Fast and Accurate Haplotype Inference with Hidden Markov Model

Yi Liu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of *Doctor of Philosophy* in the Department of Computer Science.

Chapel Hill
2013

Approved by:

Wei Wang

Yun Li

Vladimir Jojic

Fernando Pardo Manuel de Villena

William Valdar

Ethan Lange

Abstract

YI LIU: Fast and Accurate Haplotype Inference with Hidden Markov Model (Under the direction of Wei Wang and Yun Li)

The genome of human and other diploid organisms consists of paired chromosomes. The haplotype information (DNA constellation on one single chromosome), which is crucial for disease association analysis and population genetic inference among many others, is however hidden in the data generated for diploid organisms (including human) by modern high-throughput technologies which cannot distinguish information from two homologous chromosomes. Here, I consider the haplotype inference problem in two common scenarios of genetic studies:

1. Model organisms (such as laboratory mice): Individuals are bred through prescribed pedigree design.

2. Out-bred organisms (such as human): Individuals (mostly unrelated) are drawn from one or more populations or continental groups.

In the two scenarios, one individual may share short blocks of chromosomes with other individual(s) or with founder(s) if available. I have developed and implemented methods, by identifying the shared blocks statistically, to accurately and more rapidly reconstruct the haplotypes for individuals under study and to solve important related problems including genotype imputation and ancestry inference. My methods, based on hidden Markov model, can scale up to tens of thousands of individuals. Analysis

based on my method leads to a new genetic map in mouse population which reveals important biological properties of the recombination process. I have also explored the study design and empirical quality control for imputation tasks with large scale datasets from admixed population.

To my parents and my wife.

Acknowledgements

First of all I would like to express my sincere thanks to my advisors, Drs. Wei Wang and Yun Li, for their continuous guidance and support, for being approachable anytime I had a problem, for explaining to me patiently even when I was in the "memoryless" state, and for giving me much freedom in working (and playing).

I had been very lucky to have chances to explore several different areas. I feel especially fortunate to have worked with Drs. Fernando Pardo Manuel de Villena and Vladimir Jojic. Fernando has patiently taught me many basics of biology and helped me in linking computational methods to real biology problems; Vladimir introduced me to many optimization techniques and always inspired me through thoughtful questions. My thanks also go to other research collaborators and committee members for helpful discussions on research and on completing my dissertation, William Valdar, Ethan Lange, Gary Churchill, Leonard McMillan, Xiang Zhang and all students (current and past) in the CompGen group and in the Li lab. I am very grateful to Qi Zhang for mentoring and helping me in many ways during my first year and my internships, to Zhaojun Zhang and Qing Duan for carrying out research together.

Finally, I would like to thank my parents, Yongjian and Chengcui, for their endless support, and for buying me my first computer with all their savings 20 years ago. I am also deeply thankful to my wife, Ping, for trust in me, and for following me across the ocean to every place we have lived in and to the next place we are heading for.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Recent technological advances in life sciences have generated massive amounts of data which enables accurate analyses of genome ancestry, recombination properties, complex disease susceptibility, and drug response, among many others. However, it is often the haplotype information that is more powerful in such analyses than the data directly obtained from high-throughput technologies such as genotyping. Therefore, how to reconstruct haplotype information from massive amount of raw data and make related inference based on recovered haplotype information are key problems in genetic studies and pose serious computational challenge.

In this thesis, I have developed statistical methods and computational tools that, by reconstructing haplotype information, generate accurate inferences for important problems including genome ancestry and imputation. My methods, based on Hidden Markov Model (HMM), can efficiently handle large scale datasets from two common settings in modern genetic studies:

1. Model organisms (such as laboratory mice): Individuals are bred through prescribed pedigree design.

2. Out-bred organisms (such as human): Individuals (mostly unrelated) are drawn from one or more populations or continental groups.

1.1 Background

1.1.1 DNA and Haplotype

Diploid species, which include nearly all mammals, carry paired homologous chromosomes, one inherited from each parent. A haplotype refers to the DNA sequence data from one of the paired chromosomes. Within the same species, DNA sequences are largely identical differing only slightly among individuals. Thus haplotypes are often defined only at positions with sequence variations. Figure 1.1 shows a toy example of two chromosomes with 15 sites and the two haplotypes defined at sites with variations.

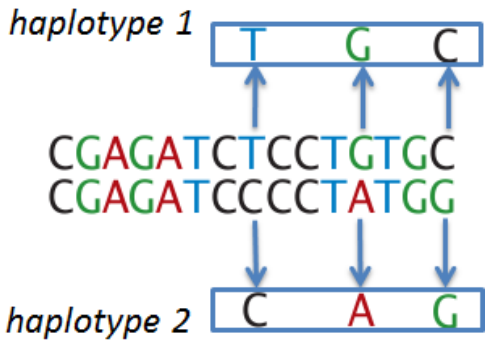


Figure 1.1: Toy example of two chromosomes with haplotypes defined on three sites containing variations

Haplotype knowledge describes how genetic materials are inherited from generation to generation. It thus provides direct knowledge of genome ancestry and historical recombination events. Furthermore, utilizing haplotype sharing information, one can fill in missing genotypes (imputation) [Li *et al.*, 2009]. Haplotypes are also important to many other fundamental problems in genetics. To name a few: (1) linkage analysis and linkage disequilibrium patterns [Stephens *et al.*, 2001; Wall *et al.*, 2003]; (2) mapping complex traits and diseases [Johnson *et al.*, 2001; Altshuler *et al.*, 2008]; (3) selection, evolution and historical migration in population genetics [Sabeti *et al.*, 2002; Merriwether *et al.*,

1995]. In these problems, even if reconstructed with uncertainty, haplotype information could lead to significantly increased power in inferences.

Even though it is possible to obtain haplotype information of diploid organisms directly from biological experiments, it is generally expensive and cannot scale to large sample size. On the contrary, modern high-throughput genotyping technologies can generate accurate genotype readings on hundreds of thousands of markers at much lower cost. It is thus valuable to conduct analysis by reconstructing haplotypes based on genotype inputs.

1.1.2 Genotype

Modern high-throughput genotyping technologies generate genotype readings on a preselected set of genetic markers. The set of markers can be defined by standard commercial platforms (e.g., Affymatrix 6.0, Illumina 1M), or customized by researchers (e.g., Yang *et al.* [2009]). Each genotype reading, or simply genotype, is an unordered combination of two alleles from paired chromosomes. In other words, genotypes are unable to distinguish between the two haplotypes of a diploid organism. It cannot tell which allele is from which haplotype.

In this dissertation, I consider how to bridge the gap between genotype data and desired genetic analyses by reconstructing haplotypes probabilistically. Here, I consider two common settings in genetic studies and related inference problems specific to settings.

1.2 Model Organisms from Prescribed Breeding

Model organisms, such as laboratory mice, are frequently *bred* or *crossed* in order to study genetic influences [Churchill *et al.*, 2004; Valdar *et al.*, 2006; Chia *et al.*, 2005]. Often, organism resources are generated using prescribed breeding system to ensure diversity and reproducibility, which leads to complex pedigree structure consisting of many generations. Through recombination, DNA sequences of founder organisms are inter-

mixed in each generation. A DNA sequence of any descendant organism is a mosaic of its founders' DNA segments.

One example of such resources is the international Collaborative Cross (CC) project which is a major effort in the mouse research community and has been under development for more than 10 years [Threadgill and Churchill, 2012]. The CC project consists of hundreds of independently bred, recombinant-inbred mouse lines generated through a funnel breeding design (Figure 2.3). Each line has more than 20 expected generations. High-density genotype data of the CC resources not only provide opportunities for fine-resolution quantitative trait locus (QTL) studies, but also facilitate exciting new research areas such as the inference of genetic networks underlying phenotypic traits in mammals.

Among many analyses of interest, a core problem is to discover the founder attribution to genomes in subsequent generations. That is to say, given a descendant organism in the resource, I want to find out which part of its DNA sequences is inherited from which founder (genome ancestry in founders). The genome ancestry information provides direct knowledge of historical recombination events and opportunities for error detection and imputation. It also enables downstream analyses such as measuring strain effect in quantitative traits.

Inference of genome ancestry involves resolving the potential inheritance flow at all markers of interest. This naturally requires the resolution of haplotype information as haplotypes correspond to the variants inherited together in the breeding process. It is straightforward to show that, in a pedigree with n non-founders and m markers of interest, there are 2^{mn} possible inheritance configurations even if one assumes known founder haplotypes and only bi-allelic markers. In a typical CC pedigree, there could be more than 40 mice and the enormous search space presents a major computational challenge.

The commonly favored pedigree-based haplotyping methods [Kruglyak *et al.*, 1996; Abecasis *et al.*, 2001; Gudbjartsson *et al.*, 2005] are all based on the Lander-Green algorithm (Lander and Green, 1987) as the running time is linear to the number of markers

which far exceeds other parameters. However, these methods are limited to pedigrees of moderate size since the running time grows exponentially with pedigree size. When they are applied to the genotype data from CC, the search space becomes extraordinarily large due to the large pedigree structure with many untyped intermediate generations. Other pedigree-based haplotyping methods include MCMC sampling methods [Sobel and Lange, 1996; Jensen and Kong, 1999], whose computing time can be substantial when applied to a large number of tightly linked markers, and rule-based methods [Qian and Beckmann, 2002; Li and Jiang, 2005], which have a crude approximation by minimizing recombinations in pedigree. More computationally efficient approaches for solving the genome ancestry problem have ignored pedigree information, including the breeding scheme. Examples include the combinatorial optimization approach by Zhang *et al.* [2008] and the HMM-based method in HAPPY [Valdar *et al.*, 2006; Mott *et al.*, 2000], a QTL mapping tool suite for association studies. All ancestry compositions are considered possible in the two methods. While breeding design does not determine the locations of recombination, it places important constraints on the possible ancestry choices at a single marker and at neighboring markers. Therefore, incorporating breeding design information would lead to more accurate inference.

1.3 Samples from Out-bred Human Populations

The ultimate goal of almost all genetic research is to understand genetic mechanisms in humans. Therefore, tremendous efforts have been spent on investigating human samples directly. In contrast to model organisms where breeding is often designed and controlled, humans are out-bred and the genetic data of founders are generally unavailable. Since Risch and Merikangas [1996] showed that association studies are more powerful than linkage studies, genetic data collected for humans in the past one and a half decades are largely from unrelated individuals. The consequence of out-breeding, lack of founder

genetic information, and use of unrelated individuals is that individuals studied tend to share only short haplotype segments (e.g., several hundred Kbs) of their chromosomes. This is further confounded by the presence of population and sub-population structure. Reconstruction of haplotype in such out-bred populations is therefore challenging but of great importance in genetic studies.

By aligning samples under study to samples in existing studies (e.g., HapMap and 1000 Genomes projects [The International HapMap Consortium, 2010; The 1000 Genomes Project Consortium, 2012]), researchers can identify the shared haplotype segments among samples. Consequently, one can not only recover the sporadic technological failures in genotypes, but also impute the markers that are untyped in individual studies but typed in reference samples. This genotype imputation technique greatly improves the marker density and analysis power of individual studies.

Moreover, as the typical small to moderate effect of individual genetic variant on complex trait entails large sample size, collaborative efforts that pool information across multiple studies are typically taken to enhance the statistical power for detecting causal variants. In these collaborative efforts, samples from different studies are typically genotyped at different sets of markers because different commercially available genotyping platforms are used. The commonly used genotyping platforms have a small fraction of markers in common ($\sim$10% is typical between platforms from two different companies). Restricting analysis to markers in common leads to much reduced marker density and huge loss of information. Imputation of markers untyped in individual studies greatly facilitates the integration of samples across studies (meta-analysis) .

Several HMM-based imputation methods [Li *et al.*, 2010a; Howie *et al.*, 2009; Browning and Browning, 2009] have previously been developed by reconstructing the haplotypes and shown to achieve good imputation performance in a number of populations [Pei *et al.*, 2008; Huang *et al.*, 2009], particularly those with high level of linkage disequilibrium (LD) or having closely matched reference population(s) from the HapMap

or the 1000 Genomes Projects [The International HapMap Consortium, 2010; The 1000 Genomes Project Consortium, 2010]. The wealth of literature using genotype imputation has focused on using external reference panels (for example, phased haplotypes from the HapMap and 1000 Genomes projects), largely in individuals of European ancestry, for inference of genotypes at common (minor allele frequency [MAF] > 0.05) genetic markers. Several important issues have not been adequately addressed including the utility of study-specific reference, accommodation of increasingly large reference panels, performance in admixed populations, and quality for less common (MAF ~ 0.005-0.05) and rare (MAF < 0.005) variants. These issues only recently became addressable with Genome-Wide Association (GWA) follow-up studies using dense genotyping or sequencing in large samples of non-European individuals.

Also, little methodological work exists for imputation in admixed populations, such as African Americans and Hispanic Americans, which comprise more than 20% of the US population. Admixed populations offer a unique opportunity for gene mapping, but also impose challenges for imputation. To efficiently benefit from emerging large reference panels, one key issue to consider is on how to traverse the reference space harboring the most probability mass with minimum computational efforts. In modern genotype imputation framework, this corresponds to the selection of effective reference panels. Existing works often focused on constructing a **pre-defined** reference panel **prior to** running the imputation engine. Such methods (e.g., a cosmopolitan panel [Hao *et al.*, 2009; Li *et al.*, 2009; Shriner *et al.*, 2010] or a weighted combination panel [Egyud *et al.*, 2009; Huang *et al.*, 2009; Pasaniuc *et al.*, 2010; Pemberton *et al.*, 2008]) have limited flexibility and aggravate the already heavy computation burden. Another approach, based on whole-haplotype closeness heuristics, has been adopted by IMPUTE2 [Howie *et al.*, 2009] and can be embedded within other existing imputation models. The above-mentioned methods have shown promising results but have not been evaluated systematically. In addition, both categories of methods can be further improved statistically and compu-

tationally, for example, through integration of the former approach within (rather than prior to) the hidden Markov model, or through more elegant heuristics.

1.4 Thesis Statement

Genetic analyses of model organism resources and out-bred populations can be achieved by reconstructing haplotype information implicitly or explicitly via HMM. By applying effective state-space pruning strategies, I present haplotype-based inference algorithms that can scale to large datasets without compromising accuracy. Application to CC mouse data leads to new biological discovery of properties of recombination events. Case study on Women's Health Initiative (WHI) metabochip data leads to generalizable quality control guidelines for imputation analysis.

1.5 Contributions

In this section, I briefly summarize the contributions presented in subsequent chapters.

1.5.1 Model Organisms from Prescribed Breeding

- In Chapter 2, I propose a method, GAIN, to infer genome ancestry in organism resources. The method can efficiently handle complex pedigrees with inbreeding which is an important process in generating organism resources. Using a pair of dependent quaternary indicators to capture all recombinations in the inbreeding history, my method achieves accurate ancestry inference without the need to explicitly model every intermediate generation. By encoding the inbreeding model into the inheritance vectors, I design a Lander-Green-like algorithm whose running time remains constant with respect to the number of inbreeding generations. GAIN is implemented and evaluated on the CC high-density *single-nucleotide polymor-*

phism (SNP) data with complex breeding design. Experiments show that, GAIN generates accurate results efficiently on data that cannot be handled by existing pedigree haplotyping software. Compared with HAPPY [Mott *et al.*, 2000], which does not model pedigree structure, GAIN substantially reduces ambiguities in ancestry inference.

- In Chapter 3, I generate a new linkage map of the laboratory mouse genome using GAIN described in previous chapter. The map is built with the recombination and ancestry information inferred from the genotypes of 237 male-female sibling pairs. Exploiting the large number of recombination events (n∼22,000), the high precision in mapping each event (∼35kb) and the unique characteristics of the CC mice, I provide a new and powerful look at the effects of sex, strain and genotypes at polymorphic loci of interest (e.g., the *Prdm9* gene) on recombination. In addition to an extended catalog of sex and strain specific hotspots, I report the presence of cold regions for recombination with striking distributions and genomic characteristics.

1.5.2 Samples from Out-bred Human Populations

- In Chapter 4, I propose and evaluate a number of methods for effective reference panel construction to improve haplotype-based imputation engines. Using a novel piecewise IBS method, my software package MaCH-Admix yields consistently higher imputation quality than existing methods/software. I evaluated the performance on individuals from recently admixed populations, including 8421 African Americans and 3587 Hispanic Americans from the Women's Health Initiative (WHI), which allow assessment of imputation quality for uncommon variants. The advantage is particularly noteworthy among uncommon variants where up to 5.1% information gain is observed with the difference being highly significant (Wilcoxon signed rank test P-value < 0.0001). This work is the first that considers

various sensible approaches for imputation in admixed populations and presents a comprehensive comparison.

- In Chapter 5, I present a case study of imputation in a large cohort of African Americans from the Women's Health Initiative (WHI) study. This study presents three under-studied aspects: (1) imputation of markers from a region-centric platform that are largely of low frequency; (2) imputation using a study-specific reference panel; and (3) imputation in admixed population. In this study, I describe a pipeline for constructing study-specific reference panels using individuals genotyped or sequenced at a larger set of genetic markers and for imputation into individuals with genotype data at a subset of markers. I demonstrate several approaches to reliably estimate imputation quality for SNPs in different MAF categories. Experiment results suggest that imputation of region-centric SNPs, including low frequency SNPs with MAF 0.005-0.05, is feasible and well worthwhile for power increase in downstream association analysis. I further provide practical guidelines regarding post-imputation quality control.

Chapter 2

Efficient Genome Ancestry Inference in Complex Pedigrees with Inbreeding

2.1 Introduction

Model organisms, such as laboratory mice, are frequently *bred* or *crossed* in order to study genetic influences [Churchill *et al.*, 2004; Valdar *et al.*, 2006; Chia *et al.*, 2005]. Often, such animal resources are generated using prescribed breeding system to ensure diversity and reproducibility, which leads to complex pedigree structure consisting of many generations. Through recombination, the DNA sequences of founder organisms are intermixed in each generation. A DNA sequence of any descendant organism is a mosaic of its founders' DNA segments. As recombinations at each breeding stage cannot be observed directly, it is of great interest to infer the ancestry of resulting DNA sequences. In other words, which part of a resulting DNA sequence is inherited from which founder.

The vast majority of the sequence variations are attributed to single base-pair mutations known as *single-nucleotide polymorphism* (SNPs), thus making SNPs ideal for resolving the genome ancestry problem. The set of SNPs on the same chromosome constitutes a *haplotype*. While any of the four nucleotides (A,T,C,G) is possible, in practice nearly all SNPs appear in only two variations. This results from the fact that SNPs

originate as mutations, which are rare events within a vast genome. It is therefore convenient to encode a SNP allele as a binary value and represent haplotypes as binary sequences. Modern high-throughput genotyping technologies are unable to distinguish between the two haplotypes of a diploid organism. Instead, a *genotype sequence* is measured where, at each SNP site, one of three possibilities is observed ($\{00, 01, 11\}$, since 10 cannot be distinguished from 01).

Using the genotype representation for DNA sequences, the genome ancestry problem estimates the origin of each genotype from a descendant's sequence given the genotype sequences of its distant founders. To achieve high resolution, dense SNP markers are used (tens of thousands on each chromosome). Knowledge of genotype's ancestry is particularly useful in many problems such as studying the structure and history of haplotype blocks [Gabriel *et al.*, 2002; Zhang *et al.*, 2002; Schwartz *et al.*, 2004], and mapping quantitative trait loci (QTLs)[Valdar *et al.*, 2006; Mott *et al.*, 2000]. In these studies, a probabilistic interpretation is favored over discrete solutions, due to the prevalence of ambiguities and measurement errors.

The genome ancestry problem is closely related to haplotype inference with pedigree data. Inferring haplotypes in a pedigree often involves solving the inheritance flow of alleles at each generation. On the other hand, given the genome ancestry information, it is straightforward to reconstruct the descendant haplotypes. As pedigree analysis is NP-hard [Piccolboni and Gusfield, 2003], existing algorithms are either approximate or suffer exponential running times. Among the maximum likelihood approaches, methods [Kruglyak *et al.*, 1996; Abecasis *et al.*, 2001; Gudbjartsson *et al.*, 2005] based on the Lander-Green algorithm [Lander and Green, 1987] are often favored because their running time is linear to the number of markers. MERLIN [Abecasis *et al.*, 2001], an implementation based on sparse binary trees, is one of the most successful pedigree analysis programs. Unfortunately, methods based on Lander-Green algorithms are limited to pedigrees of moderate size since the running time grows exponentially with pedigree

size. MCMC sampling methods [Sobel and Lange, 1996; Jensen and Kong, 1999] have been proposed to address larger pedigrees. But their computing time can be substantial when applied to a large number of tightly linked markers. Other efforts include rule-based methods [Qian and Beckmann, 2002; Li and Jiang, 2005] which approximates a solution by minimizing recombinations in the pedigree (MRHC). PedPhase [Li and Jiang, 2005], which employs an effective integer linear programming (ILP) formulation, has been widely used in solving the MRHC.

Current haplotyping methods for pedigrees are incapable of solving the genome ancestry problem in animal resources for the following reasons: 1) Pedigrees of model animal resources often contain large number of generations to ensure diversity and reproducibility. 2) None or few of the intermediate generations are genotyped due to the size of the resources. 3) A large number of dense markers are genotyped to achieve fine resolution. As a concrete example, more than one thousand lines have been started in the Collaborative Cross project [Churchill *et al.*, 2004; The Collaborative Cross Consortium, 2012]. Each line is expected to undergo at least 23 generations before reaching 99% inbred. Hundreds of mice of various generations were genotyped, but on average only few are from the same line. The missing genotypes make the search space extraordinarily large.

Other computationally efficient approaches for solving the genome ancestry problem have largely ignored the breeding scheme. While breeding design does not determine the locations of recombination, it often places constraints on the possible ancestry choices at a single site and at neighboring sites. The genome ancestry problem was modeled as a combinatorial optimization problem in [Zhang *et al.*, 2008]. By minimizing recombinations, discrete solutions are generated. Mott et al. has proposed an approach using Hidden Markov Model (HMM) for ancestry inference in HAPPY [Valdar *et al.*, 2006; Mott *et al.*, 2000], a QTL mapping tool suite for association studies. All founder pairs are considered as possible hidden states for emitting the observed genotype at each site. Besides founder genotypes, no pedigree data are used in these two approaches.

There have also been many efforts to analyze pedigree by identifying symmetries in HMM state space [Donnelly, 1983; McPeek, 2002; Browning and Browning, 2002; Geiger *et al.*, 2009]. The states are then grouped to accelerate the calculation. However, finding the maximal grouping is non-trivial. In real-world problems, only obvious symmetries such as founder phase and chain structure in pedigree can be best utilized.

Besides model organisms, the genetic ancestry problem has been studied for human individuals that have recently been admixed from a set of isolated populations, instead of a set of founders[Tang *et al.*, 2006; Sundquist *et al.*, 2008; Sankararaman *et al.*, 2008; Paşaniuc *et al.*, 2009]. In this problem, pedigree structure is usually not present (unrelated individuals) or the size of pedigree is small. Efficient methods have been developed to handle large-scale datasets[Tang *et al.*, 2006; Sundquist *et al.*, 2008; Sankararaman *et al.*, 2008].

Leveraging the observation that large animal resource pedigrees often contain repetitive sub-structures, I propose a method that can efficiently handle complex pedigrees with inbreeding which is an important process in generating animal resources. Using a pair of dependent quaternary indicators to capture all recombinations in the inbreeding history, my method achieves accurate ancestry inference without explicit modeling every generation. By encoding the inbreeding model into the inheritance vectors, I design a Lander-Green-like algorithm whose running time remains constant with respect to the number of inbreeding generations. My method is implemented and evaluated on the Collaborative Cross breeding design [Chesler *et al.*, 2008; The Collaborative Cross Consortium, 2012] with dense SNP data. Experiments show that, my approach generates accurate results efficiently on data that cannot be handled by existing pedigree haplotyping software. Compared with HAPPY, which does not consider pedigree structure, my approach significantly reduces ambiguities and errors in ancestry inference.

2.2 The Genome Ancestry Problem

Given a pair of chromosomes, consider L SNP markers ordered by their chromosomal locations. For each SNP site, we use 0 and 1 to encode the two possible values. The genotype at each site is the unordered combination of corresponding alleles from both chromosomes, which can assume one of three values: 00, 01, 11. A genotype sequence is a genome-ordered set of genotypes denoted as: $G = g_1...g_l...g_L, (g_l \in \{00, 01, 11\})$. A haplotype $H = h_1...h_l...h_L$ consists of alleles from one of the chromosomes where $h_l \in \{0, 1\}$.

Consider a pedigree containing a set of founders $FS = \{F1, ..., F_N\}$ and a descendant of interest. I denote the set of founder genotype sequences by $\{G_{F1}, ..., G_{F_N}\}$, all of which are given. Given the genotype sequence, G_D, of the descendant generated through the pedigree structure, its genome ancestry is to be determined. Every genotype g_l in G_D inherits its alleles from two founders, say F_A and F_B. I refer to the founder pair (F_A, F_B) as the genome ancestry at site l of genotype sequence G_D. I want to estimate, for every SNP site l, the probability $P(Ancestry(g_l) = (F_A, F_B))$ for every founder pair $(F_A, F_B) \in FS \times FS$. Note that founder pairs are unordered ($(F_A, F_B) = (F_B, F_A)$), and it is possible that $F_A = F_B$.

2.3 Modeling Inheritance in Pedigree

I start from the standard Lander-Green approach to model a pedigree: At each SNP site, an inheritance indicator is used to indicate the outcome of each meiosis. These inheritance indicators together form the inheritance vector. Since a child haplotype inherits its allele from either the paternal or maternal sequence, an inheritance indicator is a binary variable. For a pedigree with n non-founder animals, there are $2 \times n$ inheritance indicators at each site. Hence, the inheritance vector at site l, v_l, can be defined as a binary sequence of length $2 \times n$. An instance of v_l specifies a possible configuration of

inheritance flow at site l of all animals in the pedigree. When SNP markers are dense enough, one can assume at most one recombination between two sites in generating one haplotype. If a recombination happens between site l and $l+1$, the corresponding inheritance indicator will have different states for the two sites. Hence, to measure the number of recombinations between l and $l+1$ in the whole pedigree, one can count the difference in bits between v_l and v_{l+1}. The probability of having d recombinations between l and $l+1$ is $\theta^d(1-\theta)^{2n-d}$, where θ is the recombination fraction.

The length of inheritance vector grows linearly with the number of animals in the pedigree and this causes exponential growth in the number of possible inheritance patterns. Considering the fact that full pedigree analysis is computationally intractable, I overcome the issue by modeling important sub-structure in breeding systems as a shortcut to efficient computation. My first natural choice of sub-structure is inbreeding: 1) Inbreeding is often used in model animal resources to generate genetically diverse and/or reproducible descendants. 2) Inbreeding is often carried out for many generations and each generation elongates the inheritance vectors by 4 bits. Hence, if a pedigree involves inbreeding, the inbreeding generations often account for most of the computational complexity. I seek an aggregated inheritance indicator to replace the collection of many inheritance indicators in the inbreeding process. Such an aggregated indicator can be encoded in much shorter length and incorporated into the inheritance vector. If the state and transition probability of the aggregated indicator can be modeled efficiently, full pedigree analysis will become feasible on these animal resources. In the next section, I explain how inheritance in inbreeding generations can be modeled as an aggregated indicator.

2.3.1 Modeling Inbreeding Generations

During inbreeding, offspring are produced by sibling matings for many generations. At each generation, four new haplotypes are formed by recombining the four haplotypes from the previous generation. The inbreeding process at a single site is shown in Figure

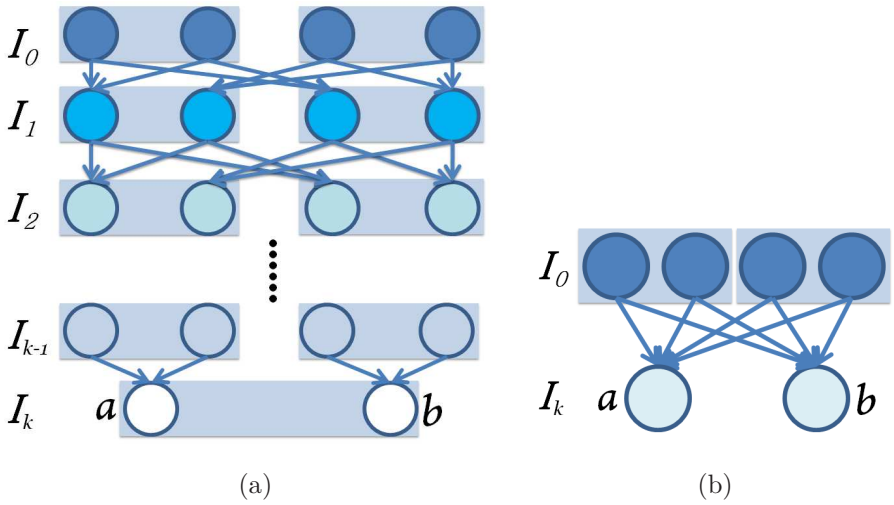


Figure 2.1: (a) Lattice of binary inheritance indicators representing the inheritance pattern of an inbreeding process at a single site. (b) An equivalent quaternary indicator representation

2.1(a). I denote the beginning generation of inbreeding as generation I_0. Observe that, at each site, because of the symmetry of inbreeding structure, the four alleles at generation I_0 have equal probabilities to be passed down to any haplotypes after I_1. Thus, for a descendant haplotype at generation I_k ($k > 2$), I can simply replace the lattice of binary inheritance indicators by a single quaternary indicator. Each choice of the quaternary indicator has 1/4 probability. Two quaternary indicators are needed for the two haplotypes of a I_k descendant (Figure 2.1(b)). However, the two quaternary indicators are not independent as the two haplotypes share the same inbreeding history until I_{k-1}. To model this dependency between the two quaternary indicators, I find out the transition events and probabilities of the pair of indicators. The grouped pair is then used as an aggregated inheritance indicator as discussed above.

I label the four I_0 haplotypes as $1, 2, 3, 4$. I then denote by a, b the two I_k descendant haplotypes and $S(a_l), S(b_l)$ are their I_0 sources at site l, i.e., $S(a_l), S(b_l) \in \{1, 2, 3, 4\}$. Their I_0 sources along the chromosome is denoted by $S(a), S(b) \in \{1, 2, 3, 4\}^L$. A transition happens in $S(a)$ between site l and $l + 1$ if $S(a_l) \neq S(a_{l+1})$. I consider, between two

adjacent sites, l and $l+1$, all the possible transitions from $S(a_l), S(b_l)$ to $S(a_{l+1}), S(b_{l+1})$ (Table 2.1).

Note that:

$$P_{EE0} + P_{EN1} + P_{EE2} + P_{EN2} = P(S(a_l) = S(b_l)) =$$

$$P_{EE0} + P_{EE2} + P_{NE1} + P_{NE2} = P(S(a_{l+1}) = S(b_{l+1}))$$

and

$$P_{NE1} + P_{NN0} + P_{NN1} + P_{NN2} + P_{NE2} = P(S(a_l) \neq S(b_l)) =$$

$$P_{EN1} + P_{EN2} + P_{NN0} + P_{NN1} + P_{NN2} = P(S(a_{l+1}) \neq S(b_{l+1}))$$

The prior probability $P(S(a_l) = S(b_l))$ at any site l is called the *inbreeding coefficient* [Wright, 1922]. To calculate the probability, let IC_k denote the inbreeding coefficient at generation I_k. IC_k can be computed recursively using $IC_k = \sum_{j=0}^{k-2} (\frac{1}{2})^{k-j} \times (1 + IC_j)$.

Site l	Possible Transitions	Site $l+1$	Denote By
$S(a_l) = S(b_l)$	Neither $S(a)$ or $S(b)$ transitions.	$S(a_{l+1}) = S(b_{l+1})$	P_{EE0}
	Either $S(a)$ or $S(b)$ transitions, but not both.	$S(a_{l+1}) \neq S(b_{l+1})$	P_{EN1}
	Both $S(a)$ and $S(b)$ transition to same value.	$S(a_{l+1}) = S(b_{l+1})$	P_{EE2}
	Both $S(a)$ and $S(b)$ transition, but to different values.	$S(a_{l+1}) \neq S(b_{l+1})$	P_{EN2}
$S(a_l) \neq S(b_l)$	Neither $S(a)$ nor $S(b)$ transitions.	$S(a_{l+1}) \neq S(b_{l+1})$	P_{NN0}
	Either $S(a)$ or $S(b)$ transitions, but not both. $S(a)$ and $S(b)$ become equal after the transition.	$S(a_{l+1}) = S(b_{l+1})$	P_{NE1}
	Either $S(a)$ or $S(b)$ transitions, but not both. $S(a)$ and $S(b)$ remain different after the transition.	$S(a_{l+1}) \neq S(b_{l+1})$	P_{NN1}
	Both $S(a)$ and $S(b)$ transition. $S(a)$ and $S(b)$ remain different after the transition.	$S(a_{l+1}) \neq S(b_{l+1})$	P_{NN2}
	Both $S(a)$ and $S(b)$ transition. $S(a)$ and $S(b)$ become the same after the transition.	$S(a_{l+1}) \neq S(b_{l+1})$	P_{NE2}

Table 2.1: All possible transitions of $S(a), S(b)$. Each type of transition is denoted by 3 characters. First two letters indicate the equality of $S(a), S(b)$ before and after the transition. Then followed by a digit indicating the number of transitions in $S(a), S(b)$.

Next, I derive the probabilities in Table 2.1. Consider that any transition in $S(a)$ or $S(b)$ is caused by one or more recombinations in the inbreeding process (Figure 2.1(a)). My calculation is based on the assumption that the recombination fraction, θ, is reasonably small. Hence, for any haplotype c at generation I_j $(1 \leq j \leq k)$, I assume that any single transition in $S(c)$ is solely caused by one recombination in generating c or its ancestor haplotypes. In other words, a single transition in $S(c)$ is not the result of multiple recombinations in the pedigree. My assumption is generally true for dense SNP markers where θ is usually well below 0.001. Under the assumption, if a transition in $S(c)$ is caused by a recombination in generating c itself, I define this to be a *lead transition*. Intuitively, a lead transition is one not inherited from its ancestors. A lead transition in c will change the I_0 source of c and all descendant haplotypes inheriting the transition. A lead transition is only possible when the two parental haplotypes of c have different I_0 sources. Hence, between two sites, a haplotype at generation j has a lead transition with probability $\theta \times (1 - IC_{j-1})$.

With the inbreeding coefficients calculated, I can derive the marginal probability of observing transition in one of the I_k haplotypes, $P_{1T} = P(S(a_l) \neq S(a_{l+1})) = P(S(b_l) \neq S(b_{l+1}))$. Without loss of generality, I consider $P(S(a_l) \neq S(a_{l+1}))$ for haplotype a. $S(a)$ will transition if a itself or any of its ancestor haplotypes has a lead transition. At generation k, the lead transition happens with probability $\theta \times (1 - IC_{k-1})$. For generation $k-1$, there are 2 possible ancestor haplotypes, each with $\frac{1}{2}\theta \times (1 - IC_{k-2})$ chance of causing a transition in $S(a)$. For each generation j from 1 to $k-2$, there are 4 possible ancestor haplotypes with probability $\frac{1}{4}\theta \times (1 - IC_{j-1})$. Consider that, at one site, any two haplotypes from the same generation cannot both be the ancestor of a. Thus, for any generation j, the expected probability of causing transition in $S(a)$ is $\theta \times (1 - IC_{j-1})$. Under my assumption, $P(S(a_l) \neq S(a_{l+1}))$ can be expressed by $1 - \prod_{j=1}^{k} (1 - \theta \times (1 - IC_{j-1}))$.

I then derive the probability P_{EE2} that $S(a)$ and $S(b)$ have equal state at site l, and both transition to another state at site $l+1$. This event happens only if a haplotype c at

some previous generation is the common ancestor of a, b and c has a lead transition. The probability of c at generation j being the common ancestor of a and b is $\frac{1}{4}IC_{k-j}$. The probability that c has a lead transition is $\theta \times (1 - IC_{j-1})$. Again, consider the fact that, at one site, any two haplotypes from the same generation cannot both be the common ancestor of a and b. Thus, the probability of EE2 event caused by lead transition at I_j ($1 \leq j \leq k-2$) is $\theta \times (1 - IC_{j-1})IC_{k-j}$. Assuming a small θ, P_{EE2} can be calculated by $1 - \prod_{j=1}^{k-2}(1 - \theta \times (1 - IC_{j-1})IC_{k-j})$.

Lastly I consider the probability P_{NN1}. To simplify my discussion, assume that the transition happens in $S(a)$ (i.e. $S(a_l) \neq S(a_{l+1})$) and it inherits a lead transition in haplotype c of generation j. Since $S(a_l)$, $S(a_{l+1})$ and $S(b_l)$ all have different I_0 ancestry, alleles from at least 3 distinct I_0 haplotypes should be observed at generation $j-1$. Let $P_{Distinct}(m, j)$ be the probability of observing exactly m distinct I_0 alleles at generation j. $P_{Distinct}(3, j)$ and $P_{Distinct}(4, j)$ can be computed recursively using:

$$P_{Distinct}(4, j) = \frac{1}{4}P_{Distinct}(4, j-1)$$

$$P_{Distinct}(3, j) = \frac{1}{2}P_{Distinct}(3, j-1) + \frac{1}{2}P_{Distinct}(4, j-1)$$

Then, P_{NN1} is the probability that (1) at least 3 distinct I_0 alleles are present at generation $j-1$ and (2) a's ancestor c at generation j has a lead transition between sites l and $l+1$ which is inherited by a (3) before and after transition, the I_0 source of c is different from that of b.

Under my assumption of a small θ, $P_{NN2}, P_{NE2}, P_{EN2}$ are all sufficiently small and can be ignored in calculating other probabilities. The intuition is as follows: if k is small, there are few animals in the inbreeding lattice and the chance of observing multiple transitions is rare; when k becomes larger, the probability $P(S(a_l) \neq S(b_l))$ approaches 0

rapidly and $P_{NN2}, P_{NE2}, P_{EN2}$ are much smaller than $P(S(a_l) \neq S(b_l))$. With P_{1T}, P_{EE2} and P_{NN1} derived, I can easily solve all the rest probabilities in Table 2.1:

$$P_{NE1} = P_{EN1} = \frac{1}{2}(2 \times (P_{1T} - P_{EE2}) - P_{NN1})$$

$$P_{EE0} = IC_k - P_{EE2} - P_{EN1}$$

$$P_{NN0} = 1 - IC_k - P_{NE1} - P_{NN1}$$

$P_{NN2}, P_{NE2}, P_{EN2}$ are approximated by a small probability $P_{NE1} \times P_{NE1}$. I use simulation to validate the probabilities derived above. The results are shown in Figure 2.2. For θ around 0.01, my method gives reasonably close approximation. For θ below 0.001, my method is very accurate. The recombination fraction between dense SNP markers is usually well below 0.001.

So far I have derived all event probabilities in Table 2.1. The transition probability from $(S(a_l), S(b_l))$ to $(S(a_{l+1}), S(b_{l+1}))$ is the corresponding probability in Table 2.1 conditioned on $P(S(a_l) = S(b_l))$ or $P(S(a_l) \neq S(b_l))$.

2.3.2 Integrating the Inbreeding Model

I have argued that each inbreeding process can be modeled by two quaternary indicators and their transition probabilities can be accurately approximated when θ is small. It is then straightforward to integrate the inbreeding model into the original Lander-Green model. I encode the two quaternary indicators using 4 binary bits in the inheritance vector. Consider a pedigree containing i inbreeding processes and n' other members not involved in inbreeding. The inheritance vector v_l at every site l now has length $2 \times n' + 4 \times i$. Each possible realization of v_l is a hidden state in HMM. The transition probability from v_l to v_{l+1} is the product of transition probabilities of all binary indicators and pairs of quaternary indicators. I can then solve the HMM using standard routine:

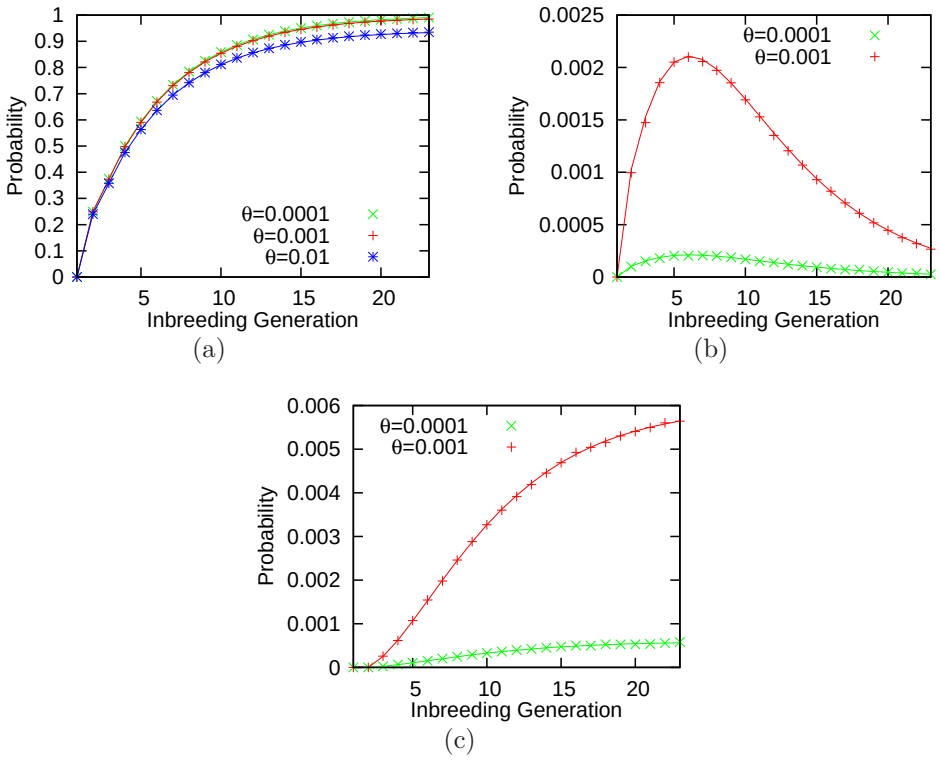


Figure 2.2: Comparison of predicted probabilities and observed probabilities from 10000000 simulations. The data points in the figures are observed probabilities from simulations. The curves are derived from my formulas. (a) Predicted and simulated P_{EE0} for $\theta = 0.01, 0.001, 0.0001$. (b) Predicted and simulated $P_{EN1} = P_{NE1}$ for $\theta = 0.001, 0.0001$. (c) Predicted and simulated P_{EE2} for $\theta = 0.001, 0.0001$. I do not plot the case of $\theta = 0.01$ in (b) and (c) because the values are much larger than that of the other two θ values.

$$
\begin{aligned}
P(v_l|G_D) &= \frac{P(G_D|v_l)P(v_l)}{P(G_D)} \\
&= \frac{P(g_1,...,g_l|v_l)P(g_{l+1},...,g_L|v_l)P(v_l)}{P(G_D)} \\
&= \frac{P(g_1,...,g_l,v_l)P(g_{l+1},...,g_L|v_l)}{P(G_D)} \\
&= \frac{\alpha(v_l)\beta(v_l)}{P(G_D)}
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha(v_l) &= P(g_1,...,g_l,v_l) \\
\beta(v_l) &= P(g_{l+1},...,g_L|v_l)
\end{aligned}
$$

$\alpha(v_l)$ and $\beta(v_l)$ can be solved recursively:

$$
\begin{aligned}
\alpha(v_{l+1}) &= \sum_{v_l} \alpha(v_l)P(v_{l+1}|v_l)P(g_{l+1}|v_{l+1}) \\
\beta(v_l) &= \sum_{v_{l+1}} \beta(v_{l+1})P(v_{l+1}|v_l)P(g_{l+1}|v_{l+1})
\end{aligned}
$$

$P(G_D)$ is obtained from the calculated $\alpha(v_l)$ and $\beta(v_l)$ at any site l:

$$
P(G_D) = \sum_{v_l} \alpha(v_l)\beta(v_l)
$$

The genome ancestry at site l is, for every founder pair (F_A, F_B),

$$
P(Ancestry(g_l) = (F_A, F_B)) = \sum_{v_l} P(v_l|G_D)
$$

for all v_l s.t. g_l is inherited from (F_A, F_B).

Note that, if I place the bits of quaternary indicators at the end of inheritance vector, the recursive calculation of α and β can still greatly benefit from the Elston-Idury algorithm [Idury and Elston, 1997].

2.4 Modeling the Collaborative Cross

The Collaborative Cross (CC) [Churchill *et al.*, 2004; Chesler *et al.*, 2008; The Collaborative Cross Consortium, 2012] is a large panel of reproducible, recombinant-inbred mouse lines proposed by the Complex Trait Consortium. Over a thousand of mouse lines have been started among which several hundred lines are kept inbreeding. All mouse lines are generated using eight genetically diverse founders via a common breeding scheme designed to randomize the genomic contribution of each founder. It provides an ideal platform for testing my approach.

2.4.1 The Breeding Scheme

CC mice are derived from 8 fully inbred founders using the 8-way funnel breeding scheme shown in Figure 2.3(a). The chromosomes of the eight founders (shown in different colors) are combined by two generations of crosses (labeled $G1$ and $G2I_0$), followed by at least 20 inbreeding generations ($G2I_1$ to $G2I_\infty$).

The positions of the 8 founders are not fixed. Permutations of the founders are used to randomize the genomes and balance the founder contributions to the resulting CC lines. This variation in initial positions imposes different ancestry constraints on each line. Without loss of generality, I assume a founder order of $F_1F_2F_3F_4F_5F_6F_7F_8$ as shown in Figure 2.3(a).

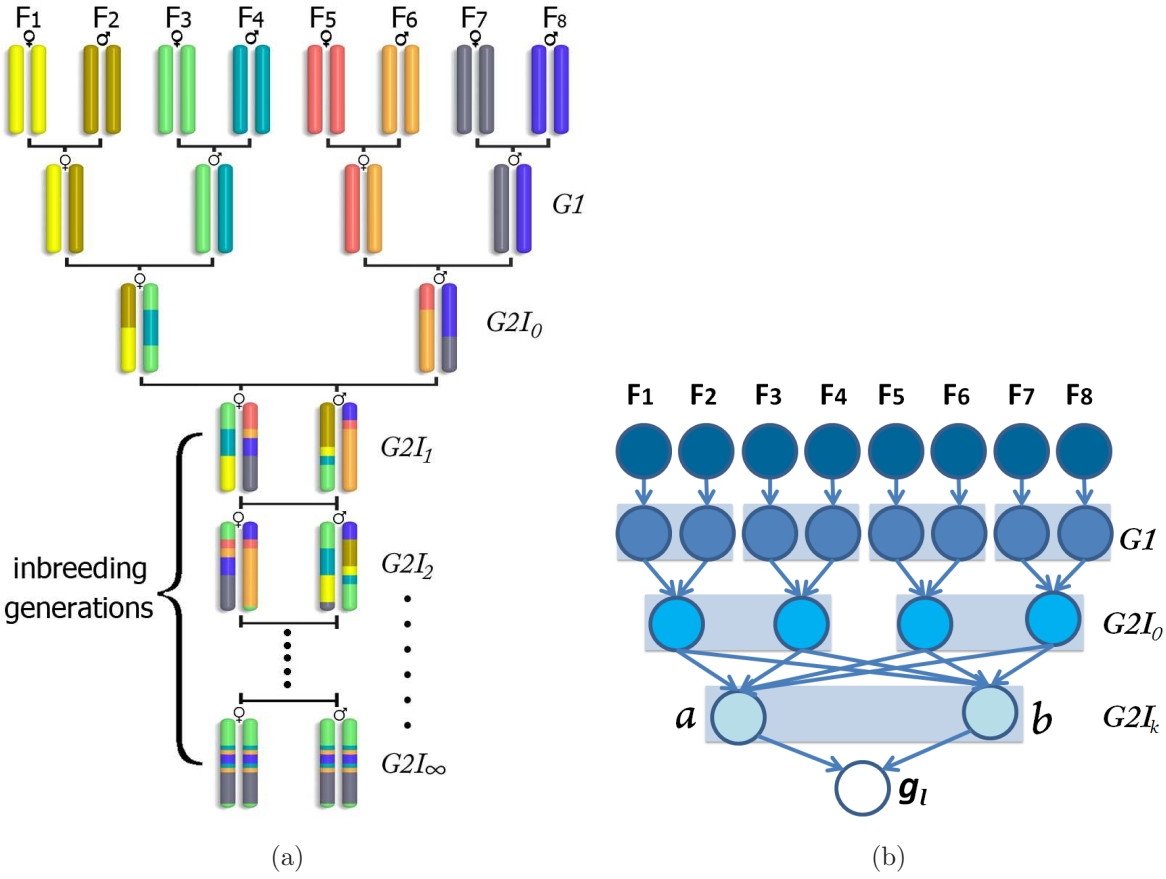


Figure 2.3: (a) Collaborative Cross breeding scheme: An example derivation of chromosomes by recombining chromosomes from 8 ordered founders. $G1$ and $G2I_0$ are two generations of crosses. $G2I_1$ to $G2I_\infty$ are multiple generations of inbreeding. (b) The inheritance indicators used to represent the inheritance flow at a SNP site.

2.4.2 Modeling the Genome of $G2I_k$ Generation

In a CC pedigree, any recombination in the formation of $G1$ haplotypes can be virtually ignored since all founders are fully inbred. Hence, at each SNP site, I only need 4 inheritance indicators for $G2I_0$ haplotypes and 2 quaternary indicators for the two haplotypes in a resulting $G2I_k$ descendant. The structure of the inheritance indicators is shown in Figure 2.3(b).

$G2I_1$ mice are an exception which only involve one generation of inbreeding. For a $G2I_1$ mouse, I simply let the two quaternary indicators revert back to binary indicators.

This becomes a standard Lander-Green model and it can be seen that the two $G2I_1$ haplotypes are restricted to be from the left and right half of the funnel respectively.

2.5 Experiments

In this section, I evaluate the proposed model on both simulated data and real CC genotype data. I implement my model GAIN (**G**enome **A**ncestry with **IN**breeding) for CC using C++. GAIN is compared with MERLIN [Abecasis *et al.*, 2001] and HAPPY [Mott *et al.*, 2000]. MERLIN is a widely used pedigree analysis software based on Lander-Green algorithm and can handle large number of markers. HAPPY is a QTL mapping tool suite and can analyze genome ancestry based on only founder and descendant genotype data, i.e., it ignores pedigree structure. Both software estimate the genome ancestry directly or indirectly.

2.5.1 Experiments on Simulated Data

As ground truth is generally unavailable for real data, I evaluate the accuracy of genome ancestry analysis using simulated data. I simulate the genotype of a $G2I_k$ mouse by recombining real CC founder haplotypes according to the CC pedigree structure. Given the founder genotypes, the founder haplotypes can be obtained trivially since all founders are fully inbred. At each generation I choose recombination position randomly. To simulate genotyping errors, I also introduce random errors to the resulting genotype sequence. When a site is selected to represent an error, I flip its value to heterozygous if it is homozygous originally. If a heterozygous site is selected, I change it to one of the homozygous state randomly. This resembles the fact that most genotyping errors are between heterozygous and homozygous states, instead of between the two homozygous states.

I simulate 20 test cases for each generation from $G2I_1$ to $G2I_{20}$. The number of markers ranges from 6 to 10 thousands. As MERLIN does not output probability distribution

for each inheritance vector, I first compare the best founder ancestry pair estimated by each method against the true answer. The error rate is measured by the percentage of sites where the estimated best founder ancestry does not match the ground truth. Figure 2.4 shows the error rate of all three methods in the simulated data with and without errors. Results of MERLIN are only available for the first 4 generations as the running time grows exponentially with the size of pedigree. No results can be generated within reasonable running time (3 hours) for generations beyond $G2I_4$. By incorporating pedigree information, both GAIN and MERLIN infer accurate estimates (error rate less than 2%). In contrast, HAPPY has much higher error rates and is more sensitive to noise.

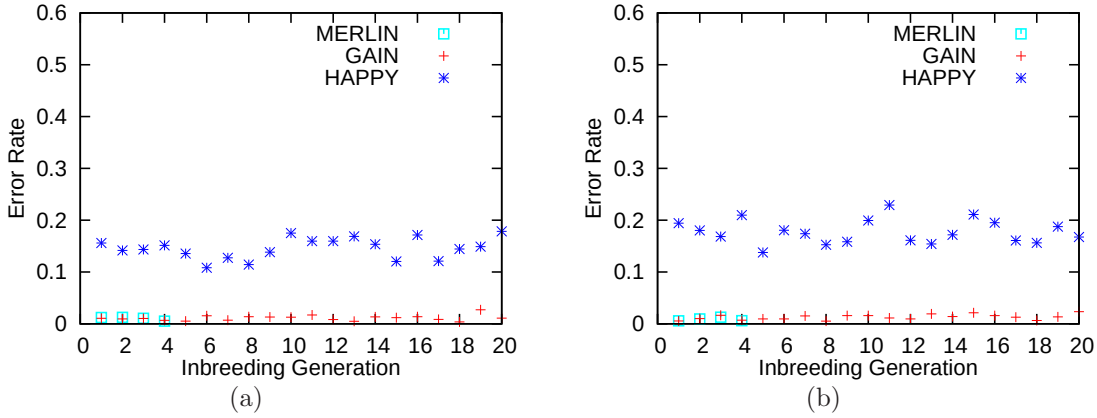


Figure 2.4: (a) Comparison of error rates of GAIN, MERLIN and HAPPY on a simulated data set with no noise. (b) Comparison on a simulated data set with 1% noise.

As mentioned previously, an accurate solution to the genome ancestry problem is important to subsequent studies such as QTL analysis. In such studies, not only the most likely genome ancestry is desired, but also the probabilities of each founder pair are wanted. Hence, it is also important to evaluate the probability distribution generated by each method. Both GAIN and HAPPY compute a probability distribution of each founder pair being the ancestry at a SNP site. I investigate the proportion of probabilities assigned to wrong founder ancestry. The result in Figure 2.5 shows that the knowledge of pedigree structure is indispensable in solving the genome ancestry prob-

lem. While HAPPY infers the most probable ancestry correctly for more than 80% of the markers, it assigns near 60% of the total probabilities to wrong ancestry choices. The mis-assigned probabilities could hamper further studies. With pedigree structure modeled, GAIN can resolve most ambiguities and assigns only less than 4% of the total probabilities to wrong ancestry.

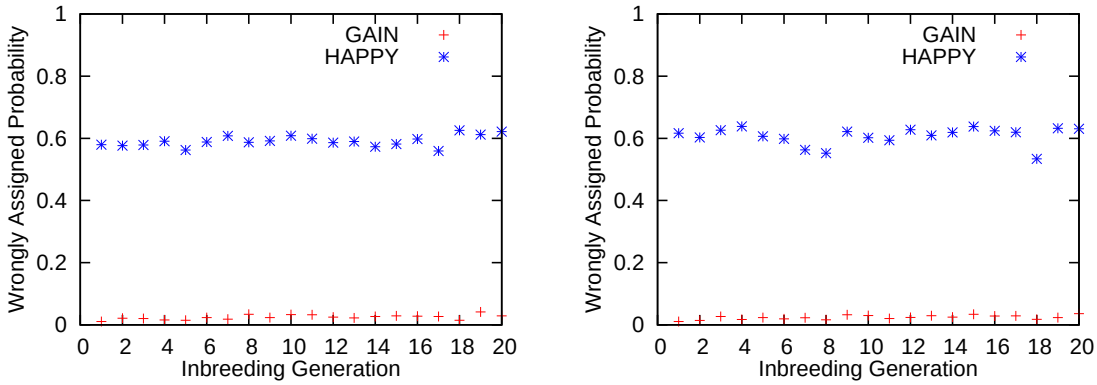


Figure 2.5: (a) Proportion of probabilities assigned to wrong ancestry by GAIN and HAPPY on a simulated data set with no noise. (b) Proportion of probabilities assigned to wrong ancestry by GAIN and HAPPY on a simulated data set with 1% noise.

2.5.2 Experiments on Real CC data

The data set consists of genotypes of all autosomes from 96 mice of generation $G2I_5$ to $G2I_{12}$. The number of SNP markers on each chromosome ranges from 4122 to 35172. Due to the running time constraint of MERLIN, I only compare GAIN with HAPPY which does not consider pedigree structure. Since the true genome ancestry is unknown, I investigate the difference between the results of the two approaches.

I compare both the best ancestry estimated and the full probability distribution of each possible ancestry. The first comparison (Figure 2.6(a)) shows the percentage of sites of which the best ancestry estimated by the two methods do not agree. The difference in best ancestry choice is very similar to that of my experiments on simulated

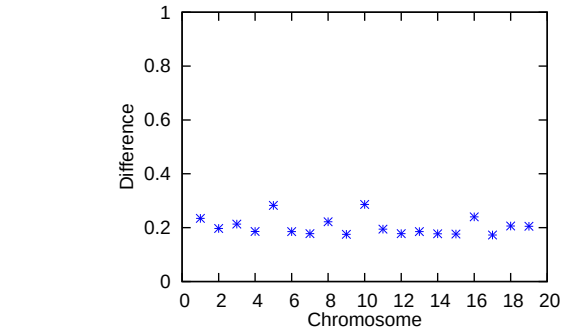
data with random error: the results from the two methods differ by 20%. I further measure the difference in probability distributions quantitatively using Jensen-Shannon(JS) Divergence [Lin, 1991] which is a smoothed and bounded divergence based on Kullback-Leibler Divergence. The JS Divergence (JSD) between two probability distributions p_1 and p_2 is defined as:

$$JSD(p_1||p_2) =$$

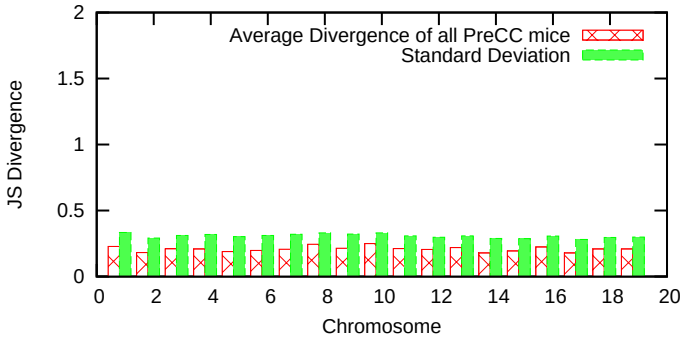
$$\sum_i p_1(i) \log_2 \frac{p_1(i)}{\frac{1}{2}p_1(i) + \frac{1}{2}p_2(i)} + \sum_i p_2(i) \log_2 \frac{p_2(i)}{\frac{1}{2}p_1(i) + \frac{1}{2}p_2(i)}$$

A low JS Divergence indicates high similarity between $p1$ and $p2$. The JS divergence ranges between 0 and 2. Figure 2.6(b) compares the mean and standard deviation of the JS Divergence between HAPPY's results and ours over all markers and all 96 mice, grouped by chromosomes.

Though I cannot compare the results against the ground truth for real CC data, the source of difference are further investigated. Consider again the CC pedigree in Figure 2.3(a). The initial four founder-mating pairs (F_1, F_2),(F_3, F_4), (F_5, F_6), (F_7, F_8) cannot serve as ancestry for any genotypes of $G2I_k$ descendants. This is because any genetic material passed from a founder mating pair is carried by a single haplotype in the $G2I_0$ generation. These four founder pairs are thus invalid ancestry choices if the pedigree structure is considered. As an example to show the improved inference due to incorporating pedigree knowledge, the ancestry of chromosome 7 of a $G2I_6$ mouse inferred by GAIN and HAPPY are shown in Figure 2.7(a) and 2.7(b) respectively. The most probable founder pair inferred by HAPPY agrees with GAIN's result at most sites. But their actual probabilities are often different. To quantify the extent to which HAPPY assigns positive probabilities to invalid ancestry, at each site l, I aggregate the probabilities of invalid ancestry and plot this "pedigree inconsistency" measure in Figure 2.7(c). I can see that, the difference between Figure 2.7(a) and 2.7(b) is largely influenced



(a)



(b)

Figure 2.6: (a) The difference in best ancestry estimated by GAIN and HAPPY (b) The average JS Divergence between results from GAIN and HAPPY on chromosome 1 to 19 of 96 real CC mice.

by the "pedigree inconsistency". Moreover, the probability distributions of ancestry choices at neighboring sites are not independent. Probabilities assigned to pedigree-inconsistent ancestry can substantially influence the choice of ancestry at neighboring sites. Such "propagated error" is sometimes the main cause of the JS Divergence between HAPPY's results and ours. As an example, Figure 2.7(d) shows a region in chromosome 1 from another $G2I_6$ mouse where the propagated error is the main cause of divergence. In this region, HAPPY does not assign significant probabilities to invalid ancestry choice, except for a few sites at both ends of this region. But, in the middle part, HAPPY favors ancestry choices that are one recombination away from these invalid ancestry choices.

To sum up, even partial pedigree knowledge causes a big difference in analyzing genome ancestry. Though HAPPY can conduct analysis rapidly, its results on complex

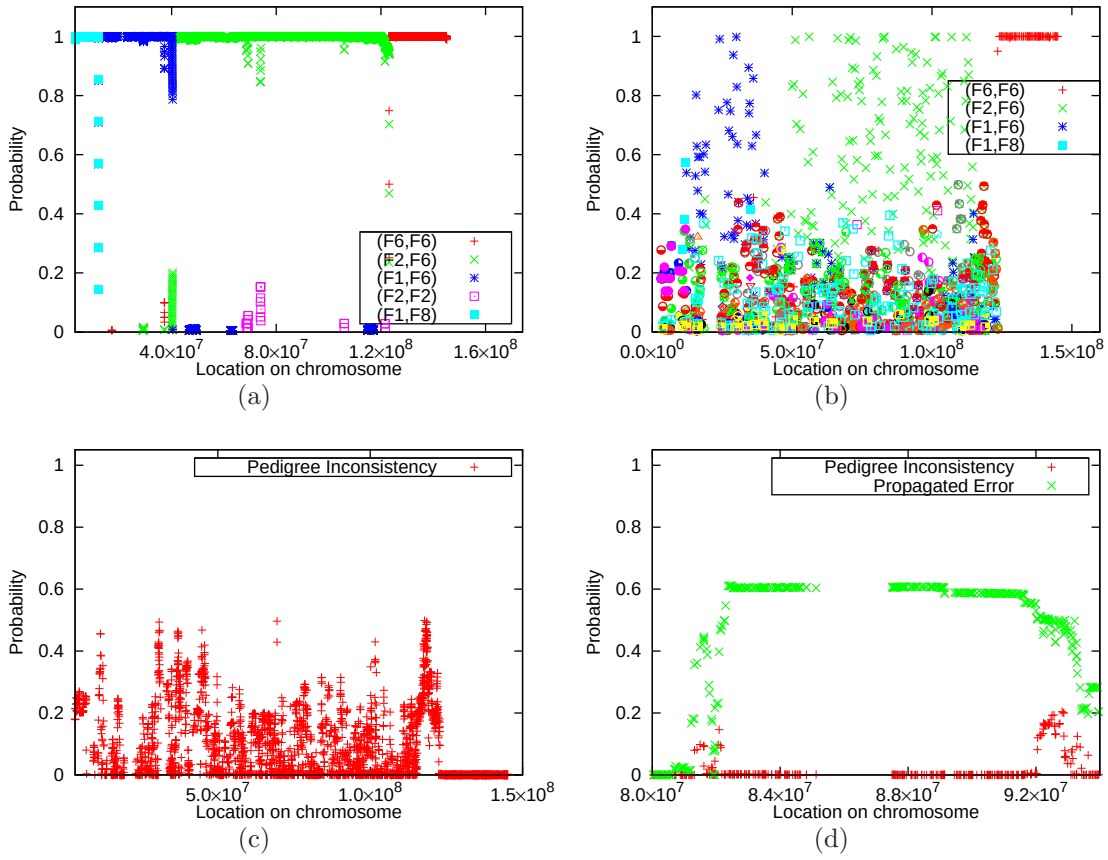


Figure 2.7: (a) Ancestry inference on chromosome 7 of a $G2I_6$ mouse by GAIN (b) Ancestry inference on chromosome 7 of the same mouse by HAPPY (c) The pedigree inconsistency in (b), i.e. the aggregated probability assigned to ancestry that violates pedigree knowledge. (d) A region in chromosome 1 from another $G2I_6$ mouse where propagated error is the main cause of divergence.

pedigrees can be biased. On the other hand, my method can provide a pedigree consistent inference in comparable running time.

2.5.3 Running Time Performance

For a pedigree containing i inbreeding processes and n' members not involved in inbreeding, the time complexity of GAIN is $O(L \times n' \times 2^{2n'} \times 2^{8i})$ where L is the number of SNP markers. For any $G2I_k$ animal in CC pedigree, the time complexity remains the same. The running time does not depend on the error rate of genotype data either. Figure 2.8 shows the running time comparison of GAIN, MERLIN and HAPPY.

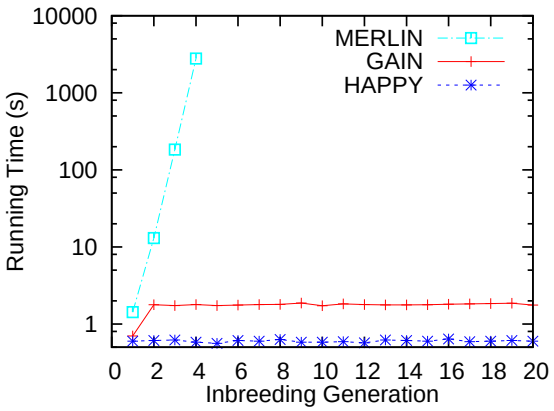


Figure 2.8: Average running time of the three methods on data set containing 6644 markers. The experiment is conducted on an Intel desktop with 2.66Ghz CPU and 8GB memory.

2.6 Discussion

The development of high density SNP technology makes model animal resources a powerful tool for studying genetic variations. It also makes any analysis on such resources computationally challenging. In this chapter, I demonstrate that modeling repetitive sub-structure of a pedigree can provide significant improvement in efficiency without compromising accuracy. I introduce a novel method for modeling the inbreeding pro-

cess. Integrated into the Hidden Markov Model framework originally introduced by the Lander-Green algorithm, my method can handle large pedigrees such as Collaborative Cross efficiently. The inbreeding sub-structure model alone does not speed up the ancestry inference for all types of pedigrees, but, as I have shown with the Collaborative Cross, the computational benefit can be crucial for analyzing many model animal resources. In analyzing such data, my method outperforms previous methods in terms of accuracy and efficiency. I believe that sub-structure modeling is a promising approach for large pedigree analysis, especially when specific types of pedigree are of interest.

Chapter 3

High Definition Recombination Map in a Highly Divergent Mouse Population

3.1 Introduction

Recombination is an essential biological process in sexual reproduction as it ensures accurate chromosome segregation during meiosis and also contributes significantly to DNA repair and genetic diversity. Abnormal recombination can result in missegregation and is associated with multiple developmental diseases [Hassold and Hunt, 2001]. Despite its importance, the regulation mechanism for the rate and pattern of recombination is largely unknown, although previous studies have shown the influence of factors, including sex, chromosome, DNA sequence and hotspots [Robinson, 1996; Smagulova *et al.*, 2011].

The Collaborative Cross (CC) provides a unique opportunity for the study of genome-wide recombination. The CC is a large panel of recombinant inbred lines (RIL) currently under development [Chesler *et al.*, 2008; The Collaborative Cross Consortium, 2012]. It is derived from eight genetically diverse founder strains, including five classical inbred strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, and NZO/H1LtJ) and three wild-derived strains (CAST/EiJ, PWK/PhJ, and WSB/EiJ). The eight founder strains were selected to capture a much greater level of genetic diversity than existing RIL panels

[Roberts *et al.*, 2007]. Each of the independently bred lines has equal contributions from all eight founder strains via a funnel breeding scheme (Figure 2.3(a)). The eight founder strains are first intercrossed to generate the $G1$ generation. The $G1$ progeny are then crossed to create the four-way $G2I_0$ generation. The first eight-way progeny, the $G2I_1$ s are then generated from a $G2I_0 \times G2I_0$ cross [1]. After this generation, CC strains become inbred by repeated generations of inbreeding through sibling mating. At the top of the funnel, the eight founder strains are arranged in order that is randomized and not repeated across lines. The left four founders contribute to the left half of the funnel and the remaining four contribute to the right half. I also denote the four pairs of founders that are crossed to produce $G1$ progeny as four quarters of the funnel.

In this study, I focus on the $G2I_1$ generation which has balanced genome contribution from both sides of the funnel pedigree. The breeding pedigree leading to $G2I_1$ generation contains eight observable meioses (Figure 3.1). I denote the four at crossing $G1$ generation as MGM, MGP, PGM, PGP and the four meioses at crossing $G2I_0$ generation as M_m, M_f, P_m, P_f. Using the genotype data of $G2I_1$ generation, I reconstructed the haplotype at $G2I_1$ generation and inferred all switching points of genome ancestry which correspond to past recombination events in the pedigree. With the design of the breeding scheme, every inferred recombination event can be assigned uniquely to one of the eight meioses. With all recombinations inferred and characterized by gender, meioses and genetic features, this study presents a high definition genome-wide recombination map and associated analysis of its properties.

[1] Other researchers have used $G2$ and $G2F_1$ to denote $G2I_0$ and $G2I_1$ generations

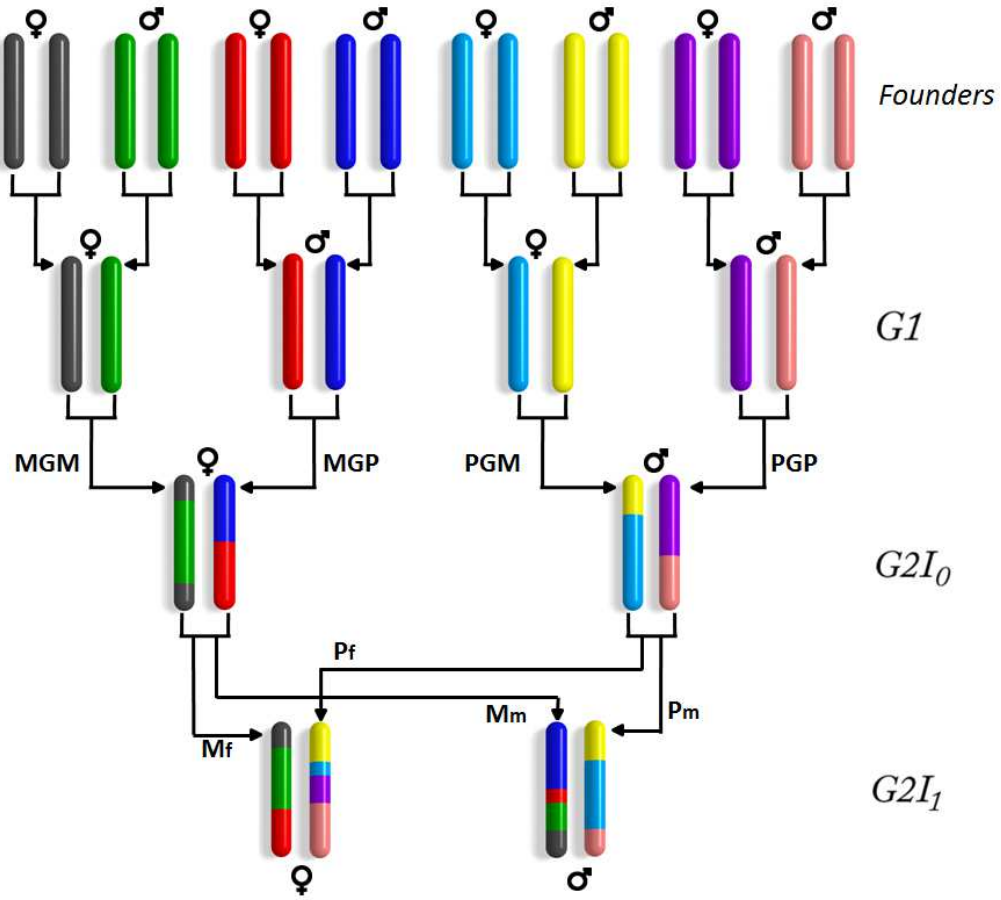


Figure 3.1: The CC funnel pedigree to $G2I_1$ generation. In total there are eight meioses in the pedigree.

3.2 Materials and Methods

3.2.1 The Genotype Data

The genotype data were obtained from 244 male-female sibling pairs at $G2I_1$ generation using a customized high-density genotyping array [Yang *et al.*, 2009]. The array contains 623,124 SNPs that capture the known genetic variation in laboratory mouse. Before I conducted haplotype reconstruction, I separated SNPs into high-quality and mid-to-low quality groups by examining:

- Genotype completeness (>0.99)

- Concordance between $G2I_1$ mice, founder mice and partially available $G1$ genotypes

I kept only 15∼25% of all SNPs on each chromosome in the high-quality group and used only these high-quality SNPs for haplotype reconstruction and recombination inference. The mid-to-low quality SNPs were used later to help refine recombination boundaries. I also excluded samples[2] and chromosomes [3] with exceptionally high discordance rate in haplotype reconstruction.

3.2.2 Haplotype Reconstruction and Recombination Inference

I utilized the method GAIN to conduct haplotype reconstruction and recombination inference. The method, as described in Chapter 2, is a hidden-Markov-model based method that can model haplotype and recombinations with all pedigree knowledge incorporated. It has been shown that GAIN can perform analysis in the CC with both high accuracy and scalability with respect to the pedigree size (proportional to number of generations). For the specific $G2I_1$ generation, the model constructed in GAIN is similar to that in an efficient implementation of Lander-Green algorithm (e.g., MERLIN [Abecasis *et al.*, 2001]) because there are no further inbreeding generations. I performed analysis on each funnel independently but jointly on the siblings in the same funnel. This is because siblings can share recombinations and joint analysis can help resolve ambiguity on recombination locations and haplotype boundaries. Recombinations, however, are not shared across funnels.

For each pair of $G2I_1$ sibling mice, GAIN took the genotypes of the eight founder mice and genotypes of the two sibling mice as input. In addition, it required the funnel order of eight founders. It then inferred the founder ancestry (in probabilities) at each SNP site by building a descendency model at each SNP and evaluating the probabilities of recombining between adjacent SNPs. The founder ancestry at each SNP describes

[2] fourteen mouse samples or seven sibling pairs

[3] six samples' chromosome 18 and four samples' chromosome X

the probability that each pair of founders (e.g., C57BL/6J and CAST/EiJ) are the two founders where the two alleles are inherited from. With pedigree knowledge considered and careful QC steps, GAIN achieved a very high level of confidence in estimating the best ancestry at most sites. More than 98% of the sites in all mice have the best ancestry choice estimated with ≥ 0.99 probability. With the ancestry probability information, I could define the haplotype blocks and recombinations trivially by tracing the most probable founder ancestry along chromosomes. Each recombination event is described by:

- a mid-point where the most probable founder ancestry changes
- proximal and distal boundaries where the probability of the most founder ancestry shrinks to a threshold
- proximal and distal ancestry founders on the recombining chromosome
- the type of meiosis it is associated to

The *recombination interval* inferred (from proximal to distal boundary) is expected to contain the recombination event with high probability. Note that there are regions where multiple founder ancestries have similar probabilities (due to lack of markers, low genotyping quality or similar DNA sequence in multiple founders). In such cases, long recombination intervals were obtained and the recombination events cannot be determined with high resolution.

Upon obtaining the recombination inference results, I further refined them with the mid-to-low-quality SNPs filtered in the QC step. This was done by examining the consistency at mid-to-low quality SNPs between founders, each $G2I_1$ mice, and all $G2I_1$ mice assigned the same ancestry. On average, this reduced the recombination intervals inferred by approximately half.

Note also that GAIN fully enforces all constraints imposed by pedigree knowledge. For example, two of the strongest constraints for $G2I_1$ mice are:

- For any SNP of any $G2I_1$ mouse, the two alleles must come from different halves of the funnel.
- Two siblings cannot inherit different alleles from one quarter funnel at any SNP site.

If the input data contained errors (genotype data or funnel order), GAIN would infer significantly more recombinations in order to satisfy the corresponding constraints. This can be used as an effective indicator to identify and remove:

- Wrongly labeled funnels and mice
- Poorly performing and/or incorrectly mapped SNPs

3.3 Overview of the Recombination Map

Table 3.1: Summary of Identified Recombination Events in $G2I_1$ Mice

			Autosomes			X chromosome			
Meiosis #	Type	Sex of $G2I_1$	Non-Shared	Shared	All	Non-Shared	Shared	All	Total
1	M	f	3282	-	3282	183	-	183	3465
2	M	m	3255	-	3255	150	-	150	3405
3	P	f	2871	-	2871	-	-	-	2871
4	P	m	2783	-	2783	-	-	-	2783
5	MGM	f	826	756	1582	35	48	83	
		m	767	756	1523	35	48	83	
		all	1593	756	2349	70	48	118	2467
6	MGP	f	733	730	1463	-	-	-	
		m	768	730	1498	-	-	-	
		all	1501	730	2231	-	-	-	2231
7	PGM	f	807	766	1573	174	-	174	
		m	782	766	1548	-	-	-	
		all	1589	766	2355	174	-	174	2529
8	PGP	f	740	745	1485	-	-	-	
		m	757	745	1502	-	-	-	
		all	1497	745	2242	-	-	-	2242

A total of 25,038 recombination events were identified in the 474 individual $G2I_1$ mice. Of these 18,948 events are observed only once and 3,045 recombination events are shared by the sib pair. Therefore, we have identified 21,993 unique recombination events in our population, 21,368 on the autosomes and 625 on chromosome X. Table 3.1 presents a summary of all types of recombination events identified.

At a high level, I examined the correctness of the events by checking the ratio between types of events (expected and observed). Firstly, the ratio of shared vs non-shared events is expected to be 1:2 based on Mendel's Law of Segregation. In the observed data, non-shared events represent 67.3% of events in the MGM, MGP, PGM and PGP meiosis (6,180 out of 9,177 events, the binomial test p-value is 0.17). This is consistently observed in each type of meiosis: MGM, 67.8%; MGP, 67.2%; PGM, 67.5% and PGP, 66.8% (binomial test p-values are 0.25, 0.54, 0.42, 0.93). Secondly, there should not be significant differences in the number of events in same type of meiosis (Mf vs Mm, Pf vs Pm, MGM vs PGM and MGP vs PGP). The ratio of events observed is highly consistent: Mf vs Mm, 1.02; Pf vs Pm, 1.03; MGM vs PGM, 0.975; and MGP vs PGP, 0.999 (binomial test p-values are 0.48, 0.25, 0.39, 0.88). Lastly, the ratio between (M+P) events and (MGM+MGP+PGM+PGP) should be 4:3 ($\frac{4}{4+3} = .57$) [4]. I observed 12,191 and 9,177 events, respectively ($\frac{12191}{12191+9177} = .57$, the binomial test p-value is 0.79).

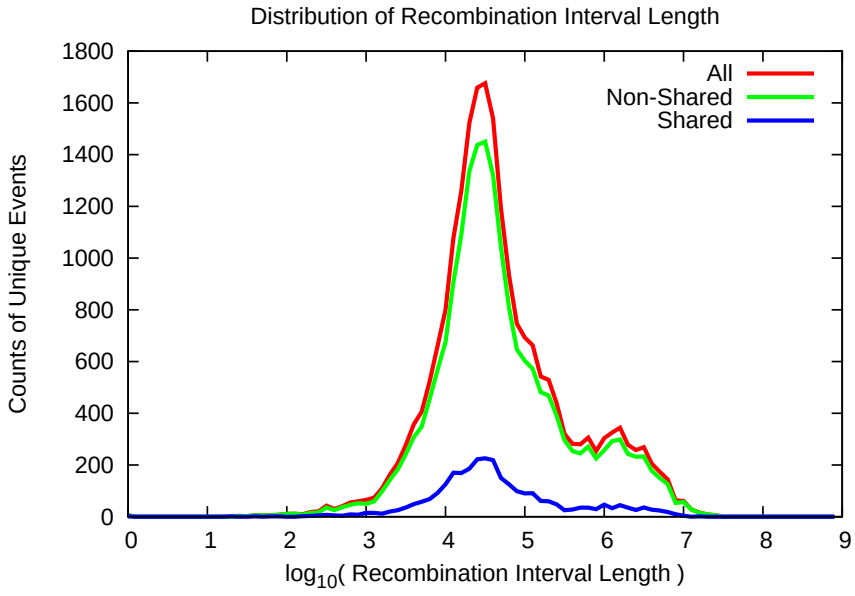


Figure 3.2: Distribution of recombination interval length in log-scale

[4]For one $G2I_1$ mouse, we expect to observe four informative independent meiosis. But if we consider two siblings, we expect to observe $4 \times 2 - 1 = 7$ meioses. Because each $G1$ meiosis has 0.25 probability to be observed in both siblings

On average, the resolution of recombination events is very high (Figure 3.2). The median size of recombination interval is 35kbp. There are, however, some recombinations that have very large uncertainty intervals (peak in Figure 3.2 between 1∼3Mbp). These are mainly due to strain dependent identical-by-descent (IBD) regions or lack of genetic markers in the interval. Based on the 21,993 identified unique recombination events, a recombination density map that can be smoothed at different scales is constructed. When smoothed with windows larger than 500kb, the $G2I_1$ map is remarkably similar to the map recently published but with much lower density of markers [Cox *et al.*, 2009].

3.4 Sex Effect on Recombination

As expected, the total number of recombination events in autosomes is significantly smaller in the male germline than in the female germline (10,127 events and 11, 241 events, respectively; binomial test p-value $\leq 3 \times 10^{-14}$; Table 3.1). This sex difference is also observed in the number of recombination events observed in each individual in both $G1$ and $G2$ meioses. To investigate the possible causes of this difference, the effect of the *Prdm9* genotype on the size of the autosomal map was determined. One of the eight founder strains of the CC, CAST/EiJ, carries the $Prdm9^a$ allele, four strains (A/J, C57BL/6J, 129S1/SvImJ and NZO/HILtJ) carry the $Prdm9^b$ allele, two strains (NOD/ShiLtJ and WSB/EiJ) carry the $Prdm9^c$ allele and the PWK/PhJ strain carries the $Prdm9^d$ allele. There is a significant expansion of the female map length and a reduction of the male map length in carriers of the $Prdm9^a$ allele (1,450 cM and 1,195 cM, respectively). There is also a significant contraction of the female map length and an expansion of the male map length in carriers of the $Prdm9^d$ allele (1,300 cM and 1,325 cM, respectively). Finally, carriers of both $Prdm9^b$ and $Prdm9^c$ alleles have similar ratio of female to male map lengths (Figure 3.3).

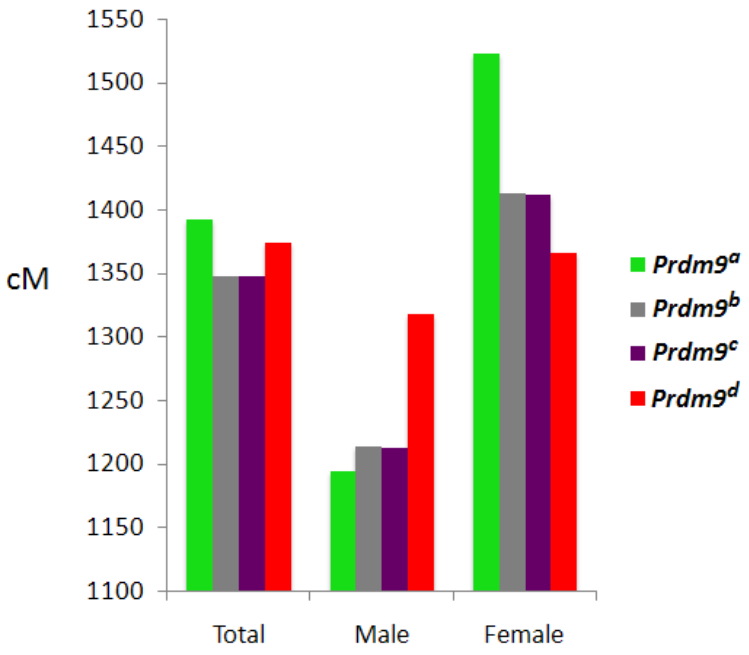
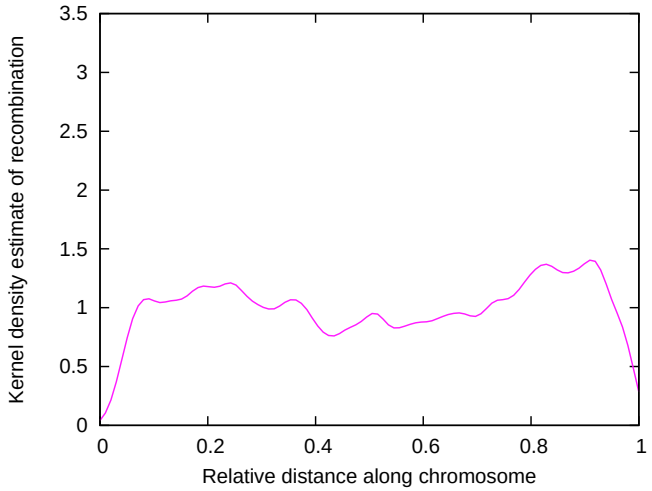


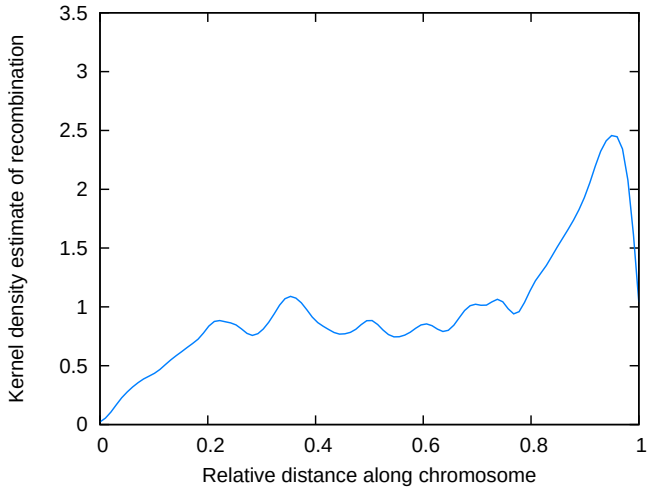
Figure 3.3: Recombination map length of autosomes by *Prdm9* allele and gender

In addition to the sex differences in overall recombination, there are dramatic sex differences in the pattern recombination events in the autosomes. Figure 3.4 shows the distribution of recombination events along the autosomes in female and male meioses. The most obvious difference is the increase in the density of recombination events in the distal ends of chromosomes in male meiosis. In female meioses, there is a more even distribution of recombination events along the autosomes (Figure 3.4(a)). In male meioses, approximately half of the recombination events occur in the distal quarter of the chromosomes and almost one third of events occur in the distal 10% of the autosomes (data not shown). Comparison of the recombination density observed in single and double recombinants reveals striking differences while preserving the increase in recombination in the distal ends of the chromosomes (Figure 3.5). The most obvious difference is that in double recombinants there are two peaks of high recombination rate separated by very low recombination rate in the middle while in single recombinants the density proximal to the distal peak remains basically constant. In double recombinants from male meioses,

the distal peak is both higher and sharper than in singles and the proximal peak is lower and much wider. This pattern suggests that recombination may progress temporally from the telomere to centromere in males.



(a) female meioses

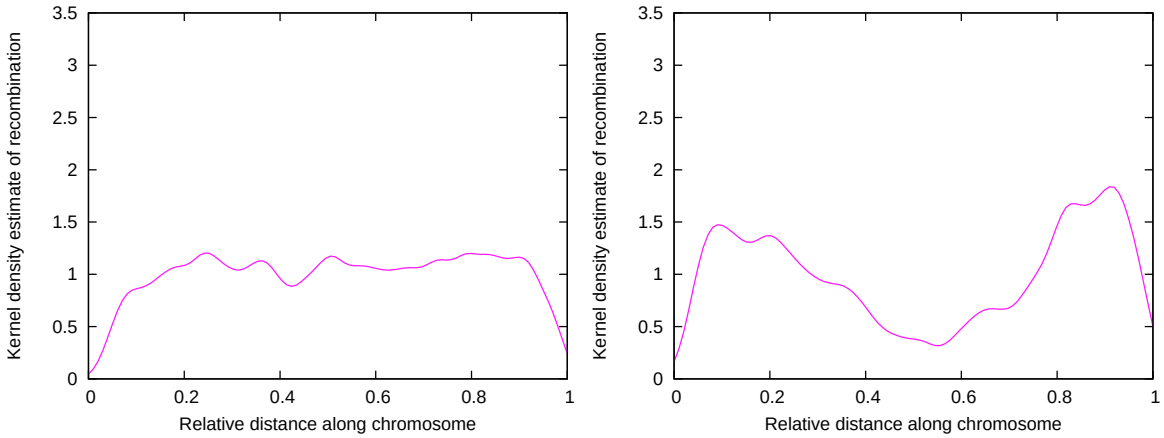


(b) male meioses

Figure 3.4: Distribution of recombination events along the autosomes in female and male meioses. The x-axis corresponds to the relative position in all autosomes. The y-axis indicates the kernel density estimates of recombinations in each type of meiosis.

3.5 Cold Regions

Regions with low levels of recombination have been reported previously [Smagulova *et al.*, 2011] and many mouse researchers have anecdotal evidence that the ability to efficiently



(a) female meioses, single recombinant

(b) female meioses, double recombinants

(c) male meioses, single recombinant

(d) male meioses, double recombinants

Figure 3.5: Distribution of single and double recombination events along the autosomes in female and male meioses. The x-axis corresponds to the relative position in all autosomes. The y-axis indicates the kernel density estimates of recombinations in each type of meiosis.

reduce the size of many candidate regions of interest is undermined by an apparent lack of recombination. However, we know very little about the size, distribution, genomic features and evolutionary stability of such regions. Thus identification and characterization of such cold regions may provide important information on the distribution of genetic variation and the level of linkage disequilibrium in the mammalian genome, the accuracy of imputation of genetic variants and may provide new models to study the molecular and cellular mechanisms of meiotic recombination.

3.5.1 Identification of Cold Regions in the $G2I_1$ Population

Cold regions are defined as long (>500 kb) continuous genomic intervals that are markedly depleted of recombination events in the $G2I_1$ population. Given the total number of recombination events in my experiment (∼22,000) I set up this 500 kb threshold in the initial identification of cold regions to reduce the number of false positives (i.e., on average I expect 8.7 recombination events per Mb).

The 50 coldest regions in male and female meioses are first identified independently to allow for possible cold regions on chromosome X. The union of these regions constituted the initial set and underwent several filtering steps. In the fist step regions in which no calls (Ns) represent a large fraction of the nominal length are excluded. After this step the boundaries of the 59 remaining cold regions are refined using the recombination intervals in the $G2I_1$ population. For 51 of these regions the new refined interval has no recombination events and they are bound by the distal boundary of proximal recombination event and the proximal boundary of the distal recombination event (Table 3.2). Overall, cold regions span 124.1 Mb (∼5% of the genome), distributed along 18 chromosomes (all chromosomes have cold regions except chromosomes 10 and 11) and with an enrichment for proximal and distal sections of the chromosomes (Table 3.2)

3.5.2 External Validation of Cold Regions

To determine whether the results in the $G2I_1$ population are replicable in other populations, I estimated the recombination rate in these regions in the heterogeneous stock used to construct the most recent linkage map of the mouse [Cox *et al.*, 2009]. On average, there is a four-fold reduction in recombination density in cold regions (0.14 cM/Mb versus the expected 0.5 cM/Mb that is observed genome wide). In fact, 57 of the 59 regions are below the genome wide average and for 16 regions the recombination density in the Cox map is zero (Table 3.2). The extent of validation is striking given the differences in genetic background (only five of the 16 strains are shared between these two studies and the non shared strains include three wild derived strains representing two subspecies that are rare or absent in the genetic makeup of the strains in the Cox study), marker density and approach to estimate recombination distances between these two populations.

Recently, several maps of recombination initiation sites in the mouse have been published [Smagulova *et al.*, 2011; Brick *et al.*, 2012]. These studies identified regions with significant enrichment of double strand breaks (DSB) in the male germline of mice of different genetic backgrounds. Smagulova *et al.* [2011] identified 21 recombination deserts larger than 3 Mb, but noted that the inability of identifying hotspots in some of these regions may be due to sequencing gaps or highly repetitive DNA. Eleven of the cold regions identified in the $G2I_1$ population overlap with those described previously in Smagulova *et al.* [2011]. This level of concordance is even more remarkable once one considers that one of the Smagulova desserts was eliminated from my analysis because of complete lack of sequence [5] and the fact that nine additional regions that fail to make the cut in my list still show low levels of recombination in the $G2I_1$ population. More importantly, data from the second study [Brick *et al.*, 2012] can be used to estimate the density of DSB in any given region. On average there is a 18X reduction in DSB density (range 14X to 24X) in cold regions compared to the genome average.

[5] chr 7: 39 Mb, see also new GRCm38 assembly of the mouse genome

Table 3.2: List of Cold Regions Identified

Chr	Start	End	Size	$G2I_1$	Cox	B6	9R	13R	F1	Smagulova	(C+G)	SD	genes
1	7552017	8916100	1364083	0	0	2.217	2.37	3.112	3.511	No	38.29	0.06	6
1	16629188	17556666	927478	0	0.234	3.065	2.903	3.502	3.573	No	40.12	0	11
1	26449654	29071537	2621883	1	0.069	2.621	2.468	2.425	2.481	Yes	36.53	0.35	14
1	47989803	50296981	2307178	1	0	2.794	2.566	2.526	2.689	No	36.58	0.25	11
1	100644783	101554230	909447	0	0.198	3.001	2.845	1.939	0	No	37.72	0	2
1	111488506	113990313	2501807	0	0.048	2.802	2.653	2.278	2.652	No	36.46	0.22	11
1	147042689	150558199	3515510	1	0.031	2.638	2.523	2.372	2.654	No	36.39	0.58	7
1	151085005	152671374	1586369	0	0.228	3.115	2.949	3.307	3.731	No	39.01	0.08	13
2	80863558	82092319	1228761	0	0.103	3.171	3.027	3.402	3.853	No	36.79	0	1
2	99031585	101085143	2053558	0	0.292	3.669	3.546	2.783	3.546	No	36.99	0.29	6
3	5879218	7942959	2063741	0	0.092	2.944	2.799	3.221	3.495	No	38.02	0.22	12
3	12781208	13766597	985389	0	0	3.378	3.178	3.661	3.938	No	37.08	0	4
3	71924425	72840203	915778	0	0	3.216	3.039	3.127	3.039	No	36.30	0	1
3	91404734	92463929	1059195	0	0	2.564	1.984	3.591	4.034	No	38.83	0.34	35
3	93110421	94190526	1080105	0	0.271	2.486	2.042	3.32	3.923	No	41.99	0.54	37
3	111864347	114650278	2785931	0	0.134	2.511	2.358	2.267	2.258	No	36.95	0.40	19
4	3872920	4995031	1122111	0	0.334	3.777	3.509	3.527	3.857	No	39.69	0	9
4	71515502	73850480	2334978	0	0	3.268	3.113	3.171	3.589	No	38.71	0.33	21
4	78445659	80238398	1792739	0	0.126	2.43	2.341	2.572	2.627	No	36.77	0.06	3
4	109873098	110836231	963133	0	0.04	1.802	0	3.721	4.03	No	38.3	0	1
4	112348832	113862435	1513603	0	0	0	0	1.744	2.235	No	37.23	0.18	9
4	123025109	124537945	1512836	0	0.332	3.105	2.871	2.644	2.703	No	45.89	0.01	23
5	7624169	8730184	1106015	0	0.055	3.684	3.543	3.315	3.4	No	39.83	0.03	10
5	10717204	11949724	1232520	0	0.126	2.316	1.985	0	0	No	37.77	0.89	25
5	86051378	89055526	3004148	0	0.083	3.682	3.384	3.392	3.784	No	38.19	0.52	59
5	93726598	96365273	2638675	0	0	0	0	0	1.995	No	41.00	0.94	46
6	129843394	131358958	1515564	0	0	0	0	2.671	2.998	No	38.22	0.63	29
7	7231821	12298098	5066277	0	0.584	1.989	1.7	1.843	1.999	Yes	38.98	0.92	157
7	20689345	24188289	3498944	0	0.073	0	0	0	1.66	Yes	40.94	0.97	157
7	31943682	34982574	3038892	0	0.051	2.747	2.438	1.5	1.743	Yes	41.14	0.90	71
7	66503334	69332558	2829224	0	0.138	3.048	2.889	2.713	3.045	No	41.74	0.80	27
7	91783502	92931816	1148314	0	0.06	0	0	0	0	No	37.85	0.06	31
7	110909130	111734201	825071	0	1.017	3.23	3.083	1.947	0	No	40.43	0.10	55
7	113744664	114326885	582221	0	0.151	2.355	2.264	2.509	2.729	No	39.04	0	29
8	75489675	77295059	1805384	1	0.241	2.977	2.82	3.264	3.557	No	40.02	0.10	3
9	4932310	6150379	1218069	0	0	2.685	2.556	3.252	3.646	No	37.72	0	7
9	8055696	9141014	1085318	0	0	2.879	2.665	0	0	No	39.83	0	8
12	21307827	25245820	3937993	0	0.048	2.257	1.962	0	1.387	Yes	44.85	0.96	47
12	47788818	49655054	1866236	0	0.169	3.338	3.178	3.054	3.181	No	36.77	0.06	4
12	61325419	62237566	912147	0	0.229	2.856	2.706	3.026	3.206	No	36.70	0	4
12	119492704	120437290	944586	0	0	2.906	2.661	2.563	2.974	No	40.08	0	6
13	16817515	18417648	1600133	0	0.318	3.713	3.296	3.39	3.691	No	39.45	0.08	12
13	85193045	87485288	2292243	1	0.021	2.747	2.604	1.635	1.74	No	37.43	0.26	8
14	13925179	19817805	5892626	0	0.191	3.727	3.568	3.323	3.68	No	40.89	0.79	47
14	42010628	43812329	1801701	0	0.063	0	0	0	0	No	41.37	0.89	34
14	50370810	51760945	1390135	0	0.423	2.947	2.695	3.467	3.692	No	40.24	0.20	54
14	83838558	86170267	2331709	0	0	1.67	0	2.3	2.394	No	36.39	0.25	5
14	123768455	124578293	809838	0	0.305	3.134	2.887	2.806	3.209	No	37.98	0	3
15	14228552	15476602	1248050	0	0.19	2.084	0	2.708	2.866	No	36.63	0	1
16	46467540	47489432	1021892	1	0.098	2.684	2.509	2.643	2.535	No	38.16	0.029	4
17	18006880	21070398	3063518	0	0.26	0	1.551	2.08	2.54	Yes	37.32	0.70	69
17	38288007	40000987	1712980	0	0.111	1.906	1.739	0	0	Yes	37.23	0.28	17
18	17782801	18654834	872033	1	0	0	0	0	0	No	36.74	0	1
19	12545972	14689226	2143254	2	0.217	3.373	3.132	3.54	3.937	No	38.89	0.19	84
X	23031354	33248619	10217265	0	0.294	1.331	0.93	2.883	2.411	Yes	39.61	0.87	0
X	73163204	76792464	3629260	0	0.023	4.164	3.967	3.749	4.161	Yes	39.15	0.58	0
X	110686456	112707404	2020948	0	0	3.226	3.081	3.282	3.753	No	37.25	0.24	0
X	115246562	118828417	3581855	0	0.057	2.365	2.381	3.319	3.706	Yes	36.69	0.63	0
X	143956636	146641734	2685098	0	0	1.349	0	2.679	2.737	Yes	40.61	0.91	0

The table provides the chromosome location of 59 putative cold regions identified in the $G2I_1$ population. In addition to the size of these regions, the table lists:

$G2I_1$, number of recombination events in the cold regions in the $G2I_1$ population.

Cox, recombination rate in Cox *et al.* [2009]

B6, $\log_1 0$(number of reads at DSB/Mb) in the C57BL/6J strain of the Brick *et al.* [2012]

9R, $\log_1 0$(number of reads at DSB/Mb) in the 9R strain of the Brick *et al.* [2012]

13R, $\log_1 0$(number of reads at DSB/Mb) in the 13R strain of the Brick *et al.* [2012]

F1, $\log_1 0$(number of reads at DSB/Mb) in the (9Rx13R) strain of the Brick *et al.* [2012]

Smagulova, recombination dessert in reported by Smagulova *et al.* [2011] in (C57BL/10.SC57BL/10.F)F1 mice

(C+G), base composition in percent

SD,fraction of the interval included in segmental duplication identified using dotplots generated using Gepard [Krumsiek *et al.*, 2007]

 # genes, number of genes

3.5.3 Genomic Analysis of Cold Regions

Several genomic features have been associated with suppressed recombination in regions such centromeres including low C+G content, frequent and complex duplications and enrichment for repeated sequences. Therefore, I determined the content of cold regions for these and additional genomic features (gene content, presence of segmental duplications (tandem and inverted)).

The overall C+G content in cold regions is significantly lower than the genome wide average (Table 3.2). When all 59 intervals are plotted together the plot resembles the aggregate of three different distributions with obvious peaks at 36%, 40% and 44%. The lower peak is the most pronounced and represents approximately half of the cold regions (26 cold regions with low C+G). This suggests that cold regions tend to be associated with local low C+G content. I also observed a highly significant enrichment for large (>15 kb) segmental duplications either in tandem or inverted in cold regions. On average, in cold regions 28% of the sequences are involved in some type of rearrangement.

3.6 Conclusion

In this chapter, I present a genome-wide recombination study based on recombination events inferred from the $G2I_1$ generation in the CC resource. The unique design of the CC allows us to fully determine the meiosis of each recombination event and attribute recombinations to gender and other genetic features. I performed careful quality control steps in constructing the recombination map. Extensive internal and external validations have been done to verify the correctness of results obtained. The sex, jointly with *Prdm9* alleles, have strong effect on the pattern of recombinations. The distribution of double recombinants in male meioses strongly suggests a temporal pattern of the recombination progression. Furthermore, the vast majority of cold regions identified in the $G2I_1$ population represent bona fide regions of suppressed recombination independent of the

genetic background. Besides establishing the association with reduction in DSB density, I investigated the relationship between cold regions and local DNA sequence.

Chapter 4

MaCH-Admix: Genotype Imputation for Admixed Populations

4.1 Introduction

Imputation of untyped genetic markers has been routinely performed in genome-wide association studies (GWAS) [Sanna *et al.*, 2010; Scott *et al.*, 2007; WTCCC, 2007] and meta-analysis [Dupuis *et al.*, 2010; Smith *et al.*, 2010; Willer *et al.*, 2008], and will continue to play an important role in sequencing-based studies [Fridley *et al.*, 2010; The 1000 Genomes Project Consortium, 2010]. Li *et al.* [2010a] have previously developed a hidden Markov model (HMM) based method for imputation and shown that it achieves high imputation accuracy in a number of populations [Huang *et al.*, 2009], particularly those with high level of linkage disequilibrium (LD) or having closely matched reference population(s) from the HapMap [The International HapMap Consortium, 2010] or the 1000 Genomes Projects (1000G) [The 1000 Genomes Project Consortium, 2010, 2012]. However, little methodological work exists for imputation in admixed populations, such as African Americans and Hispanic Americans, which comprise more than 20% of the US population (see Web Resources).

Admixed populations offer a unique opportunity for gene mapping because one could utilize admixture LD to search for genes underlying diseases that differ strikingly in preva-

lence across populations [Reich and Patterson, 2005; Rosenberg *et al.*, 2010; Tang *et al.*, 2006; Winkler *et al.*, 2010; Zhu *et al.*, 2004]. Although useful for admixture mapping, admixture LD also imposes challenges for imputation. Since an admixed individual's genome is a mosaic of ancestral chromosomal segments, to appropriately impute the genotypes, it is imperative to incorporate the underlying ancestry information. Practically, this is equivalent to selecting an appropriate reference panel that matches the corresponding ancestral population(s).

Existing studies have evaluated a wide range of choices on the construction of a reference panel **prior to** running the imputation engine. The recommendation is to use a **pre-defined** panel that either combines all reference populations (a cosmopolitan panel) [Hao *et al.*, 2009; Li *et al.*, 2009; Shriner *et al.*, 2010] or a weighted combination panel [Egyud *et al.*, 2009; Huang *et al.*, 2009; Pasaniuc *et al.*, 2010; Pemberton *et al.*, 2008]. The cosmopolitan panel may include haplotypes from populations that are irrelevant, and fails to reflect the underlying ancestry proportions and consequently the LD pattern for the target population. The weighted combination panel is generated by duplicating haplotypes according to certain weights, which substantially and unnecessarily increases computational costs [Egyud *et al.*, 2009].

An alternative approach, based on identity-by-state (IBS) sharing between the target individual and haplotypes in the reference populations, can be embedded within existing imputation models. This approach constructs individual-specific effective reference panels, by selecting the most closely related haplotypes (according to IBS score) from the entire reference pool. The IBS-based selection is intuitive and useful for reducing the size of the effective reference panel and is tailored separately for each target individual. The selection is usually conducted by finding pairwise Hamming distances which is computationally very appealing. A simple IBS-based method, which selects a subset of haplotypes into the effective reference panel according to their Hamming distance with the haplotypes to be inferred across the entire genomic region to be imputed (hereafter

referred to as whole-haplotype), has been adopted by IMPUTE2 [Howie *et al.*, 2009]. Although some promising results have been shown when compared with random selection, no work has examined alternatives to this simple whole-haplotype based matching, partly due to the heavy computational burden posed.

In this chapter, I evaluated two classes of reference selection methods: IBS-based and ancestry-weighted approaches. Among the IBS-based approaches, I propose a novel method based on IBS matching in a piecewise manner. The method breaks genomic region under investigation into small pieces and finds reference haplotypes that best represent *every* small piece, for each target individual separately. The method can be incorporated directly into existing imputation algorithms and has identical computational complexity to that of the existing whole-haplotype IBS-based method. Results from all real datasets evaluated suggest that my piecewise IBS method is highly robust and stable even when a small number of reference haplotypes are selected. Importantly, for uncommon variants, my piecewise IBS selection method manifests more pronounced advantage with large reference panels.

I have implemented all methods evaluated, including my piecewise IBS selection method, in the software package MaCH-Admix. Besides the new reference selection functionality, my software also retains high flexibility in two major aspects. First, both regional and whole-chromosome imputation can be accommodated. Second, both data independent and data dependent model parameter estimation are supported. Thus, besides standard reference panel with pre-calibrated parameters, I can elegantly handle study-specific reference panels and target samples with unknown ethnic origin.

The rest of the chapter is organized as follows. I first present the general framework of the imputation algorithm, followed by the intuition and formulation of my piecewise IBS and various other effective reference selection methods. Then I evaluate all these methods implemented in MaCH-Admix, the whole-haplotype IBS method implemented

in IMPUTE2 [Howie *et al.*, 2009], and BEAGLE[Browning and Browning, 2009] using the following datasets:

- 3587 Hispanic American individuals from the Women's Health Initiative (WHI)
- 8421 African American individuals from the WHI
- 49 HapMap III African American individuals
- 50 HapMap III Mexican individuals

All datasets are imputed with reference from the 1000 Genomes Project (2188 haplotypes). I also explored the performance with small/medium reference set from HapMap II/III. Finally, I provide practical guidelines for imputation in admixed populations in the Discussion section.

4.2 Materials and Methods

Assume that we have n individuals in the target population that are genotyped at a set of markers denoted by M_g. In addition, we have an independent set of H reference haplotypes, e.g., those from the International HapMap or the 1000 Genomes Projects, encompassing a set of markers denoted by M_r . Without loss of generality, I assume that the set of markers assayed in the target population, M_g, is a subset of M_r, the markers in the reference population. The goal of genotype imputation is to fill in missing genotypes including those missing by design (for example, genotypes at markers in M_r but not M_g, commonly referred to as *untyped markers*). As described earlier [Li *et al.*, 2010a], the hidden Markov model as implemented in MaCH fulfills the goal by inferring the haplotypes encompassing M_r markers for each target individual, from unphased genotypes at the directly assayed markers in M_g. Haplotype reconstruction is accomplished by building imperfect mosaics using some of the H reference haplotypes.

4.2.1 General Framework

Since admixed individuals have inherited genetic information from more than one ancestral population, I start with a pooled panel: a panel with haplotypes from all relevant populations, for example, CEU+YRI for African Americans and CEU+YRI+JPT+CHB for Hispanic Americans, where CEU is an abbreviation for Utah residents (CEPH) with Northern and Western European ancestry; YRI for Yoruba in Ibadan, Nigeria; JPT for Japanese in Tokyo, Japan; and CHB for Han Chinese in Beijing, China. Let $\mathcal{G} = (g_1, g_2, g_3, ..., g_{M_r})$ denote the unphased genotypes at M_r markers for a target individual. Furthermore I define a series of variables $S_m, m = 1, 2, ..., M_r$ to denote the hidden state underlying each unphased genotype g_m. The hidden state S_m consists of an ordered pair of indices (x_m, y_m) indicating that, at marker m, the first chromosome of this particular target individual uses reference haplotype x_m as the template and the second chromosome uses reference haplotype y_m as the template, where x_m and y_m both take values from $\{1, 2, ..., H\}$.

I seek to infer the posterior probabilities of the sequence of hidden states $\mathcal{S} = (S_1, S_2, ..., S_{M_r})$ for each individual as the knowledge of $\mathcal{S}$ will determine genotype at each of the M_r markers. Define $P(S_m|\mathcal{H}, \mathcal{G})$ as the posterior probability for S_m, the hidden state at marker m with $\mathcal{H}$ denoting the pool of reference haplotypes and $\mathcal{G}$ denoting the genotype vector of the target individual. To infer these posterior probabilities, I run multiple Markov iterations. Within each iteration, I calculate the conditional joint probabilities $P(S_m, \mathcal{G}|\mathcal{H})$ at each marker m via an adapted Baum's forward and backward algorithm as previously described [Li *et al.*, 2010a].

For admixed populations, as one tends to include more reference haplotypes in the pool under the philosophy of erring on the safe side, and as one attempts not to duplicate haplotypes, one key aspect of the modeling is on how to traverse the sample space harboring the most probability mass with minimum computational efforts.

4.2.2 Piecewise IBS-based Reference Selection

In piecewise IBS selection, I seek to construct a set of t effective reference haplotypes from the pool of H haplotypes within each HMM iteration for each target individual separately. Selected reference panels are therefore tailored for each target individual. For presentation clarity, I consider a single target individual. Specifically, I calculate the genetic similarity (measured by IBS, the Hamming distance between two haplotypes) in a piecewise manner between the individual and each haplotype in the reference pool, ignoring the sub-populations (e.g., CEU or YRI) within the reference.

Denote $(\mathbf{h}'_1, \mathbf{h}'_2)$ as the current haplotype guess for the target individual. I break haplotype $\mathbf{h}'_1$ into a maximum of $\frac{t}{2}$ pieces so that the typed markers are evenly placed across pieces. Each piece has a minimum length of ν typed markers to ensure that the calculated Hamming distance is informative. Denote the number of pieces by p. For each haplotype piece, I calculated the piece-specific IBS score between $\mathbf{h}'_1$ and each reference haplotype and selects the top $\frac{t}{2p}$ reference haplotypes, resulting in a total of $\frac{t}{2}$ selected for $\mathbf{h}'_1$ across all p regions. I repeat the same procedure for $\mathbf{h}'_2$ and select a second set of $\frac{t}{2}$ reference haplotypes. In my implementation, I set $\nu = 32$, which corresponds to an average length of <200Kb for commonly used genomewide genotyping platforms. To avoid creating spurious recombinations at piece boundary, I apply a random offset to the first piece in each sampling so that the boundaries differ across iterations. In the case where $\frac{t}{2p}$ is not an integer, I select $\overline{(\frac{t}{2p})}$ (the ceiling integer) reference haplotypes in each piece for each target haplotype. Then I sample randomly from the selected reference haplotypes. Note that the piecewise selection is repeated for each individual in each sampling iteration. Thus the selection will change along with the intermediate sampling results.

I have also implemented two whole-haplotype IBS-based methods, IBS Single Queue (IBS-SQ) and IBS Double Queue (IBS-DQ). The former defines IBS score with any reference haplotype as the minimum Hamming distance to $\mathbf{h}'_1$ and $\mathbf{h}'_2$, thus ordering

the H reference haplotypes in a single queue. The top t reference haplotypes will be selected accordingly. The latter defines two separate IBS scores for $\mathbf{h}'_1$ and $\mathbf{h}'_2$, thus ordering the H reference haplotypes in two queues. The top $t/2$ reference haplotypes will be selected for $\mathbf{h}'_1$ according to IBS scores for $\mathbf{h}'_1$. Similarly, another $t/2$ reference haplotypes will be selected for $\mathbf{h}'_2$.

Figure 4.1 explains the three IBS strategies under two simple scenarios. In both scenarios, there are eight markers measured in both target and reference with color indicating the allelic status where the same color at the same locus implies the same allele. In both Figures 4.1A and 4.1B, the first chromosome of the target individual shares all eight alleles with the dark-colored reference haplotypes and zero alleles with the light-shaded reference haplotypes. In Figure 4.1A, the second chromosome of the target individual shares two alleles with the dark-colored reference haplotypes and the remaining six alleles with the light-shaded reference haplotypes; whereas in Figure 4.1B, the second chromosome shares six alleles with the dark-colored reference haplotypes and the remaining two alleles with the light-shaded reference haplotypes.

Suppose $t = \frac{H}{2}$. Figure 4.1A illustrates a scenario where the whole-haplotype Single Queue strategy is not optimal because only dark-colored haplotypes will be selected into the effective reference panel. By combining two sets selected from two separate queues, the whole-haplotype Double Queue strategy is advantageous in the scenario. On the other hand, neither the whole-haplotype Single Queue nor the whole-haplotype Double Queue strategy can handle the scenario in Figure 4.1B well because both strategies would only select the dark-colored reference haplotypes. Ideally, the selected reference haplotypes should, when possible, contain information to represent every part of both chromosomes carried by the target individual. In the scenario presented in Figure 4.1B, because the target individual carries segment of the light-shaded haplotype, it is desirable to have some representation of the light-shaded haplotypes in the effective reference panel. My piecewise IBS method achieves this by breaking the whole region into pieces and selecting

some reference haplotypes according to genetic matching in each piece (illustrated in the bottom part of Figure 4.1A and 4.1B). By conducting local IBS-matching and choosing a few reference haplotypes within each piece, it is able to have some representation of the light-shaded reference haplotypes. As a result, all parts of the target chromosomes are well represented by the selected reference haplotypes. In general, I believe that selecting a small number of reference haplotypes for each piece locally performs better than selecting globally at the whole-haplotype level. Note that the piecewise IBS method has the same computational complexity as the two whole-haplotype IBS methods.

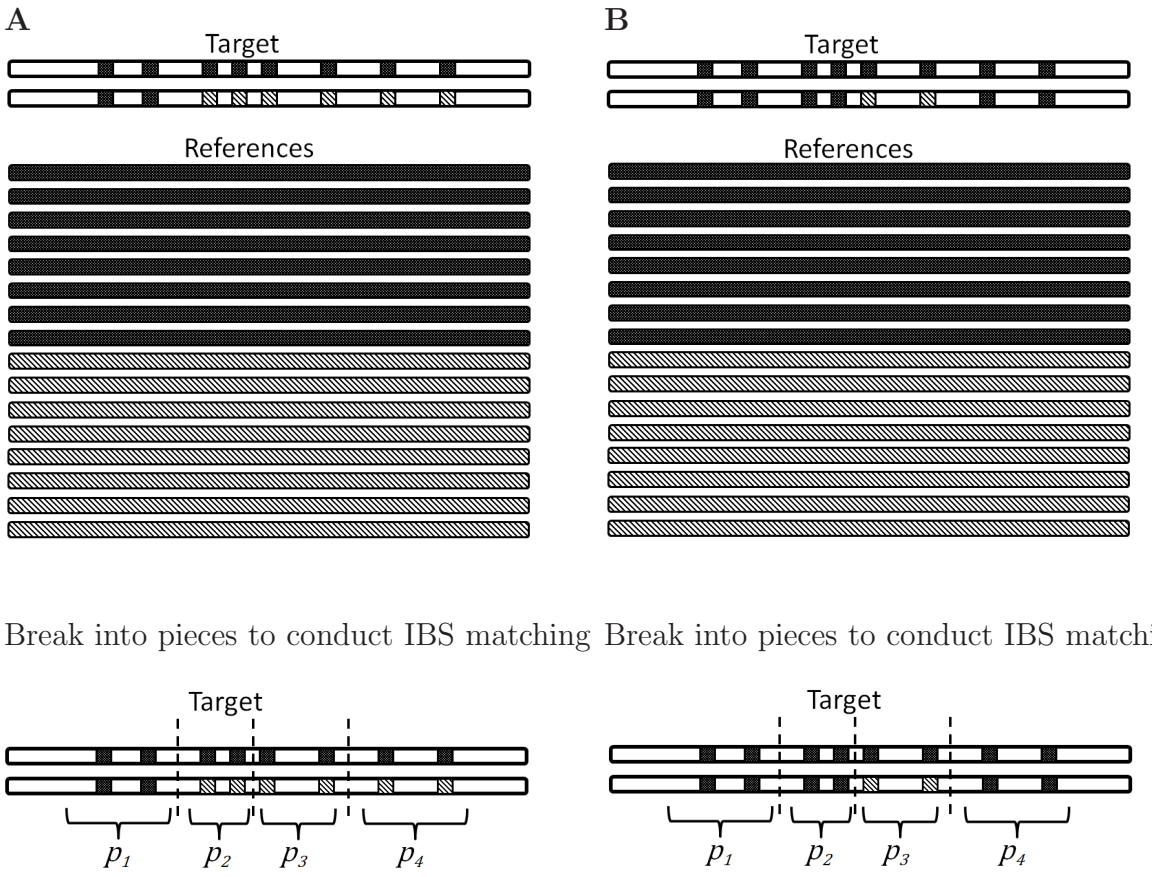


Figure 4.1: A cartoon illustration of two scenarios where three IBS-based selection methods perform differently. The two lines on the top panel represent the two chromosomes of a target individual and the lines on the bottom panel represent the pool of H=16 reference haplotypes. Color determines the allelic status such that the same color at the same locus implies the same allele. The bottom parts show how my piecewise selection method breaks the imputation region into four pieces with $t = \frac{H}{2} = 8$. Here I assume no constraint on the minimum piece size (i.e., $\nu = 0$).

4.2.3 Ancestry-weighted Approach

Besides IBS-based methods, I also evaluate an ancestry-weighted selection method, which is motivated by the idea of weighted cosmopolitan panel discussed in the Introduction Section. This method concerns the scenario where the reference panel consists of haplotypes from several populations, for instance CEU and YRI, such that the H reference haplotypes are naturally decomposed into several groups. Let Q denote the number of populations included and H_q denote the number of haplotypes from reference population $q, q = 1, 2, ..., Q$. I first consider the issue of weight determination for each contributing reference population, i.e., the fraction of reference haplotypes to be selected from that population. Intuitively, the weights should depend on the proportions of ancestry from these reference populations for the target admixed individual(s). The weights can be, on one extreme, the same for all individuals in the target population (for example, when the admixture makeup is similar across all individuals), or different for sub-populations within the target population, or on the other extreme, specific for each target individual. For presentation clarity, I suppress the individual index i and denote $\mathbf{w} = (w_1, w_2, ..., w_Q)$ as the vector of weights, under the constraint that $w_1 + w_2 + ... + w_Q = 1$. In this work, I consider the same set of weights for all target individuals. The weights are to represent the average contributions over the imputation region and for all target individuals. I choose to use such average weights over weights specific to each single individual because the average weights can be more stably estimated.

There are several natural ways to estimate the weights. One could pre-specify the weights according to estimates of ancestry proportion. For example, it is reasonable to use a ∼2:8 CEU:YRI weighting scheme for African Americans who are estimated to have about 20% Caucasian and 80% African ancestries [Lind *et al.*, 2007; Parra *et al.*, 1998; Reiner *et al.*, 2007; Stefflova *et al.*, 2011]. Alternatively, one can estimate the ancestry proportions for the target individuals under investigation. I have implemented an

imputation-based approach within MaCH-Admix to infer ancestry proportions, according to the contributions of reference haplotypes from each population to the constructed mosaics of the target individuals so that the weights can be estimated by MaCH-Admix internally. I use the software package ***structure*** [Pritchard *et al.*, 2000], specifically its Admix+LocPrior model, on LD-pruned set of SNPs to confirm my internal ancestry inference.

Having determined the weights, I am interested in constructing a set of t effective reference haplotypes within each Markov iteration from the pool of H reference haplotypes according to the ancestry proportions. I achieve this by sampling without replacement $t \times w_q$ haplotypes from the H_q haplotypes in reference population q. For each target individual, I sample a different reference panel under the same set of weights.

4.2.4 MaCH-Admix

I have implemented the aforementioned methods (three IBS-based and one ancestry-weighted) in my software package MaCH-Admix. MaCH-Admix breaks the one-step imputation in MaCH into three steps: phasing, model parameter (including error rate and recombination rate parameters) estimation and haplotype-based imputation. The splitting into phasing and haplotype-based imputation is similar to IMPUTE2. My software can accommodate both regional and whole-chromosome imputation and allows both data dependent and data independent model parameter estimation. The flexibility regarding model parameter estimation allows one to perform imputation with standard reference panels such as those from the HapMap or the 1000 Genomes Projects with pre-calibrated parameters in a data independent fashion, similar to IMPUTE2, which uses recombination rates estimated from the HapMap data and a constant mutation rate. Alternatively, if one works with study-specific reference panels, or suspects the model parameters differ from those pre-calibrated (for example, when target individuals are of

unknown ethnicity or from an isolated population), one has the option to simultaneously estimate these model parameters while performing imputation.

4.2.5 Datasets

I assessed the reference selection methods in the following six target sets:

- 3587 WHI Hispanic Americans (WHI-HA)
- 8421 WHI African Americans (WHI-AA)
- 200 randomly sampled WHI-HA individuals
- 200 randomly sampled WHI-AA individuals
- 49 HapMap III African Americans (ASW)
- 50 HapMap III Mexican individuals (MEX)

The WHI SHARe consortium offers one of the largest genetic studies in admixed populations. WHI [The WHI Study Group, 1998; Anderson *et al.*, 2003] recruited a total of 161,808 women with 17% from minority groups (mostly African Americans and Hispanics) from 1993-1998 at 40 clinical centers across the U.S. The WHI SHARe consortium genotyped all the WHI-AA and WHI-HA individuals using the Affymetrix 6.0 platform. Detailed demographic and recruitment information of these genotyped samples are previously described [Qayyum *et al.*, 2012]. Besides standard quality control (details described previously in [Liu *et al.*, 2012]), I removed SNPs with minor allele frequency (MAF) below 0.5%. To evaluate the imputation performance on target sets of smaller size, I randomly sampled 200 individuals from WHI-HA and WHI-AA separately.

For the two HapMapIII datasets, my target individuals are ASW (individuals of African ancestry in Southwest USA) and MEX (individuals of Mexican ancestry in Los Angeles, California) respectively from the phase III of the International HapMap Project

[The International HapMap Consortium, 2010]. These individuals (83 ASW and 77 MEX) were all genotyped using two platforms: the Illumina Human1M and the Affymetrix 6.0. I restricted my analysis to founders only: 49 ASW and 50 MEX.

The main focus of my work is imputation with large reference panel. Thus, I first evaluated the imputation performance of all six target sets with reference from the 1000 Genomes Project (release 20101123, $H = 2188$ haplotypes). For the WHI datasets, the number of markers overlapping between the target and reference, bounded by the number of markers typed in target samples, is smaller than that in the HapMap individuals. Therefore, I performed imputation 10 times, each time masking a different 5% of the Affymetrix 6.0 markers. This masking strategy allowed us to evaluate imputation quality at 50% of Affymetrix 6.0 SNPs. For HapMap III ASW and MEX individuals, I randomly masked 50% of the overlapping markers and evaluated the performance at these markers. I used two different masking schemes for the HapMap and WHI samples because I have ∼1.5 million typed markers in the HapMap samples and thus can still achieve reasonable imputation accuracy by masking 50% of the markers in a single trial. In the WHI samples, masking 50% of the ∼0.8 million markers in a single trial would substantially reduce imputation accuracy and using one trial with a small percentage of markers masked would lead to insufficient number of markers for evaluation. Therefore, I used multiple trials with 5% masking for the WHI datasets.

To provide a comprehensive evaluation, I also conducted imputation on all six target sets using HapMapII or HapMapIII haplotypes as the reference. I used HapMap II CEU+YRI ($H = 240$) for WHI-AA individuals and HapMapII CEU+YRI+JPT+CHB ($H = 420$) for WHI-HA individuals. The evaluation is based on masking 50% of the overlapping markers. For HapMap III ASW target set, I considered three different reference panels: HapMapII CEU+YRI ($H = 240$), HapMapIII CEU+YRI ($H = 464$), and HapMapIII CEU+YRI+LWK+MKK ($H = 930$), where LWK (Luhya in Webuye, Kenya) and MKK (Maasai in Kinyawa, Kenya) are two African populations from Kenya. For

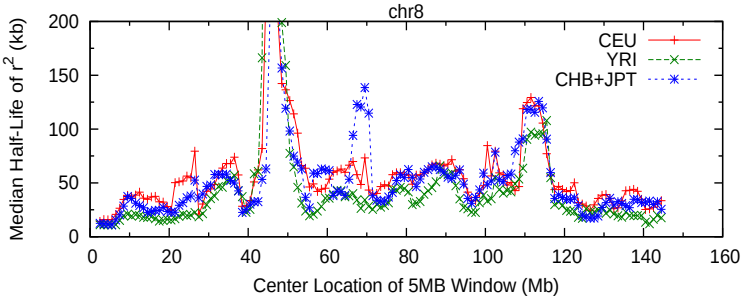
HapMap III MEX target set, I considered HapMapII CEU+YRI+JPT+CHB ($H = 420$), and HapMapIII CEU+YRI+JPT+CHB ($H = 804$). For the HapMap target sets with HapMap references, I used genotypes at SNPs on the Illumina HumanHap650 BeadChip for imputation input and reserved other genotypes for evaluation. I have posted the HapMap data and my command lines used in this work on MaCH-Admix website (see Web Resources).

I picked five 5Mb regions across the genome to represent a wide spectrum of LD levels. I first calculated median half life of r^2, defined as the physical distance at which the median r^2 between pairs of SNPs is 0.5, for every 5Mb region using a sliding window of 1Mb, in CEU, YRI, and JPT+CHB, respectively. I used HapMapII phased haplotypes for the calculation. The five regions I picked are: chromosome3:80-85Mb, chromosome1:75-80Mb, chromosome4:57-62Mb, chromosome14:50-55Mb, and chromosome8:18-23Mb in a decreasing order of LD level. The median half life of r^2 is around 90th, 70th, 50th, 30th, and 10th percentile within each of the three HapMap populations, for the five regions respectively (Table 4.1). Figure 4.2 shows the LD levels for the five residing chromosomes. For each region, I treat the middle 4Mb as the core region and the 500Kb on each end as flanking regions. Only SNPs imputed in the core region were evaluated to gauge imputation accuracy.

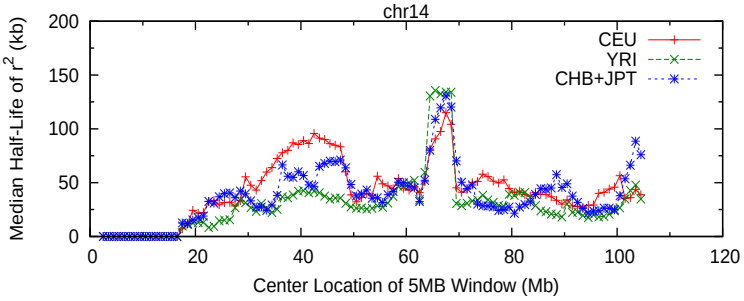
Table 4.1: Median Half Life of r^2 (in Kb)

	CEU	YRI	JPT+CHB
10th Percentile	26	16	22
30th Percentile	38	24	32
50th Percentile	48	30	41
70th Percentile	60	39	55
90th Percentile	92	57	83
chromosome3:80-85Mb	106	70	124
chromosome1:75-80Mb	69	38	80
chromosome4:57-62Mb	47	31	32
chromosome14:50-55Mb	40	25	43
chromosome8:18-23Mb	25	16	23

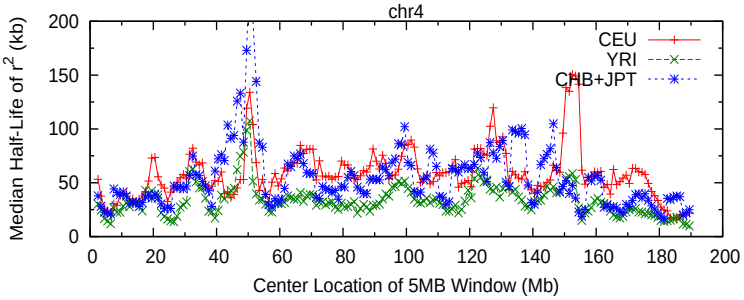
Percentiles are calculated within each population using all 5Mb windows across the genome.



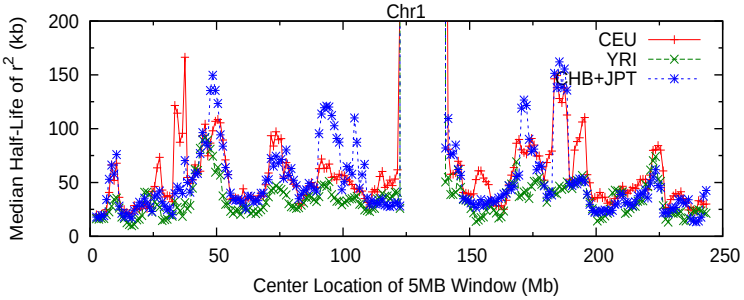
(a) Chr 8



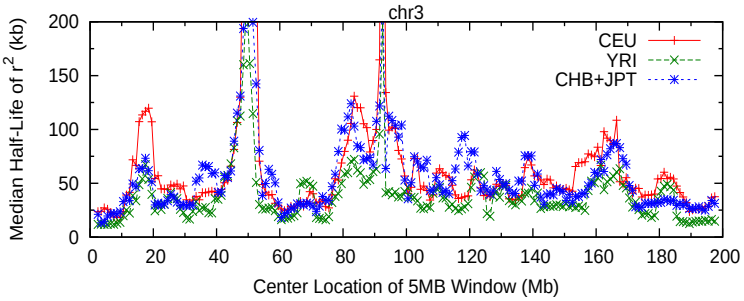
(b) Chr 14b



(c) Chr 4



(d) Chr 1



(e) Chr 3

Figure 4.2: Median r^2 half-life value of 5Mb windows on 5 chromosomes

4.2.6 Methods Compared

I evaluated the following reference selection approaches implemented in MaCH-Admix:

- random selection (MaCH-Admix Random or original MaCH)
- IBS Piecewise selection (MaCH-Admix IBS-PW)
- IBS Single-Queue selection (MaCH-Admix IBS-SQ)
- IBS Double-Queue selection (MaCH-Admix IBS-DQ)
- Ancestry-Weighted selection (MaCH-Admix AW) (for HapMapIII datasets)

I also included IMPUTE2 [Howie *et al.*, 2009] and BEAGLE [Browning and Browning, 2009] for comparison. I used IMPUTE 2.1.2 and BEAGLE 3.3.1 with default settings (*-k_hap* 500 *-iter* 30 for IMPUTE2; *niterations*=10 *nsamples*=4 for BEAGLE). As aforementioned, MaCH-Admix can conduct imputation with pre-calibrated parameters (similar to IMPUTE2); alternatively, MaCH-Admix can perform imputation together with data-dependent parameter estimation in an integrated mode. The integrated mode generates slightly better results at the cost of increased computing time. Here, I report results from the pre-calibrated mode.

4.2.7 Measure of Imputation Quality

Previous studies have proposed multiple statistics to measure imputation quality [Browning and Browning, 2009; Li *et al.*, 2009; Lin *et al.*, 2010; Marchini and Howie, 2010], measuring either the concordance rate, correlation, or agreement between the imputed genotypes or estimated allele dosages (the fractional counts of an arbitrary allele at each SNP for each individual, ranging continuously from 0 to 2) and their experimental counterpart. I opt to report the dosage r^2 values, which are the squared Pearson correlation between the estimated allele dosages and the true experimental genotypes (recoded as 0,

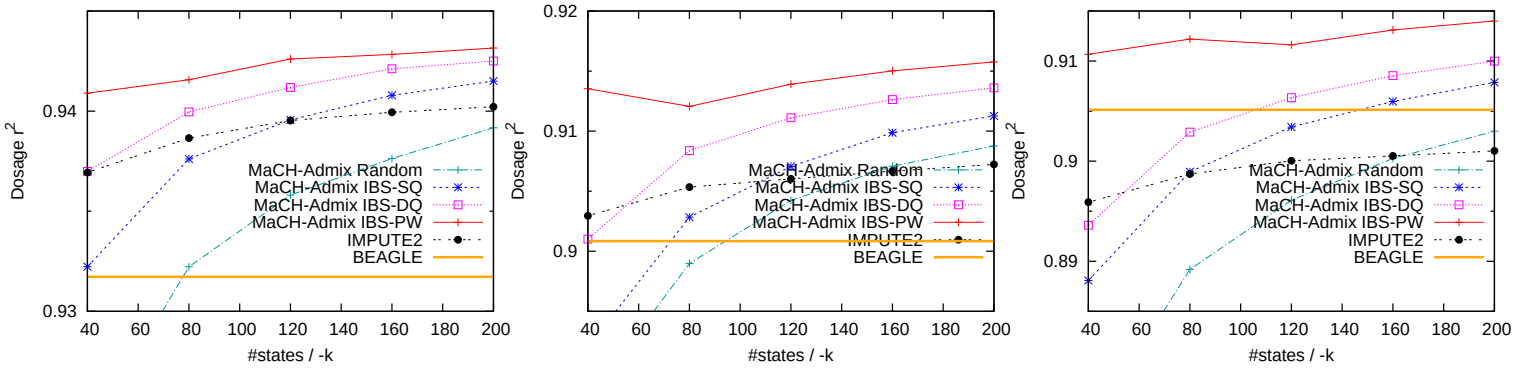
1, and 2 corresponding to the number of minor alleles), because it is a better measure for uncommon variants by taking allele frequency into account and directly related to the effective sample size for downstream association analysis (Pritchard and Przeworski, 2001). For the remainder of the work, with no special note, average dosage r^2 values will be plotted as a function of approximation level (measured by the effective reference panel size, i.e., t described in Methods section, corresponding to MaCH-Admix's --*states* option and IMPUTE2's -k option). Hereafter, I use approximation level, effective reference size, t, and #states/-k interchangeably. I note that for standard haplotypes-to-genotype imputation (that is, using reference haplotypes to imputed target individuals with genotypes), computational costs increase quadratically with the approximation level. MaCH-Admix and IMPUTE2 both also have an approximation parameter at the haplotype-based imputation step, MaCH-Admix's --*imputeStates* and IMPUTE2's -*k_hap*, which increases the computation time linearly and is by default set at a large value (500). I kept both at the default value because increasing beyond the default has rather negligible effects on imputation quality and that total computing time attributable to the haplotype-based imputation step is typically much smaller compared to --*states* and -k.

4.3 Results

4.3.1 WHI-AA and WHI-HA with the 1000G Reference

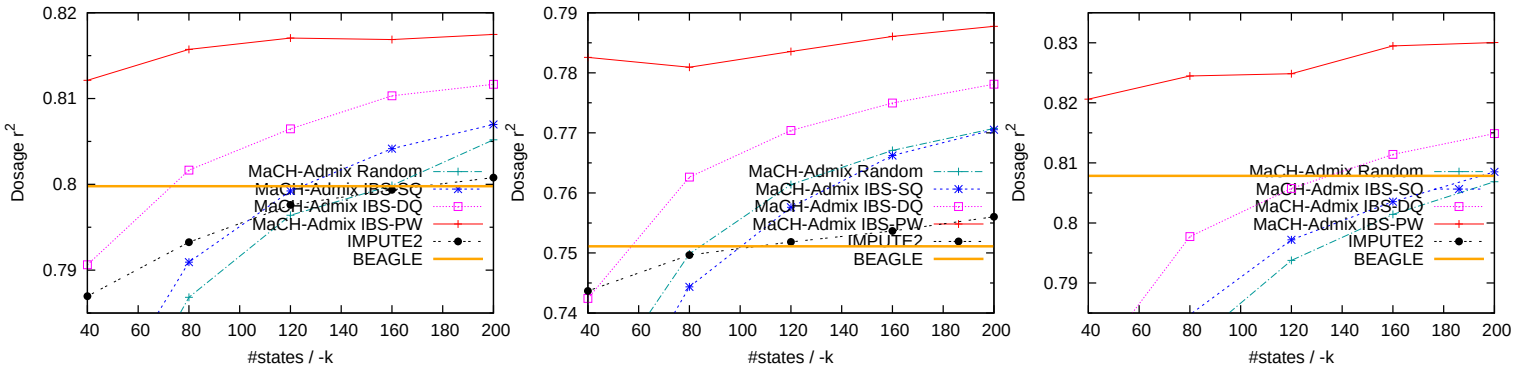
Figures 4.3 and 4.4 show results for full WHI-HA and WHI-AA sets using 2188 haplotypes from 20101123 release of the 1000 Genomes Project as the reference (selected three out of the five 5Mb regions: the 1st, 3rd, and 5th regions according to level of LD). The remaining results under the default or middle settings are presented in Tables 4.2 and 4.3 (all five regions for WHI-HA and WHI-AA respectively). Note that BEAGLE's performance remains constant because it does not have a parameter analogous to MaCH-Admix's --*states* or IMPUTE2's -k.

Generally, I observe higher imputation accuracy in regions with higher level of LD for all approaches evaluated. In addition, in regions with higher LD, imputation accuracy reaches a plateau with smaller effective reference sizes. This is because the LD pattern can be captured fairly well by a smaller number of reference haplotypes in regions with higher level of LD. In regions with lower level of LD, accuracy plateau is reached with larger effective reference sizes. But generally an effective reference size of 80 to 120 is good for MaCH-Admix to perform well at all LD levels.



(a) Chr 3, 80-85Mb (b) Chr 4, 57-62Mb (c) Chr 8, 18-23Mb

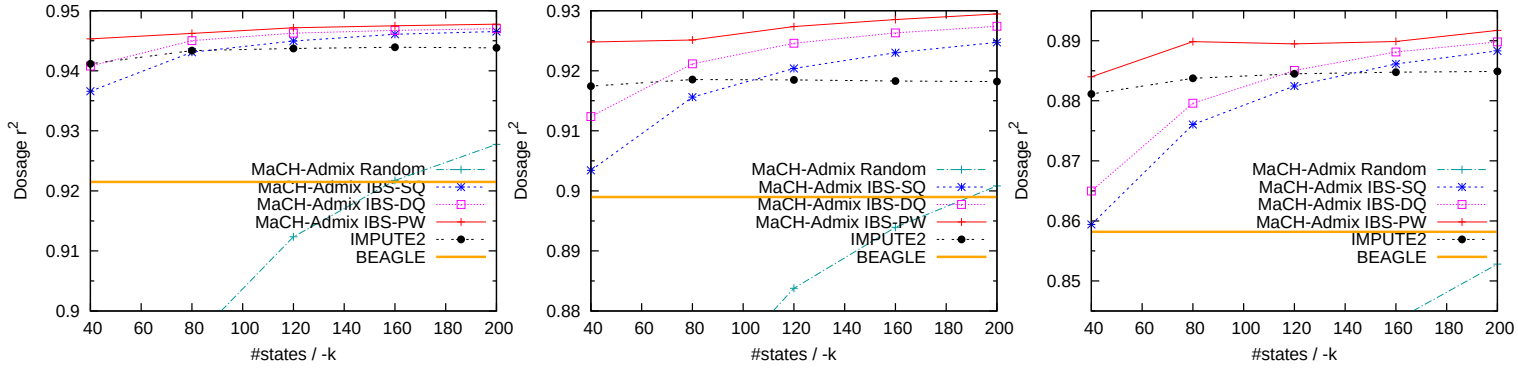
A: Imputation quality of WHI-HA with the 1000G reference panel



(d) Chr 3, 80-85Mb (e) Chr 4, 57-62Mb (f) Chr 8, 18-23Mb

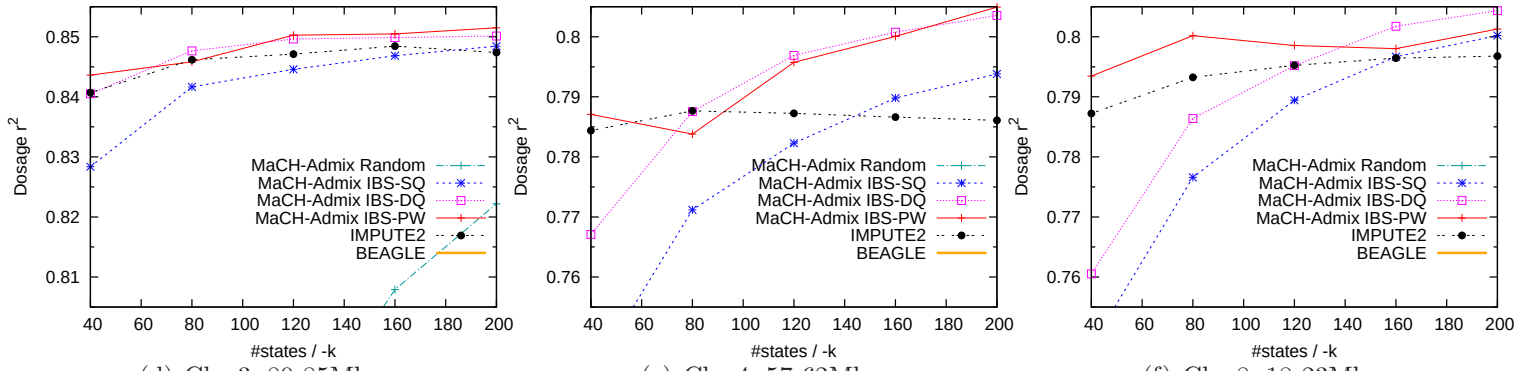
B: Uncommon SNP imputation quality of WHI-HA with the 1000G reference panel. I set the maximum plotting range on y-axis to be 5%. IMPUTE2 in (c) is below the lower bound of the plotting range.

Figure 4.3: Imputation of 3587 WHI-HA with the 1000G reference panel. Imputation quality (measured by dosage r^2) is plotted as a function of the effective reference panel size (i.e., #states), for WHI-HA individuals in three selected 5Mb regions (ordered by LD from high to low).



(a) Chr 3, 80-85Mb (b) Chr 4, 57-62Mb (c) Chr 8, 18-23Mb

A: Imputation quality of WHI-AA with the 1000G reference panel



(d) Chr 3, 80-85Mb (e) Chr 4, 57-62Mb (f) Chr 8, 18-23Mb

B: Uncommon SNP imputation quality of WHI-AA with the 1000G reference panel. Note that WHI-AA has significantly less number of SNPs in this category than WHI-HA does. Also, I set the maximum plotting range on y-axis to be 5%. MaCH-Admix Random in (b),(c) and BEAGLE in (a),(b),(c) are below the lower bound of the plotting range.

Figure 4.4: Imputation of 8421 WHI-AA with the 1000G reference panel. Imputation quality (measured by dosage r^2) is plotted as a function of the effective reference panel size (i.e., #states), for WHI-AA individuals in three selected 5Mb regions (ordered by LD from high to low).

I found that the piecewise IBS selection approach (IBS-PW) is clearly the best among the three IBS-based methods implemented in MaCH-Admix. Its performance is stable even with a small #states value. For the other two IBS-based reference selection approaches implemented in MaCH-Admix, I observed IBS-DQ performs better than IBS-SQ. The performance order of the three MaCH-Admix IBS-based methods is expected based on my reasoning in the Material and Methods Section. In addition, all three IBS-based methods show clear advantage over random selection, particularly when the effective reference size is small. IMPUTE2 has similar performance to that of IBS-DQ when the effective reference size is small. Interestingly, IMPUTE2's accuracy curve tends to stay relatively flat while those for MaCH-Admix's IBS-based methods increase with the effective reference size.

Across all five regions evaluated, with effective reference size at 120, IBS-PW has consistent performance gain over other evaluated methods. Importantly, IBS-PW and IBS-DQ, particularly IBS-PW, manifest more pronounced advantage for uncommon variants (MAF <5%) in WHI-HA. For these uncommon variants, average dosage r^2 is 0.818, 0.782, and 0.794 (0.808, 0.805, and 0.756) for WHI-HA (WHI-AA) using IBS-PW, IMPUTE2, and BEAGLE respectively. The advantage of IBS-PW in uncommon SNPs is however smaller in WHI-AA largely because of the much smaller number of uncommon variants in WHI-AA (Figure 4.5). However, the difference is highly significant (p-value $\leq 5.02 \times 10^{-5}$) in both WHI samples. My observation is consistent in both the full set and the subset of 200 individuals (Tables 4.2 and 4.3).

The variance of imputation quality by markers is heavily influenced by the MAF distribution. All methods exhibits much larger variance in imputing uncommon variants. The standard error of my IBS-PW method ranges from 0.0046 to 0.0055 for all variants, and from 0.016 to 0.0248 for uncommon variants in imputing the WHI-HA full set. In imputing the WHI-AA full set, the standard error of IBS-PW ranges from 0.0037 to 0.0047 for all variants, and from 0.02 to 0.0299 for uncommon variants.

Table 4.2: Imputation Results of WHI-HA Individuals over Five 5Mb Regions with the 1000G reference

	All 3587 individuals			Random 200 Subset		
	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time
chromosome3:80-85Mb						
MaCH-Admix Random	0.935(0.107)	0.796(0.189)	35968	0.921(0.121)	0.794(0.231)	841
MaCH-Admix IBS-PW	**0.942(0.101)**	**0.817(0.189)**	40422	0.923(0.111)	**0.814(0.210)**	1041
MaCH-Admix IBS-SQ	0.939(0.104)	0.799(0.190)	38208	0.923(0.119)	0.796(0.231)	988
MaCH-Admix IBS-DQ	0.941(0.102)	0.806(0.191)	38439	0.924(0.119)	0.799(0.232)	995
IMPUTE2	0.939(0.104)	0.797(0.191)	40722	**0.925(0.119)**	0.799(0.233)	2076
BEAGLE	0.931(0.107)	0.799(0.190)	162888	0.912(0.128)	0.779(0.231)	6614
chromosome1:75-80Mb						
MaCH-Admix Random	0.918(0.130)	0.821(0.190)	50108	0.924(0.129)	0.855(0.211)	1214
MaCH-Admix IBS-PW	**0.927(0.123)**	**0.841(0.186)**	57671	0.927(0.121)	**0.873(0.197)**	1490
MaCH-Admix IBS-SQ	0.923(0.122)	0.823(0.187)	53908	0.926(0.125)	0.861(0.209)	1443
MaCH-Admix IBS-DQ	0.926(0.121)	0.830(0.185)	57321	**0.928(0.123)**	0.866(0.207)	1452
IMPUTE2	0.921(0.121)	0.809(0.183)	51362	0.921(0.127)	0.845(0.204)	2545
BEAGLE	0.917(0.124)	0.815(0.184)	229514	0.917(0.129)	0.851(0.209)	9194
chromosome4:57-62Mb						
MaCH-Admix Random	0.904(0.148)	0.761(0.208)	53960	0.918(0.137)	0.813(0.213)	1239
MaCH-Admix IBS-PW	**0.913(0.139)**	**0.783(0.202)**	61827	**0.922(0.134)**	**0.824(0.212)**	1527
MaCH-Admix IBS-SQ	0.907(0.141)	0.757(0.195)	60806	0.918(0.135)	0.807(0.209)	1460
MaCH-Admix IBS-DQ	0.911(0.138)	0.770(0.195)	59088	0.921(0.133)	0.817(0.210)	1455
IMPUTE2	0.906(0.142)	0.751(0.198)	62272	0.908(0.147)	0.773(0.225)	2991
BEAGLE	0.900(0.150)	0.751(0.218)	360545	0.907(0.155)	0.787(0.244)	14888
chromosome14:50-55Mb						
MaCH-Admix Random	0.921(0.132)	0.800(0.202)	57082	0.936(0.122)	0.847(0.202)	1600
MaCH-Admix IBS-PW	**0.932(0.120)**	**0.826(0.184)**	60800	**0.940(0.119)**	**0.859(0.198)**	1876
MaCH-Admix IBS-SQ	0.927(0.118)	0.807(0.175)	61112	0.938(0.119)	0.849(0.199)	1877
MaCH-Admix IBS-DQ	0.930(0.115)	0.819(0.176)	61175	0.939(0.118)	0.854(0.197)	1876
IMPUTE2	0.924(0.120)	0.793(0.180)	52818	0.931(0.125)	0.828(0.216)	2579
BEAGLE	0.926(0.121)	0.806(0.189)	332586	0.929(0.130)	0.824(0.218)	14182
chromosome8:18-23Mb						
MaCH-Admix Random	0.896(0.155)	0.793(0.212)	75511	0.901(0.150)	0.821(0.225)	1899
MaCH-Admix IBS-PW	**0.911(0.143)**	**0.824(0.198)**	84885	**0.906(0.147)**	**0.833(0.221)**	2302
MaCH-Admix IBS-SQ	0.903(0.145)	0.797(0.200)	83051	0.903(0.149)	0.820(0.227)	2270
MaCH-Admix IBS-DQ	0.906(0.143)	0.805(0.200)	80794	0.904(0.147)	0.822(0.224)	2285
IMPUTE2	0.900(0.145)	0.773(0.206)	75001	0.893(0.159)	0.781(0.247)	3647
BEAGLE	0.905(0.142)	0.807(0.201)	498822	0.894(0.154)	0.800(0.232)	17146

All results were generated using default or suggested parameter values: MaCH-Admix: *--rounds* 30, *--states* 120, *--imputeStates* 500; IMPUTE2: *-iter* 30, *-k* 120, *-k_hap* 500; BEAGLE: *niterations*=10 *nsamples*=4. Running time is measured in seconds.

Table 4.3: Imputation Results of WHI-AA Individuals over Five 5Mb Regions with the 1000G reference

	All 8421 Individuals			Random 200 Subset		
	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time
chromosome3:80-85Mb						
MaCH-Admix Random	0.912(0.100)	0.782(0.150)	161637	0.932(0.091)	0.824(0.194)	897
MaCH-Admix IBS-PW	**0.947(0.073)**	**0.850(0.158)**	174083	**0.945(0.083)**	0.849(0.194)	1026
MaCH-Admix IBS-SQ	0.944(0.075)	0.844(0.161)	176147	0.942(0.086)	0.835(0.198)	1035
MaCH-Admix IBS-DQ	0.946(0.074)	0.849(0.160)	169442	0.944(0.083)	**0.851(0.198)**	1021
IMPUTE2	0.943(0.075)	0.847(0.151)	111307	0.943(0.085)	0.836(0.187)	2017
BEAGLE	0.921(0.088)	0.795(0.170)	23082*	0.915(0.107)	0.784(0.217)	6435
chromosome1:75-80Mb						
MaCH-Admix Random	0.873(0.143)	0.703(0.219)	214385	0.886(0.141)	0.726(0.241)	1240
MaCH-Admix IBS-PW	**0.921(0.106)**	0.802(0.176)	226019	**0.906(0.128)**	**0.770(0.232)**	1530
MaCH-Admix IBS-SQ	0.915(0.109)	0.794(0.174)	232880	0.900(0.130)	0.756(0.224)	1504
MaCH-Admix IBS-DQ	0.918(0.106)	0.803(0.168)	232858	0.903(0.131)	0.762(0.235)	1476
IMPUTE2	0.917(0.103)	**0.810(0.157)**	138080	0.898(0.135)	0.760(0.240)	2412
BEAGLE	0.892(0.119)	0.759(0.173)	25618*	0.875(0.145)	0.713(0.242)	8621
chromosome4:57-62Mb						
MaCH-Admix Random	0.883(0.126)	0.688(0.187)	241045	0.905(0.111)	0.749(0.169)	1290
MaCH-Admix IBS-PW	**0.927(0.092)**	0.795(0.159)	260231	**0.922(0.100)**	0.792(0.175)	1508
MaCH-Admix IBS-SQ	0.920(0.094)	0.782(0.148)	254002	0.915(0.105)	0.777(0.180)	1545
MaCH-Admix IBS-DQ	0.924(0.090)	**0.796(0.138)**	248524	0.920(0.100)	**0.793(0.175)**	1478
IMPUTE2	0.918(0.091)	0.787(0.129)	166642	0.912(0.104)	0.778(0.168)	2939
BEAGLE	0.898(0.109)	0.735(0.167)	43573*	0.892(0.131)	0.738(0.222)	14528
chromosome14:50-55Mb						
MaCH-Admix Random	0.875(0.140)	0.726(0.216)	240789	0.908(0.120)	0.807(0.198)	1663
MaCH-Admix IBS-PW	**0.921(0.105)**	**0.823(0.171)**	254530	**0.927(0.104)**	**0.852(0.167)**	1900
MaCH-Admix IBS-SQ	0.914(0.108)	0.809(0.172)	253231	0.919(0.112)	0.835(0.191)	1918
MaCH-Admix IBS-DQ	0.918(0.105)	0.818(0.168)	254555	0.924(0.107)	0.850(0.175)	1900
IMPUTE2	0.912(0.106)	0.815(0.157)	143772	0.913(0.116)	0.820(0.186)	2575
BEAGLE	0.893(0.118)	0.775(0.176)	27666*	0.899(0.127)	0.786(0.216)	14139
chromosome8:18-23Mb						
MaCH-Admix Random	0.830(0.177)	0.682(0.235)	343104	0.857(0.163)	0.735(0.235)	1977
MaCH-Admix IBS-PW	**0.889(0.142)**	**0.798(0.207)**	357858	**0.884(0.148)**	**0.800(0.218)**	2377
MaCH-Admix IBS-SQ	0.882(0.145)	0.789(0.207)	347473	0.877(0.152)	0.786(0.224)	2393
MaCH-Admix IBS-DQ	0.885(0.144)	0.795(0.205)	356928	0.881(0.149)	0.797(0.220)	2318
IMPUTE2	0.884(0.140)	0.795(0.194)	211879	0.876(0.153)	0.795(0.218)	3618
BEAGLE	0.858(0.151)	0.743(0.206)	43068*	0.856(0.158)	0.767(0.229)	16931

* In my experiments, BEAGLE cannot finish imputation with the complete 1000G references within 7 days which is the hard limit on my cluster server. I thus restrict the markers in the reference panel to be the set of Affymetrix 6.0 markers plus 2.5% of the remaining 1000G markers. The size of the restricted set in each region is about $10 \sim 15\%$ of the size of original 1000G marker set.

All results were generated using default or suggested parameter values: MaCH-Admix: *--rounds* 30, *--states* 120, *--imputeStates* 500; IMPUTE2: *-iter* 30, *-k* 120, *-k_hap* 500; BEAGLE: *niterations*=10 *nsamples*=4. Running time is measured in seconds.

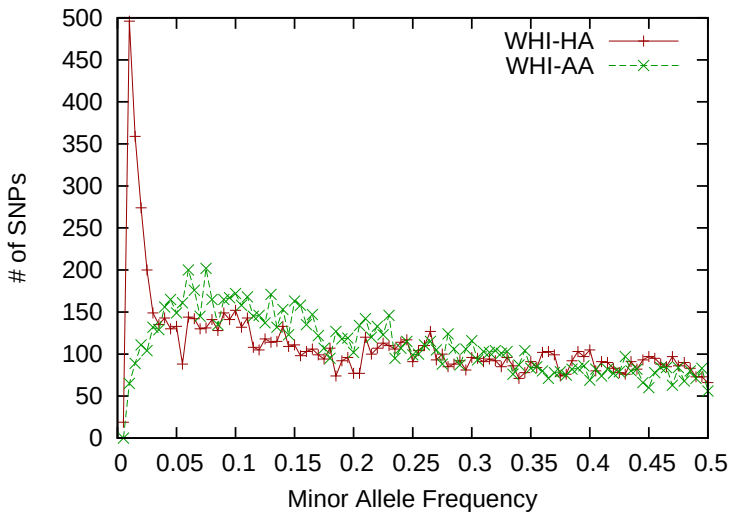
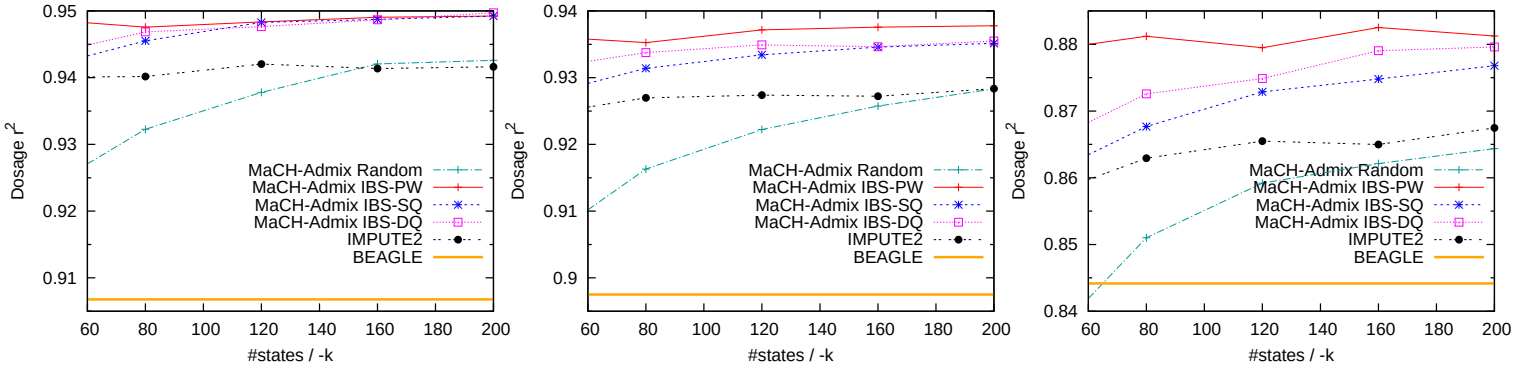


Figure 4.5: Minor Allele Frequency (MAF) distribution of SNPs in WHI-AA and WHI-HA.

4.3.2 HapMap ASW and MEX with the 1000G Reference

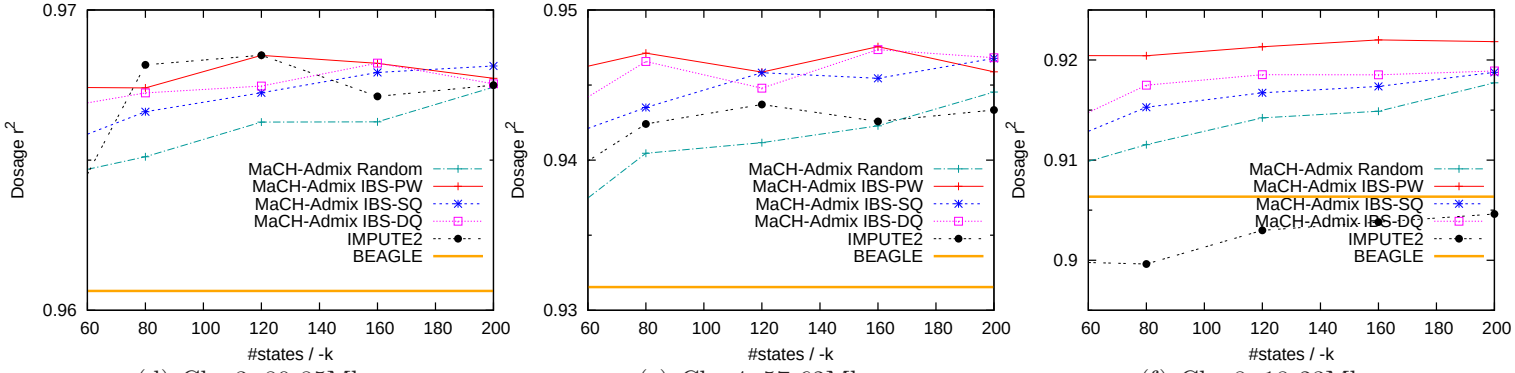
In this setting, I use a large reference panel to impute two small target sets. Figure 4.6 shows the imputation quality of three regions for both ASW and MEX. The complete results are presented in Table 4.4. Similar to previous experiments, I found that IBS-PW is very effective in finding the most relevant reference from a large panel (1000G) and clearly outperforms the other methods. IMPUTE2 again shows a flatter curve in most regions. Random selection and BEAGLE tend to perform worse than the IBS-based methods. This again proves that IBS-based selections are very effective in working with large reference panels.

In imputing ASW individuals, the standard error of my IBS-PW method ranges from 0.0034 to 0.0041 for all variants, and from 0.0189 to 0.0276 for uncommon variants. In imputing MEX individuals, the standard error of IBS-PW ranges from 0.0031 to 0.0043 for all variants, and from 0.0132 to 0.0211 for uncommon variants.



(a) Chr 3, 80-85Mb (b) Chr 4, 57-62Mb (c) Chr 8, 18-23Mb

A: Overall imputation quality of HapMap ASW with the 1000G reference panel



(d) Chr 3, 80-85Mb (e) Chr 4, 57-62Mb (f) Chr 8, 18-23Mb

B: Overall imputation quality of HapMap MEX with the 1000G reference panel

Figure 4.6: Imputation of 49 HapMap ASW and 50 HapMap MEX individuals with the 1000G reference panel. Imputation quality (measured by dosage r^2) is plotted as a function of the effective reference panel size (i.e., #states), for WHI-AA individuals in three selected 5Mb regions (ordered by LD from high to low).

4.3.3 Imputation Performance with HapMap References

First, consistent with what has been reported that imputation quality improves with reference panel size, imputation quality is indeed lower with HapMap references than with the 1000G reference. For example, average dosage r^2 is 90.0-91.3% with the 1000G reference (Table 4.2) for WHI-HA individuals in the chromosome4:57-62Mb region but drops to 84.4-86.2% with HapMapII references (Table 4.5). Second, difference among various methods is much smaller with these smaller HapMap reference sets ($H = 240 \sim 930$), which is consistent with my intuition that, given fixed computational costs, reference selection makes more pronounced difference with large reference panel since only a small portion of reference can be selected.

4.3.3.1 WHI-HA and WHI-AA with HapMap references

The complete results are presented in Tables 4.5 and 4.6. In WHI-HA (Table 4.5, $H = 420$), IBS-PW outperforms IBS-SQ and IBS-DQ slightly and the advantage disappears in WHI-AA (Table 4.6, $H = 240$). MaCH-Admix and IMPUTE2 yield similar imputation accuracy, and both outperform BEAGLE slightly.

4.3.3.2 HapMap ASW and MEX with HapMap references

For ASW, I experimented with three reference panels: HapMapII CEU+YRI, HapMapIII CEU+YRI, and HapMapIII CEU+YRI+LWK+MKK; for MEX two reference panels: HapMapII CEU+YRI+JPT+CHB and HapMapIII CEU+YRI+JPT+CHB. Results for ASW with HapMapIII CEU+YRI+LWK+MKK as the reference are shown in Figure 4.7 (the same three selected regions). The remaining results are presented in Tables 4.7, 4.8 and 4.9. Again, MaCH-Admix and IMPUTE2 yield similar imputation accuracy, both outperform BEAGLE slightly. IBS-PW is still an obvious winner in most regions and settings. But the relative difference among different methods diminishes when H is small.

Table 4.4: Imputation Results of HapMap ASW & MEX Individuals over Five 5Mb Regions with the 1000G reference ($H = 2188$)

	49 ASW Individuals			50 MEX Individuals		
	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time
chromosome3:80-85Mb						
MaCH-Admix Random	0.937(0.104)	0.854(0.210)	189	0.966(0.080)	0.960(0.149)	173
MaCH-Admix IBS-PW	**0.948(0.095)**	0.888(0.192)	252	**0.968(0.077)**	**0.968(0.148)**	212
MaCH-Admix IBS-SQ	**0.948(0.091)**	**0.898(0.176)**	220	0.967(0.079)	0.961(0.148)	203
MaCH-Admix IBS-DQ	0.947(0.095)	0.889(0.190)	220	0.967(0.079)	0.963(0.149)	221
IMPUTE2	0.942(0.106)	0.877(0.201)	457	**0.968(0.086)**	0.953(0.187)	477
BEAGLE	0.906(0.137)	0.774(0.267)	2388	0.960(0.096)	0.938(0.196)	2760
chromosome1:75-80Mb						
MaCH-Admix Random	0.915(0.135)	0.828(0.233)	273	0.937(0.132)	0.854(0.238)	257
MaCH-Admix IBS-PW	**0.930(0.123)**	**0.859(0.216)**	329	**0.940(0.134)**	0.867(0.250)	302
MaCH-Admix IBS-SQ	0.926(0.128)	0.849(0.222)	331	0.938(0.130)	**0.870(0.235)**	293
MaCH-Admix IBS-DQ	0.928(0.127)	0.852(0.227)	330	0.938(0.132)	0.866(0.243)	299
IMPUTE2	0.915(0.140)	0.842(0.229)	609	0.933(0.140)	0.847(0.270)	549
BEAGLE	0.900(0.148)	0.817(0.245)	3195	0.931(0.144)	0.839(0.264)	3779
chromosome4:57-62Mb						
MaCH-Admix Random	0.922(0.116)	0.801(0.230)	283	0.941(0.127)	0.873(0.228)	244
MaCH-Admix IBS-PW	**0.937(0.107)**	**0.852(0.220)**	325	**0.945(0.118)**	**0.896(0.203)**	298
MaCH-Admix IBS-SQ	0.933(0.110)	0.837(0.224)	322	**0.945(0.116)**	0.894(0.200)	286
MaCH-Admix IBS-DQ	0.934(0.107)	0.845(0.215)	325	0.944(0.119)	0.883(0.210)	290
IMPUTE2	0.927(0.116)	0.819(0.238)	743	0.943(0.120)	0.889(0.207)	785
BEAGLE	0.897(0.144)	0.755(0.284)	4364	0.931(0.143)	0.839(0.263)	5677
chromosome14:50-55Mb						
MaCH-Admix Random	0.899(0.144)	0.739(0.280)	392	0.947(0.119)	0.891(0.218)	366
MaCH-Admix IBS-PW	**0.914(0.134)**	0.769(0.273)	438	**0.951(0.118)**	**0.900(0.218)**	420
MaCH-Admix IBS-SQ	0.909(0.138)	0.765(0.282)	419	0.948(0.120)	0.896(0.223)	420
MaCH-Admix IBS-DQ	0.909(0.135)	0.763(0.264)	438	0.947(0.122)	0.889(0.231)	429
IMPUTE2	0.901(0.145)	**0.770(0.281)**	636	0.940(0.126)	0.874(0.234)	562
BEAGLE	0.879(0.167)	0.677(0.325)	4868	0.939(0.128)	0.872(0.232)	4643
chromosome8:18-23Mb						
MaCH-Admix Random	0.859(0.172)	0.755(0.283)	420	0.914(0.145)	0.892(0.200)	404
MaCH-Admix IBS-PW	**0.879(0.162)**	**0.792(0.280)**	523	**0.921(0.140)**	**0.908(0.186)**	487
MaCH-Admix IBS-SQ	0.872(0.164)	0.775(0.282)	537	0.916(0.145)	0.898(0.197)	485
MaCH-Admix IBS-DQ	0.874(0.166)	0.774(0.293)	526	0.918(0.143)	0.901(0.196)	495
IMPUTE2	0.865(0.173)	0.767(0.298)	818	0.902(0.164)	0.864(0.247)	854
BEAGLE	0.844(0.181)	0.760(0.285)	6295	0.906(0.156)	0.875(0.233)	6509

All results were generated using the following parameter values: MaCH-Admix: --*rounds* 30, --*states* 120, --*imputeStates* 500; IMPUTE2: -*iter* 30, -*k* 120, -*k_hap* 500; BEAGLE: *niterations*=10 *nsamples*=4. Running time is measured in seconds. Best performance in each comparison is highlighted by bold font.

Table 4.5: Imputation Results of WHI-HA Individuals over Five 5Mb Regions with the HapMapII reference ($H = 420$)

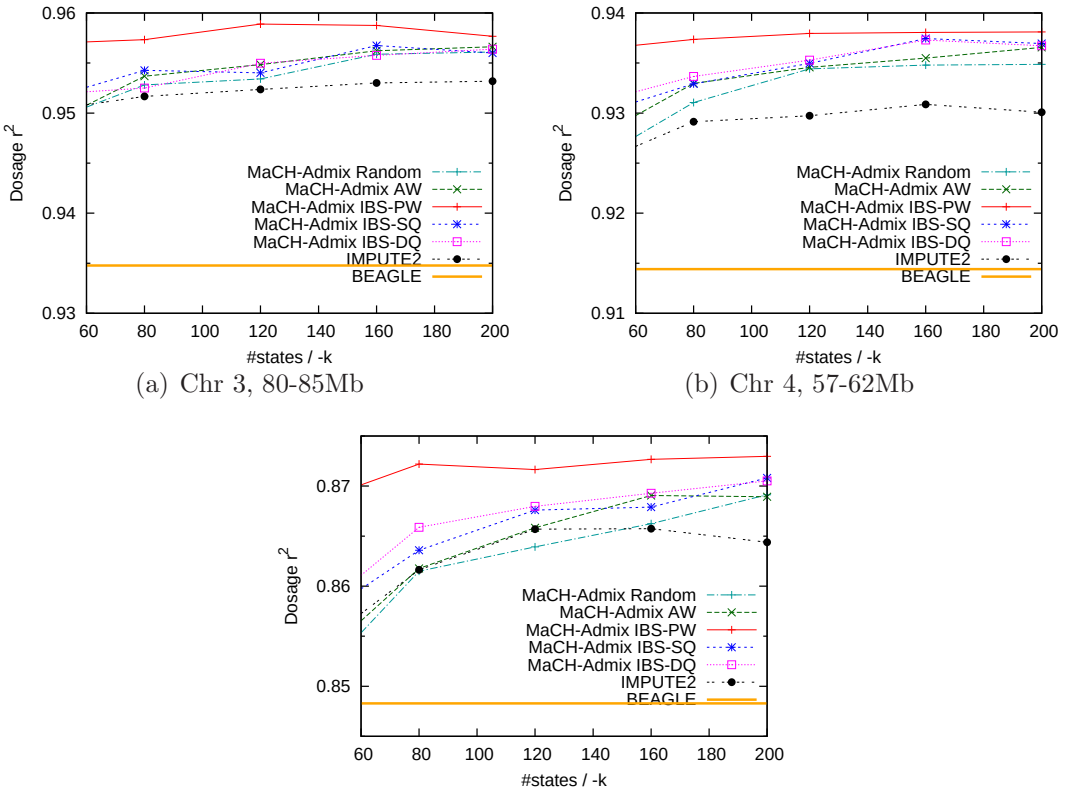
	All 3587 individuals			Random 200 Subset		
	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time
chromosome3:80-85Mb						
MaCH-Admix Random	0.897(0.157)	0.864(0.091)	9907	0.885(0.150)	0.807(0.101)	234
MaCH-Admix IBS-PW	**0.905(0.150)**	0.918(0.021)	10373	**0.888(0.150)**	0.831(0.081)	248
MaCH-Admix IBS-SQ	0.904(0.150)	0.913(0.033)	11303	0.887(0.150)	0.838(0.088)	246
MaCH-Admix IBS-DQ	0.904(0.150)	0.911(0.036)	10434	**0.888(0.147)**	**0.845(0.082)**	247
IMPUTE2	0.904(0.148)	**0.924(0.011)**	8874	0.887(0.144)	0.843(0.044)	403
BEAGLE	0.892(0.159)	0.902(0.062)	11877	0.873(0.164)	0.831(0.106)	232
chromosome1:75-80Mb						
MaCH-Admix Random	0.855(0.184)	0.752(0.222)	14227	0.857(0.185)	0.723(0.253)	350
MaCH-Admix IBS-PW	**0.863(0.176)**	0.762(0.201)	13328	0.859(0.183)	0.721(0.236)	367
MaCH-Admix IBS-SQ	0.860(0.179)	0.748(0.204)	15146	**0.860(0.181)**	0.715(0.234)	363
MaCH-Admix IBS-DQ	0.861(0.178)	0.750(0.204)	15878	0.858(0.183)	0.712(0.237)	377
IMPUTE2	0.842(0.188)	0.740(0.248)	11782	0.840(0.194)	0.701(0.282)	556
BEAGLE	0.851(0.186)	**0.792(0.230)**	15446	0.849(0.191)	**0.795(0.250)**	296
chromosome4:57-62Mb						
MaCH-Admix Random	0.852(0.169)	0.742(0.237)	14728	0.863(0.165)	0.775(0.210)	343
MaCH-Admix IBS-PW	**0.862(0.162)**	**0.764(0.217)**	17051	**0.869(0.162)**	**0.787(0.201)**	360
MaCH-Admix IBS-SQ	0.860(0.161)	0.756(0.223)	16123	0.868(0.162)	0.779(0.211)	362
MaCH-Admix IBS-DQ	0.860(0.161)	0.757(0.224)	15364	0.867(0.164)	0.786(0.205)	363
IMPUTE2	0.844(0.176)	0.717(0.231)	12369	0.847(0.180)	0.732(0.221)	541
BEAGLE	0.850(0.168)	0.740(0.234)	17503	0.851(0.174)	0.734(0.263)	348
chromosome14:50-55Mb						
MaCH-Admix Random	0.845(0.190)	0.669(0.285)	19813	0.850(0.191)	0.677(0.290)	428
MaCH-Admix IBS-PW	0.854(0.184)	**0.689(0.274)**	19214	0.854(0.186)	**0.690(0.273)**	448
MaCH-Admix IBS-SQ	0.852(0.184)	0.682(0.283)	18357	0.854(0.186)	0.678(0.289)	450
MaCH-Admix IBS-DQ	0.852(0.184)	0.686(0.278)	19201	**0.855(0.186)**	0.689(0.277)	453
IMPUTE2	**0.856(0.183)**	0.681(0.272)	14430	**0.855(0.187)**	0.686(0.286)	660
BEAGLE	0.846(0.186)	0.666(0.279)	17102	0.845(0.191)	0.641(0.327)	356
chromosome8:18-23Mb						
MaCH-Admix Random	0.826(0.216)	0.760(0.246)	22069	0.830(0.216)	0.754(0.244)	524
MaCH-Admix IBS-PW	0.838(0.211)	**0.775(0.240)**	21194	**0.838(0.213)**	**0.763(0.238)**	551
MaCH-Admix IBS-SQ	0.832(0.213)	0.765(0.241)	22098	0.833(0.213)	0.758(0.242)	551
MaCH-Admix IBS-DQ	0.833(0.213)	0.768(0.241)	22360	0.833(0.216)	0.750(0.243)	553
IMPUTE2	**0.839(0.207)**	0.772(0.236)	17910	0.835(0.214)	0.744(0.253)	875
BEAGLE	0.826(0.211)	0.742(0.245)	27236	0.822(0.215)	0.732(0.258)	543

All results were generated using the following parameter values: MaCH-Admix: *--rounds* 30, *--states* 120, *--imputeStates* 500; IMPUTE2: *-iter* 30, *-k* 120, *-k_hap* 500; BEAGLE: *niterations*=10 *nsamples*=4. Running time is measured in seconds. Best performance in each comparison is highlighted by bold font.

Table 4.6: Imputation Results of WHI-AA Individuals over Five 5Mb Regions with the HapMapII reference ($H = 240$)

	All 8421 Individuals			Random 200 Subset		
	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time
chromosome3:80-85Mb						
MaCH-Admix Random	0.877(0.140)	**0.684(0.271)**	56434	0.875(0.149)	0.636(0.354)	259
MaCH-Admix IBS-PW	0.884(0.136)	0.683(0.294)	52858	0.877(0.149)	0.641(0.369)	275
MaCH-Admix IBS-SQ	0.883(0.137)	0.678(0.294)	61142	0.877(0.148)	**0.645(0.356)**	264
MaCH-Admix IBS-DQ	0.883(0.137)	0.677(0.297)	56255	0.876(0.150)	0.627(0.349)	265
IMPUTE2	**0.885(0.135)**	0.668(0.290)	25283	**0.879(0.148)**	0.613(0.371)	388
BEAGLE	0.842(0.164)	0.575(0.259)	116113	0.841(0.173)	0.558(0.368)	234
chromosome1:75-80Mb						
MaCH-Admix Random	0.822(0.166)	0.746(0.146)	66325	0.811(0.174)	0.746(0.176)	394
MaCH-Admix IBS-PW	0.830(0.160)	0.759(0.143)	73130	**0.815(0.172)**	**0.757(0.194)**	407
MaCH-Admix IBS-SQ	0.830(0.160)	0.762(0.143)	75065	0.814(0.174)	0.746(0.200)	403
MaCH-Admix IBS-DQ	**0.831(0.160)**	**0.764(0.144)**	77968	**0.815(0.173)**	0.751(0.198)	402
IMPUTE2	0.812(0.167)	0.736(0.137)	35170	0.794(0.181)	0.712(0.167)	556
BEAGLE	0.798(0.185)	0.685(0.167)	142769	0.776(0.200)	0.656(0.222)	291
chromosome4:57-62Mb						
MaCH-Admix Random	0.832(0.150)	0.664(0.152)	77490	0.831(0.154)	0.679(0.177)	368
MaCH-Admix IBS-PW	0.841(0.144)	0.686(0.149)	74439	0.835(0.152)	0.693(0.177)	378
MaCH-Admix IBS-SQ	**0.842(0.143)**	0.689(0.143)	74604	**0.836(0.150)**	**0.704(0.169)**	400
MaCH-Admix IBS-DQ	**0.842(0.143)**	**0.691(0.142)**	76374	0.835(0.152)	0.693(0.159)	384
IMPUTE2	0.826(0.153)	0.654(0.160)	34875	0.816(0.162)	0.666(0.177)	513
BEAGLE	0.798(0.183)	0.552(0.271)	145240	0.788(0.199)	0.464(0.261)	298
chromosome14:50-55Mb						
MaCH-Admix Random	0.770(0.195)	0.628(0.288)	82618	0.780(0.199)	0.671(0.278)	427
MaCH-Admix IBS-PW	0.781(0.188)	0.645(0.279)	77589	0.784(0.195)	0.681(0.268)	442
MaCH-Admix IBS-SQ	0.780(0.187)	0.647(0.280)	82175	0.786(0.196)	0.679(0.262)	436
MaCH-Admix IBS-DQ	0.780(0.188)	0.644(0.283)	90951	0.787(0.194)	0.678(0.265)	450
IMPUTE2	**0.791(0.180)**	**0.667(0.270)**	39702	**0.789(0.194)**	**0.689(0.265)**	597
BEAGLE	0.742(0.210)	0.553(0.308)	124661	0.739(0.221)	0.579(0.315)	336
chromosome8:18-23Mb						
MaCH-Admix Random	0.754(0.222)	0.619(0.241)	99090	0.758(0.216)	0.649(0.233)	570
MaCH-Admix IBS-PW	0.764(0.217)	0.641(0.240)	104999	0.764(0.214)	0.665(0.230)	584
MaCH-Admix IBS-SQ	0.768(0.214)	0.654(0.235)	95685	0.765(0.213)	**0.677(0.232)**	593
MaCH-Admix IBS-DQ	0.768(0.213)	0.655(0.236)	104526	0.765(0.213)	0.672(0.236)	590
IMPUTE2	**0.779(0.203)**	**0.659(0.232)**	53975	**0.769(0.209)**	0.675(0.225)	869
BEAGLE	0.717(0.232)	0.535(0.243)	162132	0.709(0.237)	0.543(0.269)	452

All results were generated using the following parameter values: MaCH-Admix: --*rounds* 30, --*states* 120, --*imputeStates* 500; IMPUTE2: -*iter* 30, -*k* 120, -*k_hap* 500; BEAGLE: *niterations*=10 *nsamples*=4. Running time is measured in seconds. Best performance in each comparison is highlighted by bold font.



(a) Chr 3, 80-85Mb

(b) Chr 4, 57-62Mb

(c) Chr 8, 18-23Mb

Figure 4.7: Imputation quality of ASW with HapMapII CEU+YRI+LWK+MKK reference panel. Imputation quality (measured by dosage r^2) is plotted as a function of the effective reference panel size (i.e., #states), for ASW individuals in three selected 5Mb regions (ordered by LD from high to low).

I also included ancestry-weighted selection in evaluation in this setting because weights can be estimated stably given the relatively simple population structure in reference. Interestingly, I did not observe noticeable advantage of the ancestry-weighted selection method despite the obvious population structure within the reference panel and the target being admixed individuals. It however outperforms random selection slightly in most ASW experiments.

4.3.4 Running Time

Methods implemented in MaCH-Admix have comparable running time to that of IMPUTE2. BEAGLE has similar running time in experiments with HapMap references.

Table 4.7: Imputation Results of 49 ASW Individuals Over All Five Short Regions

	HapMapII CEU+YRI reference			HapMapIII CEU+YRI reference		
	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time
chromosome3:80-85Mb						
MaCH-Admix Random	0.937(0.106)	0.721(0.230)	168	0.942(0.121)	0.833(0.275)	138
MaCH-Admix AW	0.938(0.102)	0.766(0.191)	126	0.944(0.120)	0.837(0.275)	147
MaCH-Admix IBS-PW	**0.940(0.099)**	0.787(0.190)	125	**0.946(0.111)**	**0.860(0.249)**	158
MaCH-Admix IBS-SQ	0.939(0.100)	0.759(0.184)	128	0.943(0.121)	0.836(0.281)	158
MaCH-Admix IBS-DQ	0.937(0.106)	0.739(0.209)	132	**0.946(0.112)**	0.857(0.249)	149
IMPUTE2	0.939(0.099)	**0.803(0.155)**	288	0.942(0.119)	0.850(0.264)	316
BEAGLE	0.906(0.140)	0.702(0.276)	177	0.921(0.141)	0.796(0.296)	131
chromosome1:75-80Mb						
MaCH-Admix Random	**0.916(0.123)**	**0.862(0.201)**	197	0.921(0.135)	0.810(0.237)	250
MaCH-Admix AW	0.915(0.124)	0.853(0.209)	194	0.922(0.134)	0.812(0.236)	280
MaCH-Admix IBS-PW	0.915(0.123)	0.857(0.202)	199	**0.925(0.132)**	**0.826(0.234)**	258
MaCH-Admix IBS-SQ	0.914(0.125)	0.853(0.207)	209	0.923(0.132)	0.819(0.230)	243
MaCH-Admix IBS-DQ	0.914(0.125)	0.858(0.211)	206	0.923(0.135)	0.815(0.240)	246
IMPUTE2	0.914(0.131)	0.839(0.228)	441	0.919(0.140)	0.810(0.253)	442
BEAGLE	0.893(0.150)	0.824(0.245)	178	0.898(0.166)	0.777(0.285)	199
chromosome4:57-62Mb						
MaCH-Admix Random	0.898(0.138)	0.808(0.230)	188	0.920(0.125)	0.840(0.239)	226
MaCH-Admix AW	0.898(0.138)	0.814(0.231)	187	**0.922(0.123)**	**0.850(0.239)**	210
MaCH-Admix IBS-PW	**0.900(0.136)**	**0.821(0.231)**	203	**0.922(0.127)**	0.847(0.249)	234
MaCH-Admix IBS-SQ	0.899(0.141)	0.811(0.238)	192	0.920(0.127)	0.841(0.243)	230
MaCH-Admix IBS-DQ	0.899(0.140)	0.814(0.232)	317	0.921(0.127)	0.845(0.247)	228
IMPUTE2	0.897(0.140)	0.813(0.233)	415	0.920(0.128)	0.837(0.245)	452
BEAGLE	0.868(0.166)	0.775(0.252)	182	0.900(0.146)	0.803(0.280)	170
chromosome14:50-55Mb						
MaCH-Admix Random	0.869(0.180)	0.744(0.298)	227	0.876(0.179)	0.757(0.306)	504
MaCH-Admix AW	0.871(0.176)	**0.765(0.279)**	232	0.880(0.177)	0.766(0.304)	282
MaCH-Admix IBS-PW	**0.873(0.177)**	0.762(0.293)	249	**0.881(0.178)**	**0.769(0.304)**	296
MaCH-Admix IBS-SQ	**0.873(0.176)**	0.761(0.289)	240	0.879(0.178)	0.765(0.302)	311
MaCH-Admix IBS-DQ	**0.873(0.176)**	0.757(0.293)	249	0.878(0.180)	0.757(0.312)	310
IMPUTE2	0.870(0.180)	0.756(0.289)	497	0.879(0.180)	0.766(0.301)	523
BEAGLE	0.841(0.199)	0.688(0.332)	189	0.849(0.201)	0.694(0.340)	214
chromosome8:18-23Mb						
MaCH-Admix Random	0.861(0.170)	0.813(0.247)	329	0.849(0.189)	0.766(0.285)	392
MaCH-Admix AW	**0.863(0.171)**	**0.824(0.249)**	332	0.850(0.188)	0.765(0.288)	423
MaCH-Admix IBS-PW	0.862(0.170)	**0.824(0.247)**	332	**0.853(0.189)**	0.761(0.296)	423
MaCH-Admix IBS-SQ	0.861(0.172)	0.819(0.246)	373	0.849(0.191)	**0.778(0.290)**	508
MaCH-Admix IBS-DQ	0.862(0.171)	0.821(0.239)	344	0.849(0.190)	0.776(0.289)	418
IMPUTE2	0.860(0.175)	0.793(0.263)	658	**0.853(0.194)**	0.767(0.299)	767
BEAGLE	0.820(0.200)	0.728(0.303)	241	0.825(0.206)	0.732(0.309)	269

All results were generated using the following parameter values: MaCH-Admix: *--rounds* 30, *--states* 120; IMPUTE2: *-iter* 30, *-k* 120, *-k_hap* 500; BEAGLE: *niterations*=10 *nsamples*=4. Best performance in each comparison is highlighted by bold font.

Table 4.8: Imputation Results of 49 ASW Individuals Over All Five Short Regions

	HapMapIII CEU+YRI+LWK+MKK reference		
	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time
chromosome3:80-85Mb			
MaCH-Admix Random	0.953(0.101)	0.868(0.232)	162
MaCH-Admix AW	0.954(0.097)	0.881(0.222)	159
MaCH-Admix IBS-PW	**0.958(0.091)**	**0.898(0.208)**	167
MaCH-Admix IBS-SQ	0.954(0.100)	0.871(0.233)	179
MaCH-Admix IBS-DQ	0.954(0.100)	0.876(0.233)	173
IMPUTE2	0.952(0.100)	0.877(0.225)	291
BEAGLE	0.934(0.124)	0.811(0.271)	334
chromosome1:75-80Mb			
MaCH-Admix Random	0.932(0.122)	0.837(0.222)	236
MaCH-Admix AW	0.935(0.119)	0.847(0.217)	238
MaCH-Admix IBS-PW	**0.939(0.117)**	**0.858(0.222)**	283
MaCH-Admix IBS-SQ	0.935(0.124)	0.841(0.235)	270
MaCH-Admix IBS-DQ	0.935(0.120)	0.850(0.226)	272
IMPUTE2	0.932(0.124)	0.846(0.225)	553
BEAGLE	0.918(0.144)	0.819(0.259)	491
chromosome4:57-62Mb			
MaCH-Admix Random	0.934(0.107)	0.885(0.200)	232
MaCH-Admix AW	0.934(0.110)	0.884(0.208)	251
MaCH-Admix IBS-PW	**0.937(0.106)**	**0.892(0.200)**	253
MaCH-Admix IBS-SQ	0.934(0.110)	0.879(0.211)	247
MaCH-Admix IBS-DQ	0.935(0.109)	0.878(0.210)	267
IMPUTE2	0.929(0.120)	0.861(0.237)	426
BEAGLE	0.914(0.132)	0.833(0.256)	469
chromosome14:50-55Mb			
MaCH-Admix Random	0.883(0.170)	0.756(0.301)	318
MaCH-Admix AW	0.886(0.168)	0.772(0.295)	309
MaCH-Admix IBS-PW	0.891(0.167)	0.778(0.304)	352
MaCH-Admix IBS-SQ	0.889(0.166)	0.783(0.295)	320
MaCH-Admix IBS-DQ	0.890(0.166)	**0.786(0.294)**	335
IMPUTE2	**0.893(0.168)**	0.785(0.303)	642
BEAGLE	0.873(0.181)	0.757(0.305)	514
chromosome8:18-23Mb			
MaCH-Admix Random	0.863(0.178)	0.781(0.274)	431
MaCH-Admix AW	0.865(0.180)	0.788(0.285)	417
MaCH-Admix IBS-PW	**0.871(0.177)**	0.790(0.286)	452
MaCH-Admix IBS-SQ	0.867(0.180)	**0.800(0.281)**	479
MaCH-Admix IBS-DQ	0.867(0.178)	0.785(0.281)	462
IMPUTE2	0.865(0.186)	**0.800(0.286)**	923
BEAGLE	0.848(0.190)	0.768(0.292)	718

All results were generated using the following parameter values: MaCH-Admix: *--rounds* 30, *--states* 120; IMPUTE2: *-iter* 30, *-k* 120, *-k_hap* 500; BEAGLE: *niterations*=10 *nsamples*=4. Best performance in each comparison is highlighted by bold font.

Table 4.9: Imputation Results of 50 MEX Individuals Over All Five Short Regions

	HapMapII CEU+YRI+JPT+CHB reference			HapMapIII CEU+YRI+JPT+CHB reference		
	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time	overall dosage r^2 (std dev)	uncommon SNPs dosage r^2(std dev)	running time
chromosome3:80-85Mb						
MaCH-Admix Random	**0.965(0.083)**	0.988(0.040)	114	0.956(0.112)	0.893(0.227)	143
MaCH-Admix AW	**0.965(0.080)**	0.985(0.054)	120	**0.957(0.109)**	0.898(0.216)	144
MaCH-Admix IBS-PW	0.964(0.082)	0.989(0.037)	125	**0.957(0.110)**	**0.899(0.222)**	184
MaCH-Admix IBS-SQ	0.964(0.081)	0.987(0.046)	124	**0.957(0.110)**	0.897(0.221)	164
MaCH-Admix IBS-DQ	0.963(0.083)	0.989(0.042)	122	0.956(0.112)	0.896(0.223)	167
IMPUTE2	0.961(0.089)	0.986(0.036)	298	**0.957(0.119)**	0.898(0.237)	311
BEAGLE	0.959(0.093)	**0.995(0.012)**	225	0.947(0.130)	0.854(0.245)	232
chromosome1:75-80Mb						
MaCH-Admix Random	0.927(0.136)	0.832(0.244)	192	0.923(0.165)	0.818(0.296)	255
MaCH-Admix AW	0.929(0.134)	0.827(0.240)	186	0.924(0.168)	0.814(0.306)	248
MaCH-Admix IBS-PW	**0.930(0.134)**	**0.838(0.245)**	209	**0.926(0.169)**	0.819(0.312)	272
MaCH-Admix IBS-SQ	0.926(0.136)	**0.838(0.221)**	203	0.921(0.171)	**0.829(0.308)**	251
MaCH-Admix IBS-DQ	0.926(0.139)	0.832(0.230)	220	0.922(0.170)	0.822(0.309)	262
IMPUTE2	0.927(0.141)	0.820(0.250)	471	0.923(0.177)	0.801(0.317)	476
BEAGLE	0.915(0.146)	0.806(0.245)	239	0.908(0.191)	0.775(0.338)	299
chromosome4:57-62Mb						
MaCH-Admix Random	0.928(0.147)	0.806(0.296)	183	**0.928(0.160)**	0.840(0.286)	219
MaCH-Admix AW	**0.929(0.146)**	0.806(0.286)	189	0.927(0.162)	0.838(0.289)	214
MaCH-Admix IBS-PW	0.928(0.149)	0.802(0.304)	200	0.927(0.161)	0.844(0.287)	238
MaCH-Admix IBS-SQ	0.928(0.148)	**0.812(0.286)**	286	0.926(0.163)	**0.851(0.288)**	235
MaCH-Admix IBS-DQ	0.927(0.149)	0.809(0.292)	193	**0.928(0.161)**	0.839(0.291)	238
IMPUTE2	0.925(0.156)	0.806(0.300)	435	0.925(0.169)	0.832(0.298)	501
BEAGLE	0.920(0.160)	0.793(0.305)	230	0.919(0.172)	0.824(0.304)	320
chromosome14:50-55Mb						
MaCH-Admix Random	**0.922(0.158)**	0.895(0.167)	249	0.916(0.183)	0.823(0.290)	347
MaCH-Admix AW	0.921(0.161)	0.902(0.168)	273	0.915(0.183)	0.816(0.292)	286
MaCH-Admix IBS-PW	**0.922(0.163)**	0.900(0.171)	252	**0.918(0.182)**	0.827(0.293)	335
MaCH-Admix IBS-SQ	0.921(0.161)	**0.903(0.168)**	273	0.915(0.183)	0.828(0.286)	316
MaCH-Admix IBS-DQ	0.920(0.161)	0.898(0.166)	263	0.917(0.181)	**0.840(0.287)**	315
IMPUTE2	**0.922(0.165)**	0.901(0.169)	541	0.916(0.182)	0.827(0.290)	598
BEAGLE	0.911(0.170)	0.891(0.172)	276	0.908(0.190)	0.813(0.299)	319
chromosome8:18-23Mb						
MaCH-Admix Random	0.900(0.162)	0.852(0.233)	316	0.886(0.191)	0.824(0.284)	402
MaCH-Admix AW	0.901(0.160)	0.858(0.224)	336	0.885(0.196)	0.815(0.294)	401
MaCH-Admix IBS-PW	**0.903(0.159)**	0.867(0.218)	327	**0.888(0.197)**	**0.826(0.298)**	513
MaCH-Admix IBS-SQ	0.900(0.163)	0.863(0.223)	356	0.882(0.198)	0.817(0.298)	465
MaCH-Admix IBS-DQ	0.900(0.161)	0.864(0.212)	329	0.883(0.199)	0.813(0.301)	459
IMPUTE2	0.898(0.164)	**0.871(0.199)**	716	0.879(0.205)	0.811(0.302)	806
BEAGLE	0.889(0.169)	0.859(0.225)	340	0.870(0.211)	0.788(0.320)	434

All results were generated using the following parameter values: MaCH-Admix: *--rounds* 30, *--states* 120; IMPUTE2: *-iter* 30, *-k* 120, *-k_hap* 500; BEAGLE: *niterations*=10 *nsamples*=4. Best performance in each comparison is highlighted by bold font.

It however needs significantly more computing time than MaCH-Admix and IMPUTE2 when imputing with the 1000G reference, which I believe has to do with how consecutive untyped variants are modeled. Note that, due to the large number of experiments, I conducted all experiments on a big Linux cluster with more than 1000 CPUs. This leads to moderate fluctuations in running time over short regions due to I/O competition. But I obtain largely consistent conclusions across different experimental settings.

4.4 Discussion

In summary, the emergence of large reference panels calls for more efficient methods to utilize the rich resource. I have implemented two classes of reference-selection methods, namely IBS-based and ancestry-weighted approaches, to construct effective reference panels within previously described HMM and implemented them in software package MaCH-Admix for genetic imputation in admixed populations. I have performed systematic evaluations on large (WHI-AA and WHI-HA full sample with 8421 and 3587 individuals), medium (subset of 200 individuals from each of the two WHI admixed cohorts), and small (HapMap ASW and MEX with 49 and 50 founders respectively) target samples; using large (the latest 1000G with $H = 2188$) and small (HapMap with $H =$ 240-930) reference panels; and in five regions with different levels of LD. Compared with popular existing methods, MaCH-Admix demonstrates its advantage mostly because its piecewise algorithm takes potential changes in haplotype pattern sharing across regions into direct account (versus IMPUTE2 which adopts a whole-haplotype IBS matching approach) and because it does not reduce local haplotype complexity (versus BEAGLE which does so to gain computational efficiency). Based on my evaluations, I recommend the proposed piecewise IBS-based method, which demonstrates the best trade off between quality and computing time.

As the reference panel continues to grow rapidly (for example, the 1000 Genomes Project will generate ∼5,000 haplotypes within two years), approaches that can rapidly explore the entire reference pool will become increasingly appreciated. IBS-based approaches show such potential. As manifested by results from both WHI individuals and the HapMapIII individuals, IBS-based approaches can generate accurately imputed genotypes by preferentially selecting a small but different subset of ∼100 (corresponding to ∼5% for the current 1000G case where H=2188) haplotypes from the entire reference pool in each iteration. As computational costs increase quadratically with the effective number of haplotypes used in each iteration, such ∼95% reduction in the effective number of reference haplotypes corresponds to >99.5% reduction in computational investment.

Previous studies [Hao *et al.*, 2009; Li *et al.*, 2009; Shriner *et al.*, 2010; Zhang *et al.*, 2011; Seldin *et al.*, 2011] have recommended the use of a combined reference panel which pools haplotypes from all available reference populations (e.g., from the HapMap or the 1000 Genomes Projects), especially for populations that do not have a single best match reference population for increased imputation accuracy. Two forces working in opposite directions are introduced by including reference haplotypes from populations different from those in target samples in such a cosmopolitan panel: shared haplotype stretches (likely even shorter) that would increase imputation quality while noise added by including population-specific local haplotypes would harm imputation quality. Therefore, the recommendation of using a cosmopolitan panel to enhance imputation quality also applies to MaCH-Admix, conceptually more applicable because MaCH-Admix reduces the noise force by choosing local haplotypes that are most relevant into effective reference.

One key question concerns the optimal region size for imputation. From the perspective of including more LD information, particularly the long-range LD information that would be particularly critical for the imputation of uncommon variants, imputation over longer regions is desired. However, approaches that select reference haplotypes according to genetic matching between reference haplotypes and genotypes of target indi-

viduals across the entire region like whole-haplotype IBS-based methods will likely suffer from the change in genetic matching over a long region. For example, for both scenarios presented in Figure 4.1, there are two distinct sub-regions according to the matching pattern. Lumping them naively together, particularly using a single queue, may well lead to inferior performance as discussed earlier. I attempt to solve the problem by breaking the entire region into smaller pieces and within each piece selecting some reference haplotypes according to local genetic matching. This conceptually shares similarity with local ancestry adjustment in analysis of admixed populations [Wang *et al.*, 2011a]. Pasaniuc *et al.* [2011] also found local ancestry increases imputation accuracy. The proposed piecewise IBS based selection method is robust to imputation region size. I have evaluated the performance on whole chromosomes using ASW/MEX with HapMap references and found that both piecewise IBS and ancestry-weighted selection perform much better than whole-haplotype IBS based methods (data not shown). Between piecewise IBS and ancestry-weighted selections, the piecewise IBS method has advantage in most whole chromosome experiments and is very close to ancestry-weighted selection in the rest.

Ancestry-weighted approaches have been previously utilized to construct reference panels in admixed populations for tagSNP selection or imputation [Egyud *et al.*, 2009; Pasaniuc *et al.*, 2010; Pemberton *et al.*, 2008]. However, such reference panels created *a priori* induce two problems for imputation. First, haplotypes from contributing reference populations are literally duplicated, thus substantially increasing computational burden. Second, the same fixed pre-constructed reference haplotypes are to be used for all Markov iterations, preventing imputation algorithms from taking into account the uncertainty in creating the reference panel. My ancestry-weighted approach selects reference haplotypes probabilistically according to the estimated ancestry proportions and creates a *different* reference panel in each Markov iteration. This strategy ensures that all reference haplotypes to be selected when I run the Markov iterations long enough, thus avoiding both problems mentioned above. An attractive feature that I have added

to MaCH-Admix is a functionality to estimate ancestry proportions so that it can internally generate weights for ancestry-weighted approach without the need to install and call external programs. Although there exist many methods to infer ancestry including for example ***structure*** [Pritchard *et al.*, 2000], *HAPMIX* [Price *et al.*, 2009] and *GEDI-ADMX* [Pasaniuc *et al.*, 2009], I believe that researchers will find this build-in feature convenient. I found my estimates reasonably close to estimates from ***structure*** and working well for imputation purpose.

In this study, I have examined the performance of my proposed and other imputation methods in both Hispanics and African Americans. Between the two, Hispanics are known to have more complex LD structure because of three ancestral populations involved as opposed to two for African Americans. The more complex LD in Hispanics indeed makes it essential to more explicitly account for the larger variability in local ancestry (for example, using my proposed piecewise approach). The more complex LD and population substructure in Hispanics have prevented a lot of investigators from even attempting imputation. However, I observe similar if not slightly better imputation quality in the five regions examined, with an average dosage r^2 of 92.5% (81.8%) versus 92.1% (81.4%) for all (uncommon) SNPs in WHI-HA and WHI-AA respectively using my piecewise IBS approach. That imputation performance for Hispanics is comparable with that for African Americans is expected due to on average less African ancestry (where LD is the lowest and thus most challenging for imputation) in Hispanics compared to African Americans. Therefore, I highly encourage investigators working with Hispanics perform imputation as well.

Although in this work I propose the reference selection methods for imputation of admixed individuals, the methods can be directly applied to imputation in general for non-admixed populations by finding the best genetic match for each target individual. For the same reason, IBS-based methods tend to work better than ancestry-weighted approaches when between-individual variation among the target individuals is large (data

not shown). This is not surprising because IBS-based approaches select a different effective reference panel tailored for each target individual, rather than one uniform reference sampling setting for all target individuals as in the ancestry-weighted approach.

I have also attempted to examine common and uncommon genetic variants separately, using MAF 5% as cutoff. I observe more pronounced differences among the attempted methods with uncommon variants, suggesting that choice of reference selection methods matters more for uncommon variants. Due to the nature of the SNPs evaluated (either typed Affymetrix 6.0 markers for the WHI individuals, or HapMap markers) and the target sample size (49-50 for HapMapIII ASW and MEX), there are few really rare (MAF<1%) variants. Although several attempts have been made [Wang *et al.*, 2011b; Howie *et al.*, 2011; The International HapMap Consortium, 2010; Liu *et al.*, 2012], imputation quality for uncommon variants is far from being fully assessed and needs to be further evaluated when data from large scale sequencing efforts become available.

Last but clearly not the least point concerns computational efficiency. MaCH-Admix is very flexible in terms of the effective number of haplotypes used in each iteration and the number of iterations. Imputation accuracy depends on both parameters. Since computational cost increases quadratically with --*states* and linearly with --*rounds*, for practical purpose, I recommend using --*states* 100-120 and --*rounds* $\geq$20. I also have an option analogous to IMPUTE2's -*k_hap*, which increases computational costs linearly and even defaulting at a large value (500) contributes to only a small proportion of computing time. Between the two categories of approaches proposed, the ancestry-weighted approach requires only one-time up-front costs for the estimation of ancestry proportions. The IBS-based methods, on the other hand, require overhead costs at each iteration for calculating genetic similarities between individuals in the target population and the reference haplotypes. For both, the costs increase with the reference panel size. Finally, computational costs would increase only linearly with --*states* if I start with haplotypes of the target individuals, that is, for haplotype-to-haplotype (both reference and target

are in haplotypes) imputation as performed by software minimac. I plan to extend my proposed methods to minimac in the future.

Web Resources

Census fact for admixed populations,

http://quickfacts.census.gov/qfd/states/00000.html

The 1000 Genomes Project, http://www.1000genomes.org/

MaCH-Admix, http://www.unc.edu/~yunmli/MaCH-Admix/

MaCH, http://www.sph.umich.edu/csg/yli/mach/

IMPUTE, http://mathgen.stats.ox.ac.uk/impute/impute.html

BEAGLE, http://faculty.washington.edu/browning/BEAGLE/BEAGLE.html

structure: http://pritch.bsd.uchicago.edu/software.html

Chapter 5

Genotype Imputation of Metabochip SNPs in African Americans Using a Study Specific Reference Panel

5.1 Introduction

Genotype imputation has become standard practice to increase genome coverage and improve power in Genome-Wide Association Studies (GWAS) and meta-analysis [de Bakker *et al.*, 2008; Li *et al.*, 2009; Marchini and Howie, 2010]. The wealth of literature using genotype imputation has focused on using external reference panels (for example, phased haplotypes from the International HapMap Project [The International HapMap Consortium, 2007] or the 1000 Genomes Project [The 1000 Genomes Project Consortium, 2010]), largely in individuals of European ancestry, for inference of genotypes at common (MAF $>$ 0.05) genetic markers.

GWAS have identified $> 4,300$ genetic variants associated with human diseases and traits (`http://www.genome.gov/gwastudies/`) [Hindorff *et al.*, 2009]. Investigators across the world have begun efforts to fine map within regions where GWAS-identified SNPs reside, through dense genotyping (e.g., using region-centric or gene-centric chips like the Metabochip for metabolic related traits (`http://www.sph.umich.edu/csg/kang/MetaboChip/`), or the ITMAT-Broad-CARe [IBC]

for cardiovascular related traits, or the immunochip for immune related diseases) or sequencing. Furthermore, multiethnic genetic association studies have been recognized as potentially more powerful for both gene discovery and fine mapping [McCarthy *et al.*, 2008; Pulit *et al.*, 2010; Rosenberg *et al.*, 2010; Teo *et al.*, 2010] and some initial efforts have been carried out [He *et al.*, 2011; Keebler *et al.*, 2010; Lanktree *et al.*, 2009; Lettre *et al.*, 2011; Smith *et al.*, 2011; Waters *et al.*, 2009]. In addition, because GWAS-identified SNPs (mostly common) explain only a small proportion of overall heritability for most complex diseases and traits [Eichler *et al.*, 2010; Maher, 2008; Manolio *et al.*, 2009], whole-genome or whole-exome sequencing for rare SNPs and genetic variants other than SNPs (e.g., copy number variations, structural variants) are under way.

So far, there has been relatively little research on the performance of genotype imputation in this new context. My study provides a typical scenario where $8,421$ African Americans from the Women's Health Initiative [The WHI Study Group, 1998] SNP Health Association Resource (SHARe) were genotyped using the Affymetrix 6.0 genotyping platform. In an attempt to generalize genetic effects across racial groups, the Population Architecture using Genomics and Epidemiology (PAGE) consortium genotyped a subset of 1,962 African American WHI participants with data on multiple metabolic related phenotypes using the Metabochip [Matise *et al.*, 2011]. To increase the power to detect moderate to small genetic effects, I sought to impute the Metabochip SNPs in the remaining 6,459 individuals in WHI SHARe with Affymetrix 6.0 data only. Imputing SNPs in the fine mapping region tends to be more challenging because these SNPs tend to be rare and in low linkage disequilibrium (LD) with GWAS SNPs. Here I describe a pipeline for constructing study-specific reference panels using individuals genotyped or sequenced at a larger set of genetic markers (in this case, individuals genotyped using both Affymetrix 6.0 and Metabochip) and for imputation into individuals with genotype data at a subset of markers (in this case, individuals genotyped using Affymetrix 6.0 only). I benchmark the quality of my imputation in an African American population, for

SNPs on the Metabochip, a region-centric genotyping platform, with particular focus on low frequency SNPs (MAF down to 0.001), using a large study-specific reference panel containing 3,924 haplotypes. An African American sample poses a greater challenge for genotype imputation due to more complex LD patterns in African Americans compared with individuals of European ancestry [Egyud *et al.*, 2009; Shriner *et al.*, 2010], and in which comparatively less discovery work has been done.

I first describe how I constructed my study-specific reference panel using the 1,962 African American individuals with genotypes for both Affymetrix 6.0 and Metabochip SNPs and how I performed imputation of the Metabochip-only SNPs into the remaining 6,459 individuals. I then show several approaches through which I estimated imputation quality for SNPs in different MAF categories, with a special focus on less common (MAF: $0.01 - 0.05$) and rare (MAF < 0.01) variants. I provide practical guidelines regarding post-imputation quality control for different MAF categories, as well as for the inclusion of rare variants during imputation.

5.2 Materials and Methods

5.2.1 Pre-Imputation Quality Control

Prior to phasing and imputation, quality control was applied to both the Metabochip data and the GWAS data. Specifically, for the GWAS dataset (n = 6,459) I removed Affymetrix 6.0 SNPs with genotype call rates $< 90\%$ (m = 1,633), or Hardy-Weinberg exact test [Wigginton, et al. 2005] p-value $< 10^{-6}$ (m = 16,327), or MAF < 0.01 (m = 14,014), resulting in a 829,370 GWAS SNPs passing quality control criteria [Reiner *et al.*, 2011]. Separate quality control criteria were applied to the Metabochip SNPs, leading to 182,397 QC+ SNPs with genotype call rates $> 95\%$ and Hardy-Weinberg p-value $> 10^{-6}$ Individuals were excluded if they had a call rate below 95%, showed excess heterozygosity, were part of an apparent first-degree relative pair, or were ancestry outliers as determined

by Eigensoft [Price *et al.*, 2006]. Details can be found in the PAGE Metabochip platform paper [Buyske *et al.*, 2011].

5.2.2 General Pipeline for Reference Construction and Subsequent Imputation

Figure 5.1 shows schematically how imputation was performed. In the top left panel, I first merged genotypes from the Affymetrix GWAS panel (blue) and the Metabochip (yellow) SNPs genotyped as part of the PAGE study for the 1,962 reference individuals (i.e., individuals with genotype data from both platforms). I then reconstructed haplotypes encompassing both GWAS and Metabochip SNPs for the reference individuals, constituting the reference panel of 3,924 haplotypes. In the top right panel, haplotype reconstruction for target individuals (i.e., individuals with GWAS genotypes only) was carried out similarly, but at the GWAS markers only. Finally, a haplotype-to-haplotype (that is, data are in haplotype form for both the reference and target individuals) imputation was performed to generate estimated genotypes at the Metabochip SNPs for the 6,459 target individuals.

5.3 Results

5.3.1 Genomewide Imputation using Large Study-Specific Reference

After careful matching on strand (so that genotypes from both Affymetrix 6.0 and the Metabochip are on the same strand), SNP ID, genomic coordinates, and actual genotypes for SNPs in common, I had a merged set of 987,749 SNPs for the 1,962 reference individuals. The average concordance rate for the 23,703 SNPs in common was 99.7%. For discordant genotypes, I kept the GWAS genotypes to match those of the target indi-

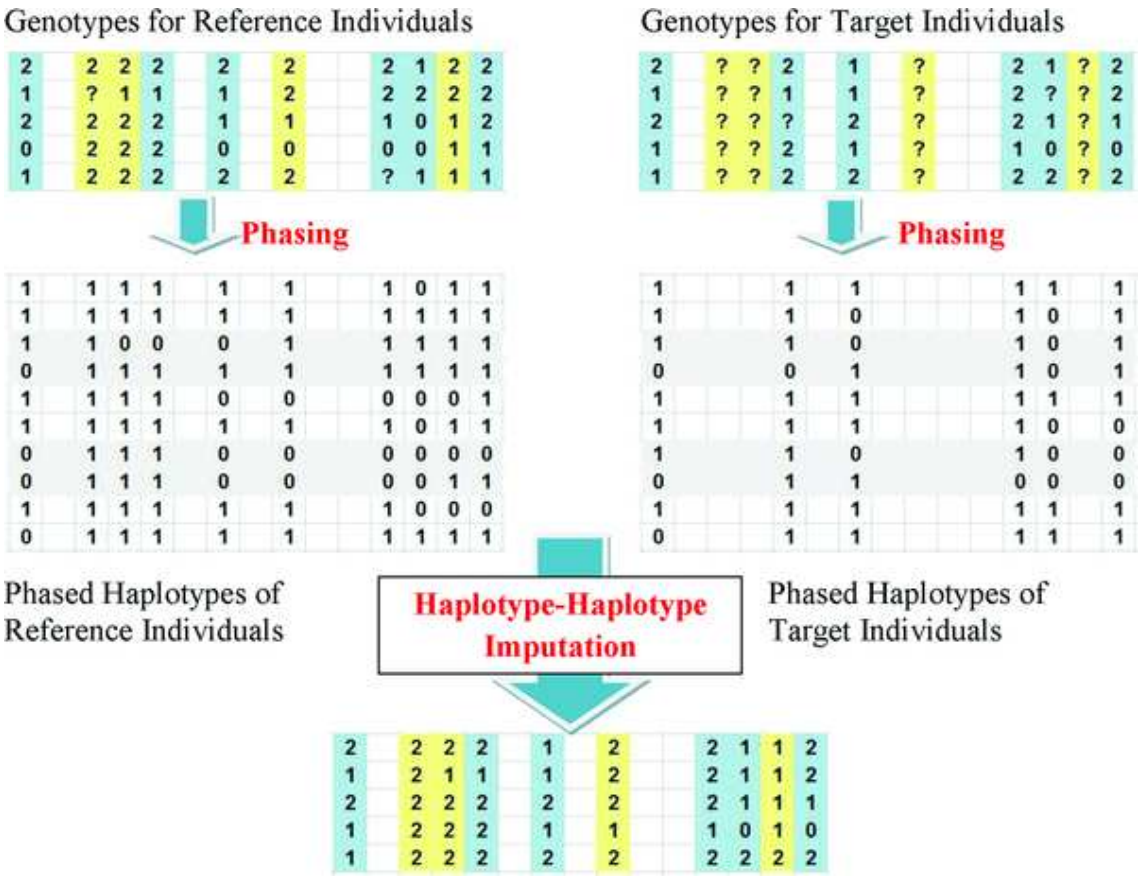


Figure 5.1: Reference construction and imputation pipeline using a study-specific reference panel. This schematic cartoon shows how I constructed my study-specific reference panel using five individuals genotyped on both the Affymetrix 6.0 and the Metabochip platform and how I performed imputation into the remaining five individuals with Affymetrix 6.0 data only.

viduals with GWAS data only. Haplotypes were reconstructed on the merged set using MaCH [Li *et al.*, 2010a]. In parallel, I constructed haplotypes across the 829,370 QC+ GWAS SNPs for all 8,421 individuals. Finally, I used the 3,924 haplotypes across the merged set of 987,749 SNPs as reference to impute into haplotypes across GWAS SNPs of the target individuals. The final haplotype-to-haplotype imputation was performed using the software package minimac, which generates the allele dosages (the fractional counts of an arbitrary allele at each SNP for each individual, ranging continuously from 0 to 2). Minimac also generates the SNP-level quality metric Rsq, which is the SNP-specific estimated r^2 between allele dosages and the unknown true genotypes. Rsq has been

recommended as an efficient post-imputation quality control metric. Rsq, estimated r^2, and estimated imputation r^2 are used interchangeably in the literature [Browning and Browning, 2009; Li *et al.*, 2009].

5.3.2 Quality Estimate by Masking Genotypes at 2% GWAS SNPs

Aside from production (actual imputation presented in the section above), I randomly masked 2% of the GWAS SNPs among the target individuals in the minimac imputation step to estimate the true imputation accuracy as well as to evaluate the utility of Rsq as a quality metric. By comparing imputed dosages with experimental genotypes, previous studies have proposed several statistics to measure true imputation accuracy [Browning and Browning, 2009; Li *et al.*, 2009; Lin *et al.*, 2010; Marchini and Howie, 2010], measuring either the concordance rate, correlation, or degree of agreement. Here, I choose to report the dosage r^2, which is the squared Pearson correlation between the estimated allele dosages and the true experimental genotypes (recoded as 0, 1, and 2 corresponding to the number of minor alleles), because it is a more informative measure for low frequency variants by taking allele frequency into account and because it is directly related to the effective sample size for subsequent association analysis [Pritchard and Przeworski, 2001]. As dosage r^2 is calculated using the true genotypes (assuming the experimental genotypes are the true genotypes), people also call it true r^2. Like Rsq, dosage r^2 is also specific to each SNP.

Figure 5.2 shows the average dosage r^2 values for the 2% masked GWAS SNPs by chromosome. Genomewide average is 93.68% (range 87.18% [chromosome 19] - 95.26% [chromosome 10]). As expected, larger chromosomes (in terms of physical length) tend to be slightly easier to impute due to slightly lower recombination rates and therefore higher level of LD [The International HapMap Consortium, 2005]. Chromosome 19, with the highest gene density, is most challenging for imputation. Table 5.1 shows the

average dosage r^2 values by MAF. Not surprisingly, lower frequency variants are harder to impute due to poorer coverage by GWAS SNPs, lower degree of LD, and more challenging haplotype reconstruction. For example, the average dosage r^2 for SNPs with MAF $>$ 0.05 is 95.08% ; while the average for SNPs with MAF 0.005-0.01 is 70.84

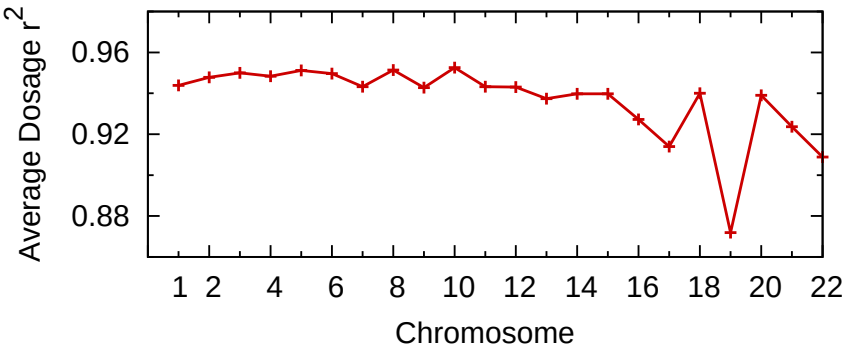


Figure 5.2: Imputation accuracy by chromosome for 2% randomly masked GWAS SNPs. Imputation accuracy (as measured by average dosage r^2) for 2% GWAS SNPs masked at random is plotted by chromosome.

Table 5.1: Average Dosage r^2 by MAF, Estimated by Masking 2% GWAS SNPs

MAF	#SNPs	Average Dosage r^2	Std Dev Dosage r^2
0.005-0.01	17	70.84%	18.23%
0.01-0.03	724	82.97%	16.07%
0.03-0.05	876	90.36%	11.03%
0.05-0.50	14983	95.08%	7.70%

While Figure 5.2 and Table 5.1 show the true imputation accuracy, in practice, researchers are more interested in how well imputation quality metrics can predict true imputation accuracy (measured by dosage r^2). Figure 5.3 assesses the quality metric Rsq by plotting it against dosage r^2. One can see that Rsq can predict dosage r^2 quite well, particularly for common SNPs and those with reasonable Rsq values. For example, the Pearson correlation is 0.938 for all SNPs (regardless of MAF and Rsq), 0.952 for SNPs with MAF $>$ 0.03 (regardless of Rsq), and 0.955 for SNPs with MAF $>$ 0.03 and Rsq $>$ 0.3.

Whereas masking GWAS SNPs is a simple approach to estimate imputation accuracy, the approach estimates imputation quality for the "wrong" set of SNPs in that I am imputing genotypes for Metabochip SNPs, not GWAS SNPs. The two set of SNPs differ in two major aspects: MAF and physical density distribution. First, in terms of allele frequency distribution: while Affymetrix 6.0 SNPs, like most commercially available genomewide genotyping platforms, contain SNPs that are mostly common, the Metabochip platform contains a much larger proportion of lower frequency variants. For example, while only 4.3% and 9.9% of the Affymetrix SNPs have MAF < 0.03 and < 0.05 respectively, the proportions are 29.8% and 37.8% for Metabochip SNPs. Figure 5.4 shows the MAF distributions of the Affymetrix 6.0 SNPs and the Metabochip SNPs. Second, the physical distribution of the SNPs is quite different. The Affymetrix 6.0 SNPs are rather evenly spread across the genome. SNPs on the Metabochip, chosen for fine mapping of regions identified through GWAS to be associated with metabolic related traits, scatter unevenly across the genome and are concentrated around GWAS-identified signals. Figure 5.5 shows two typical regions where the GWAS SNP density (green) is quite uniform across the region while Metabochip SNP density (red) peaks in a sub-region chosen for follow-up but drops sharply outside the sub-region of interest.

5.3.3 Quality Estimate by Masking Genotypes at Metabochip SNPs for a Subset of Reference Individuals

To estimate the imputation quality for the actually imputed Metabochip SNPs, I masked Metabochip genotypes for 100 reference individuals, imputed them, and compared the estimated dosages with the masked experimental genotypes. Note that I used haplotypes constructed from GWAS data only for the 100 individuals. Figure 5.6 shows the average dosage r^2 by chromosome. Again imputation quality is slightly higher for larger chromosomes and lowest for chromosome 19. Table 5.2 presents imputation accuracy by MAF, with and without post-imputation filtering according to Rsq. First, it is clear

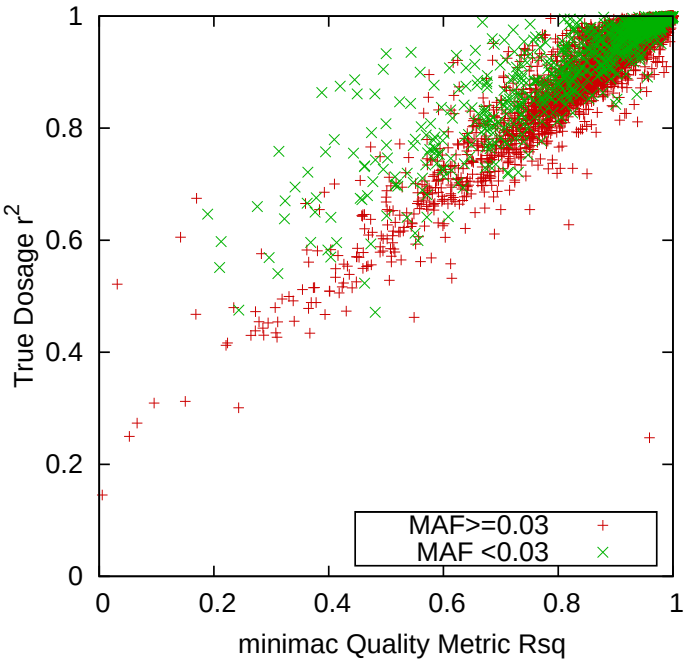


Figure 5.3: Rsq by dosage r^2 for 2% randomly masked GWAS SNPs. Estimated imputation accuracy (minimac output Rsq) is plotted against the true dosage r^2, for the 2% GWAS SNPs masked at random.

that lower frequency variants are harder to impute. Previous studies have shown earlier that imputation accuracy increases with the reference panel size, especially for the imputation of lower frequency variants [Li *et al.*, 2009; Marchini and Howie, 2010; The International HapMap Consortium, 2010]. However, even with a reference panel of 3,924 haplotypes, I am not able to obtain reasonable imputed data for SNPs with MAF under 0.001. Without post-imputation filtering, the average dosage r^2 is merely 0.39%. If I apply a post-imputation filter of Rsq > 0.3 (>0.5), only 0.4% (0.3%) of the SNPs with MAF < 0.001 pass the filter with an average dosage r^2 of 24.85% (30.45%). For this rarest category of SNPs (MAF < 0.001), even at an Rsq threshold of 0.95, which retains merely 23 out of 18,959 SNPs, I can only achieve an average dosage r^2 of 47.82% (Figure 5.7(a)). Second, SNPs with MAF > 0.01 can be imputed fairly well using a reference panel of this size. For example, even without any post-imputation quality control filter, the average dosage r^2 is 85.32%, 91.73%, and 94.62% for SNPs with MAF 0.01-0.03, 0.03-0.03, and

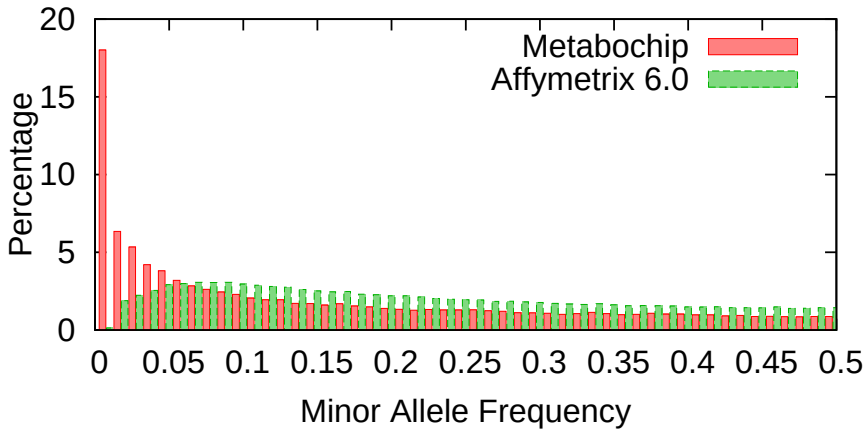
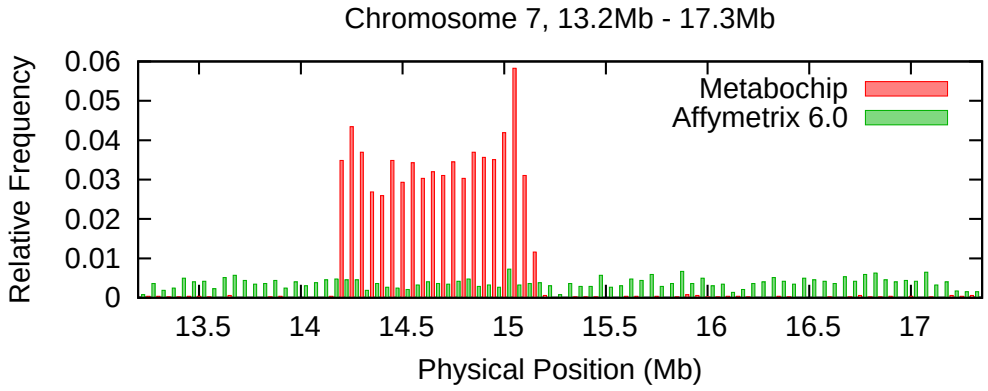
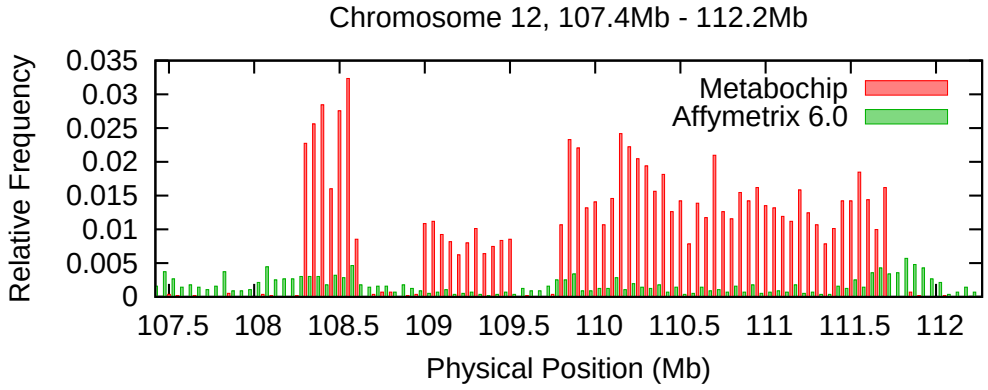


Figure 5.4: MAF distributions of Affymetrix 6.0 and Metabochip QC+ SNPs. I show the histograms for the MAFs of the 829,370 Affymetrix 6.0 QC+ SNPs (top panel) and of the 182,397 Metabochip QC+ SNPs (bottom panel).

>0.05, indicating that ∼85-95% of the information can be recovered for SNPs in these MAF categories. Third, I am able to impute a considerable proportion of less common (MAF 0.001-0.01) variants reasonably well using a reference panel of this size along with post-imputation quality filtering according to Rsq. For example, I can obtain an average dosage r^2 of 79.71% for 20.5% of the SNPs with MAF 0.001-0.005 by excluding SNPs with Rsq < 0.5; and an average dosage r^2 of 83.05% for 52.0% of the SNPs with MAF 0.005-0.01 by excluding SNPs with Rsq < 0.3, with both Rsq thresholds selected such that the average Rsq is above 80%.



(a)



(b)

Figure 5.5: Physical spreading of Affymetrix 6.0 and Metabochip QC+ SNPs. SNP frequency is plotted against genomic coordinate for two randomly chosen regions, green for Affymetrix 6.0 SNPs and red for Metabochip SNPs. The frequency is normalized so that the total frequency in each region is 1.

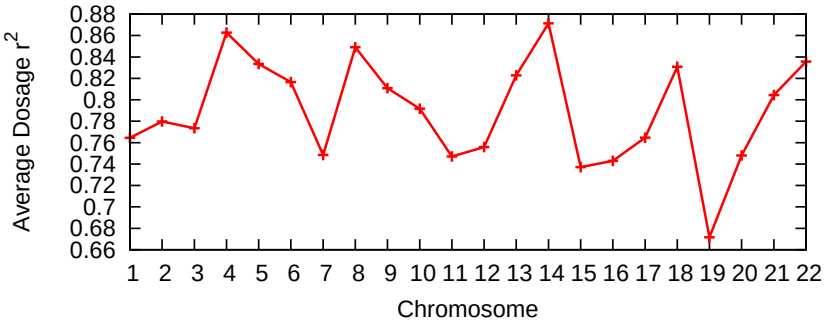


Figure 5.6: Imputation accuracy by chromosome for Metabochip SNPs (estimated by masking 100 reference individuals). Imputation accuracy (as measured by average dosage r^2) for Metabochip SNPs is plotted by chromosome, by masking 100 reference individuals

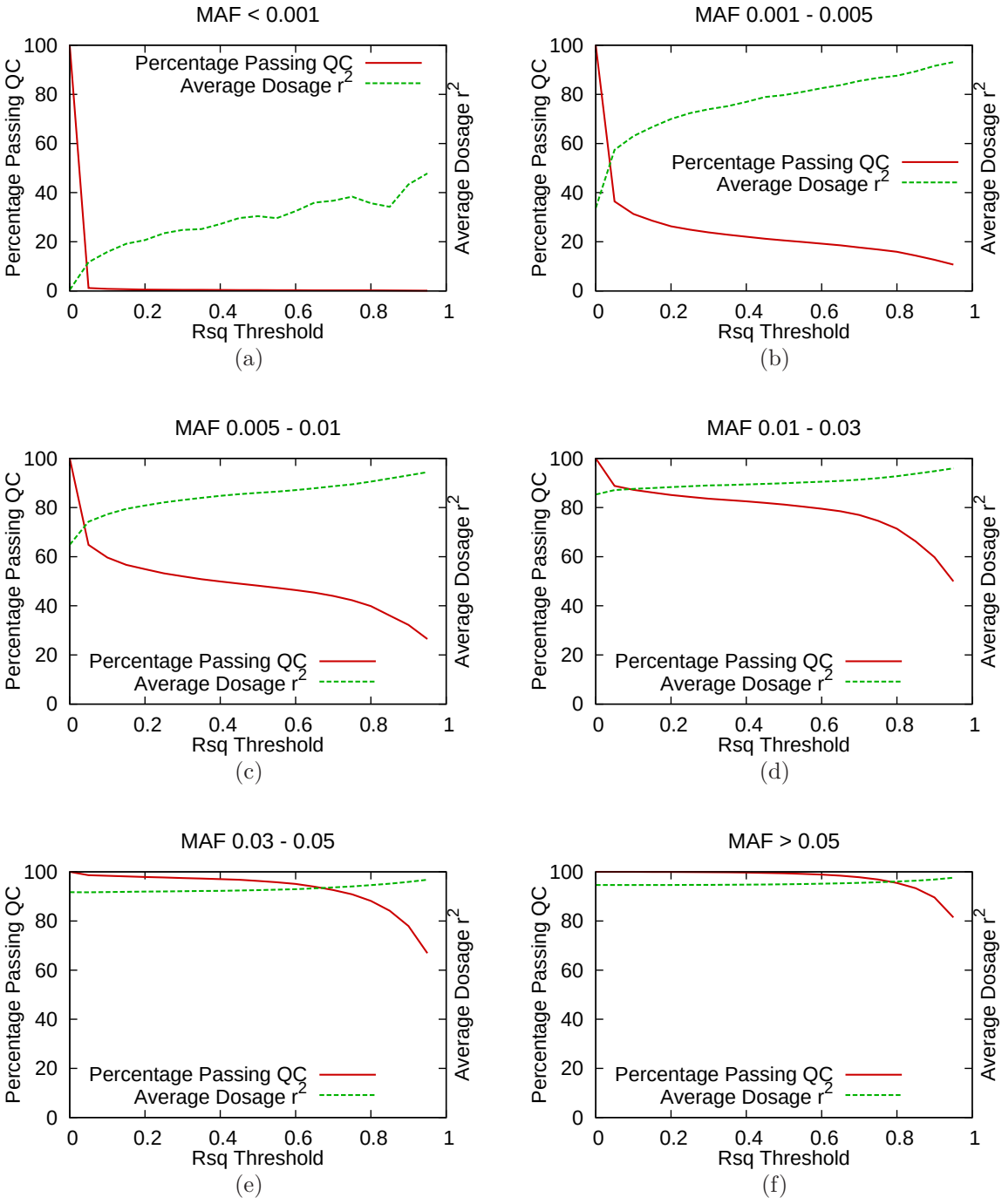


Figure 5.7: Accuracy and calibration of imputation. Percentages of SNPs passing post-imputation QC (left Y-axis) and average dosage r^2 (right Y-axis) are plotted against Rsq threshold used for post-imputation QC for SNPs in different MAF categories.

Table 5.2: Average Rsq and Dosage r^2 by MAF, Estimated by Masking 100 Reference Individuals

MAF	No Rsq Filter			Rsq > 0.3			Rsq > 0.5		
	#SNPs	average Rsq	average dosage r^2	%SNPs	average Rsq	average dosage r^2	%SNPs	average Rsq	average dosage r^2
0-0.001	18959	0.46%	0.39%	0.4%	72.31%	24.85%	0.3%	83.77%	30.45%
0.001-0.005	6925	21.80%	33.74%	23.8%	82.41%	73.94%	20.5%	89.24%	79.71%
0.005-0.01	7001	47.49%	64.87%	52.0%	87.32%	83.05%	48.2%	91.14%	86.00%
0.01-0.03	19894	77.57%	85.32%	83.6%	91.72%	88.98%	81.2%	93.21%	89.88%
0.03-0.05	13315	92.11%	91.73%	97.5%	94.27%	92.11%	96.3%	94.91%	92.57%
0.05-1.00	92597	96.94%	94.62%	99.9%	97.05%	94.71%	99.4%	97.30%	94.94%

Note: I evaluated a total of 158,691 out of the total 182,397 QC+ Metabochip SNPs because 23,706 SNPs are both on the Metabochip and the Affymetrix 6.0 panel and were excluded from quality evaluation to avoid upward bias.

5.3.4 Overall Imputation Performance and Practical Guidelines

In practice, I recommend using Rsq as the post-imputation quality control metric. Figure 5.8 attests to the high correlation between Rsq and dosage r^2. I observe that the vast majority of SNPs are both imputed well and are predicted to be well imputed, corresponding to the biggest point masses (red to yellow range according to SNP frequency/count spectrum) with both high Rsq and high dosage r^2. Overall, I find that Rsq can predict dosage r^2 fairly well, particularly for common SNPs and those with reasonable Rsq values. For example, Pearson correlation between Rsq and dosage r^2 is 0.86 for SNPs with MAF 0.005-0.01 and Rsq > 0.5; and 0.93 for SNPs with MAF 0.01-0.03 and Rsq > 0.3. I also observe a noticeable point mass at the right bottom corner, corresponding to SNPs that are predicted to be poorly imputed (low Rsq) but are actually well imputed (high dosage r^2). Closer examination revealed that most of these SNPs are of low frequency (95.4% have MAF < 0.03 and 99.7% have MAF < 0.05), for which the imputation model has low confidence in the estimated dosages that actually match the true dosages fairly well.

Furthermore, I recommend different Rsq thresholds for different MAF categories. Figure 5.7 presents the percentage of SNPs passing post-imputation QC (left Y axis) and the average dosage r^2 (right Y axis) as a function of Rsq threshold (X axis). To achieve an average dosage r^2 of at least 0.85 for example, one would have to use an Rsq threshold of 0.7 for SNPs with MAF 0.001-0.005 while an Rsq threshold of 0 suffices for SNPs with MAF > 0.03. Based on Table 5.2 and Figure 5.7, for my dataset, I chose an Rsq threshold of 0.5 for SNPs with MAF 0.001-0.005 and an Rsq threshold of 0.3 for SNPs with MAF > 0.005, resulting in a total of 127,132 SNPs (out of 158,691) passing post-imputation QC. The sample size for SNPs with MAF < 0.001 is too small for conclusions, but the pattern suggests that the few SNPs passing the post imputation filter of Rsq > 0.5 are well imputed. In general, I recommend selecting an Rsq threshold such that the average Rsq is above the desired average dosage r^2.

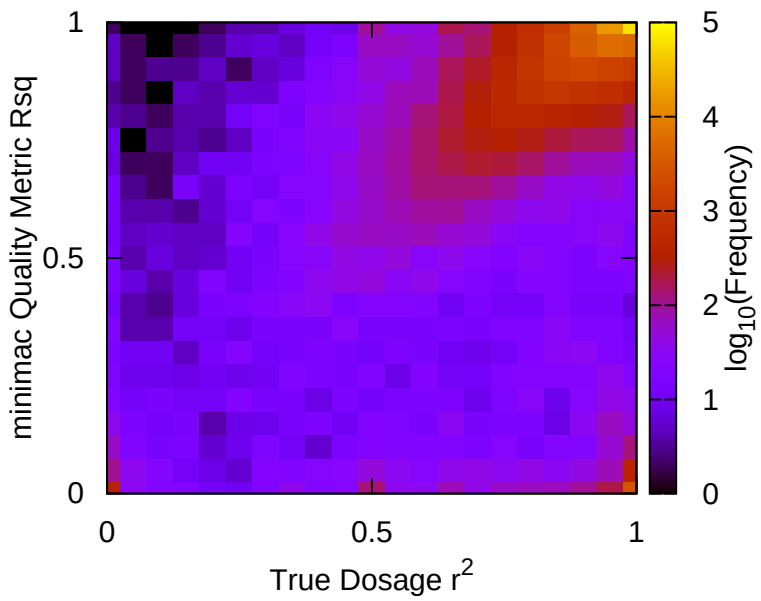


Figure 5.8: Rsq by dosage r^2 for Metabochip SNPs (estimated by masking 100 reference individuals). Estimated imputation accuracy (minimac output Rsq) is plotted against the true dosage r^2, for Metabochip SNPs by masking 100 reference individuals.

5.3.5 To Include or Not to Include: Rare SNPs during Haplotype Reconstruction

One open question concerns whether rare SNPs should be included for haplotype reconstruction, either for the reference individuals or for the target individuals. For the reference panel construction, on one hand, one would like to include as many variants as possible so that they can be subsequently imputed in the target individuals. On the other hand, inclusion of very rare SNPs may interfere with phasing (in the extreme case, for example, singletons cannot be phased), resulting in less accurately constructed haplotypes, and ultimately leading to inferior imputation quality, with little or no benefit in return because these very rare SNPs are unlikely to be accurately imputed into the target individuals. Similarly, for the target individuals, inclusion of rare SNPs may

harm phasing quality, leading to less accurate imputation. On the other hand, as rare to-be-imputed SNPs are more likely to be tagged by rare GWAS SNPs than by common GWAS SNPs, inclusion of rare GWAS variants is expected to increase imputation quality for rarer SNPs. To evaluate this, I assessed the following 20 combinations by varying two parameters: MAF threshold used for the reference panel construction and MAF threshold used for phasing target individuals. For the reference panel construction, I evaluated the following four settings: A) all MAF (i.e., no filtering by MAF); B) no singletons (i.e., removing SNPs with only one copy of the minor allele among the 8,421 individuals with GWAS data); C) MAF > 0.001; and D) MAF > 0.005. For phasing target individuals, I evaluated the following five settings: i) all MAF; ii) no singletons (i.e., removing SNPs with only one copy of the minor allele among reference); iii) MAF > 0.001; iv) MAF > 0.005; and v) MAF > 0.01. Note that for my production imputation, I used v) MAF > 0.01. I picked a medium size chromosome, chromosome 12, for evaluation.

As the comparisons among the four settings for building the reference panel show similar patterns across the five settings for target haplotype reconstruction and vice versa, I present the average of all settings defined by the other parameter. For example, Table 5.3 shows the effect of including rare variants for reference panel construction, where the statistics (number of SNPs and average dosage r^2) for each of the four settings are averaged across the five settings for reconstructing target haplotypes. Among the four settings evaluated, setting B (No Singletons) provides the best trade-off: noticeable gains for MAF categories 0.001-0.01 at little cost for common SNPs. For example, for SNPs with MAF 0.001-0.005, at an Rsq threshold of 0.3, setting B leads to 119 well-imputed SNPs with an average dosage r^2 of 84.0%, outperforming setting A which also results in119 well-imputed SNPs but with a lower average dosage r^2 of 82.8%, setting C of 123 well-imputed SNPs with dosage r^2 of 82.8%, and setting D of 0 well-imputed SNPs (by design). For common SNPs with MAF > 0.01, all four settings have similar performance. On the other hand, there is no clear winner among the five settings for phasing GWAS

data (Table 5.4). Removing SNPs with MAF < 0.001 or 0.005 (settings iii and iv) is slightly advantageous for imputing SNPs with MAF 0.001-0.01. For example, with an Rsq threshold of 0.3, average dosage r^2 for SNPs with MAF 0.001-0.005 is 85.6% and 84.0% respectively for setting iii and iv; while dosage r^2 for the other three settings are $\leq$ 83.0%. However, these settings result in slightly lower imputation quality for SNPs with MAF 0.01-0.05. For example, with an Rsq threshold of 0.3, average dosage r^2 for SNPs with MAF 0.01-0.03 is 90.4% (for 1255 SNPs) and 90.6% (for 1269 SNPs) respectively for setting iii and iv; while dosage r^2 for the other three settings are $\geq$ 91.0% for a larger number of SNPs (number of SNPs $\geq$ 1289).

Table 5.3: Effect of Including Rare Variants for Reference Panel Construction

MAF	Rsq Threshold	A: All MAF		B: No Singletons		C: MAF > 0.1%		D: MAF > 0.5%	
		#SNPs	average dosage r^2	#SNPs	average dosage r^2	#SNPs	average dosage r^2	#SNPs	average dosage r^2
0-0.001	0	22	44.0%	22	43.7%	0	NA	0	NA
0.001-0.005	0	266	70.9%	266	72.9%	266	72.3%	0	NA
0.005-0.01	0	494	85.7%	494	85.7%	494	84.8%	494	85.3%
0.01-0.03	0	1521	90.4%	1521	90.3%	1521	90.3%	1521	90.3%
0.03-0.05	0	955	93.4%	955	93.5%	955	93.4%	955	93.4%
0.05-1.00	0	5494	95.5%	5494	95.5%	5494	95.5%	5494	95.5%
0-0.001	0.3	2	100.0%	3	75.8%	0	NA	0	NA
0.001-0.005	0.3	119	82.8%	119	84.0%	123	82.8%	0	NA
0.005-0.01	0.3	333	87.6%	333	87.8%	328	87.6%	335	87.6%
0.01-0.03	0.3	1307	91.1%	1306	91.0%	1307	91.0%	1307	91.0%
0.03-0.05	0.3	941	93.6%	941	93.7%	940	93.8%	941	93.7%
0.05-1.00	0.3	5486	95.6%	5486	95.6%	5487	95.5%	5487	95.5%
0-0.001	0.5	2	100.0%	2	65.8%	0	NA	0	NA
0.001-0.005	0.5	105	85.6%	103	86.4%	105	85.9%	0	NA
0.005-0.01	0.5	310	89.1%	310	89.3%	308	89.1%	311	89.2%
0.01-0.03	0.5	1268	92.1%	1266	92.0%	1268	92.0%	1269	91.9%
0.03-0.05	0.5	931	94.2%	932	94.2%	932	94.2%	931	94.1%
0.05-1.00	0.5	5460	95.9%	5460	95.8%	5461	95.8%	5459	95.8%

Table 5.4: Effect of Including Rare Variants for Haplotype Reconstruction among Target Individuals

		i: All MAF		ii: No Singletons		iii: MAF > 0.1%		iv: MAF > 0.5%		v: MAF > 1%	
MAF	Rsq Threshold Threshold	#SNPs	average dosage r^2	#SNPs	average dosage r^2	#SNPs	average dosage r^2	#SNPs	average dosage r^2	#SNPs	average dosage r^2
0-0.001	0	22	45.3%	22	44.9%	22	44.2%	22	47.3%	22	37.5%
0-0.001	0.3	3	100.0%	2	100.0%	3	81.3%	2	75.0%	3	83.3%
0-0.001	0.5	3	100.0%	2	100.0%	3	81.3%	1	83.3%	3	83.3%
0.001-0.005	0	266	73.0%	266	72.7%	266	72.7%	266	71.8%	266	70.1%
0.001-0.005	0.3	102	83.0%	123	81.0%	122	85.6%	120	84.0%	133	82.3%
0.001-0.005	0.5	86	86.4%	104	84.5%	106	87.6%	107	86.5%	118	84.9%
0.005-0.01	0	494	85.5%	494	85.8%	494	86.6%	494	85.5%	494	83.4%
0.005-0.01	0.3	285	84.8%	332	88.8%	346	88.8%	350	88.3%	348	87.5%
0.005-0.01	0.5	264	86.4%	316	89.9%	325	90.1%	326	89.7%	317	89.8%
0.01-0.03	0	1521	90.5%	1521	90.6%	1521	90.3%	1521	90.1%	1521	90.1%
0.01-0.03	0.3	1289	91.4%	1347	91.6%	1255	90.4%	1269	90.6%	1373	91.0%
0.01-0.03	0.5	1256	92.3%	1293	92.7%	1222	91.5%	1231	91.6%	1337	92.1%
0.03-0.05	0	955	93.4%	955	93.6%	955	93.2%	955	93.5%	955	93.5%
0.03-0.05	0.3	938	93.7%	943	93.8%	943	93.4%	933	93.8%	946	93.7%
0.03-0.05	0.5	932	94.1%	932	94.4%	934	93.9%	922	94.4%	938	94.1%
0.05-0.50	0	5494	95.4%	5494	95.5%	5494	95.5%	5494	95.5%	5494	95.5%
0.05-0.50	0.3	5486	95.5%	5490	95.5%	5487	95.5%	5487	95.6%	5484	95.7%
0.05-0.50	0.5	5460	95.8%	5463	95.8%	5461	95.8%	5457	95.9%	5460	95.9%

Table 5.5: Effect of Including/Excluding the 100 Masked Reference Individuals during Reference Haplotype Reconstruction

MAF	Rsq Threshold	n = 1862 (Excluding)		n = 1962 (Including)	
		#SNPs	average dosage r^2	#SNPs	average dosage r^2
0-0.001	0	22	47.4%	22	40.3%
0-0.001	0.3	3	89.2%	2	85.2%
0-0.001	0.5	3	84.2%	2	77.1%
0.001-0.005	0	266	71.6%	266	72.6%
0.001-0.005	0.3	117	84.1%	123	82.3%
0.001-0.005	0.5	100	87.3%	108	84.7%
0.005-0.01	0	494	85.4%	494	85.4%
0.005-0.01	0.3	333	87.8%	332	87.5%
0.005-0.01	0.5	309	89.4%	310	89.0%
0.01-0.03	0	1521	90.3%	1521	90.4%
0.01-0.03	0.3	1305	91.0%	1308	91.1%
0.01-0.03	0.5	1267	92.0%	1268	92.1%
0.03-0.05	0	955	93.4%	955	93.4%
0.03-0.05	0.3	941	93.7%	940	93.7%
0.03-0.05	0.5	932	94.2%	931	94.2%
0.05-0.50	0	5494	95.5%	5494	95.5%
0.05-0.50	0.3	5487	95.5%	5487	95.6%
0.05-0.50	0.5	5459	95.8%	5462	95.8%

Table 5.6: Average Rsq and Dosage r^2 by MAF, Estimated by Masking One Reference Individual at a Time (Chromosome 12)

MAF	No Rsq Filter			Rsq > 0.5			Rsq > 0.75		
	#SNPs	average Rsq	average dosage r^2	%SNPs	average Rsq	average dosage r^2	%SNPs	average Rsq	average dosage r^2
0-0.001	1798	4.58%	2.47%	4.7%	70.58%	38.74%	1.7%	84.57%	47.31%
0.001-0.005	935	64.66%	48.41%	73.5%	76.06%	60.59%	38.8%	87.51%	77.21%
0.005-0.01	639	84.77%	79.81%	95.9%	86.67%	81.92%	80.3%	90.70%	86.66%
0.01-0.03	1586	90.86%	88.56%	99.1%	91.35%	89.13%	91.4%	93.55%	91.62%
0.03-0.05	955	94.60%	92.87%	99.6%	94.76%	93.04%	96.6%	95.63%	93.99%
0.05-1.00	5494	96.31%	94.73%	99.5%	96.64%	95.05%	97.6%	97.21%	95.68%

5.4 Discussion

As we are moving into the sequencing era, existing GWAS data provide an inexpensive opportunity to leverage expensive sequencing data. Researchers across the world are becoming increasingly keen on imputation as a tool to infer genotypes at less common

(MAF 0.01-0.05) and rare (MAF < 0.01) variants. Li *et al.* [2010b] have previously shown that larger reference panels improve imputation accuracy for less common variants. In particular, enlarging a reference panel of 60 haplotypes to 1,000 haplotypes increases dosage r^2 for SNPs with MAF < 0.05 from 74% to 93%. However, there has been little, if any research, on truly rare variants: it is not until recently that data became available to assess imputation accuracy for these truly rare variants. Here, I used a reference panel of 3,924 reference haplotypes to demonstrate that it is indeed possible to impute a considerable proportion of rare variants reasonably well, even in a challenging admixed sample of African Americans. Specifically (as indicated in bold in Table2), I was able to impute 99.9% (97.5%, 83.6%, 52.0%, 20.5%) of SNPs with MAF > 0.05 (0.03-0.05, 0.01-0.03, 0.005-0.01, and 0.001-0.005) with average dosage r^2 94.7% (92.1%, 89.0%, 83.1%, and 79.7%).

In the previous section, I presented results from masking Metabochip genotypes for 100 reference individuals during minimac imputation, whom I also included along with the other 1,862 individuals during reference panel construction. One may reasonably argue that the inclusion of the 100 individuals during phasing results in local haplotype mosaics of other individuals better matching haplotypes of these 100 individuals (because constructed haplotypes of the 100 individuals are likely to serve as template to construct haplotypes of other individuals), and therefore over-estimated imputation accuracy. I evaluated this potential over-estimation of imputation accuracy by re-constructing the reference panel only on the other 1,862 individuals. Table 5.5 compares imputation accuracy at Metabochip SNPs for the 100 masked individuals with (phasing ref n = 1,962) or without (phasing ref n = 1,862) them during phasing. I observed no obvious over-estimation: the quality is either very close; or one has slightly smaller number of well-imputed SNPs with slightly higher dosage r^2 than the other. For example, for SNPs with MAF 0.001-0.005, when using Rsq > 0.3 as the post-imputation filter, the reference constructed using 1862 individuals resulted in slightly fewer (117) SNPs passing

the filter with a slightly better average dosage r^2 (84.1%), than the reference constructed using 1962 individuals which had 123 SNPs passing the filter with an average dosage r^2 of 82.3%. The over-estimation may manifest itself if the reference panel were smaller because the 100 masked individuals would contribute more to the haplotype reconstruction of other reference individuals.

I would also like to note that masking 100 reference individuals, although allowing us to directly evaluate imputation quality at actually imputed Metabochip SNPs, still has limitations. For example, sample MAF cannot go below 0.005 and SNPs with "population" MAF (calculated based on n = 1,962 individuals) < 0.005 are either non-varying or have the minor allele over-represented among the 100 individuals (i.e., sample MAF $>$ "population" MAF). Therefore, such SNPs are either not imputable (dosage r^2 undefined and set to zero in my calculations) or tend to be easier to impute than a typical SNP in the population MAF category. The latter case leads to a winner's curse phenomenon such that the actual imputation quality tends to be over-estimated. In order to obtain more reliable estimates for the rarest MAF categories, I attempted a slightly more complicated experiment on chromosome 12 where I masked one reference individual at a time and imputed her genotypes at Metabochip SNPs using other reference individuals' haplotypes. This experiment allows us to examine a sample size of 1,962 instead of 100.

The overall recommendation of picking an Rsq threshold such that the average Rsq is at least 80% to achieve an average dosage r^2 of 80% or above still applies. However, compared with results based on 100 individuals, the actual Rsq thresholds selected for the rare MAF categories are considerably larger, but result in the passing of a larger proportion of SNPs. For example, an Rsq threshold of 0.75 (instead of 0.5 based on the 100 individuals) needs to be applied for SNPs with MAF 0.001-0.005 for the average Rsq to be above 80%, passing 38.8% (instead of 20.5% SNPs). The larger Rsq threshold and larger passing proportion are consistent with the winner's curse phenomenon I discuss above. For example, for SNPs with population MAF 0.001-0.005, the vast majority of

SNPs are monomorphic among the 100 individuals and thus have Rsq close to zero, reflected by the fact that 68.7% of SNPs have Rsq < 0.1 (Figure 5.7(b)). For the small proportion of SNPs that have reasonable Rsq (Rsq > 0.3), which is the proportion of SNPs with minor allele either over-represented or in more extensive LD with neighboring SNPs among the sample of 100 masked individuals), the distribution is highly skewed towards high values. For example, among the 20.5% SNPs with Rsq > 0.5, 16.9% (or 82.0% of the 20.5%) have Rsq > 0.75 such that the average Rsq is 89.24%. In contrast, a much larger proportion of SNPs are no longer monomorphic among the 1,962 individuals and better represent the full range of SNPs in these rare MAF categories, specifically by adding the more challenging SNPs (SNPs with less or no over-representation of the minor allele, and SNPs with less extensive LD with neighboring SNPs). For example, now only 1.8% (compared with 68.7% above based on 100 individuals) SNPs have Rsq < 0.1 for SNPs with MAF 0.001-0.005. Among the 73.5% (compared with 20.5% above) of SNPs with Rsq > 0.5, 38.8% (or 52.8% of the 73.5%) have Rsq > 0.75 (Table 5.6).

Although this study examines an African American population genotyped using Affymetrix 6.0 platform, the recommendation to use Rsq threshold such that average Rsq is around but over the desired dosage r^2 value is generalizable to other populations and other GWAS genotyping platforms, based on similar experiments conducted in several European and Asian populations using different choices of genotyping platforms. For example, in a sample of Filipinos [Wu *et al.*, 2010] genotyped using the Affymetrix 5.0 platform, I found applying a filter of Rsq > 0.6 for SNPs with MAF 0.01-0.02, the average dosage r^2 across the SNPs passing the filter was 0.8085 with an average Rsq of 0.8417. Additional assessment in other populations or using other GWAS platforms can be found in earlier studies [Li *et al.*, 2011, 2010a]. Before more data become available, however, caution needs to be taken when applying the recommendation to rare variants. For example, although imputation in general is more difficult in African populations because of more combinations of the common alleles, recent work [Fumagalli *et al.*, 2010;

Gravel *et al.*, 2011] argue that the more distinctive background of common alleles may benefit imputation of rare variants. In addition, tagSNPs on the Affymetrix 6.0 platform were selected largely based on physical positions, in contrast to those on the Illumina platforms which were selected largely to provide good coverage of the common SNPs according to HapMap-based LD. Therefore, the Affymetrix 6.0 platform may perform better for rare SNP imputation, particularly in samples of non-European ancestry.

My sample consist of females only, therefore, it is straightforward to perform imputation on chromosome X. Even for samples including males, widely used imputation methods can now perform X chromosome imputation (see `http://genome.sph.umich.edu/wiki/MaCH:_machX` and Marchini and Howie [2010]). I did not attempt chromosome X in my dataset because there are only 93 QC+ Metabochip SNPs on chromosome X.

In summary, by constructing a study-specific reference panel of 3,924 haplotypes, I found it feasible to impute SNPs on the Metabochip, a region-centric dense genotyping platform, in a sample of African Americans, including less common SNPs with MAF 0.005-0.05. In addition, I confirmed Rsq as an effective imputation quality metric for these less common variants. In particular, I recommend different Rsq thresholds for different MAF categories such that the average Rsq is above 80%. Furthermore, I found it helpful to remove singleton SNPs when constructing reference haplotypes.

I view this work useful for investigators conducting fine-mapping studies using either dense genotyping or next generation sequencing, particularly for studies in non-European populations. Many efforts to fine map, especially in non-European ancestry participants, are limited by small sample sizes. Now that there are increasing numbers of GWAS studies conducted in non-European populations, imputation can provide a good solution to this sample size problem. For admixed samples like those in this study, new methods are being developed that both leverage the admixture for phenotype-genotype

association mapping and take imputation uncertainty into account [Manolio *et al.*, 2009; Pasaniuc *et al.*, 2011].

Chapter 6

Conclusion

In this dissertation, I present efficient algorithms for genetic analyses in two common genetic study scenarios:

1. Model organisms that are bred through prescribed pedigree design.

2. Humans that are drawn from out-bred populations or continental groups.

By reconstructing haplotype information implicitly or explicitly via HMM, I address two core problems, genome ancestry and genotype imputation, for the two scenarios respectively. In the genome ancestry problem (Chapter 2), I prune the state space in HMM by contracting an important repetitive sub-structure, inbreeding. The major limit of my inbreeding model is that it does not accelerate the ancestry inference for all types of pedigrees. But the computational benefit brought can be crucial in important existing model organism resources such as the Collaborative Cross. I have demonstrated both the effectiveness and efficiency of the algorithm on synthetic and real Collaborative Cross datasets. In the genotype imputation problem (Chapter 4), I accelerate the computation using piecewise "greedy" selection of individual-tailored references. My method is most effective with the emerging sequencing-based large-scale reference panels such as the 1000G panel [The 1000 Genomes Project Consortium, 2010, 2012]. Experiments on admixed populations suggest that my method, implemented in the software package MaCH-Admix, can achieve comparable imputation accuracy by selecting 1/10 of the

total references or less, which corresponds to substantial saving in computation effort. Compared with existing methods, my method has particularly noteworthy advantage among uncommon variants.

In addition to the methodology work, I have presented subsequent analysis of Collaborative Cross data using the ancestry inferred. My analysis (Chapter 3) establishes a new linkage map of the laboratory mouse genome and reveals important properties of recombination events. I have also presented a case study of genotype imputation in a large cohort (∼4000) of African Americans (Chapter 5). My study not only examines imputation performance in under-studied aspects, but also provides practical guidelines for both conducting imputation and post-imputation quality control.

6.1 Future Directions

Below I discuss potential subsequent analysis and future research avenues of the studies presented in this dissertation.

6.1.1 Model Organisms from Prescribed Breeding

The ancestry knowledge estimated by my method GAIN has played a key role in studying the complex traits present in emerging CC lines [Aylor *et al.*, 2011; The Collaborative Cross Consortium, 2012]. In this dissertation, I have also presented analysis on recombination events in early generations of CC resource. As the CC lines continue to develop and, more importantly, become recombinant inbred, I could conduct more powerful subsequent studies in various aspects of studying genetic variations. Also, as I have discussed previously, modeling repetitive sub-structure is a promising approach not only for CC but also for complex pedigrees in many other model organism resources. For example, I may extend my method to handle the repetitive selfing process in plant resources [Cavanagh *et al.*, 2008; Kover *et al.*, 2009].

6.1.2 Samples from Out-bred Human Populations

One limitation of my imputation model is the assumption that a individual-tailored small subset of the reference haplotypes is enough in explaining each target individual. The assumption generally holds well for short or high-LD imputation regions in which the haplotype diversity is low. For longer regions with higher haplotype diversity, using a small subset of reference haplotypes could result in loss in imputation accuracy, especially for rare variants. The loss can be partially compensated by the haplotype-based imputation step in MaCH-Admix which utilizes a much larger set of reference haplotypes. I would also like to explore a more robust framework that can eliminate the dependency on the assumption. In addition, with the emergence of reference panels like 1000G [The 1000 Genomes Project Consortium, 2010, 2012] which contains samples from multiple continental groups and populations, my imputation method can be naturally adapted for population ancestry estimation. Compared with existing methods [Price *et al.*, 2009; Sundquist *et al.*, 2008], my imputation-based framework could be more flexible in handling multiple ancestral sources.

Bibliography

Abecasis, G., Cherny, S., Cookson, W., and Cardon, L. (2001). Merlin: rapid analysis of dense genetic maps using sparse gene flow trees. *Nature genetics*, **30**(1), 97–101. 4, 12, 27, 38

Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *science*, **322**(5903), 881–888. 2

Anderson, G. L., Manson, J., Wallace, R., Lund, B., Hall, D., Davis, S., Shumaker, S., Wang, C.-Y., Stein, E., and Prentice, R. L. (2003). Implementation of the women's health initiative study design. *Annals of epidemiology*, **13**(9), S5–S17. 61

Aylor, D. L., Valdar, W., Foulds-Mathes, W., Buus, R. J., Verdugo, R. A., Baric, R. S., Ferris, M. T., Frelinger, J. A., Heise, M., Frieman, M. B., *et al.* (2011). Genetic analysis of complex traits in the emerging collaborative cross. *Genome research*, **21**(8), 1213–1222. 115

Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R. D., and Petukhova, G. V. (2012). Genetic recombination is directed away from functional genomic elements in mice. *Nature*, **485**(7400), 642–645. 47, 48

Browning, B. L. and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics*, **84**(2), 210–23. 6, 54, 65, 94

Browning, S. and Browning, B. (2002). On reducing the statespace of hidden markov models for the identity by descent process. *Theoretical population biology*, **62**(1), 1–8. 14

Buyske, S., Wu, Y., Ambite, J., Assimes, T., Boerwinkle, E., Buzkova, P., Carlson, C., Carty, C., Cheng, I., Cochran, B., *et al.* (2011). Use and performance of the metabochip genotyping array in african americans: The page study. *In preparation*. 92

Cavanagh, C., Morell, M., Mackay, I., and Powell, W. (2008). From mutations to magic: resources for gene discovery, validation and delivery in crop plants. *Current opinion in plant biology*, **11**(2), 215–221. 115

Chesler, E. J., Miller, D. R., Branstetter, L. R., Galloway, L. D., Jackson, B. L., Philip, V. M., Voy, B. H., Culiat, C. T., Threadgill, D. W., Williams, R. W., *et al.* (2008). The collaborative cross at oak ridge national laboratory: developing a powerful resource for systems genetics. *Mammalian Genome*, **19**(6), 382–389. 14, 25, 35

Chia, R., Achilli, F., Festing, M., and Fisher, E. (2005). The origins and uses of mouse outbred stocks. *Nature genetics*, **37**(11), 1181–1186. 3, 11

Churchill, G., Airey, D., Allayee, H., Angel, J., Attie, A., Beatty, J., Beavis, W., Belknap, J., Bennett, B., Berrettini, W., *et al.* (2004). The collaborative cross, a community resource for the genetic analysis of complex traits. *Nature genetics*, **36**(11), 1133–1137. 3, 11, 13, 25

Cox, A., Ackert-Bicknell, C. L., Dumont, B. L., Ding, Y., Bell, J. T., Brockmann, G. A., Wergedal, J. E., Bult, C., Paigen, B., Flint, J., *et al.* (2009). A new standard genetic map for the laboratory mouse. *Genetics*, **182**(4), 1335–1344. 42, 47, 48

de Bakker, P. I., Ferreira, M. A., Jia, X., Neale, B. M., Raychaudhuri, S., and Voight, B. F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human molecular genetics*, **17**(R2), R122–R128. 89

Donnelly, K. (1983). The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology*, **23**(1), 34–63. 14

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A., Wheeler, E., Glazer, N., Bouatia-Naji, N., Gloyn, A., and et al. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics*, **42**(2), 105–116. 51

Egyud, M. R. L., Gajdos, Z. K. Z., Butler, J. L., Tischfield, S., Le Marchand, L., Kolonel, L. N., Haiman, C. A., Henderson, B. E., and Hirschhorn, J. N. (2009). Use of weighted reference panels based on empirical estimates of ancestry for capturing untyped variation. *Human Genetics*, **125**(3), 295–303. 7, 52, 85, 91

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, **11**(6), 446–450. 90

Fridley, B. L., Jenkins, G., Deyo-Svendsen, M. E., Hebbring, S., and Freimuth, R. (2010). Utilizing genotype imputation for the augmentation of sequence data. *PLoS ONE*, **5**(6), e11018. 51

Fumagalli, M., Cagliani, R., Riva, S., Pozzoli, U., Biasin, M., Piacentini, L., Comi, G. P., Bresolin, N., Clerici, M., and Sironi, M. (2010). Population genetics of ifih1: ancient population structure, local selection, and implications for susceptibility to type 1 diabetes. *Molecular biology and evolution*, **27**(11), 2555–2566. 111

Gabriel, S., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., *et al.* (2002). The structure of haplotype blocks in the human genome. *Science*, **296**(5576), 2225–2229. 12

Geiger, D., Meek, C., and Wexler, Y. (2009). Speeding up hmm algorithms for genetic linkage analysis via chain reductions of the state space. *Bioinformatics*, **25**(12), i196–i203. 14

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., Bustamante, C. D., Altshuler, D. L., *et al.* (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, **108**(29), 11983–11988. 112

Gudbjartsson, D., Thorvaldsson, T., Kong, A., Gunnarsson, G., and Ingolfsdottir, A. (2005). Allegro version 2. *Nature genetics*, **37**(10), 1015–1016. 4, 12

Hao, K., Chudin, E., McElwee, J., and Schadt, E. E. (2009). Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genetics*, **10**(1), 27. 7, 52, 84

Hassold, T. and Hunt, P. (2001). To err (meiotically) is human: the genesis of human aneuploidy. *Nature Reviews Genetics*, **2**(4), 280–291. 35

He, J., Wilkens, L. R., Stram, D. O., Kolonel, L. N., Henderson, B. E., Wu, A. H., Le Marchand, L., and Haiman, C. A. (2011). Generalizability and epidemiologic characterization of eleven colorectal cancer gwas hits in multiple populations. *Cancer Epidemiology Biomarkers & Prevention*, **20**(1), 70–81. 90

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, **106**(23), 9362–9367. 89

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics*, **1**(6), 457–470. 87

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, **5**(6), 15. 6, 7, 53, 54, 65

Huang, L., Li, Y., Singleton, A. B., Hardy, J. A., Abecasis, G., Rosenberg, N. A., and Scheet, P. (2009). Genotype-imputation accuracy across worldwide human populations. *The American Journal of Human Genetics*, **84**(2), 235–250. 6, 7, 51, 52

Idury, R. and Elston, R. (1997). A faster and more general hidden markov model algorithm for multipoint likelihood calculations. *Human heredity*, **47**(4), 197–202. 25

Jensen, C. and Kong, A. (1999). Blocking gibbs sampling for linkage analysis in large pedigrees with many loops. *The American Journal of Human Genetics*, **65**(3), 885–901. 5, 13

Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordel, H., Eaves, I. A., and Dudbridge, F. (2001). Haplotype tagging for the identification of common disease genes. *Nature genetics*, **29**(2), 233–237. 2

Keebler, M. E., Deo, R. C., Surti, A., Konieczkowski, D., Guiducci, C., Burtt, N., Buxbaum, S. G., Sarpong, D. F., Steffes, M. W., Wilson, J. G., *et al.* (2010). Fine-mapping in african americans of 8 recently discovered genetic loci for plasma lipidsclinical perspective the jackson heart study. *Circulation: Cardiovascular Genetics*, **3**(4), 358–364. 90

Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., Durrant, C., and Mott, R. (2009). A multiparent advanced generation inter-cross to fine-map quantitative traits in arabidopsis thaliana. *PLoS genetics*, **5**(7), e1000551. 115

Kruglyak, L., Daly, M., Reeve-Daly, M., and Lander, E. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *American journal of human genetics*, **58**(6), 1347. 4, 12

Krumsiek, J., Arnold, R., and Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, **23**(8), 1026–1028. 48

Lander, E. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*, **84**(8), 2363–2367. 12

Lanktree, M. B., Anand, S. S., Yusuf, S., Hegele, R. A., *et al.* (2009). Replication of genetic associations with plasma lipoprotein traits in a multiethnic sample. *Journal of lipid research*, **50**(7), 1487–1496. 90

Lettre, G., Palmer, C. D., Young, T., Ejebe, K. G., Allayee, H., Benjamin, E. J., Bennett, F., Bowden, D. W., Chakravarti, A., Dreisbach, A., *et al.* (2011). Genome-wide association study of coronary heart disease and its risk factors in 8,090 african americans: the nhlbi care project. *PLoS genetics*, **7**(2), e1001300. 90

Li, J. and Jiang, T. (2005). Computing the minimum recombinant haplotype configuration from incomplete genotype data on a pedigree by integer linear programming. *Journal of Computational Biology*, **12**(6), 719–739. 5, 13

Li, L., Li, Y., Browning, S. R., Browning, B. L., Slater, A. J., Kong, X., Aponte, J. L., Mooser, V. E., Chissoe, S. L., Whittaker, J. C., *et al.* (2011). Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PloS one*, **6**(9), e24945. 111

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annual Review of Genomics and Human Genetics*, **10**(1), 387–406. 2, 7, 52, 65, 84, 89, 94, 97

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010a). Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, **34**(8), 816–834. 6, 51, 54, 55, 93, 111

Li, Y., Byrnes, A. E., and Li, M. (2010b). To identify associations with rare variants, just whait: Weighted haplotype and imputation-based tests. *The American Journal of Human Genetics*, **87**(5), 728–735. 109

Lin, J. (1991). Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, **37**(1), 145–151. 30

Lin, P., Hartz, S. M., Zhang, Z., Saccone, S. F., Wang, J., Tischfield, J. A., Edenberg, H. J., Kramer, J. R., M Goate, A., Bierut, L. J., and et al. (2010). A new statistic to evaluate imputation reliability. *PLoS ONE*, **5**(3), 10. 65, 94

Lind, J. M., Hutcheson-Dilks, H. B., Williams, S. M., Moore, J. H., Essex, M., Ruiz-Pesini, E., Wallace, D. C., Tishkoff, S. A., O'Brien, S. J., and Smith, M. W. (2007). Elevated male european and female african contributions to the genomes of african american individuals. *Human Genetics*, **120**(5), 713–722. 59

Liu, E. Y., Buyske, S., Aragaki, A. K., Peters, U., Boerwinkle, E., Carlson, C., Carty, C., Crawford, D. C., Haessler, J., Hindorff, L. A., Marchand, L. L., Manolio, T. A., Matise, T., Wang, W., Kooperberg, C., North, K. E., and Li, Y. (2012). Genotype imputation of metabochipsnps using a study-specific reference panel of ∼4,000 haplotypes in african americans from the women's health initiative. *Genetic Epidemiology*, **36**(2), 107–117. 61, 87

Maher, B. (2008). The case of the missing heritability. *Nature*, **456**(7218), 18–21. 90

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., *et al.* (2009). Finding the missing heritability of complex diseases. *Nature*, **461**(7265), 747–753. 90, 113

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, **11**(7), 499–511. 65, 89, 94, 97, 112

Matise, T. C., Ambite, J. L., Buyske, S., Carlson, C. S., Cole, S. A., Crawford, D. C., Haiman, C. A., Heiss, G., Kooperberg, C., Le Marchand, L., *et al.* (2011). The next page in understanding complex traits: design for the analysis of population architecture using genetics and epidemiology (page) study. *American journal of epidemiology*, **174**(7), 849–859. 90

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, **9**(5), 356–369. 90

McPeek, M. (2002). Inference on pedigree structure from genome screen data. *Statistica Sinica*, **12**(1), 311–336. 14

Merriwether, D. A., Rothhammer, F., and Ferrell, R. E. (1995). Distribution of the four founding lineage haplotypes in native americans suggests a single wave of migration for the new world. *American Journal of Physical Anthropology*, **98**(4), 411–430. 2

Mott, R., Talbot, C., Turri, M., Collins, A., and Flint, J. (2000). A method for fine mapping quantitative trait loci in outbred animal stocks. *Proceedings of the National Academy of Sciences*, **97**(23), 12649–12654. 5, 9, 12, 13, 27

Parra, E. J., Marcini, A., Akey, J., Martinson, J., Batzer, M. A., Cooper, R., Forrester, T., Allison, D. B., Deka, R., Ferrell, R. E., and et al. (1998). Estimating african american admixture proportions by use of population-specific alleles. *American Journal of Human Genetic*, **63**(6), 1839–1851. 59

Pasaniuc, B., Kennedy, J., and Mandoiu, I. (2009). Imputation-based local ancestry inference in admixed populations. *Lecture Notes in Computer Science*, **5542**, 221–233. 86

Paşaniuc, B., Sankararaman, S., Kimmel, G., and Halperin, E. (2009). Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, **25**(12), i213–i221. 14

Pasaniuc, B., Avinery, R., Gur, T., Skibola, C. F., Bracci, P. M., and Halperin, E. (2010). A generic coalescent-based framework for the selection of a reference panel for imputation. *Genetic Epidemiology*, **34**(8), 773–782. 7, 52, 85

Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G. K., Tandon, A., Kao, W. H. L., Ruczinski, I., Fornage, M., Siscovick, D. S., Zhu, X., and et al. (2011). Enhanced statistical tests for gwas in admixed populations: Assessment using african americans from care and a breast cancer consortium. *PLoS Genetics*, **7**(4), 15. 85, 113

Pei, Y.-F., Zhang, L., Li, J., and Deng, H.-W. W. (2008). Analyses and comparison of imputation-based association methods. *PloS one*, **5**. 6

Pemberton, T. J., Jakobsson, M., Conrad, D. F., Coop, G., Wall, J. D., Pritchard, J. K., Patel, P. I., and Rosenberg, N. A. (2008). Using population mixtures to optimize the utility of genomic databases: linkage disequilibrium and association study design in india. *Annals of Human Genetics*, **72**(Pt 4), 535–546. 7, 52, 85

Piccolboni, A. and Gusfield, D. (2003). On the complexity of fundamental computational problems in pedigree analysis. *Journal of Computational Biology*, **10**(5), 763–773. 12

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, **38**(8), 904–909. 92

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, **5**(6), e1000519. 86, 116

Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, **69**(1), 1–14. 94

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**(2), 945–59. 60, 86

Pulit, S. L., Voight, B. F., and de Bakker, P. I. (2010). Multiethnic genetic association studies improve power for locus discovery. *PloS one*, **5**(9), e12600. 90

Qayyum, R., Snively, B. M., Ziv, E., Nalls, M. A., Liu, Y., Tang, W., Yanek, L. R., Lange, L., Evans, M. K., Ganesh, S., Austin, M. A., Lettre, G., Becker, D. M., Zonderman, A. B., Singleton, A. B., Harris, T. B., Mohler, E. R., Logsdon, B. A., Kooperberg, C., Folsom, A. R., Wilson, J. G., Becker, L. C., and Reiner, A. P. (2012). A Meta-Analysis and Genome-Wide Association Study of Platelet Count and Mean Platelet Volume in African Americans. *PLoS Genet*, **8**(3), e1002491+. 61

Qian, D. and Beckmann, L. (2002). Minimum-recombinant haplotyping in pedigrees. *The American Journal of Human Genetics*, **70**(6), 1434–1445. 5, 13

Reich, D. and Patterson, N. (2005). Will admixture mapping work to find disease genes? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**(1460), 1605–7. 52

Reiner, A. P., Carlson, C. S., Ziv, E., Iribarren, C., Jaquish, C. E., and Nickerson, D. A. (2007). Genetic ancestry, population sub-structure, and cardiovascular disease-related traits among african-american participants in the cardia study. *Human Genetics*, **121**(5), 565–75. 59

Reiner, A. P., Lettre, G., Nalls, M. A., Ganesh, S. K., Mathias, R., Austin, M. A., Dean, E., Arepalli, S., Britton, A., Chen, Z., *et al.* (2011). Genome-wide association study of white blood cell count in 16,388 african americans: the continental origins and genetic epidemiology network (cogent). *PLoS genetics*, **7**(6), e1002108. 91

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, **273**(5281), 1516–1517. 5

Roberts, A., Pardo-Manuel de Villena, F., Wang, W., McMillan, L., and Threadgill, D. W. (2007). The polymorphism architecture of mouse genetic resources elucidated using genome-wide resequencing data: implications for qtl discovery and systems genetics. *Mammalian Genome*, **18**(6), 473–481. 36

Robinson, W. P. (1996). The extent, mechanism, and consequences of genetic variation, for recombination rate. *American journal of human genetics*, **59**(6), 1175. 35

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., and Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nature Reviews Genetics*, **11**(5), 356–366. 52, 90

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., *et al.* (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**(6909), 832–837. 2

Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008). Estimating local ancestry in admixed populations. *American journal of human genetics*, **82**(2), 290. 14

Sanna, S., Pitzalis, M., Zoledziewska, M., Zara, I., Sidore, C., Murru, R., Whalen, M. B., Busonero, F., Maschio, A., Costa, G., and et al. (2010). Variants within the immunoregulatory cblb gene are associated with multiple sclerosis. *Nature Genetics*, **42**(6), 495–497. 51

Schwartz, R., Clark, A., and Istrail, S. (2004). Inferring piecewise ancestral history from haploid sequences. *Computational Methods for SNPs and Haplotype Inference*, pages 615–616. 12

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., Chines, P. S., Jackson, A. U., and et al. (2007). A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, **316**(5829), 1341–1345. 51

Seldin, M. F., Pasaniuc, B., and Price, A. L. (2011). New approaches to disease mapping in admixed populations. *Nature Reviews Genetics*, **12**(8), 523–528. 84

Shriner, D., Adeyemo, A., Chen, G., and Rotimi, C. N. (2010). Practical considerations for imputation of untyped markers in admixed populations. *Genetic Epidemiology*, **34**(3), 258–265. 7, 52, 84, 91

Smagulova, F., Gregoretti, I. V., Brick, K., Khil, P., Camerini-Otero, R. D., and Petukhova, G. V. (2011). Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature*, **472**(7343), 375–378. 35, 44, 47, 48

Smith, J. G., Magnani, J. W., Palmer, C., Meng, Y. A., Soliman, E. Z., Musani, S. K., Kerr, K. F., Schnabel, R. B., Lubitz, S. A., Sotoodehnia, N., *et al.* (2011). Genome-wide association studies of the pr interval in african americans. *PLoS genetics*, **7**(2), e1001304. 90

Smith, N. L., Chen, M.-H., Dehghan, A., Strachan, D. P., Basu, S., Soranzo, N., Hayward, C., Rudan, I., Sabater-Lleal, M., Bis, J. C., and et al. (2010). Novel associations of multiple genetic loci with plasma levels of factor vii, factor viii, and von willebrand factor: The charge (cohorts for heart and aging research in genome epidemiology) consortium. *Circulation*, **121**(12), 1382–1392. 51

Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American journal of human genetics*, **58**(6), 1323. 5, 13

Stefflova, K., Dulik, M. C., Barnholtz-Sloan, J. S., Pai, A. A., Walker, A. H., and Rebbeck, T. R. (2011). Dissecting the within-africa ancestry of populations of african descent in the americas. *PLoS ONE*, **6**(1), e14495. 59

Stephens, J. C., Schneider, J. A., Tanguay, D. A., Choi, J., Acharya, T., Stanley, S. E., Jiang, R., Messer, C. J., Chew, A., Han, J.-H., *et al.* (2001). Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, **293**(5529), 489–493. 2

Sundquist, A., Fratkin, E., Do, C., and Batzoglou, S. (2008). Effect of genetic divergence in identifying ancestral origin using hapaa. *Genome research*, **18**(4), 676–682. 14, 116

Tang, H., Coram, M., Wang, P., Zhu, X., and Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, **79**(1), 1–12. 14, 52

Teo, Y.-Y., Small, K. S., and Kwiatkowski, D. P. (2010). Methodological challenges of genome-wide association analysis in africa. *Nature Reviews Genetics*, **11**(2), 149–160. 90

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319), 1061–1073. 7, 51, 89, 114, 116

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 1. 6, 51, 114, 116

The Collaborative Cross Consortium (2012). The genome architecture of the collaborative cross mouse genetic reference population. *Genetics*, **190**, 389–401. 13, 14, 25, 35, 115

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, **437**(7063), 1299–1320. 94

The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million snps. *Nature*, **449**(7164), 851–861. 89

The International HapMap Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**(7311), 52–8. 6, 7, 51, 62, 87, 97

The WHI Study Group (1998). Design of the womens health initiative clinical trial and observational study. *Controlled Clinical Trials*, **19**(1), 61 – 109. 61, 90

Threadgill, D. W. and Churchill, G. A. (2012). Ten years of the collaborative cross. *Genetics*, **190**(2), 291–294. 4

Valdar, W., Solberg, L., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W., Taylor, M., Rawlins, J., Mott, R., and Flint, J. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature genetics*, **38**(8), 879–887. 3, 5, 11, 12, 13

Wall, J. D., Pritchard, J. K., *et al.* (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, **4**(8), 587–597. 2

Wang, X., Zhu, X., Qin, H., Cooper, R., Ewens, W., Li, C., and Li, M. (2011a). Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics*, **27**(5), 670–7. 85

Wang, Z., Jacobs, K., Yeager, M., Hutchinson, A., Sampson, J., Chatterjee, N., Albanes, D., Berndt, S., Chung, C., Diver, W., Gapstur, S., Teras, L., Haiman, C., Henderson, B., Stram, D., Deng, X., Hsing, A., Virtamo, J., Eberle, M., Stone, J., Purdue, M., Taylor, P., Tucker, M., and Chanock, S. (2011b). Improved imputation of common and uncommon snps with a new reference set. *Nature Genetics*, **44**(1), 6–7. 87

Waters, K. M., Le Marchand, L., Kolonel, L. N., Monroe, K. R., Stram, D. O., Henderson, B. E., and Haiman, C. A. (2009). Generalizability of associations from prostate cancer genome-wide association studies in multiple populations. *Cancer Epidemiology Biomarkers & Prevention*, **18**(4), 1285–1289. 90

Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J., Najjar, S. S., Stringham, H. M., and et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics*, **40**(2), 161–169. 51

Winkler, C. A., Nelson, G. W., and Smith, M. W. (2010). Admixture mapping comes of age. *Annual Review of Genomics and Human Genetics*, **11**, 65–89. 52

Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist*, **56**(645), 330–338. 18

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**(7145), 661–78. 51

Wu, Y., Li, Y., Lange, E. M., Croteau-Chonka, D. C., Kuzawa, C. W., McDade, T. W., Qin, L., Curocichin, G., Borja, J. B., Lange, L. A., *et al.* (2010). Genome-wide association study for adiponectin levels in filipino women identifies cdh13 and a novel uncommon haplotype at kng1–adipoq. *Human molecular genetics*, **19**(24), 4955–4964. 111

Yang, H., Ding, Y., Hutchins, L. N., Szatkiewicz, J., Bell, T. A., Paigen, B. J., Graber, J. H., de Villena, F. P.-M., and Churchill, G. A. (2009). A customized and versatile high-density genotyping array for the mouse. *Nature methods*, **6**(9), 663–666. 3, 37

Zhang, B., Zhi, D., Zhang, K., Gao, G., Limdi, N. N., and Liu, N. (2011). Practical Consideration of Genotype Imputation: Sample Size, Window Size, Reference Choice, and Untyped Rate. *Statistics and its interface*, **4**(3), 339–352. 84

Zhang, K., Deng, M., Chen, T., Waterman, M., and Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences*, **99**(11), 7335–7339. 12

Zhang, Q., Wang, W., McMillan, L., Prins, J., Pardo-Manuel de Villena, F., and Threadgill, D. (2008). Genotype sequence segmentation: Handling constraints and noise. *Algorithms in Bioinformatics*, pages 271–283. 5, 13

Zhu, X., Cooper, R. S., and Elston, R. C. (2004). Linkage analysis of a complex disease through use of admixed populations. *American Journal of Human Genetics*, **74**(6), 1136–53. 52