha, Henry Hui, Geoffrey Yang, Giulia C. Kennedy, Teresa A. Webster, Simon Cawley, P. Sean Walsh, Keith W. Jones, Stephen P. A. Fodor, and Rui Mei. Genotyping over 100,000 snps on a pair of oligonucleotide arrays. *Nat Methods*, 1(2):109–11, 11 2004.

[82] F. Melo, R. Sanchez, and A. Sali. Statistical potentials for fold assessment. *Protein Science: A Publication of the Protein Society*, 11(2):430, 2002.

[83] A. Montpetit, M. Nelis, P. Laflamme, R. Magi, X. Ke, M. Remm, L. Cardon, T. J. Hudson, and A. Metspalu. An evaluation of the performance of tag snps derived from hapmap in a caucasian population. *PLoS Genet*, 2(3):e27, 2006.

[84] N. Murugan and Y. Dai. Prediction of mhc class ii binding peptides based on an iterative learning model. *Immunome research*, 1(1):6, 2005.

[85] Dan L. Nicolae. Testing untyped alleles (tuna)-applications to genome-wide association studies. *Genet Epidemiol*, 30(8):718–27, 12 2006.

[86] M. Nielsen, C. Lundegaard, P. Worning, C. Sylvester-Hvid, K. Lamberth, S. Buus, S. Brunak, and O. Lund. Improved prediction of mhc class i and ii epitopes using a novel gibbs sampling approach. *Bioinformatics*, 20:1388–97, 2004.

[87] T. Niu, S. Qin, X. Xu, and J. Liu. Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. *Am. J. Hum. Genet.*, 70:157–169, 2002.

[88] G. W. Oehlert. A note on the delta method. *American Statistician*, pages 27–29, 1992.

[89] N Patil, AJ Berno, DA Hinds, WA Barrett, JM Doshi, CR Hacker, CR Kautzer, DH Lee, C Marjoribanks, DP McDonough, BT Nguyen, MC Norris, JB Sheehan, N Shen, D Stern, RP Stokowski, DJ Thomas, MO Trulson, KR Vyas, KA Frazer, SP Fodor, and DR Cox. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–23, Nov 23 2001.

[90] Itsik Pe'er, Paul I. W. de Bakker, Julian Maller, Roman Yelensky, David Altshuler, and Mark J. Daly. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet*, 38(6):663–7, 6 2006.

[91] Innate Immunity PGA. Nhlbi program in genomic applications.

[92] Knot Pipatsrisawat and Adnan Darwiche. Clone: Solving weighted max-sat in a reduced search space. In *Proceedings of Twentieth Australian Joint Conference on Artificial Intelligence*, 2007.

[93] J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, 69(1):1–14, 7 2001.

[94] Z. S. Qin, S. Gopalakrishnan, and G. R. Abecasis. An efficient comprehensive search algorithm for tagsnp selection using linkage disequilibrium criteria. *Bioinformatics*, 22(2):220–225, 2006.

[95] G. P. Raghava. Mhcbench–evaluation of mhc binding peptide prediction algorithms. *URL http://www. imtech. res. in/raghava/mhcbench*, 2006.

[96] H. G. Rammensee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor, and S. Stevanovi. Syfpeithi: database for mhc ligands and peptide motifs. *Immunogenetics*, 50(3):213–219, 1999.

[97] P. A. Reche, J. P. Glutting, H. Zhang, and E. L. Reinherz. Enhancement to the rankpep resource for the prediction of peptide binding to mhc molecules using profiles. *Immunogenetics*, 56(6):405–419, 2004.

[98] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–7, 9 1996.

[99] N. S. Roy, S. Farheen, N. Roy, S. Sengupta, and P. P. Majumder. Portability of tag snps across isolated population groups: an example from india. *Annals of Human Genetics*, 72(1):82–89, 2008.

[100] Chiara Sabatti, Susan K. Service, Anna-Liisa L. Hartikainen, Anneli Pouta, Samuli Ripatti, Jae Brodsky, Chris G. Jones, Noah A. Zaitlen, Teppo Varilo, Marika Kaakinen, Ulla Sovio, Aimo Ruokonen, Jaana Laitinen, Eveliina Jakkula, Lachlan Coin, Clive Hoggart, Andrew Collins, Hannu Turunen, Stacey Gabriel, Paul Elliot, Mark I. McCarthy, Mark J. Daly, Marjo-Riitta R. Jrvelin, Nelson B. Freimer, and Leena Peltonen. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet*, 41(1):35–46, 1 2009.

[101] Tian Sang, Paul Beame, and Henry Kautz. Solving Bayesian networks by weighted model counting. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05)*, volume 1, pages 475–482. AAAI Press, 2005.

[102] O. Schueler-Furman, Y. Altuvia, A. Sette, and H. Margalit. Structure-based prediction of binding peptides to mhc class i molecules: application to a broad range of mhc alleles. *PRS*, 9(09):1838–1846, 2000.

[103] P. C. Sham, S. S. Cherny, S. Purcell, and J. K. Hewitt. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet*, 66(5):1616–30, 5 2000.

[104] Sajjad Siddiqi and Jinbo Huang. Hierarchical diagnosis of multiple faults. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[105] S. Southwood, J. Sidney, A. Kondo, M. F. Del Guercio, E. Appella, S. Hoffman, R. T. Kubo, R. W. Chesnut, H. M. Grey, and A. Sette. Several common hla-dr types share largely overlapping peptide binding repertoires. *Journal of immunology (Baltimore, Md.: 1950)*, 160(7):3363, 1998.

[106] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68(4):978–89, 4 2001.

[107] J. N. H. Stern, Z. Ills, J. Reddy, D. B. Keskin, M. Fridkis-Hareli, V. K. Kuchroo, and J. L. Strominger. Peptide 15-mers of defined sequence that substitute for random amino acid copolymers in amelioration of experimental autoimmune encephalomyelitis. *Proceedings of the National Academy of Sciences*, 102(5):1620–1625, 2005.

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

[109] Daniel O. Stram. Tag snp selection for association studies. *Genet Epidemiol*, 27(4):365–74, 12 2004.

[110] Daniel O. Stram, Celeste Leigh Pearce, Phillip Bretsky, Matthew Freedman, Joel N. Hirschhorn, David Altshuler, Laurence N. Kolonel, Brian E. Henderson, and Duncan C. Thomas. Modeling and e-m estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum Hered*, 55(4):179–90, 2003.

[111] Felip Manya Teresa Alsinet and Jordi Planes. A max-sat solver with lazy data structures. In *Proceedings of Ninth Ibero-American Conference on Artificial Intelligence (IBERAMIA 2004)*, 2004.

[112] J. D. Terwilliger and T. Hiekkalinna. An utter refutation of the fundamental theorem of the hapmap. *European Journal of Human Genetics*, 14(4):426–437, 2006.

[113] Federico Heras Viaga, Javier Larrosa, and Albert Oliveras. Minimaxsat: a new weighted max-sat solver. In *Proceedings of SAT'07*, 2007.

[114] R. Vita, K. Vaughan, L. Zarebski, N. Salimi, W. Fleri, H. Grey, M. Sathiamurthy, J. Mokili, H. H. Bui, P. E. Bourne, and others others. Curation of complex, context-dependent immunological data. *BMC bioinformatics*, 7(1):341, 2006.

[115] M. Wachter and R. Haenni. Logical compilation of bayesian networks. Technical Report iam-06-006, University of Bern, Switzerland, 2006.

[116] DG Wang, JB Fan, CJ Siao, A Berno, P Young, R Sapolsky, G Ghandour, N Perkins, E Winchester, J Spencer, L Kruglyak, L Stein, L Hsie, T Topaloglou, E Hubbell, E Robinson, M Mittmann, M Morris, N Shen, D Kilburn, J Rioux, C Nusbaum, S Rozen, TJ Hudson, and Lander ES. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280:1077–82, 1998.

[117] Hansong Wang, Duncan C. Thomas, Itsik Pe'er, and Daniel O. Stram. Optimal two-stage genotyping designs for genome-wide association scans. *Genet Epidemiol*, 30(4):356–68, 5 2006.

[118] Mike E. Weale, Chantal Depondt, Stuart J. Macdonald, Alice Smith, Poh San Lai, Simon D. Shorvon, Nicholas W. Wood, and David B. Goldstein. Selection and evaluation of tagging snps in the neuronal-sodium-channel gene scn1a: implications for linkage-disequilibrium gene mapping. *Am J Hum Genet*, 73(3):551–65, 9 2003.

[119] Maria-Esther Vidal Yolif Arvelo, Blai Bonet. Compilation of query–rewriting problems into tractable fragments of propositional logic. In *Proceedings of AAAI National Conference*, 2006.

[120] F. Yu, P. C. Sabeti, P. Hardenbol, Q. Fu, B. Fry, X. Lu, S. Ghose, R. Vega, A. Perez, S. Pasternak, and others others. Positive selection of a pre-expansion cag repeat of the human sca2 gene. *PLoS Genet*, 1(3):e41, 2005.

[121] Noah Zaitlen, Hyun Min Kang, Eleazar Eskin, and Eran Halperin. Leveraging the hapmap correlation structure in association studies. *Am J Hum Genet*, 80(4):683–91, 4 2007.

[122] Noah A. Zaitlen, Hyun Min Kang, Michael L. Feolo, Stephen T. Sherry, Eran Halperin, and Eleazar Eskin. Inference and analysis of haplotypes from

combined genotyping studies deposited in dbsnp. *Genome Res*, 15(11):1594–600, 11 2005.

[123] Eleftheria Zeggini, Laura J. Scott, Richa Saxena, Benjamin F. Voight, Jonathan L. Marchini, Tianle Hu, Paul I. W. de Bakker, Gonalo R. Abecasis, Peter Almgren, Gitte Andersen, Kristin Ardlie, Kristina Bengtsson Bostrm, Richard N. Bergman, Lori L. Bonnycastle, Knut Borch-Johnsen, Nol P. Burtt, Hong Chen, Peter S. Chines, Mark J. Daly, Parimal Deodhar, Chia-Jen J. Ding, Alex S. F. Doney, William L. Duren, Katherine S. Elliott, Michael R. Erdos, Timothy M. Frayling, Rachel M. Freathy, Lauren Gianniny, Harald Grallert, Niels Grarup, Christopher J. Groves, Candace Guiducci, Torben Hansen, Christian Herder, Graham A. Hitman, Thomas E. Hughes, Bo Isomaa, Anne U. Jackson, Torben Jrgensen, Augustine Kong, Kari Kubalanza, Finny G. Kuruvilla, Johanna Kuusisto, Claudia Langenberg, Hana Lango, Torsten Lauritzen, Yun Li, Cecilia M. Lindgren, Valeriya Lyssenko, Amanda F. Marvelle, Christa Meisinger, Kristian Midthjell, Karen L. Mohlke, Mario A. Morken, Andrew D. Morris, Narisu Narisu, Peter Nilsson, Katharine R. Owen, Colin N. A. Palmer, Felicity Payne, John R. B. Perry, Elin Pettersen, Carl Platou, Inga Prokopenko, Lu Qi, Li Qin, Nigel W. Rayner, Matthew Rees, Jeffrey J. Roix, Anelli Sandbaek, Beverley Shields, Marketa Sjgren, Valgerdur Steinthorsdottir, Heather M. Stringham, Amy J. Swift, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Nicholas J. Timpson, Tiinamaija Tuomi, Jaakko Tuomilehto, Mark Walker, Richard M. Watanabe, Michael N. Weedon, Cristen J. Willer, Wellcome Trust Case Control Consortium, Thomas Illig, Kristian Hveem, Frank B. Hu, Markku Laakso, Kari Stefansson, Oluf Pedersen, Nicholas J. Wareham, Ins Barroso, Andrew T. Hattersley, Francis S. Collins, Leif Groop, Mark I. McCarthy, Michael Boehnke, and David Altshuler. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*, 40(5):638–45, 5 2008.

[124] K Zhang, M Deng, T Chen, MS Waterman, and F Sun. A dynamic programming algorithm for haplotype block partitioning. *Proc. Nat. Acad. Sci. U.S.A.*, 99(11):7335–9, May 28 2002.