IN SILICO HAPLOTYPING, GENOTYPING AND ANALYSIS OF RESEQUENCING DATA USING MARKOV MODELS

by

Yun Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2009

Doctoral Committee:

Professor Gonçalo R. Abecasis, Co-Chair
Professor Michael Lee Boehnke, Co-Chair
Professor Margit Burmeister
Professor Roderick J.A. Little
Assistant Professor Noah A. Rosenberg

To Mom, Dad, Iory and Duan.

Acknowledgements

I am grateful to the myriad of people who have made this work possible.

First of all, I would like to thank Gonçalo and Mike for teaching me about science, for letting me discover how much I can enjoy research, for being encouraging and enthusiastic both because of and despite what we found, for guiding me to grow up as an independent researcher, for being exceptional role models both career-wise and family-wise, and for being approachable anytime I need help.

I owe gratitude to my committee: Rod, Margit and Noah for their help and guidance beyond their duties as committee members. I sincerely appreciate their time, advice, and encouragement. Special thanks to Rod for brining in the missing data perspectives; to Margit for always asking questions that I have an answer in the next few slides; and to Noah for access to the HGDP data.

I would like also to thank the CSG members: they make everyday research joyful and keep my spirits up even during my down times. I am also fortunate to be around a group of good friends including Qixuan, Jin, Rui, Jun, and Liming, who have made my 5-year graduate studies more colorful.

I thank my parents for their support and encouragement throughout the years, Iory for putting up with sometimes a different me at home, and Duan for brining me endless of joys.

TABLE OF CONTENTS

Chapter

List of Figures

List of Tables

ABSTRACT

IN SILICO HAPLOTYPING, GENOTYPING AND ANALYSIS OF RESEQUENCING DATA USING MARKOV MODELS

by

Yun Li

Co-Chairs: Gonçalo R. Abecasis and Michael Lee Boehnke

Searches for the elusive genetic mechanisms underlying complex diseases have long challenged human geneticists. Recently, genome-wide association studies (GWAS) have successfully identified many complex disease susceptibility loci by genotyping a subset of several hundred thousand common genetic variants across many individuals. With the rapid deployment of next-generation sequencing technologies, it is anticipated that future genetic association studies will be able to more comprehensively survey genetic variation, both to identify new loci that were missed in the original round of genome-wide association studies and to finely characterize the contributions of identified loci. GWAS, whether in the current genotyping-based form or in the anticipated sequencing-based form, pose a range of computational and analytical challenges.

I first propose and implement a computationally efficient hidden Markov model that can rapidly reconstruct the two chromosomes carried by each individual in a study. To achieve this goal, the methods combine partial genotype or sequence data for each individual with additional information on additional individuals. Comparisons with standard haplotypers in both simulated and real datasets show that the proposed method is at least comparable and more computational efficient.

I next extend my method for imputing genotypes at untyped SNP loci. Specifically, I consider how my approach can be used to assess several million common variants that are not directly genotyped in a typical association study but for which data are available in public databases. I describe how the extended method performs in a wide range of simulated and real settings.

Finally, I consider how low-depth shot-gun resequencing data on a large number of individuals can be combined to provide accurate estimates of individual sequences. This approach should speed up the advent of large-scale genome resequencing studies and facilitate the identification of rare variants that contribute to disease susceptibility and that cannot be adequately assessed with current genotyping-based GWAS approaches.

My methods are flexible enough to accommodate phased haplotype data, genotype data, or re-sequencing data as input and can utilize public resources such as the HapMap consortium and the 1000 Genomes Project that now include data on several million genetic variants typed on hundreds of individuals.

Chapter 1

Introduction

1.1 Complex diseases

The study of complex diseases is complex. Unlike most simple Mendelian disorders, complex diseases have a significant impact on human health due to their high population incidence, and they have therefore received great attention (Lander and Schork 1994; Hoh and Ott 2003; Dewan *et al.* 2007, MaCarthy *et al*. 2008). Despite great advances made and continuous efforts put into understanding the genetic basis of these complex diseases, we still have limited knowledge regarding causal genetic variants underlying common genetically complex diseases such as cancer, diabetes, cardiovascular diseases, and psychological disorders. For example, although over 200 regions of the genome have been identified to be associated with various complex diseases to date (Hindroff *et al.* 2009, also see www.genome.gov/gwastudies), the proportion of heritability explained by the identified variants is modest at best (Visscher *et al*. 2008).

Unraveling the genetic basis of complex diseases is a challenge for geneticists. The main reasons include but are not limited to multiple common alleles with modest effects, complex gene-gene and gene-environment interactions, rare variants that contribute to disease susceptibility, confounding non-inherited genetic effects, and diagnostic difficulties of (sub)phenotypes.

Due to the relatively limited knowledge of the contributing genetic variants, and even less of the underlying etiology, early efforts to identify genes conferring susceptibility to complex diseases mostly took the form of linkage and positional cloning (Botstein and Risch 2003, McCarthy *et al.* 2008), where disease genes are identified solely based on their relative position within a known map of genetic markers.

One traditional type of positional closing is linkage analysis, which has been used for studying both Mendelian and complex diseases. A typical genome-wide linkage study examines several hundred microsatellites or thousands of single nucleotide polymorphisms (SNPs) across the genome on families with multiple affected individuals. Initial linkage scans may identify regions of interest, typically 10-20 Mb in size if the phenotypic trait is complex. Subsequent fine mapping and positional candidate gene studies, where a denser map of genetic variants is scrutinized across the linked regions, are carried out to refine the linked regions and to more precisely pinpoint the contributing causal variants. Such a paradigm has achieved phenomenal successes in the study of disorders that have a relatively simple relationship between the phenotypic trait and underlying genetic variants. Such disorders are mostly Mendelian; examples include cystic fibrosis (Eiberg *et al*. 1984), haemochromatosis (Feder *et al*. 1996) and lactose intolerance (Enattah et al. 2002). However, linkage-based approaches have been much less effective in localizing genes for common complex diseases, owing mainly to the diseases' complex inheritance patterns, the coarse resolution of linkage mapping, and the relatively small magnitude of the contribution to disease risk for most common variants.

1.2 Genome-wide association studies (GWAS)

In the last ten years, genome-wide case-control association studies have been proposed as a potentially more powerful alternative to linkage studies for complex diseases (Risch and Merikangas 1996). The association-based approach typically starts with the collection of a large number of genetically unrelated individuals, including both individuals affected with the disease of interest and unaffected controls. Association studies sample “unrelated” individuals with and without the disease phenotype of interest. The individuals under study are “unrelated” only to the extent of not being related genetically in the most recent three to five generations. Since these “unrelated” individuals share shorter stretches of their chromosomes than among family members, a much denser set of SNPs is typically selected, assayed and tested for disease-marker association. SNPs have gradually become the marker of choice, replacing microsatellites, owing to their great abundance and to the availability of high-throughput analysis technologies for SNPs.

GWAS typically involve the examination of several hundred thousand SNPs across the entire genome on thousands of individuals. GWAS is the gene mapping strategy that has delivered on the promise of detecting genetic variants whose individual contributions to complex disease susceptibility are small (for examples, see www.genome.gov/gwastudies). Mapping resolution can be much finer than that of traditional linkage approaches, where the linked regions are typically 10-20 Mb in size and thus are still largely intractable for effective localization of the causal variants. Such greatly refined mapping resolution is attainable both because unrelated individuals share much shorter stretches of their chromosomes than family members due to historical recombination events, and because of the large number of SNPs examined.

The number of common (minor allele frequency [MAF] > 1%) SNPs in the human genome is believed to be approximately ten million (Kruglyak and Nickerson 2001; Hinds *et al.* 2005; The International HapMap Consortium 2007). Despite the large number of SNPs (typically several hundred thousand) assayed in GWAS, the effects of most of the ten million SNPs must be evaluated indirectly using either genotyped SNPs or haplotypes thereof as proxies. Phasing and imputation are therefore of great importance for effective identification of disease-causing variants, where phasing is the inference of haplotypes from unphased genotypes and imputation is the estimation of the allelic states of SNPs that are not directly genotyped.

1.3 The next-generation: Resequencing-based approaches

With the advent and rapid advances in very high throughput resequencing technologies (Bentley 2006), it is believed by some that genotyping-based approaches will soon become obsolete. One advantage of resequencing-based approaches is that they naturally capture variants that are currently absent from public databases including, potentially, population specific variants. Thus, resequencing is one natural and important next step toward elucidating the underlying functional mechanisms in the gene regions discovered in current association studies, and it may be used as a technique for association studies eventually.

The cost of ultra high-throughput resequencing, although it has dropped tremendously, remains daunting for application to a large number of individuals at high depth. Alternatively, for a large number of individuals, the design of low-pass short-read shotgun resequencing is particularly attractive. However, there are few, if any, existing tools to combine efficiently partial resequencing information across individuals. Novel computational and statistical tools are essential to stimulate the advent of large-scale genome resequencing-based approaches and to facilitate the identification of rare variants that contribute to disease susceptibility but that cannot be adequately assessed with current genotyping-based approaches.

1.4 The scope of this dissertation

GWAS and next-generation resequencing studies pose a wide range of computational and statistical challenges that have not yet been adequately addressed. In this dissertation, I propose and implement computationally efficient hidden Markov models that can analyze data from both GWAS and resequencing studies, involving hundreds of thousands to several million markers in thousands of individuals.

These methods can (a) rapidly reconstruct the two chromosomes (haplotypes) that are carried by each individual in a study (current high-throughput genotyping assays only measure small fragments of each chromosome and do not provide long sequences as output); (b) combine data from individual studies (which typically examine several hundred thousand up to about one million genetic variants) with data from public resources (which include information on millions of genetic variants); (c) combine data from different studies that examine different sets of genetic variants (this is especially important to achieve the large sample sizes required for detecting variants that make only small contributions to disease risk); or (d) combine low-depth shotgun resequencing data on a large number of individuals to provide accurate estimates of individual sequences. These methods are expected to speed up the advent of large-scale genome-wide resequencing studies and enable the identification of rare disease variants whose effects cannot be satisfactorily evaluated in current genotyping-based studies.

Chapter 2 of this dissertation introduces the basic form of the underlying hidden Markov models and evaluates its utility in haplotype reconstruction from unphased genotypes. Applications to both simulated datasets and datasets from real studies show that my method is comparable, if not superior to current state-of-the-art haplotypers. In addition, I evaluate in various scenarios the benefits of incorporating additional data from public databases such as the International HapMap Consortium, which now include data on millions of genetic markers on hundreds of individuals.

Chapter 3 focuses on imputing missing genotypes for GWAS. Missingness is defined broadly to include genotypes of unassayed markers in an association study. In particular, I consider how to impute genotypes for and how to assess the effects of several million common SNPs (mostly not directly genotyped) in each individual GWAS by efficiently combining with data from public databases (e.g., HapMap). Using both simulated and real studies, I show how the evaluation of unmeasured variants can effectively increase sample sizes, leading ultimately to the identification of genetic variants whose effects are moderate and that cannot be powerfully detected without combining data across several large scale GWAS, which might well examine different sets of genetic markers.

Chapter 4 presents and demonstrates the utility of an extended hidden Markov model that handles shotgun resequencing data. Enabling effective combination of partial information (i.e., low-depth sequencing data in this particular context) from a larger number of individuals, my method allows more cost-efficient allocation of limited sequencing resources. The performance of my method, measured by a wide range of statistics including proportion of polymorphisms detected, genotype-specific imputation accuracy at detected sites, and information generated for statistical analysis, is extensively evaluated through simulations. I also performed analysis with preliminary data from real whole-genome resequencing studies, results from which are consistent with my predictions from simulations.

In summary, I have developed computationally efficient models for the analysis of large-scale genetic data derived from GWAS or resequencing based studies. I believe that more advanced laboratory techniques and further successes in the area of complex disease gene identification will require even better computational tools and improved statistical methods to enable us to tackle the large datasets of SNP and structure variants now being identified and genotyped. I aim to extract subtle signals from large and complex data sets. The subtle signals together may explain a larger amount of variation in phenotypic traits. Their identification will help in our understanding of the genetic nature of common human diseases in a genomic era, where multiple loci and their interactions can be examined simultaneously.

Chapter 2
Haplotype Reconstruction

2.1 Introduction

For autosomes, the genetic material carried by each diploid individual is composed of two chromosomes (or haplotypes). Haplotype information carried by individuals in a sample may inform many genetic analyses, including linkage-disequilibrium mapping of disease genes and inference about evolutionary processes such as selection and recombination. However, obtaining haplotype information directly from diploid organisms in laboratories remains expensive, laborious and time-consuming. On the other hand, advances in high-throughput genotyping technologies have enabled the generation of accurate genotypes on hundreds of thousands of genetic markers rapidly and inexpensively. In this chapter, I consider how to reconstruct haplotypes from unphased genotypes in samples of genetically unrelated individuals.

Haplotyping in population samples shares the same underlying rationale with phase inference in samples of related individuals. That is, individuals share local stretches of their chromosomes derived from their common ancestors. The difference lies in the relative size of the locally shared stretches, which are much longer in related individuals (parent-offspring pairs, for example, tend to share stretches in size of tens of centi-Morgans, resulting from typically one to two recombination events per chromosome in meiosis). In other words, each of our chromosomes is a mosaic of others' chromosomes, with the lengths of mosaic pieces varying depending on the degree of genetic relatedness. Linkage-disequilibrium (LD), the nonrandom association of alleles among linked loci due to lack of sufficient historical recombination events accumulated, encapsulates such information and is exploited by almost all computational and statistical approaches proposed for haplotype inference. My method, described in detail in the following section, models multi-marker LD and accounts appropriately for its decay over distance and for the block-like patterns of haplotypes. In addition, the model benefits from the incorporation of additional data from public databases by augmenting the pool of reference chromosomes, mosaics of which construct the desired haplotypes of individuals under study. Figure 2.1 provides a simplified illustration of the underlying mechanism.

2.2 Methods

My approach was inspired by the Markov models commonly used for pedigree analysis (for examples, see Lander and Green 1987; Kruglyak *et al.* 1996; Abecasis *et al*. 2002) and shares several features with other hidden Markov models (HMM) used to describe sampled haplotypes as a mosaic of a set of reference haplotypes (Scheet and Stephens 2006; Mott and Flint 2002; Mott *et al*. 2000; Li and Stephens 2003; Daly *et al*. 2001). My method produces high-quality estimates of individual haplotypes given phase unknown genotypes and can also provide useful measure of the quality of inferred haplotypes.

To estimate haplotypes, my approach starts by randomly generating a pair of haplotypes that is compatible with the observed genotypes for each sampled individual. These initial haplotype estimates are then refined through a series of iterations. In each iteration, a new pair of haplotypes is sampled for each individual in turn using a hidden Markov model (HMM) that describes the haplotype pair as an imperfect mosaic of a set of reference haplotypes. The reference haplotypes can be phased haplotypes from external sources or internally constructed haplotypes of other individuals in the sample. Model parameters that characterize the probability of change in the mosaic pattern between every pair of consecutive markers and the probability of observing an imperfection in the mosaic at each specific point, are also updated using a hybrid of approximate Gibbs' sampler and Expectation-Maximization (EM) algorithm. After many iterations (typically 20-100), a consensus haplotype can be constructed by merging the haplotypes sampled in each iteration (one merging algorithm is described in Appendix 2.1).

We have implemented the model outlined in the paragraph above in a software package MACH 1.0 (MACH abbreviated for Markov Chain Haplotyping). Paragraphs below describe the underlying statistical model.

Hidden Markov Model. My model resolves a set of unphased genotypes **G** into an imperfect mosaic of several reference haplotypes. Assume that H template haplotypes are each genotyped at L loci and let $\mathrm{T}_j(\mathrm{i})$ denote the allele observed at locus j in reference haplotype i. Furthermore I define a series of indicator variables $S_1, S_2, \ldots, S_L$ that denote a hypothetical (and unobserved) mosaic state underlying the unphased genotypes. At a specific position *j* there are H^2 possible states. A specific state, such as $S_j = (x_j, y_j)$, indicates that the first chromosome uses reference haplotype x_j as a template whereas the second chromosome uses reference haplotype y_j as a template at the particular locus *j*.

The key interest is in making inferences about the sequence of mosaic states **S** that best describe the observed genotypes. Knowledge of **S** will implicitly order alleles at heterozygous sites and suggest an allele for each untyped location. I calculate the joint probability of the observed genotypes and an underlying haplotype state as:

$$P(\mathbf{G},\mathbf{S} \mid \theta,\varepsilon) = P(S_1)\prod_{j=2}^{L} P(S_j \mid S_{j-1},\theta)\prod_{j=1}^{L} P(G_j \mid S_j,\varepsilon)$$

In the model above, $P(S_1)$ denotes the prior probability of the initial mosaic state and is usually assumed to be equal for all possible configurations, $P(S_j|S_{j-1})$ denotes the transition probability between two mosaic states and reflects the likelihood of historical recombination events in the interval between *j-1* and *j*, $P(G_j|S_j)$ denotes the probability of observed genotypes at each position conditional on the underlying mosaic state and reflects the combined effects of gene conversion, mutation and genotyping error. Detailed description of the model parameters θ and ε can be found in the Parameter Estimation section.

One key assumption made for the above model to hold is no interference. Otherwise, the first-order Markov does not suffice for modeling the transition probabilities. Biologically speaking this assumption is wrong. Statistically speaking the degree of departure is negligible since the probabilities of interference are in an even lower order than the already tiny state-changing (crossing over, or recombination) probabilities. Second, the model assumes that gene conversion, mutation and genotyping error events are not context specific such that $P(G_j|S, G) = P(G_j|S_j)$.

Monte-Carlo Haplotyping Procedure. To estimate haplotypes in a sample of genotyped individuals my model first assigns a random pair of haplotypes to each individual, consistent with the observed genotypes. To do so, the model randomly orders alleles at each heterozygous site and sample alleles at untyped sites according to population frequencies. Then, it updates the haplotypes for each individual in turn by using the current set of haplotype estimates for all individuals as templates and sampling **S** proportional to the $P(\mathbf{S}|\mathbf{G}, \theta, \varepsilon) \propto P(\mathbf{G},\mathbf{S}|\ \theta, \varepsilon)$. Note that since the S_j's define a Markov Chain this sampling can be done conveniently using Baum's forward and backward algorithm (Baum 1972). A new set of haplotypes for an individual is then defined according to $P(\mathbf{S}|\mathbf{G}, \theta, \varepsilon)$, allowing imperfect mosaics and respecting observed genotypes in case of mismatches with the sampled set of haplotypes. The update procedure is repeated several times, looping over all individuals (more updates result in gradual refinement of the estimated haplotypes, but very accurate haplotype estimates can often be obtained in ~20 iterations, see Table 2.1). After a pre-specified number of iterations are completed, a consensus haplotype solution is generated by identifying a set of haplotypes that can be transformed into any of the reconstructed haplotypes across iterations with a minimal number of switches according to an algorithm described in

Appendix 2.1.

Parameter Estimation. Parameters in the above procedure are the transition probabilities *P(S$_j$|S$_{j-1}$, θ)* and emission probabilities *P(G$_j$|S$_j$,ε)*. Transition probabilities are defined as a function of the crossover parameter *θ$_j$*:

$$P(S_j \mid S_{j-1},\theta) = \begin{cases} \theta_j^2 / H^2 & \text{if } x_j \neq x_{j-1} \text{ and } y_j \neq y_{j-1} \\ (1-\theta_j)\theta_j / H + \theta_j^2 / H^2 & \text{if } x_j \neq x_{j-1} \text{ or } y_j \neq y_{j-1} \\ (1-\theta_j)^2 + (1-\theta_j)\theta_j / H + \theta_j^2 / H^2 & \text{if } x_j = x_{j-1} \text{ and } y_j = y_{j-1} \end{cases}$$

The possible values of *P(S$_j$|S$_{j-1}$)* reflect both the overall rate of change in the mosaic for the interval, given by *θ$_j$*, and the fact that when a change occurs a new mosaic state is selected at random among all H possible states.

Now let *T(S$_j$) = [T(x$_j$) , T(y$_j$)]* denote the genotype implied by state *S$_j$* and define the emission probabilities *P(G$_j$|S$_j$)* as a function of the locus-specific error parameter ε$_j$:

$$P(G_j \mid S_j, \varepsilon) = \begin{cases} (1-\varepsilon_j)^2 + \varepsilon_j{}^2 & T(S_j) = G_j \text{ and } G_j \text{ is heterozygote} \\ 2(1-\varepsilon_j)\varepsilon_j & T(S_j) \neq G_j \text{ and } G_j \text{ is heterozygote} \\ (1-\varepsilon_j)^2 & T(S_j) = G_j \text{ and } G_j \text{ is homozygote} \\ 2(1-\varepsilon_j)\varepsilon_j & T(S_j) \text{ is heterozygote and } G_j \text{ homozygote} \\ \varepsilon_j{}^2 & T(S_j) \text{ and } G_j \text{ are opposite homozygotes} \end{cases}$$

Initially, my model sets $\theta_j = \theta = 0.01$ and $\varepsilon_j = \varepsilon = 0.01$ or some other suitable constant. While sampling a new mosaic state for each individual, my algorithm keep track of the number and location of change points in the mosaic and also of the number of times that the genotype implied by the sampled mosaic state matches or does not match the observed genotype. Let $CO_{i,j}$ be the number of changes in mosaic states from marker j to marker $j+1$ for individual I, and $MM_{i,j}$ the number of mismatched alleles between the observed genotypes and genotype implied by the sampled mosaic state at marker j for individual I, both taking values 0, 1 or 2. These quantities are then used to update the θ_j and ε_j parameters for the next iteration:

$$\tilde{\theta}_j = \frac{\sum_{i=1}^{N} CO_{i,j}}{N} \text{ and } \tilde{\varepsilon}_j = \frac{\sum_{i=1}^{N} MM_{i,j}}{N}, \text{ where N is the number of individuals}$$

It is important to avoid setting either $\theta_j = 0$ or $\varepsilon_j = 0$, as that could make it difficult for the Markov sampler to investigate different mosaic configurations. To avoid this, a combined crossover parameter is estimated for intervals with a small number of sampled changes in mosaic state and an analogous procedure is employed for markers with a small number of observed mismatches between the constructed mosaic and observed genotypes. The formula below provides the recipe for dealing with small θ_j's. The same rule applies to ε_j's.

$$\theta_{baseline} = \frac{\sum_{j=1}^{M}[\theta_j * I(\theta_j < \theta_{cutoff})]}{\sum_{j=1}^{M} I(\theta_j < \theta_{cutoff})} \text{ and } \theta_j = \theta_{baseline} \text{ if } \theta_j < \theta_{cutoff}\text{, where } \theta_{cutoff} = \frac{1}{N}$$

Overall, I expect the θ_j's will reflect a combination of population recombination rates, and the relatedness between the haplotypes being resolved and the true underlying haplotypes. For example, if chromosomes carried by individuals of European descent are used as templates to resolve genotypes of Asian individuals, I expect, on average, higher θ estimates than when chromosomes of other Asian individuals are used as templates. I also tried using distance between flanking markers to inform estimates of θ_j (since θ's are generally larger over larger intervals), but did not find noticeable improvements. I expect that ε_j will reflect the combined effects of genotyping error, gene conversion events, recurrent mutation and – when genotype data from multiple platforms or laboratories is used – assay inconsistencies between different platforms.

Computational Efficiency. A number of optimizations are possible to increase the computational efficiency. For example, since haplotype states are unordered, only *H(H+1)/2* distinct states must considered at each location, rather than H^2 distinct states.

Below, I summarize some of the other efficiencies that we identified and how these are implemented in MACH 1.0.

Transition Matrices. When sampling a mosaic state **S** conditional on the observed genotypes **G**, we rely on Baum's forward and backward algorithm. The algorithm requires a series of left and right conditioned probability vectors which provide an indication of the relative probability of a specific state at a given location conditional on observed genotypes at markers to its left (or right). For example, the probability of observing state *(x,y)* at location *j* conditional on all preceding genotypes is simply:

$$\begin{aligned} Left_j(x,y) &= P(S_j=(x,y)\,|\,G_1,G_2,...G_{j-1}) \\ &= \sum_{(a,b)} Left_{j-1}(a,b)P(S_j=(x,y)\,|\,S_{j-1}=(a,b))P(G_{j-1}\,|\,S_{j-1}=(a,b)) \end{aligned}$$

$Left(S_1=i)=\frac{1}{H^2}$ or simply $Left(S_1=i)=1$ since the factor of $\frac{1}{H^2}$ applies to all left probabilities.

The calculation of these probabilities can be sped up by taking advantage of the regular patterns in the transition matrices. Specifically, I define the following quantities:

$$C(a)=\sum_b Left_{j-1}(a,b)P(G_{j-1}\,|\,S_{j-1}=(a,b))$$

$$C=\sum_a C(a)$$

Then, the previous definition becomes:

$$\begin{aligned} Left_j(x,y) = {} & Left_{j-1}(x,y)P(G_{j-1} \mid S_{j-1} = (x,y))(1-\theta_j)^2 + \\ & \frac{C(x)(1-\theta_j)\theta_j}{H} + \\ & \frac{C(y)(1-\theta_j)\theta_j}{H} + \\ & C\theta_j^{\ 2} / H^2 \end{aligned}$$

$$Left(S_1 = i) = 1$$

My algorithm calculates *C(a)* and *C* along the way and uses this updated definition to calculate left conditional probabilities for each possible state. Thus, computational requirements become *O(H)* rather than *O(H^2)* using the original definition. An analogous speed up is available for right conditioned probabilities.

Memory Efficiency. One large computational constraint when applying such an algorithm on a genomic scale is the storage required to track left conditioned probabilities. Typically, this requires storage of *L* vectors each with H^2 elements (or, as noted above *H(H+1)/2* elements). This requirement becomes cumbersome as the number of polymorphic sites increases. We devised a solution that requires storage of only *2*sqrt(L)* vectors. For notational convenience let *K = sqrt(L)*. My algorithm pre-allocates 2*K* vectors and organizes these into two groups: a framework set of *K* vectors, and a working set of another *K* vectors. When left conditional probabilities are first calculated, proceeding left to right, we store every K^{th} vector in the framework set and discard other intermediate results. Then, as these vectors are used in the second pass of the chain (which combines left and right conditional probabilities, proceeding right to left), we recalculate *K* of these vectors at a time (starting from the nearest vector in the framework set) and store them in the working set of *K* vectors. Completing the full chain requires calculation of all *L* vectors of left conditional probabilities, recalculation of *K* of these vectors *L*/*K* times, and calculation of *L* vectors of right conditional probabilities. Overall, this solution no more than doubles computing time (since each vector of left conditional probabilities must be calculated twice), but reduces memory requirements from O(*L*) to O($L^{1/2}$). The solution is general and can be applied to many other Hidden Markov Models.

Reducing the Number of Templates. If all available chromosomes are used as templates, the computational complexity of my algorithm increases cubically with sample size because of the need to explore the $(2N\text{-}2)^2$ configurations (or again, as noted above $H(H+1)/2$ configurations where $H=2N\text{-}2$) for *N* individuals. One way to avoid this is to restrict the size of the template pool. When there are more than a pre-specified number of potential templates (say H = 200 or 300), I typically select a random subset of these for each update. With this restriction, the complexity of my algorithm increases only linearly with sample size since we examine a fixed number ($H(H+1)/2$) of configurations) for *N* individuals. Furthermore, even though each update is based on only a random sample of the available haplotypes, the overall quality of solutions still increases with sample size. When the focus is on genotype imputation (Chapter 3), rather than haplotyping, an alternative is to use as templates individuals who have been genotyped for the markers being imputed (e.g. the HapMap reference samples). Both of the above solutions are heuristics that trade off some accuracy for computational efficiency due to exploiting partial information: a random subset in the first case and a selected set (haplotypes of individuals with the most genotype information) in the second case. An alternative strategy for reducing the size of the template pool is to group haplotypes locally by exploiting local similarities and redundancies among the haplotypes in the pool with no (if only identical haplotypes are grouped locally) or little (if similar haplotypes are grouped locally) loss of information. These redundancies have already been exploited to increase computational efficiency in the handling of other Markov models (Abecasis *et al*. 2002; Markianos *et al*. 2001), and our preliminary implementations (Chen *et al*. unpublished data) suggest that speed-ups of 5-10x are possible.

2.3 Data

Simulated datasets. To evaluate the performance of my approach, I simulated two sets of 100 1Mb regions that mimic the degree of LD in the HapMap CEU or YRI samples (Schaffner *et al*. 2005). In each region, I simulated genotypes for ~200 SNP markers, ascertained to mimic HapMap allele frequency patterns (details described in Marchini *et al*. 2006), in 90 individuals with 2% of the genotypes masked at random to model genotype missingness for genotyped markers.

Real datasets from FUSION. The Finland-United States Investigation of Non-Insulin-Dependent Diabetes Mellitus Genetics (FUSION) Study aims to identify genes that predispose to type 2 diabetes (T2D). In a recent genome-wide association scan, 1,161 Finnish individuals with type 2 diabetes (T2D) and 1,174 normal glucose tolerant Finnish controls were genotyped at 317,503 SNPs using the Illumina HumanHap300 BeadChip. In addition, 122 offspring were genotyped using the same chip, yielding 119 mother-father-offspring trios, one mother-father-two-offspring quartet and one parent-offspring pair (Scott *et al.* 2007). Excluding the one-parent-offspring pair and one of the two offspring from the quartet, I obtained a dataset consisting of 120 trios. Assuming no genotyping error and no parent-to-offspring recombination events, I established the "true" haplotypes of the 240 parents at every SNP locus except where all three people were heterozygous or where there was missing data. I then used my method to infer haplotypes of these parents based solely on their genotypes.

2.4 Results

Accuracy of reconstructed haplotypes

Simulated datasets. I applied my method to the simulated datasets, reconstructed individual haplotypes and tallied three measures of haplotyping quality (Marchini *et al.* 2006): (1) the number of incorrectly imputed missing genotypes; (2) among heterozygous sites, the number of consecutive sites that are phased incorrectly with respect to each other (this is the number of "flips" required to transform estimated haplotypes into the true haplotypes), and (3) the number of correctly imputed haplotypes across the entire 1Mb region. The three measures were averaged over all 100 regions and the results are summarized in Table 2.1. For comparison, the table also includes results from Beagle (Browning and Browning 2007), PHASE (Stephens and Scheet 2005; Stephens *et al.*

2001), and fastPHASE (Scheet and Stephens 2006), the latter being the two state of the art haplotyping algorithms according to Marchini *et al*. (2006). Table 2.1 clearly shows that my method is competitive in all three measures: it results in slightly fewer incorrectly imputed genotypes, requires slightly fewer flips to transform imputed haplotypes into the true haplotypes, and produces slightly more correctly imputed haplotypes over the entire 1Mb stretch. Furthermore, estimates of haplotypes and missing genotypes obtained in 5-20 minutes using my method are comparable in quality to those produced by PHASE runs averaging ~1 day.

Real data of FUSION trios. I applied my haplotyper to the 240 FUSION parents, ignoring genotype information from their offspring. Reconstructed haplotypes were compared to the "true" haplotypes inferred from trio data at loci with no uncertainty. For each autosomal chromosome, two regions of length ~2Mb were picked, inferred and evaluated. The average number of flips and correctly inferred haplotypes across the 44 regions are tabulated in Table 2.2, along with results from PHASE (Stephens and Scheet 2005; Stephens *et al*. 2001) and fastPHASE (Scheet and Stephens 2006). The number of correctly imputed genotypes was not applicable since I did not attempt to mask any parental genotypes for imputation. My method is clearly comparable to the two state-of-the-art haplotypers in terms of accuracy and computational resources invested. For example, haplotypes reconstructed in 10 iterations were already of reasonable accuracy compared with those inferred by fastPHASE. In addition, my method is very flexible regarding computational investment, with the overall quality of inferred haplotypes improving with the number of iterations invested. Haplotypes of similar accuracy could be obtained in ~10 hours (200-300 iterations) using my method, while taking over three days for PHASE.

Quality measures.

Results shown above have focused on assessing the accuracy of point estimates of reconstructed haplotypes. Quality measures, or measures of uncertainty, provide important auxiliary information that can facilitate proper and more powerful downstream analyses. My method generates two quality measures quickly and conveniently: (1) the estimated number of incorrectly inferred genotypes, and (2) the estimated number of flips needed to transform inferred haplotypes into the correct haplotypes. Figure 2.2 shows that these quality measures provide reasonably accurate evaluation of proposed solutions and that they slightly underestimate quality in cases where the error rates are high.

Benefits from External Information.

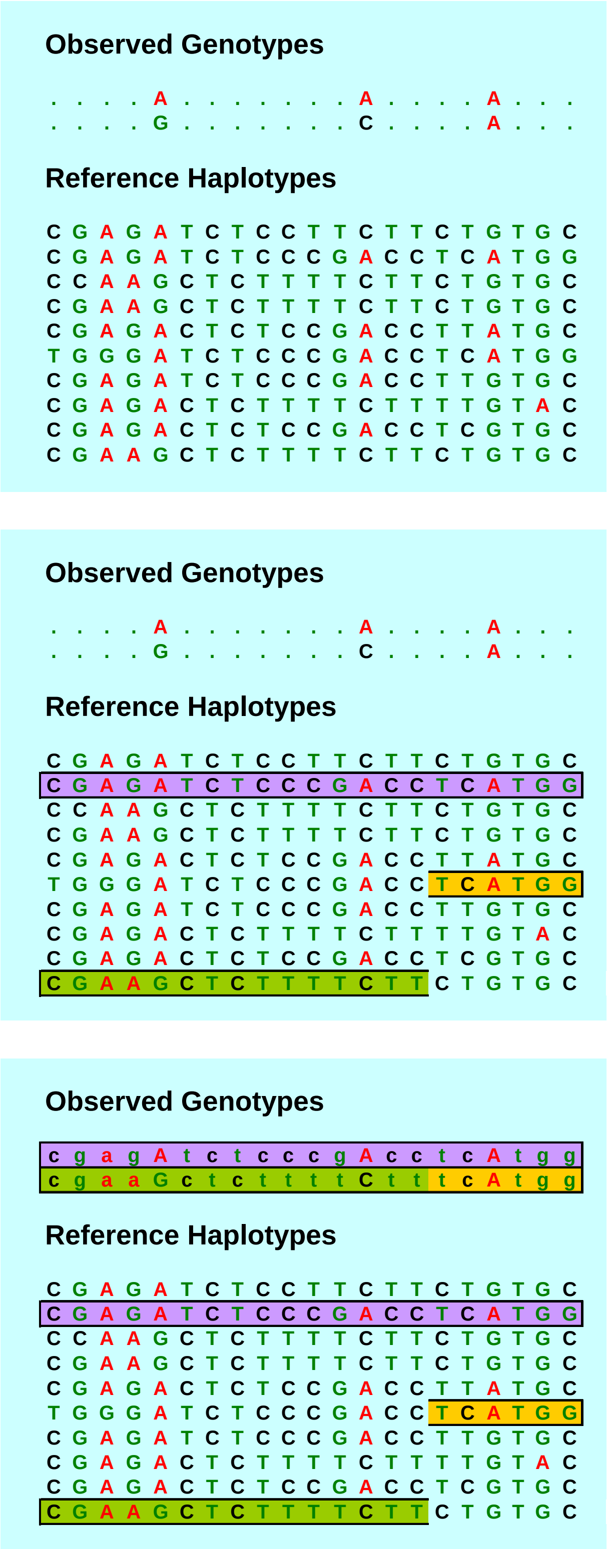
One useful feature of the proposed method is that it allows incorporation of external information such as data from the International HapMap Project. I consider it worthwhile to evaluate if, when, and how much haplotype inference benefits from external information. In particular, I simulated scenarios where the number of sample individuals to be haplotyped ranges from relatively small (60) to relatively large (500). In addition, I simulated an external set of 120 known haplotypes, from the same underlying population as the individuals under study. I then ran my haplotyper with or without joint modeling with the external set of 120 haplotypes, holding computational investment constant. Table 2.3 summarizes the accuracy of inferred haplotypes. Again, I simulated two populations, mimicking HapMap CEU and YRI respectively. I noticed haplotyping is slightly harder in the African population than in the European population as expected because of the lower level of LD in the African population. More importantly, Table 2.3 shows that higher quality haplotypes can be obtained with the aid of external information, especially when the number of sample individuals is relatively small. For example, the average number of per-person flips decreases by more than 20% (from 1.98 to 1.53) with the aid of "HapMap" chromosomes in simulated datasets mimicking HapMap CEU. Benefits from the incorporation of external data drop gradually when the number of individuals under study increases. As shown in the table, when the number of sampled individuals is relatively large (500), discarding external data generates constructed haplotypes that are as accurately as (if not even slightly more accurate than) utilizing external data. This table attempts to compare the two approaches (with and without joint modeling with external data) when investing the same amount of computation. Obviously, if one exploits all available information (that is, using the combined pool of external haplotypes and haplotypes of other sample individuals being reconstructed), a better solution would be obtained than if one ignores external information. Similar patterns carry over to simulated datasets mimicking HapMap YRI. The consequences of borrowing information from inappropriate external data (for example, borrow HapMap YRI information for the analysis of Caucasian individuals) are evaluated in Chapter 3.

2.5 Discussion

In this chapter, I have presented a hidden Markov model for haplotyping a sample of unrelated individuals and have demonstrated, through applications to both simulated and real datasets, that it is competitive with (indeed outperforms) other methods including Beagle, PHASE, and fastPHASE. Major advantages of my methods include its ability to fully exploit multilocus LD information, its computational efficiency and flexibility, and its ability to perform joint modeling with external data from public databases.

My method is computationally feasible for large datasets, which will be further demonstrated in the following chapters. My algorithm is approximately linear in the number of markers and approximately quadratic in the number of reference haplotypes. The pool of reference haplotypes can be reconstructed haplotypes of all other individuals in the same sample, known haplotypes from public databases, the combined pool of reconstructed and known haplotypes, or a random subset of the combined pool. Furthermore, when different levels of missing data are present, my model allows assigning more weights to haplotypes of individuals with more information when constructing the pool of reference haplotypes. Finally, because each chromosome is modeled as a mosaic of the reference chromosomes, the proposed model naturally handles samples of potentially mixed ethnic and genetic origins by searching for the closest-match local stretches.

Figure 2.1 Cartoon Illustrations of Haplotype Reconstruction and Genotype Inference.



Observed Genotypes

```
. . . . A . . . . . . . A . . . . A . . .
. . . . G . . . . . . . C . . . . A . . .
```

Reference Haplotypes

```
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
```

Observed Genotypes

```
. . . . A . . . . . . . A . . . . A . . .
. . . . G . . . . . . . C . . . . A . . .
```

Reference Haplotypes

```
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
```

Observed Genotypes

```
c g a g A t c t c c c g A c c t c A t g g
c g a a G c t c t t t t C t t t c A t g g
```

Reference Haplotypes

```
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
```

Figure 2.2 Quality Measures for Accuracy of Reconstructed Haplotypes.

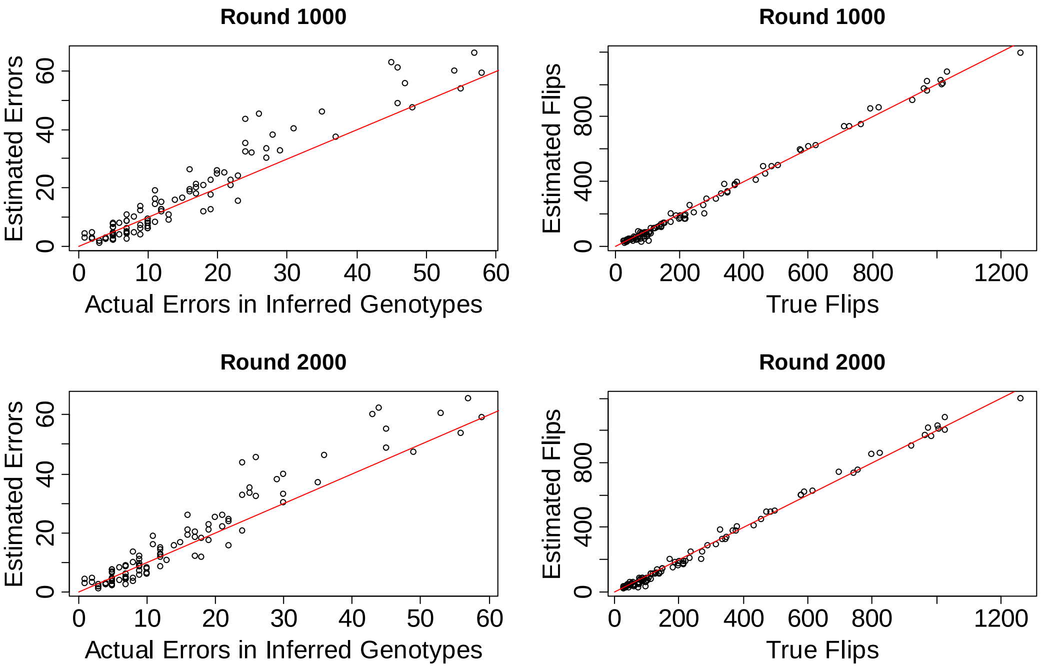


Table 2.1 Quality of Haplotype and Missing Genotype Estimates in Simulated Datasets.

			Dataset Mimicking HapMap CEU			Dataset Mimicking HapMap YRI		
Method	# Iterations	Computation Time	# Errors	# Flips	# Perfect	# Errors	# Flips	# Perfect
MACH	20	~2 m	11.6	216	26.5	17.9	256	22.6
	60	~5 m	10.8	200	28.4	16.6	232	24.1
	200	~15 m	10.6	192	29.1	16.3	222	25.1
	1000	~1.4 h	10.6	182	29.3	16.3	218	25.5
	3000	~ 3.9h	10.5	178	29.7	16.1	214	25.7
PHASE	-	~25 h	12.6	201	25.3	19.8	270	19.9
fastPHASE	-	~17 m	12.9	220	20.1	22.9	331	11.7
BEAGLE	-	~2 sec	13.9	230	21.1	23.1	332	13.1

The table summarizes results from the analysis of two sets of 100 simulated 1 Mb regions. The two sets reflect the degree of LD in the HapMap CEU and YRI samples, respectively. In each region, ~200 markers were ascertained to mimic HapMap allele frequency spectra and 2% missing data was introduced at random.

The data were then analyzed with one of four haplotypers (MACH, PHASE, fastPHASE, and BEAGLE) and the quality of haplotype solutions and imputed genotypes was evaluated. The number of missing genotypes imputed incorrectly (#Errors), the number of switches in haplotype phase required to convert the estimated haplotypes into the simulated haplotypes (#Flips) and the number of perfectly estimated haplotypes (#Perfect) were recorded. Averages of these three quantities are tabulated.

Mach 1.0 was run with default settings and different numbers of iterations. PHASE version 2.1.1(Stephens and Scheet 2005; Stephens *et al*. 2001) was run with default settings, as recommend by Matthew Stephens. Parameters for fastPHASE version 1.3 (Scheet and Stephens 2006) was run with default settings, as recommended by Paul Scheet. All timings refer to a 2.33 GHz Pentium Xeon.

Table 2.2 Quality of Reconstructed Haplotypes in FUSION Trio Dataset.

Method	# Iterations	Computation Time (in hours)	# Flips	# Perfect
MACH	10	~0.4	355	111
	60	~2.2	249	135
	200	~6.9	238	138
	300	~11.9	235	138
	1000	~43.1	226	139
	5000	~194.0	216	140
PHASE	-	~76.8	236	137
fastPHASE	-	~0.7	544	45

The table summarizes results from reconstructing haplotypes for 240 parents in a real dataset: FUSION.

For comparison, the table also lists results from PHASE and fastPHASE. Computational time (in hours), the number of switches in haplotype phase required to convert the estimated haplotypes into the simulated haplotypes (#Flips) and the number of perfectly estimated haplotypes (#Perfect) are tabulated. Numbers shown are averages from 44 autosomal regions.

Mach 1.0 was run with default settings and different numbers of iterations. PHASE version 2.1.1(Stephens and Scheet 2005; Stephens *et al*. 2001) was run with default settings, as recommend by Matthew Stephens. Parameters for fastPHASE version 1.3 (Scheet and Stephens 2006) was run with default settings, as recommended by Paul Scheet. All timings refer to a 2.33 GHz Pentium Xeon.

Table 2.3 Benefits of Incorporating External Information for Haplotyping.

#sampled individuals	Dataset Mimicking HapMap CEU				Dataset Mimicking HapMap YRI			
	with “HapMap”		without “HapMap”		with “HapMap”		without “HapMap”	
	# Flips	# Perfect	# Flips	# Perfect	# Flips	# Perfect	# Flips	# Perfect
60	1.49	0.37	1.98	0.29	1.85	0.31	2.53	0.23
100	1.37	0.40	1.68	0.35	1.61	0.36	2.06	0.30
200	1.26	0.43	1.35	0.41	1.40	0.41	1.51	0.39
500	1.21	0.44	1.18	0.45	1.33	0.42	1.26	0.44

Haplotypes of individuals in study sample were inferred with or without the aid of an external (“HapMap”) set of known haplotypes. Computational investment was held constant: a random subset of 118, 200 or 300 haplotypes was used as reference haplotypes when the number of sample individuals was 60, 100, or 200/500, regardless of the incorporation of “HapMap”.

Both statistics, namely the number of switches in haplotype phase required to convert the estimated haplotypes into the simulated haplotypes (#Flips) and the number of perfectly estimated haplotypes (#Perfect), are summarized per-person and are averaged over 100 datasets.

Appendix 2.1 Algorithm to Merge Haplotypes for Consensus.

After running a pre-defined number of iterations of the proposed hidden Markov model, the set of consensus haplotypes for each diploid individual is generated according to the following algorithm:

Step 1: Change the relative order of the pair of haplotypes sampled from each iteration so that the first heterozygous site is AL1/AL2. The two alleles AL1 and AL2 are defined for each marker consistently (albeit arbitrarily) across iterations.

Step 2: For each subsequent SNP, find the most frequently occurring haplotype configuration across iterations.

Step 3: If the most frequent configuration is in heterozygous state (i.e., AL1/AL2 or AL2/AL1), flip the relative order of the haplotype pair in iterations where the configuration is the other heterozygous state.

Step 4: Repeat Step 2 and Step 3 until the last SNP is reached.

Chapter 3

Genotype Imputation (*In Silico* Genotyping)

in Candidate Gene and Genome-Wide Association Studies

3.1 Introduction

It has been estimated that there are ~ 10 million common (minor allele frequency [MAF] > 1%) SNPs in the human genome (Kruglyak and Nickerson 2001; Hinds *et al.* 2005; The International HapMap Consortium 2007). Most ongoing genome-wide association studies (GWAS) rely on a commercial SNP genotyping panel that directly assays only a small fraction of SNPs in the human genome (Carlson *et al*. 2003; The International HapMap Consortium 2007). In these scans, the majority of SNPs in the genome must be evaluated indirectly using one or more of the genotyped SNPs as proxies (Barrett and Cardon 2006; Pe'er *et al*. 2006). Overall, GWAS provide a powerful method for successful identification of susceptibility loci in complex diseases.

For complex diseases, individual genome-wise association scans allow us to identify common alleles that make large contributions to disease risk, and a subset of the loci with smaller effects (Hirschhorn and Daly 2005). Meta-analysis of multiple genome-wide scans is needed to yield sufficient power to identify alleles that make smaller contributions to disease risk. In 2007, Scott *et al.* (2007), Zeggini *et al.* (2007), and

Diabetes Genetics Initiative (2007) provided an early example of the power of the combined analysis of multiple scans. Genotype imputation was used to combine GWAS for blood lipid levels (Kathiresan *et al.* 2008, Willer *et al.* 2008), height (Sanna *et al.* 2008), type-2 diabetes (Zeggini *et al.* 2008), body-mass index (Loos *et al.* 2008), and Crohn's disease (Barrett *et al.* 2008). The success of these meta-analyses can be dramatic: in the case of blood lipid levels (Kathiresan *et al.* 2008, Willer *et al.* 2008), a meta-analysis of three studies with relatively modest findings (each identifying one to three strongly associated loci), resulted in a total of 19 strongly associated loci including 7 loci not previously implicated in regulating cholesterol and lipoprotein levels in humans.

Although it should be possible to use one or more of the SNPs genotyped in each study as proxies for SNPs genotyped in the other studies (de Bakker *et al*. 2005; Carlson *et al*. 2004; Lin *et al*. 2004; Nicolae 2006; Zaitlen *et al*. 2007), meta-analyses of GWAS can be cumbersome because of the limited overlap between the different commercial panels and the fact that different choices of proxies for a particular SNP can lead to somewhat different conclusions. In my view, one particularly attractive approach for cross study analysis is to combine genotypes generated by the International HapMap Consortium (2007) with genotypes from individual studies, using a haplotyping algorithm that can handle genome scale data to impute genotypes at untyped markers. This strategy results in a situation where all studies are "genotyped" for all the markers examined by the HapMap consortium (albeit some markers would be genotyped using conventional means and other would be genotyped *in silico* [Burdick *et al*. 2006]). The approach relies again on the intuition introduced in Chapter 2 that even two apparently “unrelated” individuals share short stretches of haplotype derived from their common ancestors. Once one of these stretches is identified using genotypes for a few SNPs, alleles for intervening SNPs that are measured in some of the individuals, but not the others, can be imputed. Provided shared haplotype stretches are identified correctly, imputed genotypes will be accurate unless they have been disrupted by gene conversion or mutation events.

In this chapter, I provide a unified hidden Markov model for genotype imputation, also implemented in our software package MACH 1.0. The proposed Markov model describes sampled chromosomes as mosaics of each other and potentially external phase known haplotypes of additional individuals in a manner that efficiently uses all available genotype and haplotype data to impute each missing genotype. In particular, I show that genotype imputation using HapMap haplotypes as a reference is very accurate whether we have large amounts of data from genome-wide association scans or smaller amounts of data typical in fine-mapping studies. Furthermore, I show my approach is applicable to a variety of populations. I assess the performance of the genotype imputation for several currently available genotyping panels and illustrate how it might benefit from future advances, such as the 1000 Genomes Project (see www.1000genomes.org).

3.2 Methods

Genotype Imputation. Genotype imputation analyses proceed similarly to the haplotyping analyses described in Chapter 2, but do not require each sampled haplotype configuration to be stored. Instead, at the end of each iteration after burn in, a series of counters is updated to indicate the number of times each genotype was sampled at a particular position. Once all iterations are completed, these counters give are used to estimate the relative probability of observing each possible genotype, to impute the most likely genotype, to estimate the fractional allelic count, and to calculate various measures of the quality of imputed genotypes.

An alternative to sampling from a series of iterations is the most likely estimate (MLE) approach. The MLE approach is particularly advantageous when the model parameters (namely the crossover parameter θ's and the error parameter ε's) are known or are previously inferred. In such situations, Markov chains are not necessary since parameters no longer need to be updated. Instead, I take an external reference panel of haplotypes as "truth" and then find the probabilities of each of the three possible genotype guesses (at any biallelic SNP locus) for an uncertain (missing) genotype. The posterior probabilities are calculated by summing over normalized probabilities of all potential configurations of the mosaic states **S** compatible with the particular genotype guess, using only the external reference panel of known haplotypes.

The posterior probabilities of each potential mosaic state are key quantities of interest. I adopt Baum's (1972) forward and backward algorithm to obtain them. Specifically, I define the forward and backward probabilities as follows:

$$f_m(x,y) \equiv \Pr(g_1, g_2, ..., g_m, S_m = (x,y))$$

$$b_m(x,y) \equiv \Pr(g_{m+1},...,g_M \mid S_m = (x,y))$$

Define the starting point for the forward probability as:

$$f_1(x,y) \equiv \Pr(g_1, S_1 = (x,y))$$
$$= \Pr(g_1 \mid S_1 = (x,y)) \cdot \Pr(S_1 = (x,y)) = e_{(x,y)}(g_1) \cdot 1/H^2$$

where $x, y \in \{1,2,...,H\}$ and H is the number of reference haplotypes.

The remaining forward probabilities can be obtained through the following recursive formulation:

$$f_m(x,y) = \Pr(g_1, g_2,...,g_m; S_m = (x,y))$$
$$= \Pr(g_m \mid S_m = (x,y)) \cdot \sum_{(a,b)} \left[f_{m-1}(a,b) \cdot \Pr(S_m = (x,y) \mid S_{m-1} = (a,b)) \right]$$

where $x, y, a, b \in \{1,2,...,H\}$ and again H is the number of reference haplotypes.

Similarly the starting point and recursive formulation of the backward probabilities are:

$$b_{M-1}(x,y) \equiv \Pr(g_M \mid S_{M-1} = (x,y))$$
$$= \sum_{(a,b)} [\Pr(g_M \mid S_M = (a,b)) \cdot \Pr(S_M = (a,b) \mid S_{M-1} = (x,y))]$$
$$= \sum_{(a,b)} [e_{(a,b)}(g_M) \cdot \Pr(S_M = (a,b) \mid S_{M-1} = (x,y))]$$

$$
\begin{aligned}
&b_m(x,y) = \Pr(g_{m+1},\ldots,g_M \mid S_m = (x,y)) \\
&= \sum_{(a,b)} \left[b_{m+1}(a,b) \cdot e_{(a,b)}(g_{m+1}) \cdot \Pr(S_{m+1} = (a,b) \mid S_m = (x,y)) \right]
\end{aligned}
$$

Lastly, the posterior probabilities of each potential mosaic state are obtained through the following formula:

$$
\begin{aligned}
&\Pr(S_m = (x,y) \mid G) \propto \Pr(G, S_m = (x,y)) \\
&= \Pr(g_1, g_2, \ldots, g_m, S_m = (x,y)) \cdot \Pr(g_{m+1},\ldots,g_M \mid S_m = (x,y)) \\
&\equiv f_m(x,y) \cdot b_m(x,y) \\
&\textit{where } m = 1,2,\ldots,M-1
\end{aligned}
$$

$$
\Pr(S_M = (x,y) \mid G) \propto \Pr(G, S_M = (x,y)) \equiv f_M(x,y)
$$

Given the posterior probabilities, I can either sample the hidden state or obtain the most likely estimate accordingly. Without loss of generality, consider a SNP with alleles A and B. Let $n_{A/A}$, $n_{A/B}$, and $n_{B/B}$ be the number of times each possible genotype was sampled after I = $n_{A/A}$ + $n_{A/B}$ + $n_{B/B}$ iterations. For downstream analysis of imputed alleles, I typically consider either the most likely genotype or the expected number of copies of allele A. The most likely genotype is simply the genotype that was sampled most frequently. The expected number of counts of allele A is the genotype score g = ($2n_{A/A}$ + $n_{A/B}$) / I. With the MLE alternative, the most likely genotype is the genotype guess with the largest posterior probability among the three. Let $p_{A/A}$, $p_{A/B}$, and $p_{B/B}$ denote the posterior probabilities of the three possible genotype guesses with the obvious constraint $p_{A/A}$ + $p_{A/B}$ + $p_{B/B}$ = 1. The expected number of copies of allele A is simply $2p_{A/A}$ + $p_{A/B}$.

Both of the genotype score and posterior probabilities can be conveniently incorporated into a variety of analyses, including regression-based association analyses of discrete and quantitative traits. See Chapter 5 for detailed discussions on post-imputation analysis.

Estimates of Imputation Quality. To measure the accuracy of imputation for a single imputed genotype for individual *I* at marker *j* ($IG_{i,j}$), I define the genotype quality score $Q_{i,j} = n_{Igi,j} / I$. Alternatively, $Q_{i,j}$ = max ($p_{A/A}$, $p_{A/B}$, and $p_{B/B}$) in the MLE approach. This quantity can be averaged over all genotypes for a particular marker to quantify the average accuracy of imputation for that marker:

$$Q_j = \frac{\sum_{i=1}^{N} Q_{ij}}{N}$$ where *N* is the total number of individuals.

I have found that a better measure of imputation quality for a marker is the estimated r^2 between true allele counts and estimated allele counts, which will be further discussed in results section of this chapter. This quantity can be estimated by comparing the variance of the estimated genotype scores with what would be expected if genotype scores were observed without error. For a given SNP, let Var*(g)* be the variance of estimated genotype and let *p* = mean*(g)/2* be the estimated frequency of allele A. Assuming Hardy-Weinberg equilibrium (HWE), the following quantity measures the observed dispersion of genotype scores over its expected value and can be used as an estimate of r^2 with true genotypes.

E(r^2 with true genotypes) = Var(g) / [2p(1-p)], where

$$\text{Var(g)} = \frac{1}{N-1}\sum_{i=1}^{N}(g_i - \frac{\sum_{i=1}^{N} g_i}{N})^2 \quad \text{where}$$

N is the total number of individuals and

g_i is the genotype score for individual i.

The intuition behind this estimator is that mis-calling the major allele homozygotes, the dominating type of imputation error, results in under-dispersion. Relaxing the HWE assumption, an alternative estimator is defined:

$$\text{E}(r^2 \text{ with true genotypes}) = \text{I} * \text{Var(g)} / ((n_{A/A} + n_{B/B})/\text{I} - [(n_{A/A} - n_{B/B})/\text{I}]^2)$$

Empirically, I have found that while both definitions lead to similar conclusions, the first definition appears to be marginally better (refer to Figure 3.4 and Results section for details).

Association Analysis Using Imputed Genotypes. For downstream analysis for disease-marker association testing, I recommend using imputed genotype scores g (ranging continuously between 0 and 2) to properly account for imputation uncertainty. Specifically, in the examples described below, for FUSION GWAS, I used the imputed genotype scores as covariates in a logistic regression that also included age, sex and geographic origin as covariates; for analyzing simulated case control data, I fitted a logistic regression model where the imputed genotype score is the sole predictor.

3.3 Data

Age-Related Macular Generation (AMD) Candidate Gene Study. In the Michigan AMD study (Li *et al*. 2006), 544 unrelated Michigan individuals affected with AMD and 268 unaffected controls were genotyped at 84 SNP loci in a ~123Kb region overlapping AMD-predisposing gene *CFH* on chromosome 1. In addition, the study genotyped the same 84 SNP loci on the 60 HapMap CEU founders.

FUSION GWAS. As introduced in Chapter 2, the FUSION study genotyped 1,161 Finnish individuals with type 2 diabetes (T2D) and 1,174 Finnish controls at 317,503 SNPs using the Illumina HumanHap300 BeadChip in stage one of a two-stage genome-wide scan for T2D susceptible genes (Scott *et al.* 2007). Subjects collected are from the FUSION (Valle *et al*. 1998; Silander *et al*. 2004) and Finrisk 2002 (Saaristo *et al*. 2005) studies. Among the 317,503 GWA SNPs, 290,690 autosomal SNPs had minor allele frequency (MAF) >= 5% and passed quality-control (QC) criteria.

HapMap Data. The International HapMap consortium (2007) generated genotype data on over three million polymorphic SNPs for 270 individuals. Individuals genotyped include 30 father-mother-adult child trios of northern and western European ancestry living in Utah from the Centre d'Etude du Polymorphisme Humain (CEPH) collection (CEU); 30 trios from the Yoruba in Ibadan, Nigeria (YRI); 45 unrelated Han Chinese individuals in Beijing, China (CHB); and 45 unrelated Japanese individuals in Tokyo, Japan (JPT). The CEU and YRI samples each form an analysis panel, representing two major continental and ethnic groups. The CHB and JPT samples together form an analysis panel, representing the East Asian population.

The International HapMap consortium estimated haplotypes for the 210 unrelated individuals (60 CEU parents, 60 YRI parents, 45 unrelated CHB individuals and 45 unrelated JPT individuals), separately for each analysis panel using coalescent based statistical methods that take the relatedness of the CEU and YRI trios into proper account. For each analysis panel, over two million polymorphic SNPs were identified, successfully genotyped and subsequently phased: ~2.56 million SNPs in CEU, ~2.86 million in YRI and ~2.42 million in CHB+JPT.

Human Genome Diversity Project (HGDP) Data. In a recent global survey for haplotype variation and evaluation of tagSNP portability (Conrad *et al.* 2006), the Human Genome Diversity Project has collected genotypes for SNPs spread across 36 genomic regions on 927 unrelated individuals from 52 worldwide populations. The 52 populations encompass samples from all major continental groups across the world. Specifically, sampled individuals were from Africa, Europe, Middle East, Central and South Asia, East Asia, Oceania and the Americas. The 36 genomic regions were selected across a wide spectrum of local gene densities and LD levels to maximize the extent to which the regions were representative of the human genome. Each region spanned ~330 Kb including a central “core” region of ~90 Kb, where genotyping of ~60 SNPs was attempted, and two ~120 Kb flanking regions on either side, where ~12 SNPs were attempted. In total, 1,864 SNPs in 32 autosomal chromosomal regions (average minor allele frequency 15% - 24%, depending on population).

Simulation Datasets for Power Assessment. To assess whether my imputation-based approach might improve power in individual association scans, I simulated 10,000 chromosomes for a series of 1 Mb regions, using a coalescent model that mimics LD in real data, accounts for variations in local recombination rates, and models population history consistent with Europeans or Africans. In other words, I simulated LD patterns mimicking the HapMap CEU or YRI within each of 1 Mb regions (Schaffner *et al*. 2005).

I then took a random subset of 120 simulated chromosomes to generate a region specific pseudo-“HapMap”. Out of the pool of the remaining 9,880 chromosomes, I simulated a series of datasets each with 500 cases and 500 controls. Specifically, the case-control datasets were generated by picking one of the polymorphic sites at random as a “disease susceptibility locus” and subsequent sampling the two alleles at the disease locus probabilistically according to their corresponding frequencies expected among cases and controls (see Appendix 3.1). The susceptibility allele varies in frequency between 2.5% and 50% and larger effect sizes were simulated for rarer disease alleles to ensure comparable power in situations where the disease locus was available for direct testing. In addition, I simulated 2,000 datasets where the “disease allele” had no effect to calibrate region-wide type I error rates.

For the simulated HapMap data, polymorphic sites were ascertained and thinned to match the corresponding (CEU or YRI) Phase II HapMap (The International HapMap Consortium 2007) marker density, allele frequency spectrum and LD pattern, leading in the end to ~1,000 SNPs in each region for the panel of 120 HapMap chromosomes. Based on the thinned HapMap panel, I selected a set of 100 tagSNPs for each region that included the 90 tagSNPs with the largest number of proxies and 10 additional SNPs picked at random among the remaining tags. The tagSNP selection approach taken above resulted in tagSNP sets that captured (at a conventional r^2 cutoff of 0.8) ~78% of the common variants (MAF > 5%) in the simulated CEU HapMap, similar to the real life performance of the Illumina HumanHap300 BeadChip SNP genotyping platform.

Finally, each of the simulated datasets was analyzed using the selected tagSNP panel and one of the four analysis strategies: (a) single marker chi-squared association tests, (b) single and multi-marker association tests as suggested by the PLINK (Purcell *et al*. 2007) program based on LD in the simulated HapMap, (c) tests using imputed allele counts for all the markers in the simulated HapMap, or (d) multiple-imputation (MI) version of (c). For (d), I performed 10 multiple independent imputations by starting from different random initial setup and combined results according to the standard Rubin's (1978) rule where the MI point estimator is simply the sample average of those from multiple imputed complete datasets and MI variance estimator is a weighted sum of between-imputation and average within-imputation variances .

3.4 Results

Age-Related Macular Generation (AMD) Candidate Gene Study. To mimic the common practice of picking, genotyping and subsequently testing tagSNPs, I selected eleven tagSNPs to cover the 84 SNPs with an r^2 threshold of 0.8, based on LD calculated from the HapMap founders. I then masked genotypes at the remaining 73 non-tagSNPs in the Michigan individuals and inferred missing genotypes with the aid of CEU genotypes at all 84 SNP loci. Table 3.1 shows that imputed alleles differ from experimental ones only ~1.0% (~0.9%) of the time taking <1 minute's (~3 hours') computing time. In comparison, fastPHASE (Scheet and Stephens 2006) takes ~5 minutes to achieve an allelic discordance rate of ~1.0% and PHASE takes approximately one day to reach an allelic discordance rate of ~1.4%. In this particular example, PHASE (Stephens and Scheet 2005; Stephens *et al*. 2001) does not perform as well probably because it is not particularly designed and thus suitable for imputing a large proportion of missing genotypes (notice in this experiment, over 85% of genotypes are missing). The highly accurate imputed genotypes generated evidence for association that was very similar to that initially observed (Figure 3.1). The only outlier is the last SNP in the region and is in relatively low LD with any of the eleven tagSNPs.

FUSION GWAS. I applied my method to impute genotypes for untyped markers in the FUSION GWAS. Since a previous analysis suggested LD patterns in the HapMap CEU and in FUSION are similar (Willer *et al*. 2006), I used genotypes for the 290,690 autosomal markers (with MAF >= 5% and passing FUSION QC criteria) in the Illumina HumanHap300 BeadChip and for ~2.5 million polymorphic markers in the phased HapMap CEU chromosomes as input. After running the Markov chain procedure described above, I estimated the most likely genotype at each position (taking a majority vote across all iterations) and the expected number of copies of the minor allele at each position (a fractional value between 0 and 2) for each individual. I obtained similar results running 50-100 iterations of the Markov chains or using a smaller number of iterations (10-20) to estimate model parameters and then calculating most likely estimates for the missing genotypes and allele counts. The latter (MLE) approach requires less computational investment, especially when model parameters are estimated using a representative subset (say, several hundred) of individuals from the large pool of several thousand or more individuals examined in a large scale GWAS.

Different chromosomes were conveniently analyzed in parallel and, overall, imputing genotypes for all 2,335 unrelated individuals took <2 days for each of the largest chromosomes on a 2.40 GHz Pentium Xeon processor. In total, I imputed genotypes for 2,266,562 SNPs per individual. On average, my method used stretches of ~150 Kb from the HapMap CEU panel to reconstruct haplotypes for individuals in the FUSION sample.

To evaluate the quality of imputed genotypes, I contrasted my estimates of the most likely genotypes and the expected number of copies of the minor allele with actual genotype data for three sets of markers: (1) 521 SNP markers in a ~20 Mb region of chromosome 14 previous examined to fine-map a candidate linkage region (Willer *et al.*

2006), (2) 1,234 SNP markers selected to augment coverage of the Illumina HumanHap300 BeadChip in regions surrounding 222 candidate genes (Gaulton *et al.* 2008), and (3) 12,702 markers (also passing FUSION QC criteria) with MAF < 5% that were excluded from the set of 290,690 used for imputation. I expected the last two panels of markers to be harder to impute, because they represent SNPs that are not well tagged by the Illumina HumanHap300 BeadChip or that have lower MAF. I observed that 98.60% of the imputed alleles match actual genotyped alleles in the fine-mapping panel of 521 SNPs, 96.24% in the candidate gene panel of 1,234 SNPs and 98.73% in the lower MAF SNP panel. Furthermore, the average r^2 between imputed genotypes and actual genotypes was 90.4%, 79.1% and 74.0% in the three SNPs panels, respectively. This represents an improvement of 14-39% compared to the best available single marker tagSNPs, which provided an average r^2 of 76.5%, 52.8% and 35.5% in the three SNP panels, respectively. Figure 3.2 shows the improvement in r^2 for the first fine-mapping panel of 521 SNPs. I observed the overall distributional upward shift after imputation and that coverage (defined at an r^2 threshold of 0.8) increases from 62.0% to 87.1%.

As introduced in the Methods section, my model produces three estimates of imputation quality and these can be used to focus subsequent analyses on subsets of high quality genotypes. First, it produces a quality score that estimates the accuracy of each imputed genotype and is simply the proportion of iterations where the most likely genotype was selected (instead of an alternative solution). In an MLE approach, the quality score is the posterior probability of the most likely genotype. It produces an overall measure of the accuracy of imputation for each marker, which is the genotype quality score averaged across all individuals. By comparing the distribution of sampled genotypes in each iteration with the estimated allele counts that results from averaging over all iterations (or MLE estimates), it produces an estimate of the r^2 between imputed and true genotypes. Quality measures for individual genotypes were good predictors of imputation accuracy (Figure 3.3, Right Panel) and show that most imputed genotypes are called with a high degree of confidence (Figure 3.3, Left Panel). For example, as measured by their quality scores, the top 95% of genotypes had average quality scores of 98.9% and actually matched experimental genotypes 98.6% of the time. Most of the errors affect a single allele so that, when measured on a per allele basis (rather than per genotype basis), concordance increases to 99.3%.

To avoid preferential removal of rare genotypes or alleles, I recommend using the per marker quality scores to select a subset of imputed SNPs for analyses. The per marker quality measures provide an accurate aggregated estimate of the quality of imputed genotypes. Overall, I saw a correlation of 0.77 between the estimated and actual accuracy of imputed genotypes for each marker. I also saw a correlation of 0.84 between the r^2 estimated by my method and the actual r^2 that resulted from comparing allele counts with their imputed estimates. Figure 3.4 shows the ROC curve (Pepe 2003) for the two quality measures, showing that the estimated r^2 measure is more effective than the estimated accurary to discriminate poorly imputed markers from well imputed ones. In the FUSION GWAS scan, I used an r^2 threshold of 0.30 to decide which markers were well imputed and should be included in further analyses, and which were not. At this threshold, ~70% of poorly imputed markers (those where r^2 with experimental genotype is < 20%) were removed at the cost of only ~ 0.50% of better imputed markers (those where r^2 with experimental genotype is > 50%).

The results summarized so far compare a variety of imputed genotypes with experimentally derived counterparts. However, a more interesting comparison focuses on imputed genotypes that appear to show strong evidence for association, as those might motivate further downstream experiments. To evaluate the accuracy of imputed genotypes for these “strongly associated SNPs”, I compared imputed and experimental genotypes for markers that were selected for follow-up genotyping (for example, because imputed genotypes resulted in strong evidence for association but nearby directly genotyped markers did not). Table 3.2 summarizes the comparison of allele frequencies, association test statistics, and individual genotype calls between imputed genotypes and actual genotypes later determined by laboratory genotyping. Overall, it is clear that even among these strongly associated SNPs imputation provided accurate estimates of the true association test statistics and thus of the true p-values. The largest observed discrepancies were for rs17384005, rs11646114 and rs4812831 which were also the three markers for which my imputation approach estimated lower r^2 with actual genotypes. Figure 3.5 plots the –log(P-values) from imputation against those from the actual genotypes. We can see that (1) Largest departures typically have low estimated r^2 (<.55) values; (2) SNPs with high estimated r^2 values (>.97 shown) show little departure from the 45 angle line; and (3) Imputation generates reasonably accurate analyzing results even for SNPs with low r^2 values (<.3) with any tagSNP.

Remarkably, I observed that imputed genotypes could also be used to obtain very accurate estimates of LD between pairs of untyped markers, or of LD between a genotyped marker and an untyped marker. As shown in Figures 3.6 (r^2) and 3.7 (D'), estimates of LD between two SNPs obtained using imputed data are much closer to the results obtained by actually genotyping the two SNPs than estimates obtained by looking up the two markers in the HapMap CEU database.

Experiences with the FUSION GWAS, summarized above, show that imputation can be an effective way to estimate unobserved genotypes and/or allele counts. These genotypes can then be used in a variety of downstream analyses, including logistic regression analyses for discrete trait association and linear regression analyses for quantitative trait association, and to facilitate meta-analysis with studies genotyped on different platforms (Willer *et al.* 2008, Zeggini *et al.* 2008).

Experiment on HapMap Data. I set out this experiment to assess whether my method can generate imputed genotypes of similar quality when different commercial genotyping panels are used or in different populations. Answers to this question have important implications for imputation based meta-analyses that combine studies using different genotyping platforms. In this experiment, I used genotype data generated by the International HapMap Consortium (2007). I considered each of the HapMap samples in turn and masked available genotypes so as to mimic an experiment using one of the several commercially available DNA genotyping chips. For example, to evaluate Affymetrix 500K SNP chip, I masked genotypes for all markers that are not on the chip as missing for the individual being considered. I then used haplotypes for the remaining individuals on the same HapMap analysis panel (either YRI, CEU or CHB+JPT) to impute the missing genotypes. The results are summarized in Table 3.3 and clearly show that a large number of SNPs can be imputed very accurately using any of the commercially available panels (e.g., with $r^2 > 0.80$ to experimental genotypes) and that, compared to relying on single marker tagging, imputation results in improved coverage of the genome.

Depending on the commercial panel and population being investigated, coverage of the genome (proportion of SNPs with $r^2 > 0.80$) increased by 8-46% for low MAF alleles (MAF < 5%) and by 6-34% for more common alleles (MAF >= 5%). In agreement with this result, the average r^2 between each untyped SNP and its imputed counterpart was up to 40% higher on average when using imputed genotypes than when using the best available single marker proxy. The results shown in Table 3.3 are likely to represent an upper bound on the performance of my method in real settings, because additional errors will result from discrepancies in genotyping protocols between individual laboratories and the HapMap and from differences in LD patterns between the HapMap and the samples being studied. Nevertheless, they suggest my method is likely to be helpful for a variety of currently available commercial SNP panels and in different populations.

Experiment on the HGDP data. The preceding experiment on the HapMap samples has demonstrated the utility of my imputation method in all three HapMap populations. Since I advocate imputation incorporating haplotypes of the HapMap individuals as templates, I am interested in the performance of my method in a wide range of world populations, with potentially larger deviations in allele frequencies and LD patterns from the HapMap populations. In this experiment, I evaluated the performance of my imputation method in the 927 samples from 52 populations in the Human Genome Diversity Project (HGDP). To evaluate the performance of genotype imputation across these diverse populations, I selected a thinned marker set out of the 1,864 SNPs in the 32 autosomal regions. The thinned marker set had 872 SNPs spaced ~10 Kb apart across all 32 regions. I then used these SNPs to impute genotypes for the remaining 992 unselected SNPs and evaluated my approach.

Figure 3.8 shows the proportion of incorrectly imputed alleles in each of the populations. Results are presented using either a single HapMap analysis panel as a reference (either CEU, YRI, or CHB+JPT) or using all HapMap samples together as a larger reference panel. For each of the 52 populations, the reference panel that resulted in the smallest overall imputation error rate is highlighted. Overall, African samples were the most difficult to impute, with allelic error rates ranging between 5.13% for the Yoruba and 11.86% for the San when the HapMap YRI panel was used as a reference. In other parts of the world, I generally observed that the HapMap CEU provided a good reference panel for European populations and that the HapMap CHB+JPT provided a good reference panel for East Asian populations, resulting in error rates of <3.34% and <2.89% respectively. Outside Europe and East Asia, when imputation was applied to populations from the Middle East, Central and South Asia, the Americans and Oceania, it was generally better to use the combined HapMap samples as a reference than to use any single HapMap analysis panel as a reference. It is interesting to note that, in all cases, combining the three HapMap analysis panels into a single reference set was either the best option or the second best option. Furthermore, in situations where this combined reference panel reduced imputation accuracy, it resulted in an average increase of only 0.15% in error rates. The figure also illustrates that, when a larger number of individuals are genotyped for both the panel markers and additional markers to be imputed, it is possible to bypass the HapMap reference panel altogether. In the last panel of the figure, rather than using the HapMap data as reference to impute missing genotypes, I used a combined dataset including all other HGDP populations.

Figure 3.9 focuses on the estimated r^2 between imputed and observed allele counts. In each stripe, accuracy of imputation is assessed using a different reference panel. Superimposed in pink is the coverage that would be provided by single marker tagging approaches. Broadly, it is clear that imputation using an appropriate reference panel will improve coverage. Using an inappropriate reference panel (for example using the HapMap CEU to impute genotypes for one of the African populations), can result in imputed genotypes and allele counts that are not as strongly correlated with the true genotypes as the best available single marker tagSNP but, even then, the loss appears to be small. Importantly – in all cases – combining the three HapMap analysis panels resulted in substantial improvements in coverage over single marker tagging – suggesting that this might be a cautious approach when the choice of the reference panel is unclear.

Simulation Experiment for Power Assessment. Results comparing power of the four analysis approaches are summarized in Table 3.4. The first row in the table shows the empirically determined significance thresholds used for each analysis. More precisely, the thresholds are the 5th percentiles obtained from the 2,000 sets simulated under the null hypothesis. Since both the multi-marker and imputation approaches increase the total number of tests, note that the p-value threshold increases slightly when multi-marker tests are adopted and increases further when imputation is used. Subsequent rows summarize power for disease markers of different allele frequencies. In populations with strong LD, it is clear that for common susceptibility alleles the single marker tests provide high power and imputation or multi-marker analyses provide only small gains in power. However, for rarer alleles (such as those with frequencies < 5%) or in regions of more modest LD, imputation can provide dramatic increases in power. For instance, power increased from 24.4% to 56.2% when the disease allele frequency was 2.5% and imputation was used in the panel with CEU-like LD.

Multiple imputation (MI) procedure had little effect on either the empirical significance threshold or the power. The finding is not surprising because estimated allele dosages are quite stable across multiple imputations (such that average within-imputation variances accounts for 99% of the MI variance estimates), owing to the ability of my method to accurately calibrate model parameters and subsequently to identify shared chromosome stretches. My method shares features with both “hot deck” and “cold deck” imputation methods. For instance, the deck is “hot” because individuals to be imputed are also used for model parameter inference. On the other hand, the deck is “cold” because the fixed set of individuals collected separately by the International HapMap project is used as donors in most of the GWAS settings described in this chapter. My method differs from both in a number of ways. First, the imputed value is not "copied' from the "hot" deck of individuals but rather from the "cold" deck of external individuals. Second, the imputed value is not an observed value from some donor but rather a weighted sum of observed values from multiple potential donors, with weights proportional to the haplotype similarity to the recipient. In addition, my method allows both individuals under study and external individuals to serve as donors. Although it has unique features, my method may still benefit from more sophisticated hot and cold deck imputation methods, given its commonalities with both approaches. More research borrowing strength from the extensive multiple imputation literature is warranted. In the meantime, results from standard MI are reassuring. Perhaps more importantly, they suggest that multiple imputations are not absolutely necessary when (1) imputation uncertainty is taken into account by analyzing estimated allele dosages; and (2) the significance threshold is determined empirically, based on null sets in my simulations and by permuting phenotypes in real studies.

3.5 Discussion

In summary, the evaluation of imputed genotypes in the FUSION, HapMap, and HGDP samples clearly shows that imputation can be very accurate in a wide range of populations using a variety of currently available commercial SNP genotyping panels. Furthermore, investigation in the simulation experiment demonstrates the promising potential of my imputation method ultimately to improve power of detecting disease causing variants in individual association studies. Particularly important, imputation jointly analyzing data from individual studies with the external HapMap reference panel produces genotypes (experimental or *in silico*) on the same set of several million HapMap SNPs, across studies that examine individuals from different geographical or ethnic origins and that use different genotyping platforms. In this way, I believe it will continue to be an important tool for combining data across studies to achieve the large sample sizes for detecting variants whose individual contributions to disease risk are small.

A key ingredient for any imputation based approach is to ensure that alleles are consistently labeled across studies. In my evaluation of the FUSION and HGDP samples, using the HapMap as a reference, I was fortunate that a subset of the HapMap individuals were also genotyped in each study for quality control. Contrasting the genotypes for these quality control HapMap samples with those generated by the HapMap Consortium made the usually laborious process of ensuring consistent allele labeling across laboratories much easier, and I strongly recommend that all labs conducting genome-wide association studies genotype a small number of the HapMap individuals for this purpose.

So far, I illustrated the accuracy of genotype imputation that relies on existing resources (such as the PHASE II HapMap) and genotyping technologies (including a variety of currently available commercial genotyping chips). It is likely that both these resources and technologies will continue to evolve rapidly and it is interesting to consider how these developments might impact imputation based approaches. For example, it is clear that genotyping chips of the future will be able to examine an ever larger number of tagSNPs in a cost-effective manner. Extrapolating from Table 3.3, it is clear these should provide improved genomic coverage, eventually allowing investigators to impute nearly all HapMap SNPs with high accuracy. Nevertheless, it is also clear from Table 3.3 that when coupled with imputation based analyses current genotyping chips are already likely to provide excellent coverage of the genome in populations with LD patterns similar to CEU, JPT, and CHB. Thus, I expect the main advantages of new higher density chips will be in the examination of populations with less extensive LD, such as the YRI.

Another interesting possibility to consider is the impact of a larger HapMap reference panel on imputation, or similarly, the utility of using extra genotype data on a subset of individuals in a study to aid imputation in the remaining individuals in the study. To evaluate these possibilities, I generated a reference panel with varying numbers of Finnish individuals (between 60 and 500, see Table 3.5) and used these reference panels to impute genotypes for 521 SNPs in an independent set of 500 individuals from the FUSION studies of type 2 diabetes. Imputation accuracy and genomic coverage increase noticeably with the larger reference panels, with overall discrepancy rates between typed and untyped alleles as low as 0.40% when a reference panel of 500 unrelated individuals is available. One of the reasons for this increase in accuracy is that the length of haplotypes shared between individuals in the reference panel and those in the study sample increases gradually as the size of the reference panel increases. For example, mosaic fragments used to reconstitute the FUSION samples using the 500-sample reference panel were slightly > 1 Mb long on average. These long stretches are easier for my Markov model to identify and are also likely to descend from a more recent commonancestor. This means they will have undergone fewer rounds of gene conversion and mutation, which gradually erode haplotype similarities and reduce the quality of imputed genotypes. Overall, I expect that either genotyping a number of the study samples for markers of interest or increasing the size of the public reference panels will greatly improve the quality of genotype imputation.

Table 3.1 Quality of Imputed Genotypes in the AMD Candidate Gene Study.

Algorithm	#iterations	Genotype Matching Error (%)	Allele Matching Error (%)	Computation Time
Mach 1.0 Approximation (100)	20	1.80	0.97	< 1 min
	200	1.79	0.96	~5 min
	2,000	1.77	0.95	~1 hr
Mach 1.0 Approximation (200)	20	1.77	0.95	~4 min
	200	1.75	0.94	~20 min
	2,000	1.70	0.91	~3 hr
PHASE	--	2.60	1.37	24 hr
fastPHASE	--	1.81	0.97	~5 min

Figure 3.1 Disease-SNP Association Chi-Square Test: Imputed versus Experimental in the AMD Candidate Gene Study.

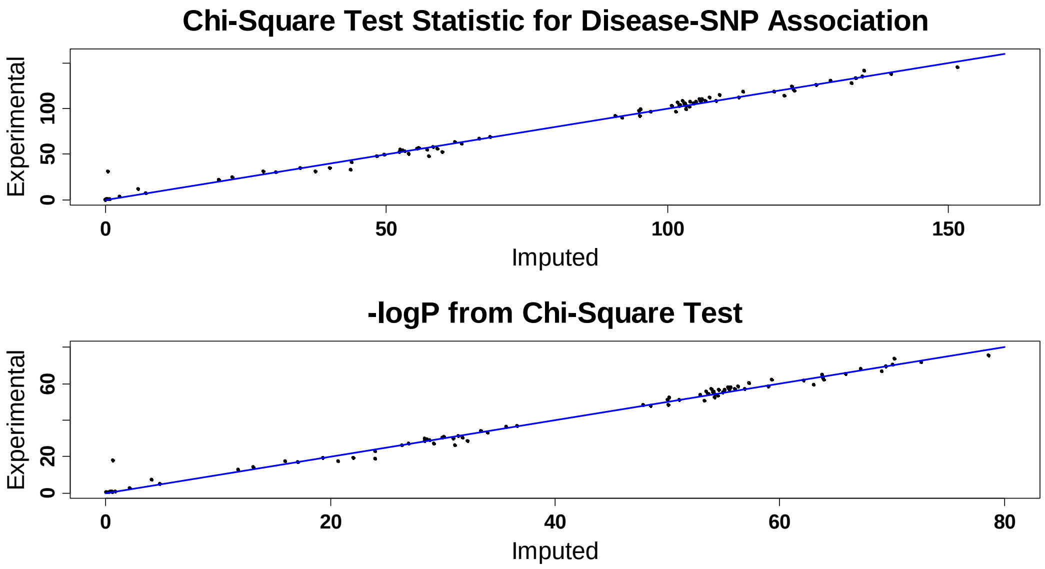


Figure 3.2 Improvement in r^2 for the Fine-mapping Panel of 521 SNPs in FUSION.

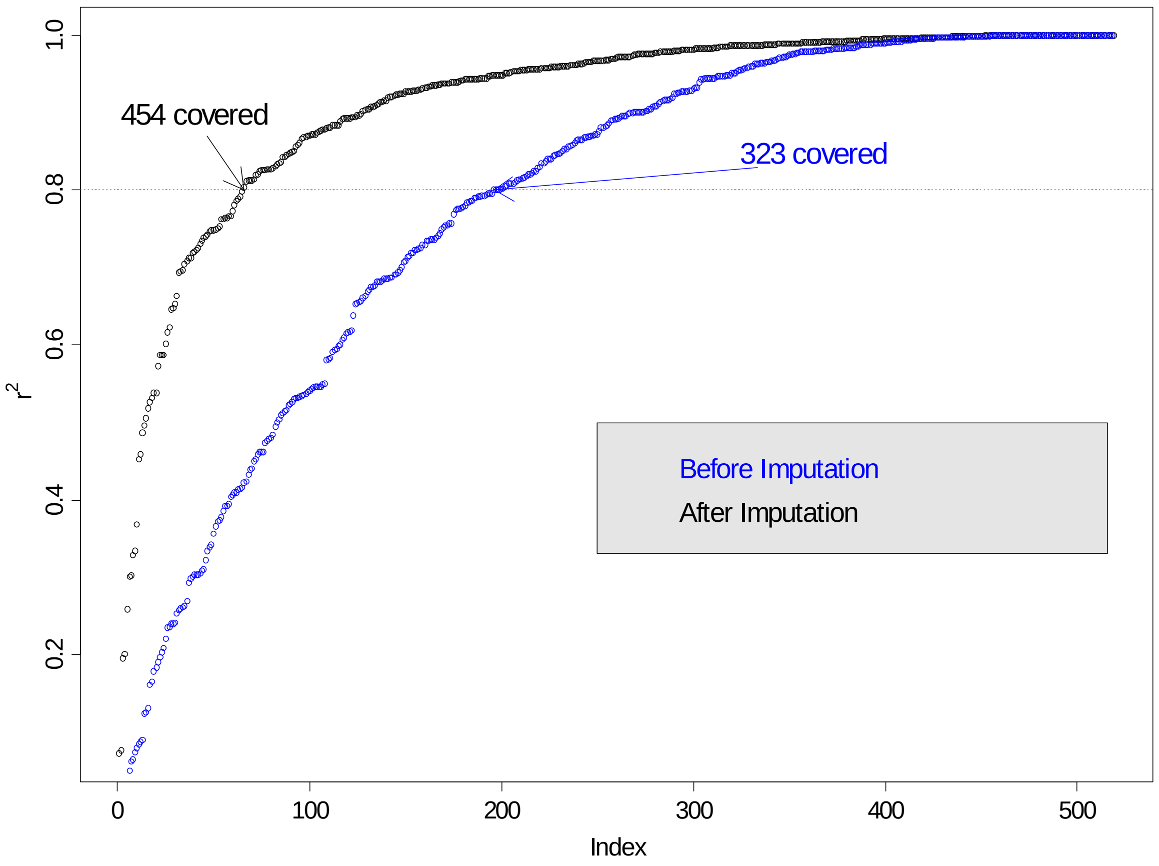


Figure 3.3 Assessment of Quality Measures for Individual Imputed Genotypes.

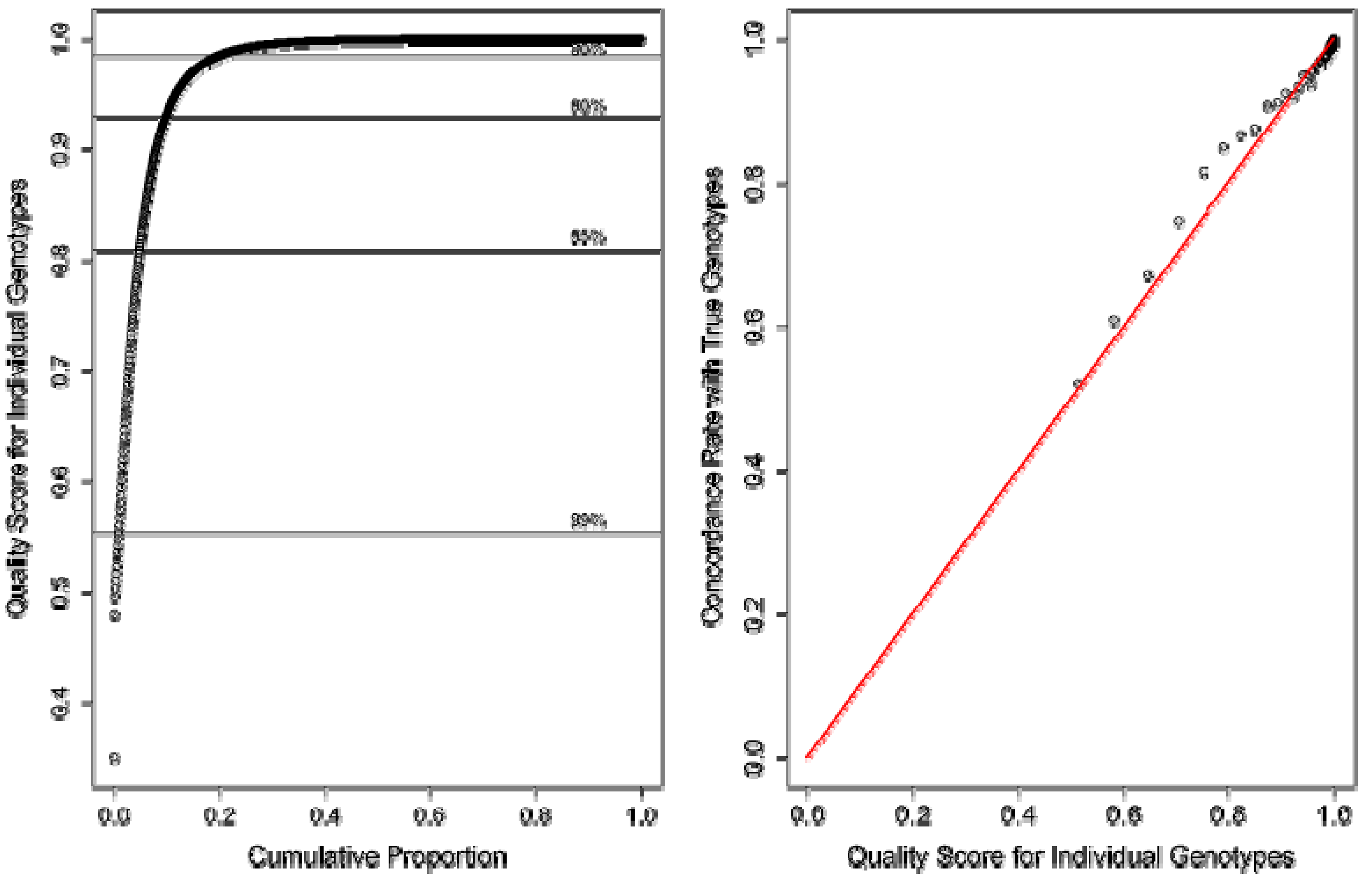
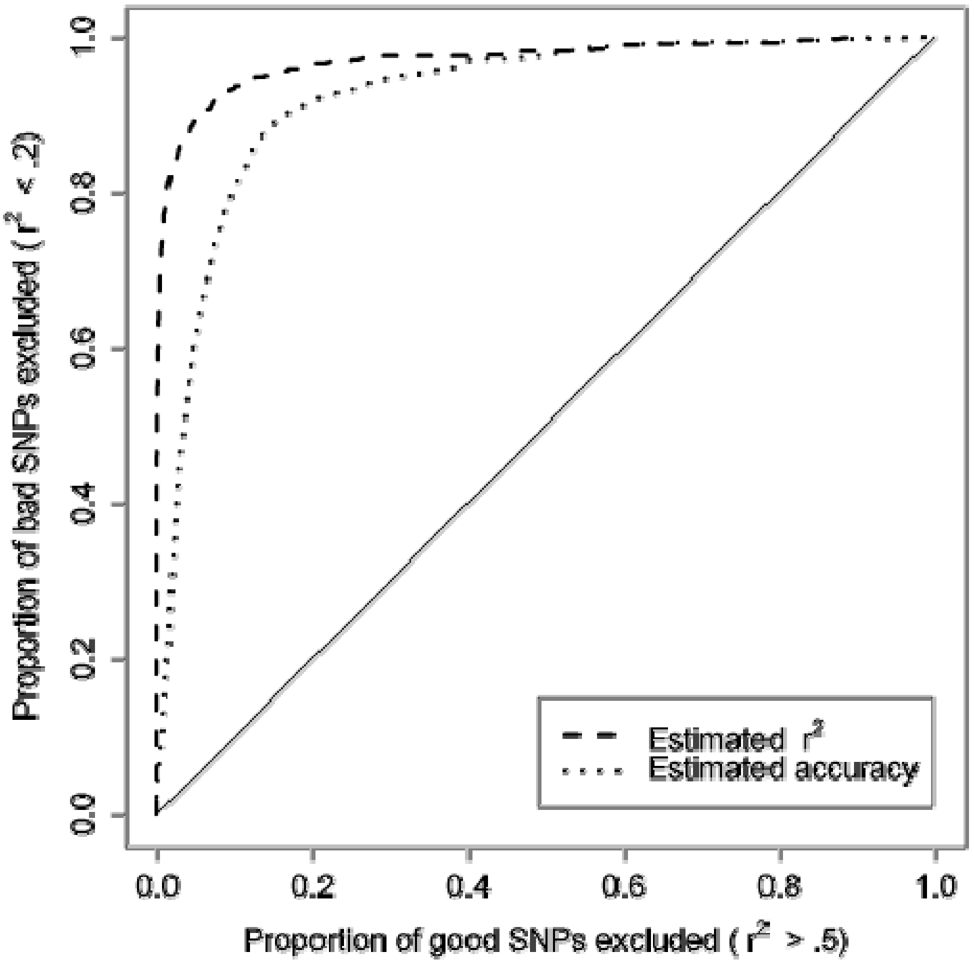


Figure 3.4 ROC Curve Comparing Two Measures of Data Quality.



For imputed SNPs on chromosome 14 among 1,190 FUSION individuals, for which both imputed and actual genotypes were available I evaluated the ability of two different measures of data quality (the estimated concordance between imputed and true genotypes and the estimated r^2 between imputed and true genotypes) to discriminate between poorly and well imputed SNPs. Both estimates of imputation quality are calculated without using the actual observed genotypes.

Table 3.2 Comparison of Imputed and Experimental Genotypes for a Subset of SNPs Showing Strong Association in FUSION.

	FUSION Allele Frequency		p-value		OR			Imputed vs. Actual genotypes, r^2		Observed allelic
SNP	Imputed	Genotyped	Imputed	Actual	Imputed	Actual	Max. R^2 w/ GWAS SNPs	Actual	Estimated	concordance
rs1738400	0.175	0.149	1.9 x 10^{-5}	0.011	1.84	1.15	0.11	0.241	0.309	0.874
rs1735641	0.580	0.715	3.0 x 10^{-5}	8.0 x	1.30	1.25	0.34	0.562	0.920	0.878
rs1161618	0.502	0.545	1.5 x 10^{-5}	4.8 x	1.40	1.27	0.27	0.755	0.585	0.919
rs2466291	0.579	0.618	6.3 x 10^{-4}	0.0016	1.26	1.22	0.47	0.829	0.830	0.935
rs2021966	0.609	0.603	9.1 x 10^{-5}	2.6 x	1.32	1.25	0.46	0.811	0.769	0.937
rs4812831	0.165	0.129	1.6 x 10^{-4}	0.0055	1.53	1.28	0.45	0.587	0.516	0.944
rs1164611	0.119	0.092	9.1 x 10^{-5}	0.002	1.66	1.38	0.13	0.687	0.512	0.956
rs8079544	0.091	0.106	8.9 x 10^{-4}	0.013	1.50	1.27	0.22	0.707	0.731	0.961
rs1409184	0.671	0.646	8.2 x 10^{-4}	0.0011	1.26	1.22	0.58	0.865	0.873	0.963
rs9402346	0.669	0.646	4.5 x 10^{-4}	0.0014	1.26	1.22	0.62	0.881	0.915	0.965
rs1800774	0.664	0.696	3.9 x 10^{-5}	7.3 x	1.39	1.35	0.29	0.861	0.617	0.972
rs1083776	0.138	0.152	1.5 x 10^{-5}	8.6 x	1.49	1.40	0.46	0.822	0.930	0.975
rs7750445	0.138	0.158	2.0 x 10^{-5}	4.1 x	1.47	1.41	0.50	0.836	0.965	0.977
rs1103662	0.080	0.071	1.7 x 10^{-5}	1.9 x	1.67	1.66	0.75	0.876	0.901	0.987
rs1270874	0.231	0.224	1.4 x 10^{-4}	3.9 x	1.33	1.30	0.24	0.933	0.954	0.988
rs1449725	0.579	0.573	5.3 x 10^{-6}	1.1 x	1.33	1.31	0.90	0.965	0.977	0.990
rs2267339	0.640	0.643	2.8 x 10^{-5}	4.5 x	1.33	1.34	0.72	0.951	0.873	0.990
rs1291082	0.035	0.033	2.5 x 10^{-6}	6.3 x	2.57	2.20	0.39	0.843	0.720	0.994
rs175200	0.476	0.479	6.6 x 10^{-5}	5.5 x	1.28	1.28	0.85	0.989	0.976	0.997
rs1329726	0.059	0.062	7.5 x 10^{-5}	9.0 x	1.72	1.65	0.28	0.973	0.916	0.998
rs4402960	0.683	0.681	1.7 x 10^{-4}	1.2 x	1.27	1.28	1.00	0.994	1.026	0.998
rs1001998	0.629	0.619	4.8 x 10^{-4}	4.2 x	1.25	1.25	0.66	0.99	0.953	0.998
rs6103716	0.371	0.371	7.3 x 10^{-5}	4.8 x	1.28	1.29	0.33	0.996	0.978	0.999
rs3802177	0.372	0.371	9.9 x 10^{-4}	0.0012	1.23	1.22	1.00	0.996	1.015	0.999
rs1708135	0.075	0.078	7.3 x 10^{-6}	5.5 x	1.70	1.68	0.87	0.989	0.954	1.000
rs1801282	0.165	0.165	9.5 x 10^{-4}	0.0011	1.31	1.30	1.00	0.999	1.002	1.000

The table shows a comparison of the results from analysis of imputed data with results from actual genotyping for a subset of the

SNPs that reached a p-value of $< 10^{-3}$ in my analysis of the FUSION data. Successive columns include SNP name, estimated allele frequency in FUSION cases and controls, using either imputed data or actual genotype data, p-value and odds ratio for association test comparing allele frequencies in cases and controls using imputed genotypes, p-value and odds ratio for association test comparing allele frequencies in cases and controls using experimentally derived genotypes, r^2 between the best single marker tag in the GWAS panel and this SNP, r^2 between imputed and observed genotypes (actual r^2 and estimated from my method as a measure of imputation quality) and, finally, proportion of alleles matched between imputed and actual genotypes.

Note that because these are all imputed SNPs that show strong association in the FUSION data, they are subject to a "winner's curse" effect. Thus, SNPs where imputation resulted in inflated p-values were more likely to be selected for follow-up in this analysis.

Not all imputed SNPs showing association at this significance level were genotyped experimentally. Rather, a subset of SNPs was selected for genotyping either because (a) they showed substantially stronger evidence for association than other nearby genotyped SNPs and stronger evidence for association than nearby imputed SNPs or (b) they were selected to improve coverage of the genome in and around 222 candidate genes (Gaulton *et al.* 2008). All SNPs with a p-value $< 10^{-3}$ in the imputed data and which were subsequently genotyped are tabulated.

Figure 3.5 FUSION T2D Association P-values: Imputed vs Genotyped.

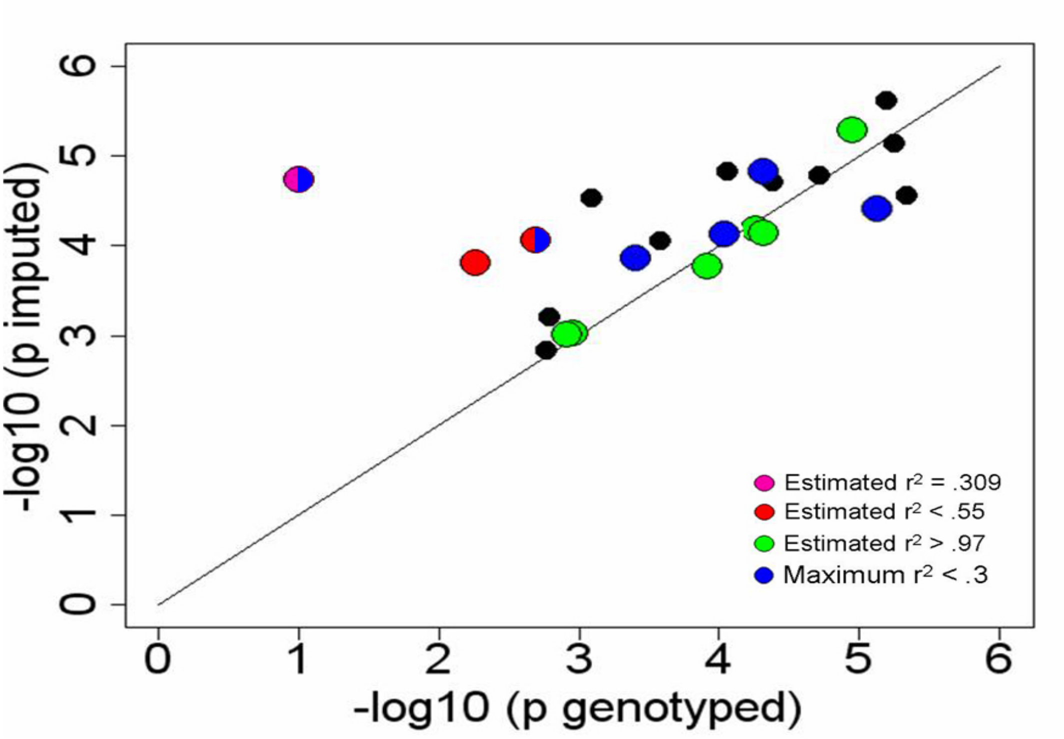
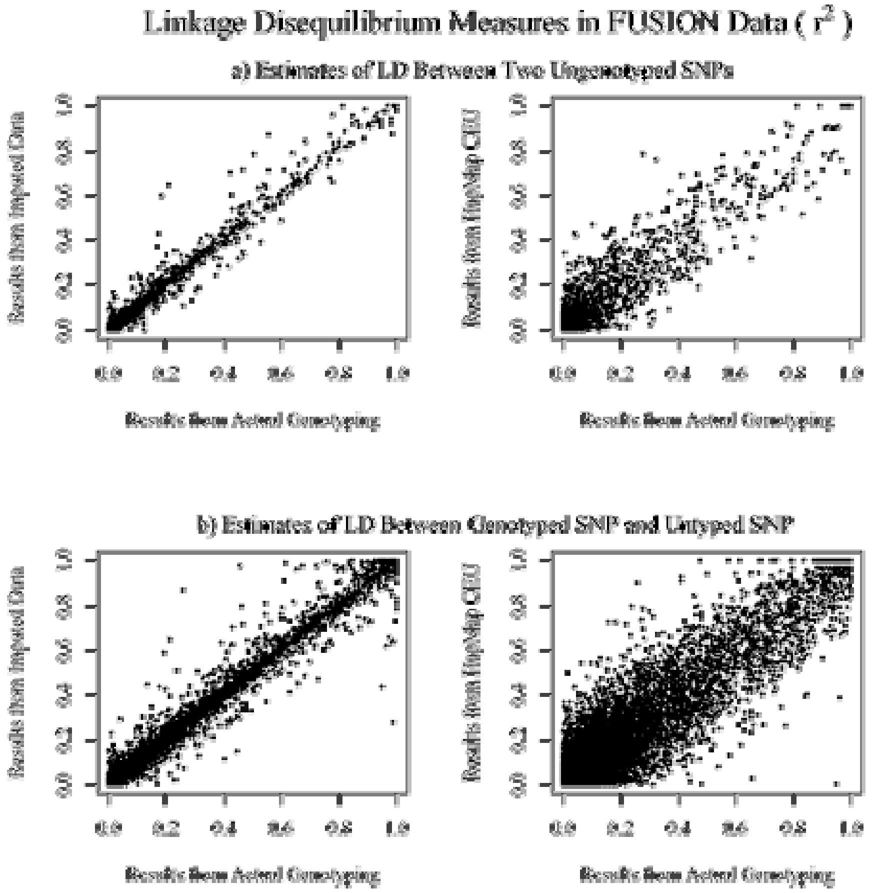
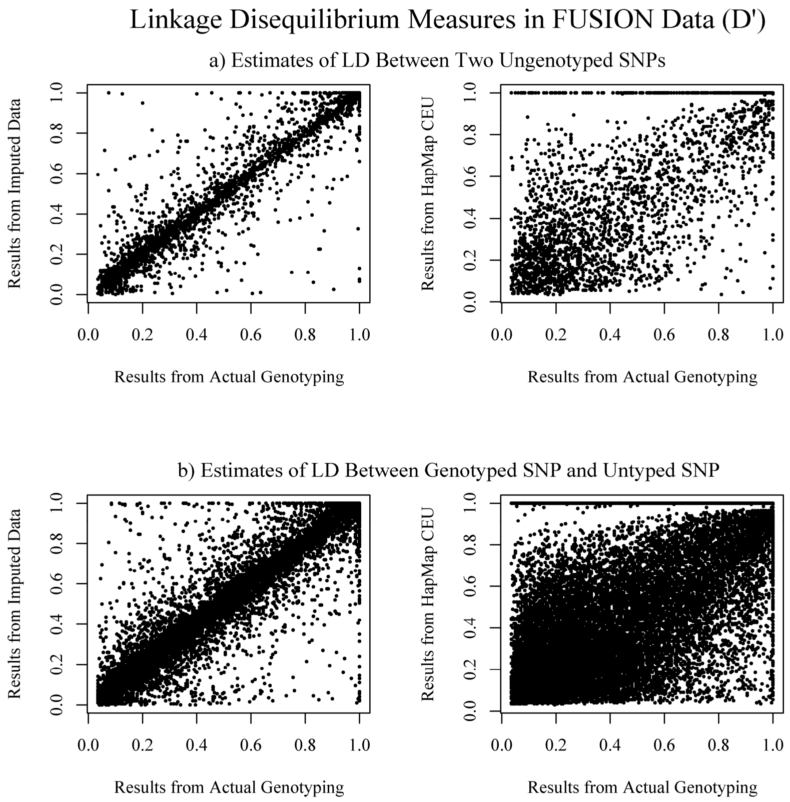


Figure 3.6 Imputation Improves Quality of LD Estimates: r^2.



For imputed SNPs on chromosome 14, the figure compares estimates of LD obtained by genotyping both SNPs (“Results from Actual Genotyping”, X axis) with estimates of LD obtained by imputing genotypes for both SNPs using markers on the 317K marker chip (“Results from Imputed Data”, Y axis, Top left), obtained by imputing genotypes for one of the SNPs (“Results from Imputed Data”, Y axis, Bottom Left) or obtained from the HapMap CEU panel (“Results from HapMap CEU”, Y axis, Top and Bottom Right).

Figure 3.7 Imputation Improves Quality of LD Estimates: D’.



Linkage Disequilibrium Measures in FUSION Data (D')

a) Estimates of LD Between Two Ungenotyped SNPs

b) Estimates of LD Between Genotyped SNP and Untyped SNP

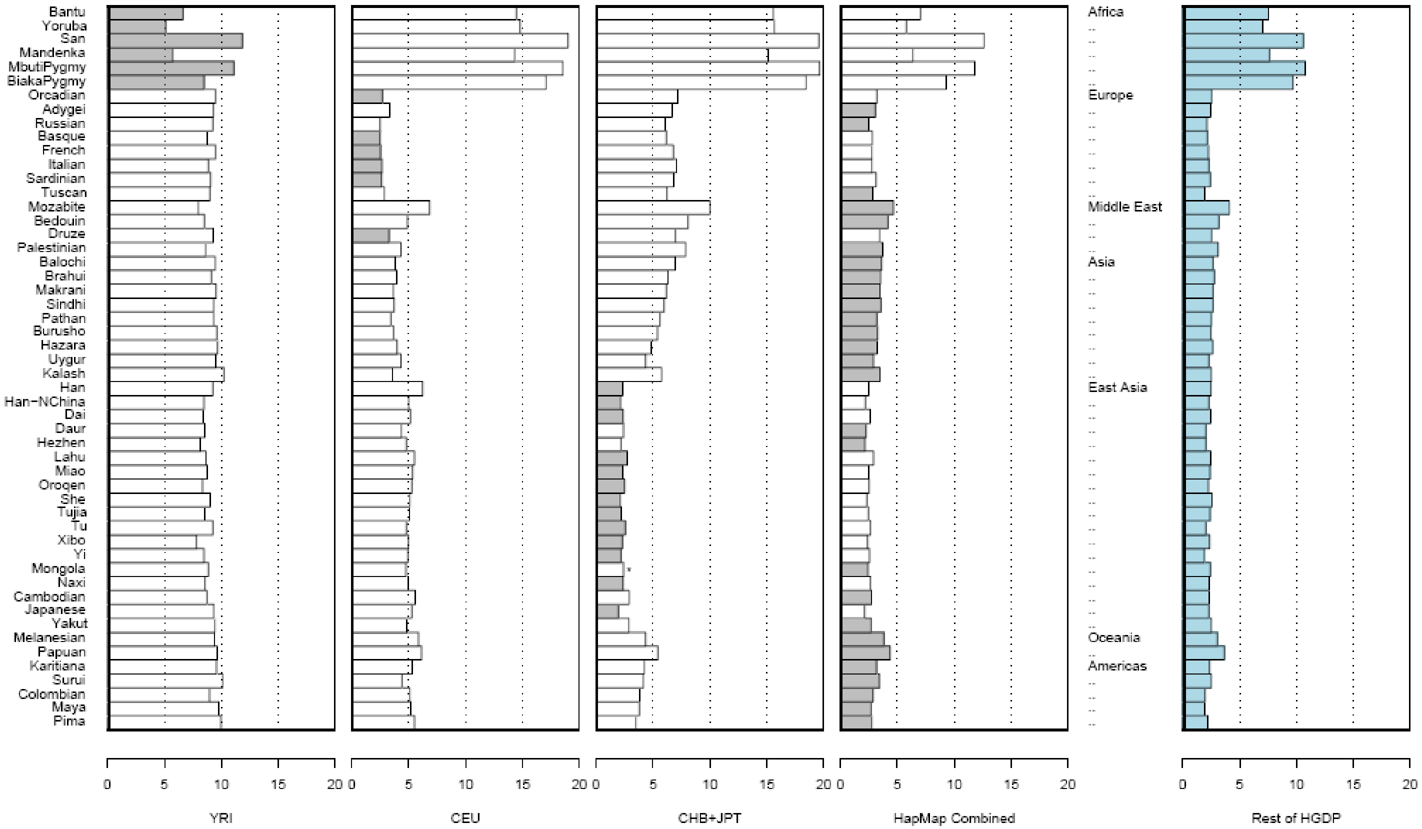
Table 3.3 Coverage of the Phase II HapMap with Commercial Genotyping Panels, before and after Imputation.

					Coverage by Single-Marker Tags				Coverage by Imputed SNPs					
	# Panel SNPs		# Imputed SNPs		MAF< 5%		MAF >= 5%		MAF< 5%			MAF >= 5%		
	Used	Lost	MAF<5%	MAF>=5%	average r^2	r^2 > 0.8	average r^2	r^2 > 0.8	average r^2	r^2 > 0.8	Error	average r^2	r^2 > 0.8	Error
CEU														
A100	100,844	1,609	259,261	2,086,690	0.36	0.22	0.50	0.31	0.47	0.32	1.80%	0.63	0.46	7.85%
A250S	195,864	4,393	251,807	2,002,214	0.48	0.31	0.65	0.44	0.61	0.46	1.33%	0.79	0.65	4.12%
A250N	216,747	4,836	250,364	1,983,146	0.50	0.34	0.67	0.48	0.63	0.49	1.26%	0.80	0.68	3.94%
A500	412,611	9,229	234,049	1,809,352	0.61	0.44	0.77	0.61	0.73	0.60	0.93%	0.89	0.82	2.12%
I300	305,050	3,115	267,573	1,871,586	0.30	0.08	0.84	0.74	0.70	0.54	1.08%	0.93	0.90	1.39%
I550	513,779	238	254,183	1,681,501	0.59	0.40	0.90	0.85	0.79	0.67	0.76%	0.95	0.94	0.90%
I650	578,864	14,627	244,431	1,630,298	0.66	0.48	0.91	0.86	0.80	0.68	0.72%	0.95	0.94	0.88%
A1000	676,182	87,766	209,636	1,580,321	0.71	0.57	0.86	0.76	0.79	0.68	0.73%	0.93	0.91	1.23%
I1000	779,800	130,014	225,439	1,456,134	0.70	0.54	0.93	0.89	0.81	0.71	0.67%	0.96	0.95	0.73%
YRI														
A100	100,627	3,223	326,772	2,320,439	0.21	0.08	0.33	0.14	0.35	0.18	2.22%	0.50	0.26	10.32%
A250S	210,242	4,698	318,680	2,220,904	0.30	0.13	0.47	0.22	0.50	0.29	1.73%	0.69	0.44	6.05%
A250N	231,026	4,971	317,321	2,201,821	0.32	0.15	0.49	0.26	0.53	0.33	1.64%	0.71	0.49	5.68%
A500	441,268	9,669	300,455	2,013,203	0.41	0.21	0.60	0.36	0.65	0.46	1.24%	0.83	0.69	3.30%
I300	271,991	15,346	315,631	2,163,803	0.33	0.15	0.52	0.26	0.60	0.39	1.42%	0.79	0.60	3.97%
I550	474,049	19,355	301,391	1,981,088	0.42	0.20	0.68	0.46	0.70	0.51	1.09%	0.88	0.80	2.13%
I650	573,953	28,487	300,785	1,881,962	0.46	0.23	0.75	0.56	0.73	0.55	0.98%	0.90	0.85	1.72%
A1000	737,369	91,811	275,794	1,749,271	0.54	0.31	0.73	0.54	0.74	0.58	0.92%	0.90	0.83	1.91%
I1000	788,503	149,152	274,766	1,702,039	0.53	0.30	0.78	0.60	0.76	0.59	0.88%	0.92	0.88	1.47%
JPT+CHB														
A100	95,521	1,994	299,643	1,919,001	0.35	0.22	0.47	0.28	0.44	0.32	1.68%	0.60	0.42	8.86%
A250S	186,411	4,368	290,265	1,840,510	0.49	0.33	0.63	0.42	0.58	0.45	1.23%	0.76	0.61	5.00%
A250N	205,274	4,713	288,661	1,823,236	0.51	0.36	0.65	0.46	0.59	0.48	1.17%	0.77	0.64	4.70%
A500	391,685	9,081	268,427	1,663,552	0.62	0.47	0.76	0.60	0.69	0.59	0.84%	0.87	0.80	2.60%
I300	274,751	12,851	287,456	1,755,289	0.54	0.38	0.75	0.58	0.69	0.57	0.88%	0.88	0.82	2.25%
I550	467,073	13,322	269,299	1,587,153	0.67	0.52	0.87	0.78	0.75	0.67	0.64%	0.93	0.91	1.26%
I650	531,807	23,155	259,962	1,534,915	0.71	0.57	0.88	0.80	0.76	0.69	0.61%	0.94	0.92	1.19%
A1000	638,817	86,838	239,528	1,455,644	0.72	0.59	0.85	0.75	0.75	0.67	0.65%	0.92	0.89	1.50%
I1000	728,837	136,560	239,252	1,365,519	0.73	0.61	0.90	0.85	0.78	0.70	0.57%	0.95	0.94	0.91%

For each platform, the table lists the number of SNPs in the platform that overlap with the phased HapMap chromosomes (release 21a). The number of SNPs that were not in the phased HapMap (Lost) is also listed, most of these were monomorphic. This number is followed by the number of SNPs that I attempted to impute, either with minor allele frequency <5% or >5%. I did not attempt to impute singletons for which the minor allele is observed only once. Coverage statistics using conventional single-marker tagging are provided and refer to the maximum r^2 between a HapMap SNP not on the panel and its best tag on the panel. Coverage statistics using imputation are also tabulated, and refer to the relationship between imputed allele counts for each SNP and true allele counts for the same SNP.

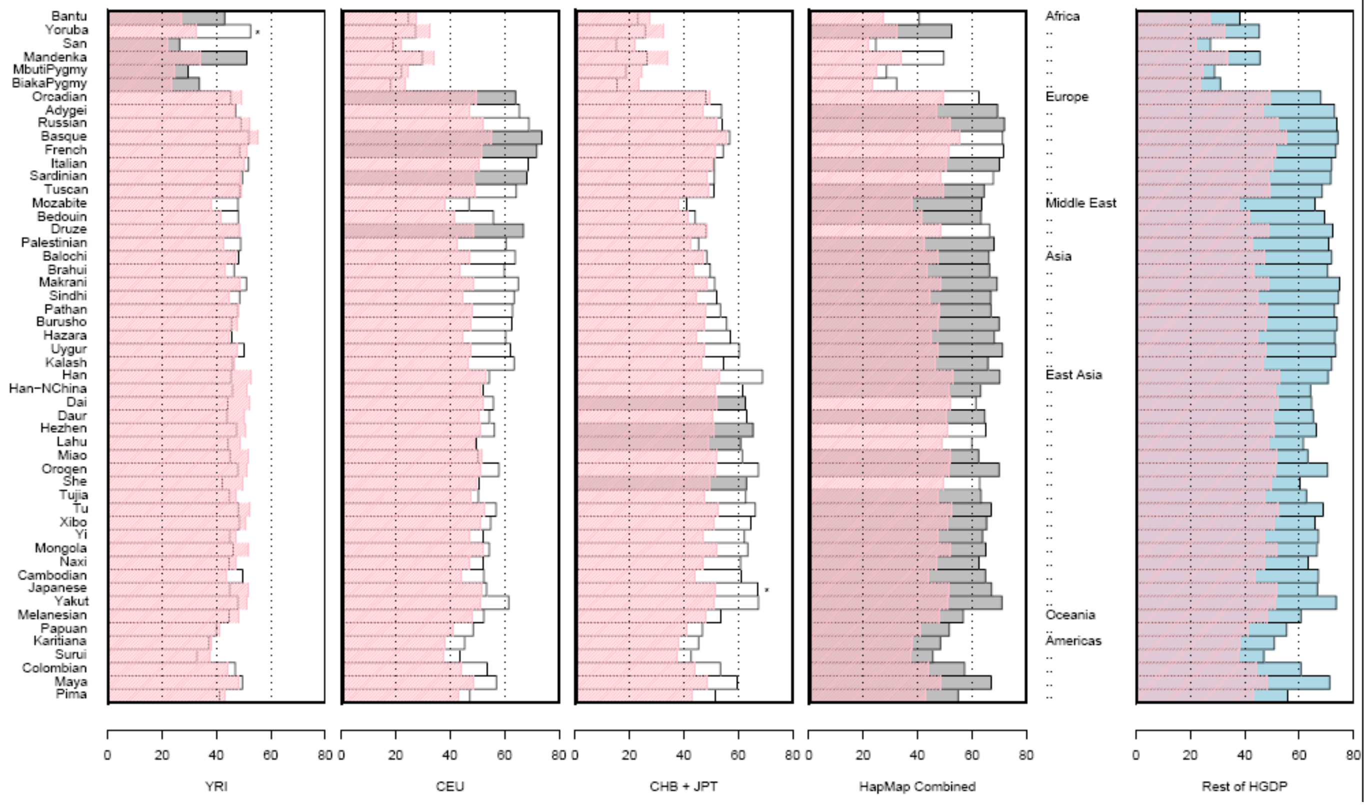
To evaluate the coverage of each genotyping platform using imputation, I focused on the markers that overlapped between the platform and the Phase II HapMap. I then considered each HapMap founder in turn and masked all genotypes for all markers not present in the commercial platform being evaluated. Finally, I used the remaining (unmasked) genotypes together with haplotypes for the other HapMap founders to impute the masked genotypes. The proportion of alleles that were imputed incorrectly, together with the correlation between imputed allele counts and actual allele counts, are tabulated for each platform.

Figure 3.8 Evaluation of Imputation Accuracy across HGDP Panels: Percentage of Alleles Imputed Incorrectly.



For each of 52 populations in the Human Genome Diversity Project (HGDP) a set of 872 SNPs distributed evenly across 32 regions, each ~330 kb in length, was used to impute 992 other SNPs. The 992 imputed SNPs were located near the middle of each imputed region. Imputation was done using either the HapMap YRI, CEU, CHB+JPT, or a combination of 3 HapMap panels as a reference (first 4 panels, best panel is shaded in gray) or using the remaining HGDP samples as a reference. In each case, the proportion of correctly imputed alleles is tabulated. The figure is based on a re-analysis of data from Conrad *et al*. (2006).

Figure 3.9 Evaluation of Imputation Accuracy across HGDP Panels: Coverage at r^2 Threshold 0.8



Genotypes for a set of 992 SNPs were imputed in the HGDP and were then compared with actual genotypes. For each SNP, an r^2 coefficient was calculated in each populations between true genotypes and imputed genotype scores. These r^2 values were then averaged across all SNPs for each population. The best set of HapMap reference individuals for each population is shaded. The coverage obtained by using the best available tagSNP (rather than imputed genotypes) is overlaid in pink. Coverage is defined as the percentage of SNPs having $r^2 > .8$ with the imputed counterpart or with the best available tagSNP respectively. See Figure 3.7 legend for further details.

Table 3.4 Imputed Genotypes Result in Increased Power.

	Power (LD mimics CEU)				Power (LD mimics YRI)			
	Single Marker Tags	Multi Marker Tags	Imputed Allele Counts	Multiple Imputation	Single Marker Tags	Multi Marker Tags	Imputed Allele Counts	Multiple Imputation
Empirical P-value Threshold	0.00081	0.00071	0.0003	0.00029	0.00067	0.00067	0.00017	0.00017
MAF = 2.5%	24.40%	25.00%	56.20%	56.00%	21.20%	22.60%	43.60%	43.60%
MAF = 5%	55.80%	56.40%	74.00%	74.00%	35.60%	36.00%	55.00%	54.80%
MAF = 10%	77.40%	78.40%	87.80%	87.80%	62.40%	63.80%	73.00%	73.00%
MAF = 20%	85.60%	86.20%	91.40%	91.60%	68.80%	70.60%	78.20%	78.20%
MAF = 50%	93.00%	93.60%	96.40%	96.40%	75.40%	77.40%	86.60%	86.60%

The table summarizes results from the analysis of two sets of 100 simulated 1 Mb regions. For each region, I generated a simulated HapMap including ~1,000 SNPs and used this panel to pick 100 tagSNPs that provided good coverage of the region (average coverage at an r^2 threshold of 0.8 of the ~800 common "HapMap" SNPs ~78% in CEU and ~61% in YRI). I then simulated and analyzed a series of case control studies, each with 500 cases and 500 controls. Association tests were carried out at each tagSNP ("Single Marker Tags"), initially. I then augmented these results with the analysis of multi-marker tags as suggested by PLINK[29] ("Multi Marker Tags"), with the analysis of imputed allele counts ("Imputed Allele Counts"), or with a multiple-imputation version of the imputed allele count analysis based on 10 independent imputations ("Multiple Imputation").

In each case, I first simulated and analyzed 2000 null (20 per region) datasets by assigning random chromosomes to each case and control. These analyses were used to establish the empirical p-value threshold that, when applied to the top signal in each region, resulted in a type I error rate of 5%. Then, for each tabulated minor allele frequency (MAF), I simulated 500 case-control datasets (5 per region, 500 cases and 500 controls each) where a variant with the specified MAF was associated with susceptibility. Power refers to the proportion of replicates where the top p-value exceeds the empirical p-value threshold.

Note that the susceptibility variant was picked at random among all simulated SNPs with the requisite MAF and was not necessarily included in the tagSNP set or in the markers ascertained for each region specific HapMap. To ensure comparable power across varying MAF, I increased genotype relative risk for rarer SNPs. Specifically, I set GRR = 2.500, 2.020, 1.715, 1.530 and 1.440 for SNPs with MAF = 2.5%, 5%, 10%, 20% and 50%, respectively. These settings correspond to ~85% for single marker tests of the susceptibility variant and a p-value threshold of 0.0005 is used (0.05 / 100, corresponding to a Bonferroni threshold that assumes 100 independent SNPs are tested).

Table 3.5 Effect of Increasing Reference Panel Size on Imputation Accuracy.

# Reference Panel Size	Genotype Matching Error	Allele Matching Error	Mean r^2	Median r^2
60	2.54%	1.31%	91.5%	97.5%
100	1.73%	0.88%	93.6%	98.2%
200	1.03%	0.52%	96.1%	98.8%
500	0.79%	0.40%	97.1%	99.1%

To evaluate the impact of a larger reference panel on the accuracy of genotype imputation, I used different numbers of individuals from the FUSION study genotyped for markers on the Illumina 317K SNP chip and also 521 SNPs on a candidate region of chromosome 14 (Willer *et al.* 2006) to impute genotypes for an independent set of 500 FUSION individuals on whom only the Illumina 317K SNP chip genotypes were available. The imputation procedure converged after ~300 iterations with panel size = 60, ~200 iterations with panel size = 100, and <100 iterations for panel sizes = 200 or 500 individuals.

Imputed genotypes were compared with experimental genotypes to determine accuracy at the genotype and allele level and to evaluate the r^2 between true and imputed genotypes.

Appendix 3.1 Calculations for Simulated Case Control Datasets.

Define the following notations:

K: disease prevalence;
A: risk allele at the disease SNP locus;
a: non-risk allele at the disease SNP locus;
p_d: frequency for allele A at the disease SNP site;
GRR: genetic relative risk
Genotype penetrances:

$$f_0 \equiv \Pr(disease \mid genotype \text{ aa}) \equiv \Pr(\text{D}|\text{aa})$$

$$f_1 \equiv \Pr(\text{D}|\text{Aa})$$

$$= f_0 * GRR, \text{ under multiplicative model}$$

$$f_2 \equiv \Pr(\text{D}|\text{AA})$$

$$= f_0 * GRR^2, \text{ under multiplicative model}$$

Applying Bayes' Rule, one can easily obtain the following probabilities:

$$\Pr(\text{AA}|\text{D}) = \frac{\Pr(\text{AA,D})}{\text{P(D)} \equiv \text{K}}$$

$$= \frac{p_d^2 * f_2}{p_d^2 * f_2 + 2p_d(1-p_d) * f_1 + (1-p_d)^2 * f_0}, \text{ under HWE}$$

Similarly,

$$\Pr(Aa \mid D) = \frac{2p_d(1-p_d) * f_1}{p_d^2 * f_2 + 2p_d(1-p_d) * f_1 + (1-p_d)^2 * f_0}$$

$$\Pr(aa \mid D) = \frac{(1-p_d)^2 * f_0}{p_d^2 * f_2 + 2p_d(1-p_d) * f_1 + (1-p_d)^2 * f_0}$$

$$\Pr(AA \mid \bar{D}) = \frac{p_d^2 * (1-f_2)}{p_d^2 * (1-f_2) + 2p_d(1-p_d) * (1-f_1) + (1-p_d)^2 * (1-f_0)}$$

$$\Pr(Aa \mid \bar{D}) = \frac{2p_d(1-p_d) * (1-f_1)}{p_d^2 * (1-f_2) + 2p_d(1-p_d) * (1-f_1) + (1-p_d)^2 * (1-f_0)}$$

$$\Pr(aa \mid \bar{D}) = \frac{(1-p_d)^2 * (1-f_0)}{p_d^2 * (1-f_2) + 2p_d(1-p_d) * (1-f_1) + (1-p_d)^2 * (1-f_0)}$$

Chapter 4
Analysis of Resequencing Data

4.1 Introduction

With the rapid development of very high throughput shotgun re-sequencing technologies (Bentley 2006), it is often proposed that genotyping based approaches will soon become outdated. One obvious advantage of re-sequencing is its ability to capture variants that are currently not recorded in public databases including, potentially, population specific variants that contribute to disease susceptibility.

I therefore further extended my hidden Markov model so that it can use whole genome re-sequencing data as input. In this setting, it uses information from individuals with similar haplotypes to reconstruct patterns of variation in regions where deep coverage is not available for a specific sample. Re-sequencing data I consider are of varying depths with realistic per base-pair error rates such that per-base pair genotype calls based on a smaller number of reads from each individual separately can hardly reach reasonable accuracy. For example, given the per base-pair error rate is more than 0.1% we cannot call any SNP with confidence using single-read coverage in the human genome where the average pairwise polymorphism rate is on the order of one per kilobase (kb). In addition, I consider situations where the read length can be very short (in simulated experiments below, reads were only 32 base pair in length). The use of such short-length reads is typical for current array-based sequencing technologies because it allows massive parallelization, reduces costs substantially, and enables super high throughput. However, data obtained from short reads are more challenging to assemble for reasons including typically higher experimental error rates and little information about phasing.

4.2 Methods

A Heuristic Introduction. Input re-sequencing data are summarized as counts of each allele at each base-pair position for each individual. For example, with depth four, we might see three traces with allele “A” and one with allele “G” at a particular base pair. Obviously, this could result from a true polymorphism at the locus or a sequencing error. More reads available for each individual (that is, a deeper re-sequencing) aids the discrimination of true polymorphisms from false positives due to sequencing errors. However, when individuals in a study are re-sequenced at low-depth, it is worthwhile to borrow information from other individuals who share similar haplotypes. My algorithm models the observed counts conditional on sequencing depth, starts with a solution of haplotypes compatible with the observed counts, and proceeds by updating one individual at a time until convergence.

Hidden Markov Model for Shotgun Sequence Data. When shotgun re-sequencing, or another single molecule re-sequencing technology, is used on diploid individuals, genotypes are not directly observed. In this case, I assume the data consists of counts A_j

and B$_j$ indicating how many times base A (or B) was observed at site *j*. Therefore, on top of the model introduced in Chapter 3, I define my hidden Markov model as:

$$P(\mathbf{A},\mathbf{B},\mathbf{S}\,|\,\theta,\varepsilon,\delta)=P(S_1)\prod_{j=2}^{L}P(S_j\,|\,S_{j-1})\prod_{j=1}^{L}\left\{\sum_{G_j}P(G_j\,|\,S_j)P(A_j,B_j\,|\,G_j)\right\}$$

Here, I sum over possible genotypes at each site and calculate the probability of the observed traits for each possible genotype set. In addition, I define the probability of observing a specific set of traces given the underlying genotype as:

$$P(A_j,B_j\,|\,G_j)=\begin{cases}Binomial(A_j,A_j+B_j,1-\delta) & G_j=A/A\\ Binomial(A_j,A_j+B_j,0.5) & G_j=A/B\\ Binomial(A_j,A_j+B_j,\delta) & G_j=B/B\end{cases}$$

The parameter δ denotes the per base pair sequencing error rate and can be separated from the effects of mutation and gene conversion captured in ε (the locus-specific error parameter introduced in Chapter 2), unless the re-sequencing depth is very low.

Consistent with notations from the previous chapters, *P(S$_1$)* denotes the prior probability of the initial mosaic state and is usually assumed to be equal for all possible configurations, *P(S$_j$|S$_{j-1}$)* denotes the transition probability between two mosaic states and reflects the likelihood of historical recombination events in the interval between *j-1* and *j*, *P(G$_j$|S$_j$)* denotes the probability of *true* genotypes at each position conditional on the underlying mosaic state.

In principle, the method could be applied to all sites where an alternative base call is observed at least once. However, since I simulated many short reads and at an error rate of 0.2%, the minor allele was observed at least once at nearly every position in many of my simulation experiments. For reasons of computational efficiency, I applied my method only to positions were the minor allele was observed in multiple traces. Specifically, I defined m_{kj} as the number of traces where the minor allele was observed at position j in individual k. Then, I defined the score $w_j = \sum_k m_{kj}(m_{kj}+1)/2$ and applied my haplotyping algorithm to all sites where w_j exceeded a predefined threshold (other sites were assumed to contain the major allele). The score gives higher weight to sites where the minor allele is observed multiple times in the same individual. I used thresholds for w_j of 5, 7, 9, 11, and 13 depending on whether the total coverage (defined as depth * individuals) was 200, 400, 800, 1200, or 1600x. When the number of individuals sequenced was 400, these thresholds were reduced to 4, 6, 8, 10, and 12 respectively. This means that, for example, when 400 individuals were re-sequenced at 4x depth (total depth = 1600x) I considered only sites where the minor allele was observed in at least 12 traces from different individuals or slightly fewer traces concentrated in one or more individuals.

4.3 Results and implications

To evaluate the possibilities, I simulated sequence data for ten 1 Mb regions. I simulated reads that were only 32 base pairs long and with a per base-pair error rate of 0.2%. Very roughly, these correspond to the expected performance of the next generation re-sequencing technologies from companies such as Illumina Solexa. I then re-sequenced between 100 and 400 individuals at different depths and used my approach to reconstruct haplotypes and genotypes for each individual. Note that the simulated reads are typically too short to include useful information on phase (because they will generally include only zero or one site that differs from the reference sequence). In addition, given the large number of bases examined, they will also suggest a large number of false-positive polymorphic sites so that it is important not only to confirm true polymorphic sites, by examining overlapping similar reads from the same individual or, potentially, from other individuals who share a similar haplotype.

For each site, I counted the number of times that the reference base or an alternative base was sequenced for each individual. For computational convenience, I only considered sites where both bases were observed several times (as described in the Methods section above) in downstream analyses and assigned the most frequently sampled base to all other sites. On this scale, the shotgun re-sequencing approach typically characterized ~4,209 polymorphic sites across the sampled individuals – ~4x the SNP density of the Phase II HapMap. Even relatively light shotgun re-sequencing provided very accurate haplotypes for each individual. For example, when 400 individuals were sequenced at 4x depth, there were only 18.97 errors per individual on average (over 1,000,000 base-pairs). Across ~980,000 sites that were monomorphic in the population only 82 false polymorphisms were called on average. Accuracy was also excellent at sites that were polymorphic in the population. For example, 3,558 of the 3,641 (97.72%) simulated polymorphic sites with MAF > 0.5% were identified and, at these sites, bases were estimated with an accuracy of 99.93% (see Table 4.1).

For any given depth, imputed genotype accuracy increased with the number of sequenced individuals (for example, accuracy at sites with MAF > 0.5% was ~98.8% when 100 individuals were sequenced at 2x coverage but increased to ~99.7% when 400 individuals were sequenced at the same depth; the number of errors per individual decreased similarly from 106.3 per individual to 40.3 per individual). In addition, the depth required to achieve a given accuracy decreased as the number of sequenced individuals increased: achieving 99.9% accuracy for sites with population MAF > 0.5% requires ~8x depth for 100 individuals, ~6x depth in 200 individuals and only 4x depth in 400 individuals. Again, advantages of re-sequencing larger numbers of individuals reflect the fact that as more individuals are sequenced the mosaic fragments identified by my haplotyper increase in length. This is also reflected in the accuracy of estimated haplotypes, which – when compared with simulated haplotypes – have ~1 switch per 50 kb when 100 individuals are examined, but ~1 switch per 500kb when 400 individuals are examined.

Proportion of variants detected. One major objective of re-sequencing analysis is to detect genetic variations that were not recorded in existing databases. Therefore, it is important to examine the proportions of variants discovered with different allocations of limited resources. Table 4.2 tabulates the proportions of SNPs identified using my re-sequencing data analyzer when different numbers of individuals are re-sequenced at different depths. For more meaningful and informative comparisons, I arrange SNPs into different minor allele frequency categories according to simulated population allele frequency. Several illuminating observations can be made from this table.

First, rarer SNPs are more difficult to identify, particularly with non-ignorable re-sequencing error. For instance, when 100 individuals are re-sequenced with an average of merely two reads per base pair, nearly all (~99%) the relatively common (MAF > 5%) SNPs can be identified while only ~1% of the very rare (MAF < .5%) SNPs can be identified. Although re-sequencing technologies have advanced greatly over the past decade, they are not yet perfect and techniques allowing vast parallelization and super high-throughput further incur errors. As noted previously, observed polymorphisms can derive either from a true underlying genetic variation, or from a re-sequencing error. Naïve approaches can hardly discriminate the two sources when re-sequencing error and the true allele frequency are comparable. My approach improves the discriminating power by taking into account multiple similar reads (of the rarer allele) from the same individual and by borrowing information from other individuals sharing similar haplotypes.

These particular features of my re-sequencing data analyzer lead to a second observation: the proportion of variants detected increases with the number of individuals re-sequenced, holding re-sequencing depth constant. This is particularly true for rarer SNPs for the obvious reason that common SNPs are easily detected with a rather small number of individuals at even low depth. For example, the detected proportion of the 15,336 very rare SNPs (MAF < 0.5%) more than doubles (increases from 3% to 7%) when 400 individuals are re-sequenced at 4x depth compared with when only 100 individuals are re-sequenced at the same depth.

Finally, larger proportions of rarer SNPs can be detected when a smaller number of individuals are re-sequenced at high depth than when a larger number of individuals are re-sequenced at low depth, holding the overall re-sequencing investment constant. For example, 59% of the 1,074 SNPs with MAF between 0.5% and 1% are detected with 100 individuals re-sequenced at 8x coverage. The detection proportion decreases to 45% when 400 individuals are re-sequenced at 2x coverage. While spreading out the same amount of investment to a larger number of individuals helps the identification of common SNPs, it does *not* help in the rare SNP category. On the other hand, with a larger number of re-sequenced individuals at the same overall investment, we obtain overall larger amount of information and have a larger reference pool which potentially better elucidates the LD structure. Therefore, such a more-individual-lower-coverage re-sequencing design can be more cost efficient for the establishment of a large public database, serving as reference for individual association studies. The benefits of a larger low-pass reference panel are quantified in section “Using resequencing data as imputation reference”.

Table 4.3 provides additional perspectives regarding the proportions of variants detected by exploring different re-sequencing error rates. Specifically, this table shows the number of false discoveries and detected SNPs in different MAF categories when 400 individuals are re-sequenced at different depths (1x, 2x and 4x) with different simulated re-sequencing error rates (error = 0.3%, 0.5% and 1.0%). As expected, common SNPs are not influenced much. For instance, the 2,947 SNPs with MAF > 5% are always 100% detected regardless of re-sequencing depth and error rate except for 1x depth at 1.0% re-sequencing error rate, where two SNPs are missed and 99.93% are identified. However, larger re-sequencing error rate makes it more challenging to effectively discriminate very rare SNPs from false positives. For example, 351 of the 15,336 (2.29%) very rare (MAF < 0.5%) SNPs can be detected at the cost of 46 false polymorphisms with a re-sequencing error of 0.3% while only 139 (0.91%) are detected at the cost of tripling the number of false positives (140 false positives) with a re-sequencing error of 1.0%. Therefore, re-sequencing accuracy within a certain range (within 1%) is essential for shotgun re-sequencing technologies to be useful for rare SNPs with MAF < 0.5% even with sophisticated analysis tools.

Imputation accuracy for detected variants. Identification of polymorphic sites is merely the starting point for marker characterization. I also want to provide accurate allelic states, with the correct corresponding haplotype backgrounds. It is important to distinguish being found polymorphic and being correctly imputed. In an extreme case of no re-sequencing error, assume we re-sequenced 100 individuals at 1x coverage and obtained two and only two different reads at a particular SNP locus. Such data under the assumption of no re-sequencing error allow us to declare polymorphic at the locus but without imputation, we have little information regarding the remaining 198 missing alleles. A constructive re-sequencing database needs not only to correctly catalog the polymorphic site, but also to accurately document alleles/haplotypes at cataloged polymorphisms. Table 4.4 and 4.5 demonstrate the capability of my re-sequencing data analyzer to provide high quality imputed genotypes at discovered loci under different settings.

First and obviously, imputation quality improves with re-sequencing depth, holding the number of re-sequenced individuals constant. For example, imputation accuracy increases from ~97% (95.01% - 97.91%) when 100 individuals are re-sequenced at 1x depth to over 99.99% when the same individuals are re-sequenced at 16x depth.

Secondly, when more individuals are re-sequenced at the same depth, imputation accuracy improves because the increased chances of finding more closely related individuals lead to more accurately reconstructed haplotypes and improved genotype imputation.

Thirdly, given total investment fixed, a larger number of individuals re-sequenced at lower coverage provides a larger pool of chromosomes at rather small reduction in accuracy. For example, at a total investment of 800x, we can have: (1) 100 individuals with an average ~99.9% (99.85% - 99.94%) accuracy; or (2) 200 individuals with an average ~99.8% (99.68% - 99.90%) accuracy; or (3) 400 individuals with an average ~99.6% (99.37% - 99.76%) accuracy. I have shown in Chapter 3 that a larger reference panel increases imputation quality. Simply put, the reason is: one is more likely to find a person more closely related (genetically) to him/her in a larger sample of people.

Therefore, the low-pass sequencing design of a large number of individuals, with such ignorable reduction in accuracy, is a promising alternative for the establishment of a large reference database.

Although elevated re-sequencing error makes it more challenging for SNP discoveries (especially for rarer SNPs), Table 4.5 shows an agreeable tendency that imputation accuracy is only minimally affected by re-sequencing error. For example, average imputation accuracy for 400 individuals re-sequenced at 2x depth is ~99.55% (99.37% - 99.76%) when re-sequencing error rate is 0.3%; ~99.54% (99.35% - 99.75%) with an error rate of 0.5%; and ~99.50% (99.33% - 99.72%) with an error rate of 1.0%.

Designing a resequence-based study. The above comparisons of different allocations of sequencing efforts have clearly suggested the benefits of a low-pass sequencing design. In this section, I compare the performances of two designs scaling up to realistic studies of complex traits: one deep sequencing of 400 individuals at 30x coverage and one low-pass sequencing of 3000 individuals at 4x coverage. I first simulated 10 replicates of 45,000 chromosomes extending 100Kb with realistic LD patterns mimicking those of HapMap CEU (Schaffner *et al*. 2005). I proceeded to sequence 400 or 3000 individuals at an average depth of 30x or 4x with short reads of length 32 bases and a per-base error rate of 0.5%. I then took sites potentially polymorphic (through trials and errors to keep the number of false polymorphisms below 100) to the imputation engine, borrowing information across individuals.

Figure 4.1 compares the capability of two designs for polymorphism detection. Both designs provide excellent power (~100%) to detect variants with MAF > 0.5%. For rare variants, the proportion of SNPs detected is influenced by two factors: the proportion of variants present in the sample and the power of detection conditional on the sequencing depth. The two factors exert their impacts in opposite directions given fixed sequencing investment. For example, ~12% SNPs with population frequency <0.1% show at least one copy of the minor allele among a sample of 400 individuals while ~49% of such rare variants are polymorphic among a sample of 3,000 individuals. However, due to the rather shallow coverage of 4x for the 3,000 individuals, only a small fraction (~8%) of the ~49% can be detected, resulting in ~4% such rare SNPs being detectable with the particular low-pass design. The deep coverage design continues to have better performance in the MAF category of 0.1-0.2%, detecting 65% of the variants, whereas the low-pass design is able to detect 58%. We see the benefits of the low-pass design in the MAF category of 0.2-0.5%, with the low-pass design detecting 94% of the variants and the deep-depth design detecting 87%.

The quality of genotype calls (Table 4.6) from the low-pass design at the detected variants, while not as good that from deep sequencing, is still impressive. For example, for variants with frequency >1%, low-pass sequencing call accuracy is always >99.9% and for heterozygous sites it is always >99.84%. High rates of polymorphism detection and accurate calling of genotypes are possible because my model effectively combines information across individuals with similar haplotypes, so that the coverage of each haplotype is, effectively, quite deep.

If rare variant detection and accurate calls are the main goals, the deep coverage design is indeed valuable. In real studies, we are most likely equally (if not more) interested in finding genetic variants associated with trait(s) of interest. Now consider the following setting: a disease with prevalence 10%, a rather high genetic relative risk of 2, and a relatively large sample size of 3,000 cases and 3,000 controls. The power to detect association is 0% when the disease allele frequency is below 0.2%. The power increases to 2% and 32% respectively when the disease allele frequency is 0.5% and 1% (Skol *et al.* 2006). Therefore, losing SNPs with MAF below 0.2% has essentially no effect on the identification of genetic variants influencing complex traits. Of course, a deep coverage design may be appreciated more when more sophisticated methods are applied for the analysis of rare variants (Li and Leal 2008, Madsen and Browning 2009).

Even sophisticated methods are still constrained by available information. Conceptually, the information for association testing provided by low-pass sequencing of large numbers of samples is much greater than that for deep sequencing of fewer samples because of the much larger sample size. One standard measure closely related to the statistical power is r^2, the squared correlation between a genetic variant and its proxy (Pritchard and Przeworski 2001). In the context of imputation, r is the correlation between the true variant genotype and its imputed counterpart. Assuming information for association testing scales as nr^2 where n is the sample size, low-pass sequencing provides much greater information than does deep sequencing given the same overall sequencing effort (Figure 4.2). For example, for variants of frequency 0.1-0.2%, 0.2-0.5%, 0.5-1.0%, 1.0-2.0% or 2.0-5.0%, 4x sequencing of 3,000 individuals provides 4.8, 5.2, 6.0, 6.9 and 7.2 times as much information as 30x sequencing of 400 individuals. Design-wise, therefore, low-pass sequencing is a logical strategy to maximize information for gene discovery.

Using resequencing data as imputation reference. Having shown the capabilities of my method to detect SNPs and accurately impute SNP genotypes from re-sequencing data over a broad set of simulation conditions, I expect that my method will allow economical association studies that evaluate SNP variation in large numbers of individuals even more exhaustively than is currently possible, by using shotgun re-sequencing of whole genomes as reference. To evaluate the possibility, I simulated a smaller and a larger "re-sequencing HapMap". The smaller "re-sequencing HapMap" reference panel consists of 120 perfect haplotypes (that is, true haplotypes from simulation) from 60 individuals re-sequenced at > 16x coverage with a potential trio design (thus a total re-sequencing investment of >960x). The larger "re-sequencing HapMap" reference panel has 800 haplotypes from analyzing 400 individuals re-sequenced at 2x depth (thus a total re-sequencing investment of 800x). I then simulated an independent random sample of 500 individuals from the same underlying population. For the study sample of 500 individuals, I genotyped only 100 or 200 tagSNPs selected randomly from SNPs discovered both in the smaller and in the larger "re-sequencing HapMap". Imputation of "re-sequencing HapMap" SNPs in the study sample proceeded in the same fashion as described in Chapter 3, using haplotypes from either the smaller or larger reference panel. Obviously, we are analyzing a much larger number of variants with a "re-sequencing

HapMap" as reference. The SNP density in the current HapMap phase II is ~1,000 SNPs in 1Mb and my "re-sequencing HapMap" has ~4,000 SNPs in 1Mb for simulated CEU and ~6,000 for simulated YRI. After imputing all the "re-sequencing HapMap" SNPs in the study sample of 500 individuals, I recorded average imputation errors per person across the entire 1Mb region, number of false positives, SNPs discovered and imputation accuracy in each minor allele frequency class (categorized according to simulated sample allele frequency [i.e., calculated from true genotypes of the 500 sample individuals]). Results are summarized in Table 4.7.

Table 4.7 clearly shows the benefits of a larger "re-sequencing HapMap". The average number of per-individual errors across the entire 1Mb region decreases by more than 15%: from 416 (302) when the smaller "re-sequencing HapMap" is used as reference and 100 (200) tagSNPs are genotyped in the study sample to 349 (255) when the larger "re-sequencing HapMap" is used.

In addition, in terms of false discoveries and true polymorphisms detected, I observe obvious gains in all but the very rare SNP (MAF < 0.5%) category with the usage of the larger reference panel. For example, I have 200 false polymorphisms when using the smaller reference panel and genotyping 200 tagSNPs. The number reduces to 155 when the larger reference panel is used. The total number of discovered SNPs with MAF at least 0.5% increases from 4,586 (4,802) to 4,688 (4,953) with the use of the larger "re-sequencing HapMap" when 100 (200) tagSNPs are genotyped. Particularly noteworthy, the gains can be substantial in some "marginal" MAF categories (i.e., categories where MAF is high enough eventually to reach reasonable statistical power with a huge sample size under meta-analyses of already large-scale studies; and at the same time low enough to be easily missed by genotype based association designs). For example, in the MAF 1-2% category, ~15% more SNPs can be detected with the larger "re-sequencing HapMap": 581 (653) up to 665 (745). Loss in SNP detection in the very rare category (MAF < 0.5%) is exaggerated in two senses. First, I have true/perfect haplotypes in the smaller reference panel. The reality, however, is never perfect and as discussed extensively in previous sessions rarer SNPs are more likely to be influenced. Secondly, the comparisons are not totally fair given the number of false discoveries differ substantially. For a fairer comparison of the rare SNPs, one should perform a separate analysis where the numbers of false positives are comparable.

Furthermore, comparable or better quality imputed genotypes can be obtained using the larger "re-sequencing HapMap" panel. The improvement in quality manifests itself more clearly in the more common SNP categories where imputation tends to be harder because of more uncertainties. For example, average imputation accuracy in the common SNP category (MAF > 5%) increases from 95.80% (97.88%) to 97.07% (98.68%) using the larger "re-sequencing HapMap" when 100 (200) tagSNPs are genotyped in the study sample. For the rarer SNP categories (MAF < 1% groups), imputation qualities appear slightly lower with the larger reference panel, but are still well over 99% and probably will not make any noticeable differences for downstream analyses. Table 4.8 show similar patterns with the focus on estimated r^2 between imputed and observed allele counts.

4.4 Discussion

Genome sequencing technologies are improving extremely rapidly. Whereas the first two human whole-genome assemblies took years to complete (Lander *et al.* 2001, Venter *et al.* 2001), several additional genomes have been assembled in the past 18 months (Bentley *et al.* 2008, Levy *et al.* 2008; Wheeler *et al*. 2008). These advances in whole genome sequencing have resulted from the development of massive throughput sequencing technologies, which differ from standard Sanger-based sequencing (Sanger *et al.* 1977) in many ways. For example, the data produced by these new technologies typically have somewhat higher error rates (on the order of 1% per base). Since these technologies produce a very large amount of data, one typically accommodates these error rates by sequencing every site of interest many times to achieve a high-quality consensus.

I expect that the continued development of these technologies will significantly change how genotype imputation is used. An example is given by the 1000 Genomes Project (see www.1000genomes.org), which aims to deliver whole genome sequences for >2,000 individuals from several different populations by the end of year 2009. To do this in a cost-effective manner, the project is using a strategy that combines massively parallel shotgun sequencing with the same statistical machinery that underlies genotype imputation. Specifically, a relatively modest amount of shotgun sequence data is being collected for each individual: Each of the target bases will be sequenced only 2-4x on average (statistical fluctuations around this average mean that many bases will not be covered even once), rather than 20-40x used in previous applications of these technologies to whole-genome resequencing. To call polymorphisms accurately in each genome, the Project will then use imputation-based approaches as described in this chapter to combine information across individuals who share a particular haplotype stretch. As shown in the simulations described in this chapter, I have predicted that when 400 diploid individuals are sequenced at only 2x depth (1x per haploid genome) and the data are analyzed though my imputation engine that combines data across individuals sharing similar haplotype stretches, polymorphic sites with a frequency of >2% can be imputed with >99.5% accuracy with a sequencing error of 1% per base (Table 4.5).

The ability to combine relatively modest amounts of sequence data across many individuals to generate high-quality sequence data for all may become one of the most common uses of imputation technologies in the near future. For a given sequencing effort, genotype imputation-based analyses may allow an increase in the number of individuals to be sequenced by five- to tenfold with minimal loss of accuracy in individual genotypes. Such an increase in sample size is critical when attempting to map the genes for complex diseases. Of course, even before massively parallel sequencing technologies are deployed more widely, one immediate change will occur with the completion of the 1000 Genomes Project. Specifically, I expect these data will provide accurate genotype information on >10 million common variants and will quickly replace the HapMap Consortium genotypes as the reference panel of choice for imputation studies. Thus imputation-based analyses will be able to examine even more genetic markers, and each of these markers will, on average, be imputed much more accurately.

To evaluate the feasibility in practice, I analyzed preliminary shotgun sequence data generated by the 1000 Genomes Project on 52 CEU individuals sequenced at ~4.7x average depth per individual. I used estimated haplotypes for these individuals to fill in missing genotypes in a case-control study of type 2 diabetes. Even with this small reference panel, ~6 million SNPs are estimated to be imputed of good quality and missing alleles in our type 2 diabetes case-control study could be imputed with an average error rate of ~2.6%. As the size and quality of the reference panel increase, I expect this average error rate to improve rapidly (as illustrated previously in Table 3.5). Recall from Chapter 3 that using HapMap as reference, ~2million imputed SNPs were estimated to be of good quality and alleles were imputed with an average error of ~1.4%. Therefore, while the number of individuals sequenced by the 1000 Genomes Project is similar to that in HapMap, there is value in using both as references because the former provides information at more markers and the latter provides information of higher quality.

Table 4.1 Accuracy of Imputed Genotypes Using Shotgun Re-sequence Data as Input.

		1,000,000 bases	979,642 Monomorphic Sites		20,358 Polymorphic Sites, Segregated According to Population Frequency									
					16716 sites with MAF<.5%		510 sites with MAF .5-1%		425 sites with MAF 1-2%		590 sites with MAF 2-5%		2116 sites with MAF >5%	
Sequencing Depth	Total Investment	Average Errors per Individual	False Positives SNPs	Allelic Accuracy	Detected SNPs	Allelic Accuracy	Detected SNPs	Allelic Accuracy	Detected SNPs	Allelic Accuracy	Detected SNPs	Allelic Accuracy	Detected SNPs	Allelic Accuracy
n = 100 individuals resequenced using a shotgun approach														
2x	200x	106.32	99	97.11%	176	98.79%	90	98.78%	176	98.81%	465	98.78%	2109	98.84%
4x	400x	45.01	43	98.42%	440	99.51%	188	99.62%	286	99.60%	550	99.62%	2115	99.66%
8x	800x	12.90	59	99.30%	995	99.92%	309	99.94%	369	99.94%	582	99.94%	2115	99.94%
12x	1200x	4.60	42	99.41%	1,310	99.98%	357	99.98%	395	99.99%	585	99.99%	2116	99.98%
16x	1600x	2.19	33	99.49%	1,432	99.99%	368	100.00%	397	100.00%	585	100.00%	2115	99.99%
n = 200 individuals resequenced using a shotgun approach														
2x	200x	57.52	219	98.87%	365	99.56%	186	99.52%	295	99.46%	565	99.41%	2116	99.59%
4x	400x	25.18	52	99.47%	734	99.84%	310	99.85%	378	99.85%	587	99.86%	2116	99.90%
6x	1200x	14.36	45	99.67%	1,270	99.95%	386	99.95%	405	99.94%	590	99.94%	2116	99.96%
8x	1600x	9.29	34	99.68%	1,654	99.98%	425	99.97%	415	99.97%	590	99.97%	2116	99.97%
n = 400 individuals resequenced using a shotgun approach														
1x	400x	84.95	212	99.02%	183	99.50%	149	99.32%	296	99.15%	570	98.94%	2116	99.14%
2x	800x	40.34	243	99.44%	532	99.72%	307	99.68%	393	99.64%	589	99.65%	2116	99.77%
3x	1200x	25.98	143	99.65%	906	99.86%	389	99.84%	413	99.83%	590	99.84%	2116	99.89%
4x	1600x	18.97	82	99.77%	1,258	99.92%	431	99.90%	421	99.91%	590	99.91%	2116	99.94%

I simulated 1Mb regions in individuals with HapMap CEU-like degrees of LD. Then, I generated shotgun sequence data for a subset of individuals (n = 100, 200 or 400) at varying depths (1x – 16x). The depths were selected to represent a total investment of between 400x and 1600x coverage of the region (200x coverage was also examined for n=100). Simulated reads were 32-bp long and had a per base error rate of 0.2%. Read counts at sites where multiple copies of each alternative base were observed were then provided as input to our software package.

The "Average Errors per Individual" column summarizes the overall haplotyping accuracy, across polymorphic and monomorphic sites. For example, when 400 individuals were sequenced at 4x depth, an average of 18.97 imputed genotypes differed from the actual simulated genotypes in each individual. The next several columns summarize results for positions where the haplotyper called a polymorphism. The number of false positive sites is listed together with the accuracy of bases at those sites. Typically, only a few false positive polymorphisms were called (in 400 individuals at 4x depth, 82 false positive polymorphisms were observed). The next columns summarize results for sites that were truly polymorphic in the population and these are grouped by frequency (calculated from a sample of N = 10,000 chromosomes). For each frequency class, information is provided on the number of polymorphic sites identified and the overall base calling accuracy at those sites. Note that especially for rare SNPs, many sites are not scored as polymorphic simply because they are invariant in the set of individuals selected for sequencing.

Table 4.2 Proportion of Variants Discovered, by Re-sequencing Investment.

	21,491 Polymorphic Sites, Segregated According to Population Frequency				
	15336 sites with MAF<.5%	1074 sites with MAF .5-1%	934 sites with MAF 1-2%	1200 sites with MAF 2-5%	2947 sites with MAF >5%
Sequencing Depth	% Detected SNPs	% Detected SNPs	% Detected SNPs	% Detected SNPs	% Detected SNPs
n = 100 individuals resequenced using a shotgun approach					
1x	0%	4%	11%	40%	94%
2x	1%	16%	34%	74%	99%
4x	3%	34%	61%	92%	100%
8x	7%	59%	83%	98%	100%
16x	10%	73%	91%	99%	100%
n = 200 individuals resequenced using a shotgun approach					
1x	1%	9%	26%	71%	99%
2x	2%	25%	55%	93%	100%
4x	5%	55%	84%	99%	100%
8x	10%	80%	96%	100%	100%
n = 400 individuals resequenced using a shotgun approach					
1x	1%	17%	49%	93%	100%
2x	2%	45%	83%	100%	100%
4x	7%	79%	97%	100%	100%

I simulated 1Mb regions in individuals with HapMap YRI-like degrees of LD. Then, I generated shotgun sequence data for a subset of individuals (n = 100, 200 or 400) at varying depths (1x – 16x). The depths were selected to represent a total investment of between 100x and 1600x coverage of the region. Simulated reads were 32-bp long and had a per base error rate of 0.3%. Read counts at sites where multiple copies of each alternative base were observed were then provided as input to our software package.

SNPs in this table are classified according to minor allele frequency (calculated from a sample of N = 10,000 chromosomes). For each minor allele frequency group, percentages of detected SNPs are tabulated.

Table 4.3 Proportion of Variants Discovered, by Re-sequencing Error Rate.

		21,491 Polymorphic Sites, Segregated According to Population Frequency				
	978,509 Monomorphisms	15336 sites with MAF<.5%	1074 sites with MAF .5-1%	934 sites with MAF 1-2%	1200 sites with MAF 2-5%	2947 sites with MAF >5%
Sequencing Depth	False Positives SNPs	# Detected SNPs	# Detected SNPs	# Detected SNPs	# Detected SNPs	# Detected SNPs
n = 400 individuals resequenced using a shotgun approach, error = 0.3%						
1x	53	105	187	462	1112	2947
2x	46	351	485	776	1194	2947
4x	55	1021	847	910	1200	2947
n = 400 individuals resequenced using a shotgun approach, error = 0.5%						
1x	276	113	182	449	1100	2947
2x	95	257	401	716	1188	2947
4x	99	842	793	901	1200	2947
n = 400 individuals resequenced using a shotgun approach, error = 1.0%						
1x	102	38	70	249	953	2945
2x	140	139	259	593	1166	2947
4x	61	430	610	846	1198	2947

I simulated 1Mb regions in individuals with HapMap YRI-like degrees of LD. Then, I generated shotgun sequence data for a subset of 400 individuals at varying depths (1x – 4x). The depths were selected to represent a total investment of between 400x and 1600x coverage of the region. Simulated reads were 32-bp long and had varying per base error rates (0.3%, 0.5% and 1.0%). Read counts at sites where multiple copies of each alternative base were observed were then provided as input to our software package.

SNPs in this table are classified according to minor allele frequency (calculated from a sample of N = 10,000 chromosomes). For each minor allele frequency group, numbers of detected SNPs are tabulated. In addition, numbers of false discoveries are included.

Table 4.4 Accuracy of Genotype Predictions, by Re-sequencing Investment.

	21,491 Polymorphic Sites, Segregated According to Population Frequency				
	15336 sites with MAF<.5%	1074 sites with MAF .5-1%	934 sites with MAF 1-2%	1200 sites with MAF 2-5%	2947 sites with MAF >5%
Sequencing Depth	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
n = 100 individuals resequenced using a shotgun approach					
1x	97.91%	97.46%	97.02%	96.71%	95.07%
2x	98.36%	98.26%	98.28%	98.27%	98.07%
4x	99.17%	99.30%	99.38%	99.48%	99.47%
8x	99.85%	99.91%	99.92%	99.93%	99.94%
16x	99.99%	99.99%	99.99%	100.00%	99.99%
n = 200 individuals resequenced using a shotgun approach					
1x	98.58%	98.35%	98.29%	98.38%	98.39%
2x	99.02%	99.11%	99.20%	99.36%	99.52%
4x	99.68%	99.76%	99.81%	99.85%	99.90%
8x	99.96%	99.96%	99.97%	99.98%	99.98%
n = 400 individuals resequenced using a shotgun approach					
1x	98.94%	98.88%	98.87%	99.00%	99.16%
2x	99.37%	99.47%	99.53%	99.64%	99.76%
4x	99.83%	99.87%	99.89%	99.92%	99.96%

I simulated 1Mb regions in individuals with HapMap YRI-like degrees of LD. Then, I generated shotgun sequence data for a subset of individuals (n = 100, 200 or 400) at varying depths (1x – 16x). The depths were selected to represent a total investment of between 100x and 1600x coverage of the region. Simulated reads were 32-bp long and had a per base error rate of 0.3%. Read counts at sites where multiple copies of each alternative base were observed were then provided as input to our software package.

SNPs in this table are classified according to minor allele frequency (calculated from a sample of N = 10,000 chromosomes). For each minor allele frequency group, imputation accuracies are tabulated.

Table 4.5 Accuracy of Genotype Predictions, by Re-sequencing Error Rate.

	21,491 Polymorphic Sites, Segregated According to Population Frequency				
	15336 sites with MAF<.5%	1074 sites with MAF .5-1%	934 sites with MAF 1-2%	1200 sites with MAF 2-5%	2947 sites with MAF >5%
Sequencing Depth	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
n = 400 individuals resequenced using a shotgun approach, error = 0.3%					
1x	98.94%	98.88%	98.87%	99.00%	99.16%
2x	99.37%	99.47%	99.53%	99.64%	99.76%
4x	99.83%	99.87%	99.89%	99.92%	99.96%
n = 400 individuals resequenced using a shotgun approach, error = 0.5%					
1x	98.99%	98.93%	98.92%	98.99%	99.11%
2x	99.35%	99.46%	99.53%	99.63%	99.75%
4x	99.82%	99.86%	99.89%	99.91%	99.95%
n = 400 individuals resequenced using a shotgun approach, error = 1.0%					
1x	99.01%	98.71%	98.75%	98.86%	98.99%
2x	99.33%	99.37%	99.47%	99.59%	99.72%
4x	99.77%	99.83%	99.86%	99.90%	99.94%

I simulated 1Mb regions in individuals with HapMap YRI-like degrees of LD. Then, I generated shotgun sequence data for a subset of 400 individuals at varying depths (1x – 4x). The depths were selected to represent a total investment of between 400x and 1600x coverage of the region. Simulated reads were 32-bp long and had varying per base error rates (0.3%, 0.5% and 1.0%). Read counts at sites where multiple copies of each alternative base were observed were then provided as input to our software package.

SNPs in this table are classified according to minor allele frequency (calculated from a sample of N = 10,000 chromosomes). For each minor allele frequency group, imputation accuracies are tabulated.

Figure 4.1 Low-pass vs. Deep Sequencing: Polymorphism Discovery.

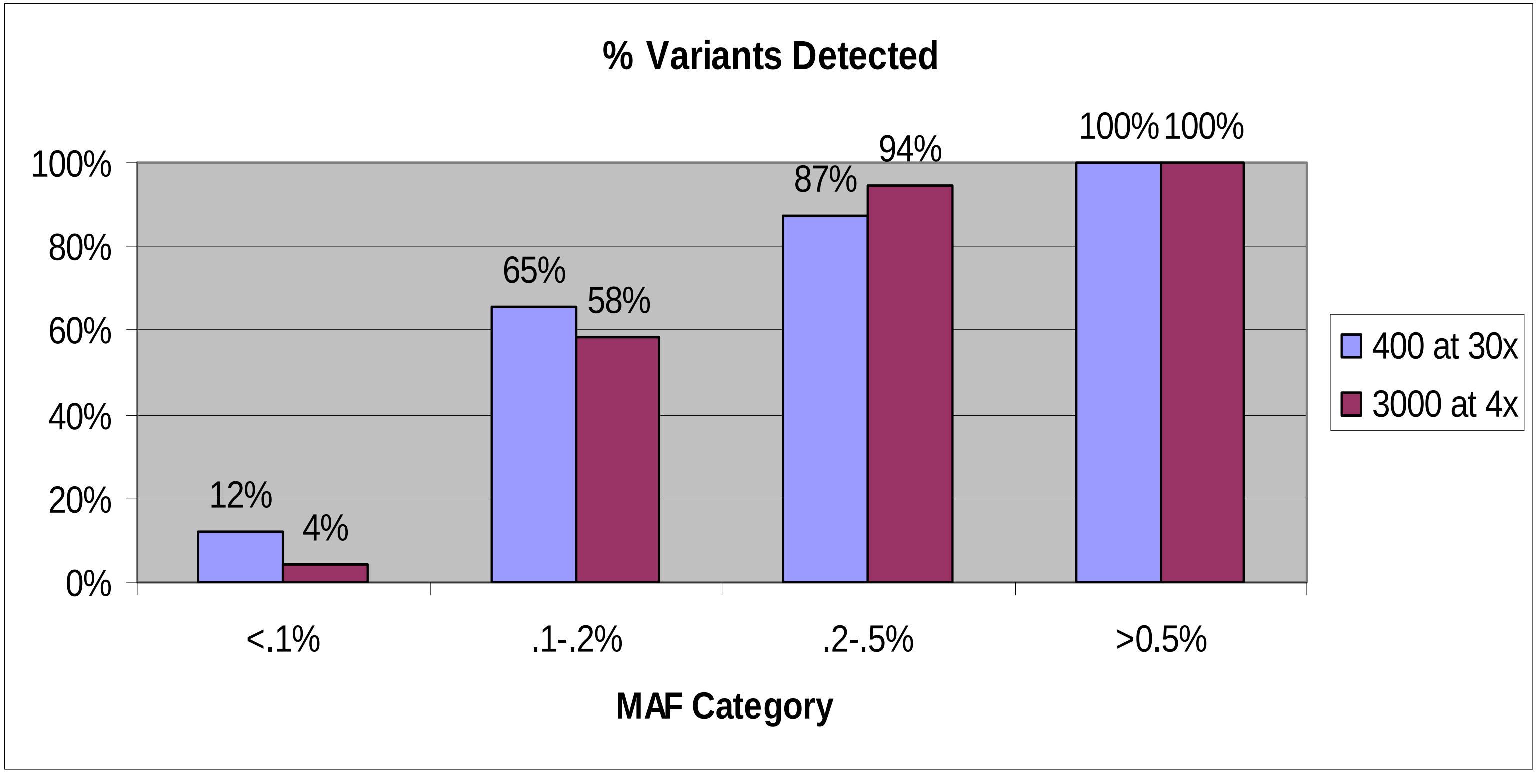


Table 4.6 Low-pass vs. Deep Sequencing: Imputation Accuracy at Discovered Loci.

Statistic	Design	<.1%	.1-.2%	.2-.5%	.5-1%	1-2%	2-5%	>5%
Accuracy (all)	400 at 30x	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	3000 at 4x	.9997	.9994	.9988	.9985	.9988	.9984	.9990
Accuracy (het)	400 at 30x	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	3000 at 4x	.8711	.8248	.8193	.9039	.9726	.9884	.9985
r^2	400 at 30x	.9930	.9949	.9961	.9974	.9981	.9988	.9998
	3000 at 4x	.6635	.6390	.6897	.8021	.9192	.9577	.9927

Figure 4.2 Low-pass vs. Deep Sequencing: Information/Effective Sample Size.

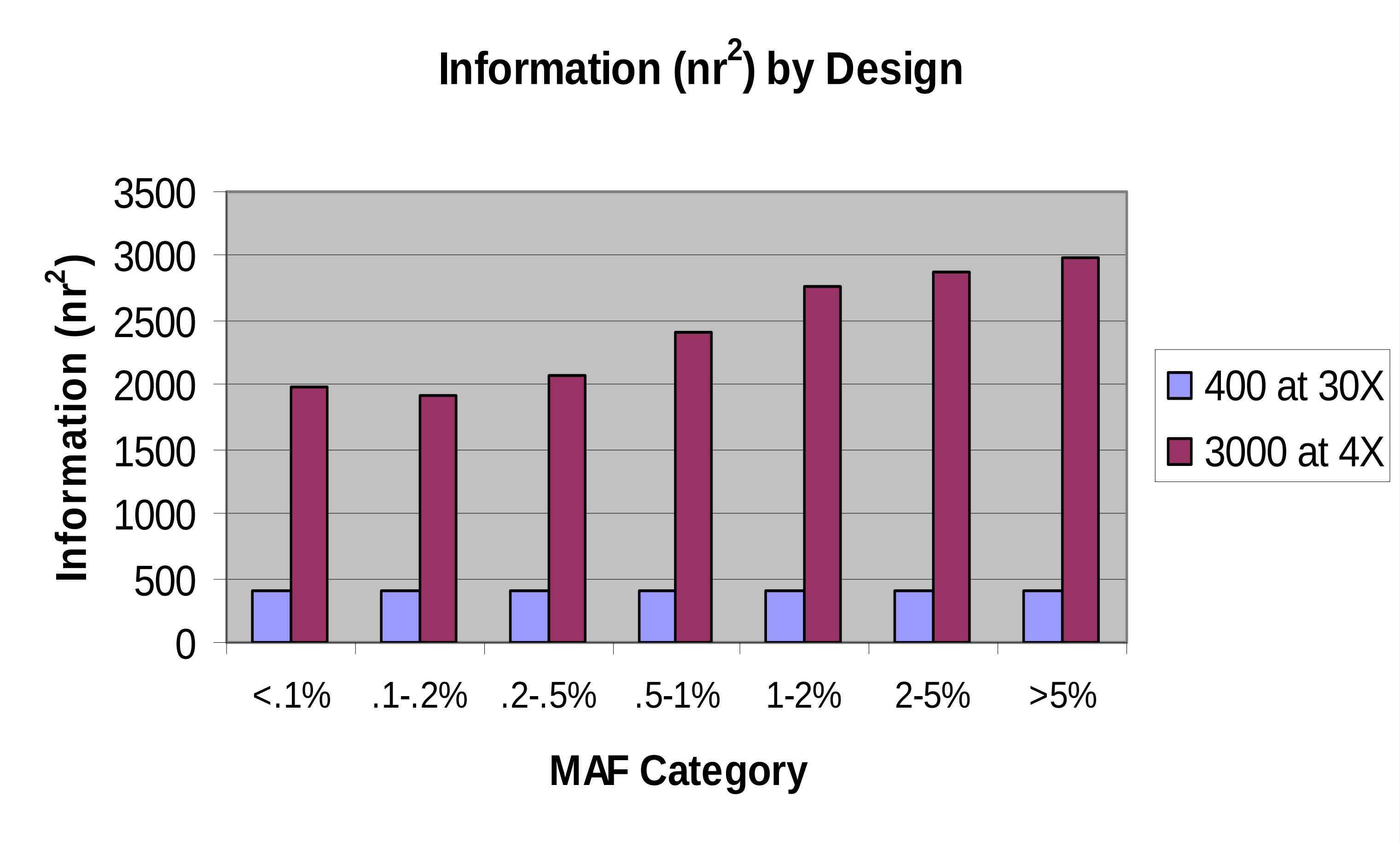


Table 4.7 Imputation Quality using a Smaller or Larger "Re-sequencing HapMap": Accuracy.

		1,000,000 bases	988,524 Monomorphic Sites		11,476 Polymorphic Sites, Segregated According to Sample Frequency									
					5125 sites with MAF<.5%		1156 sites with MAF .5-1%		1021 sites with MAF 1-2%		1209 sites with MAF 2-5%		2965 sites with MAF >5%	
Reference Sample Sequencing Depth	Reference Sample Total Investment	Average Errors per Person	False Positives SNPs	Accuracy	Detected SNPs	Accuracy	Detected SNPs	Accuracy	Detected SNPs	Accuracy	Detected SNPs	Accuracy	Detected SNPs	Accuracy
Imputation of 500 individuals based on the Smaller Sequencing Hapmap: 60 individuals sequenced at > 16X, 120 perfect reference haplotypes														
>16x	>960x	416	167	99.47%	452	99.64%	419	99.44%	581	99.11%	952	98.32%	2635	95.80%
		302	200	99.40%	527	99.60%	475	99.46%	653	99.21%	1017	98.75%	2657	97.88%
Imputation of 500 individuals based on the Larger Sequencing Hapmap: 400 individuals sequenced at 2X, 800 imperfect reference haplotypes														
2x	800x	349	138	98.92%	289	99.23%	423	99.31%	665	99.19%	982	98.76%	2618	97.07%
		255	155	98.82%	342	99.22%	486	99.34%	745	99.27%	1061	99.09%	2660	98.68%

Simulated "re-sequencing HapMap" reference panels mimic HapMap YRI-like LD pattern. Simulated reads were 32 base pair in length. Sequencing error rate was set at 0.1% for a random 90% of the region while the remaining 10% was considered non-sequencable. The smaller "re-sequencing HapMap" reference panel consists of 120 true/simulated haplotypes of 60 individuals and the larger one consists of imputed haplotypes from analyzing 400 individuals sequenced at 2x coverage. To approximate the true haplotypes of the 60 individuals in the smaller reference panel, a coverage of 16x or more is required (probably also with the aid of information from family members, for instance, using a trio design as for the current HapMap CEU and YRI). Thus, the larger panel represents a total sequencing investment of 800x and the smaller over 960x.

A study sample of 500 individuals was simulated from the same underlying population of the "re-sequencing HapMap". A set of 100, or 200 tagSNPs were selected randomly from the pool of SNPs found in both the larger and smaller "re-sequencing HapMap" and genotyped in the study sample. Genotypes of all "re-sequencing HapMap" SNPs were then imputed by jointly modeling tagSNP genotypes of the study sample individuals and haplotypes in the "re-sequencing HapMap" reference panel.

SNPs in this table are classified according to minor allele frequency (calculated from the sample of 500 individuals). For each minor allele frequency group, numbers of detected SNPs and imputation accuracies are tabulated. In addition, average imputation errors per individual across the entire 1Mb region, the number of false polymorphisms and their corresponding imputation accuracies are also included.

Table 4.8 Imputation Quality using a Smaller or Larger "Re-sequencing HapMap": r^2.

		11,476 Polymorphic Sites, Segregated According to Sample Frequency									
		5125 sites with MAF<.5%		1156 sites with MAF .5-1%		1021 sites with MAF 1-2%		1209 sites with MAF 2-5%		2965 sites with MAF >5%	
Reference Sample Sequencing Depth	Reference Sample Total Investment	Detected SNPs	r2	Detected SNPs	r2	Detected SNPs	r2	Detected SNPs	r2	Detected SNPs	r2
Imputation of 500 individuals based on 120 perfect reference haplotypes: 60 individuals sequenced at > 16x											
>16x	>960x	452	46.02%	419	45.86%	581	47.23%	952	51.15%	2635	66.89%
		527	46.73%	475	51.43%	653	55.46%	1017	63.18%	2657	81.33%
Imputation of 500 individuals based on 800 imperfect reference haplotypes: 400 individuals sequenced at 2x											
2x	800x	289	33.55%	423	48.18%	665	55.76%	982	63.18%	2618	76.17%
		342	37.81%	486	53.27%	745	62.24%	1061	73.13%	2660	87.78%

Simulated "re-sequencing HapMap" reference panels mimic HapMap YRI-like LD pattern. Simulated reads were 32 base pair in length. Sequencing error rate was set at 0.1% for a random 90% of the region while the remaining 10% was considered non-sequencable. The smaller "re-sequencing HapMap" reference panel consists of 120 true/simulated haplotypes of 60 individuals and the larger one consists of imputed haplotypes from analyzing 400 individuals sequenced at 2x coverage. To approximate the true haplotypes of the 60 individuals in the smaller reference panel, a coverage of 16x or more is required (probably also with the aid of information from family members, for instance, using a trio design as for the current HapMap CEU and YRI). Thus, the larger panel represents a total sequencing investment of 800x and the smaller over 960x.

A study sample of 500 individuals was simulated from the same underlying population of the "re-sequencing HapMap". A set of 100, or 200 tagSNPs were selected randomly from the pool of SNPs found in both the larger and smaller "re-sequencing HapMap" and genotyped in the study sample. Genotypes of all "re-sequencing HapMap" SNPs were then imputed by jointly modeling tagSNP genotypes of the study sample individuals and haplotypes in the "re-sequencing HapMap" reference panel.

SNPs in this table are classified according to minor allele frequency (calculated from the sample of 500 individuals). For each minor allele frequency group, numbers of detected SNPs and squared correlations between imputed and true allele counts are tabulated.

Chapter 5
Conclusions and Discussions

Identifying and characterizing the genetic variants that impact human traits, ranging from disease susceptibility to variability in personality measures, is one of the central objectives of human genetics. Ultimately, this aim will be achieved by examining the relationship between interesting traits and the whole-genome sequences of many individuals. Although whole-genome resequencing of thousands of individuals is not yet feasible, geneticists have long recognized that good progress can be made by measuring only a relatively modest number of genetic variants in each individual. This type of "incomplete" information is useful because data about any set of genetic variants in a group of individuals provides useful information about many other unobserved genetic variants in the same individuals. In this dissertation, I have proposed a series of hidden Markov models to maximally utilize "incomplete" genetic information garnered by individual studies as well as by large public efforts such as the International HapMap project and the 1000 Genomes Project.

5.1 Review of Previous Chapters

In Chapter 2, I introduce the basic form of the underlying hidden Markov models in the context of phase inference, along with a number of techniques to gain computational efficiency. Besides providing the theoretical underpinnings, I give intuition why the proposed method works. I proceed to evaluate its performance both in simulated settings and in real studies. By comparing with a number of standard haplotyping methods, I demonstrate that the proposed method is at least comparable in the quality of reconstructed haplotypes, and is more computationally efficient, if not both. In addition, I introduce the concept of utilizing data from public databases for a more efficient joint analysis.

Following the introduction of the concept of a joint analysis with publicly available data, Chapter 3 focuses on the inference and assessment of untyped variants that are not directly examined in individual genetic studies of certain trait(s) of interest but for which information is available in public databases. I present an extended hidden Markov model generating most likely estimates of genotypes at the untyped loci to more efficiently achieve the goal of genotype imputation (as opposed to phase inference in Chapter 2). My model conveniently generates several estimates for each missing genotype (including most likely genotype guess, the expected count of a reference allele and the posterior probabilities of each potential genotype guess) so that imputation uncertainty can be taken into account in subsequent analysis. Moreover, two measures at the SNP level (across individuals) are proposed as quality filters.

I showcase the merits of the proposed genotype imputation method in a broad range of simulated and real settings, involving up to millions of genetic markers in tens of thousands of individuals. In brief, my model generates highly accurate estimates of the missing genotypes and provides quality measures than can well discriminate well-imputed markers from badly-imputed ones. The statement holds true for a random set of markers across the genome, and perhaps more importantly, for subsets of markers that show strong association with trait(s) of interest.

Chapter 3 examines the applicability of my method to non-Caucasian populations, the choice of appropriate public reference panels particularly for study populations that do not have an obvious matching population from public databases, the extent of information gain when started with different sets of directly assayed genetic markers across different populations, the impact of larger reference databases, and the boost in statistical power achieved by imputation. Simulated datasets I generated for power evaluation have been shared with a large group of collaborators as a benchmark to compare different imputation methods on untyped markers.

In Chapter 4, I further extend my model to accommodate massive parallel sequence data. I preview the role of the imputation-based approach in the era of sequencing-based studies. Although sequencing costs have dropped drastically, whole-genome deep sequencing of a large number of individuals is still not practical. I therefore propose low-pass designs where a relatively large number of individuals are sequenced at low depth and information is combined across individuals using the proposed imputation-based model. I demonstrate the merits of such designs through extensive simulation studies in terms of polymorphism discovery and genotype calling among the sequenced individuals, as well as the utility of the reconstructed chromosomes as reference data for individual genotyping-based studies. Preliminary analysis on real data from the 1000 Genomes Project generates results consistent with expectations based on simulations. The proposed method is influencing the design of several large-scale genetic studies by enabling an alternative that results in a much larger effective sample size. The increase in sample size is critical for mapping genes influencing complex traits.

5.2 Remarks on Subsequent Analysis

Genotype imputation has been adopted by a larger and larger number of GWAS and has become a routine for meta-analysis. The imputation-based gene-mapping is a two-step process. First, genotypes at untyped markers are imputed. Then imputed genotypes are tested for association with phenotypes. This dissertation has so far focused on the first step.

At the end of the first step, my software generates at each marker locus three sets of summary statistics that can be potentially used for subsequent association analysis: (1) an imputed “best-guess” genotype for each individual, which corresponds to the marginal mode of the posterior distribution of the underlying genotype integrated over all possible haplotype configurations; (2) an expected allelic count, or dosage for each individual; and (3) the marginal posterior probabilities of the three potential underlying genotypes. (3) contains the most information while (1) the least. For example, for a marker with two alleles A and B, if the posterior probabilities for A/A, A/B, and B/B are 0.9, 0.1, and 0 respectively, the best-guess genotype would be A/A and the dosage for allele A would be 0.9*2 + 0.1 = 1.9.

With the three summary statistics, one can use the following three strategies for subsequent analysis: (1) least-squares regression on the “best-guess” imputed genotype; (2) regression on the expected genotype score or “dosage”; or (3) mixture regression models that more fully incorporate posterior probabilities of genotypes at untyped SNPs

Using (1) for subsequent analysis ignores the uncertainty in the imputed genotypes. When imputation is accurate, the correspondence between the true and imputed genotypes is strong and analyzing the best-guess genotypes might result in little bias and power loss compared with an analysis of the true genotypes. However, if imputation accuracy is low, there could be substantial bias and power loss. I therefore recommend *NEVER* using the best-guess genotypes for subsequent gene-mapping analysis.

On the other extreme, one can use mixture regression models to take full advantage of the individual posterior probabilities. This approach should be superior when imputation uncertainty is not reflected by allelic dosages. For example, this may occur when the posterior probabilities are high for the two homozygotes, and the allelic dosage would indicate a heterozygous underlying genotype.

We have used simulations to assess the relative performance of three approaches across a range of sample sizes, minor allele frequencies, and imputation accuracies to compare the performance of the different methods under multiple genetic models (Zheng et al. unpublished data). The mixture models performed the best in the setting of a small sample size (below 200) and low imputation accuracies (Rsq below 0.3).

For most realistic settings of GWAS, such as modest genetic effects, large sample sizes, and high average imputation accuracies, dosage-based analysis (i.e., regressing the phenotype of interest on the dosages) provides adequate performance. In fact, for these settings, small gains from using the full mixture models are rendered negligible by the increased model complexity and associated cost of estimating additional parameters.

5.3 Limitations and Future Directions

As one of the first attempts to deal with data of this scale, my method can be further improved. First, in my hidden Markov models, the two sets of parameters are not sampled according to conditional probabilities but rather are estimated by counting the proportion of relevant event (crossover or mismatch event). I have explored the performance of sampling the parameters from a Beta distribution with parameters (#crossovers/#mismatches + 2, #non-crossovers/#matches + 2), derived from a non-informative conjugate prior of Beta (1, 1). In the FUSION chromosome 14 data, the performance was slightly worse. Specifically the average allelic discordance and average squared correlation between imputed genotype scores and true genotypes are 1.41% and 0.916, instead of 1.37% and 0.919 when using the original model. The slightly worse performance is likely due to more noise introduced. More work is warranted to explore how such fully Bayesian approaches influence imputation quality, and more importantly how they influence subsequent analysis. By doing the proper drawing under a Gibbs' sampling framework, a proper multiple imputation procedure can be adopted for subsequent marker-trait analysis.

Secondly, I have not assessed the theoretical properties of the proposed heterogeneous hidden Markov models. Although having good performances in real and simulated datasets, such models need to be rigorously examined in terms of desirable statistical prosperities. Specifically, I would like to evaluate the following two aspects: (1) conditions that lead to model convergence; and (2) whether the estimated paratmers are consistent.

In addition, I would like to further explore other measures of imputation quality measures. Similar in nature to the r^2 measure introduced in Chapter 3, others (Huang *et* al. 2009; Browning and Browning 2009) have proposed alternative predictors of imputation quality. Their performance needs to be evaluated, particularly for rare variants.

5.4 Conclusion

In summary, I have developed computationally efficient models for the analysis of large-scale genetic association studies. My methods are flexible enough to accommodate phased haplotype or genotype data. My approach is one of the first attempts to deal with data of this scale in a manner that is statistically and computationally efficient. My models can handle millions of genetic markers measured on tens of thousands of individuals. My methods have played a key role in mapping genes associated with the risk of complex diseases and genes influencing complex non-disease traits. Table 5.1 provides a partial list of recent genome-wide association scans that use my imputation method (Li *et al.* 2009). Our software MACH has been downloaded by more than 100 research groups. My implemented post-imputation analysis software mach2dat and mach2qtl (for binary and quantitative traits respectively) have also been widely used. I believe the methods will continue to facilitate the identification of genes that contribute to the risk of complex diseases, in particular, through the combined analysis of large-scale studies that examine different sets of genetic markers, and through maximally exploiting information from resequencing-based studies.

Table 5.1 Examples of GWAS that Have Used MACH for Genotype Imputation.

First Author	Journal	Publication Date	Title
Aulchenko	Nature Genetics	2008/12	Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts
Barrett	Nature Genetics	2008/06	Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease
Broadbent	Hum Mol Genet	2007/11	Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p
Chambers	Nature Genetics	2008/05	Common genetic variation near MC4R is associated with waist circumference and insulin resistance
Chen	J Clin Invest	2008/07	Variations in the G6PC2/ABCB11 genomic region are associated with fasting glucose levels
Dehghan	The Lancet	2008/10	Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study
Ferreira	Nature Genetics	2008/07	Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder
Hung	Nature	2008/04	A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25
Kathiresan	Nature Genetics	2008/12	Common variants at 30 loci contribute to polygenic dyslipidemia
Lettre	Nature Genetics	2008/04	Identification of ten loci associated with height highlights new biological pathways in human growth
Loos	Nature Genetics	2008/05	Common variants near MC4R are associated with fat mass, weight and risk of obesity
Rafiq	Diabetologia	2008/10	Gene variants influencing measures of inflammation or predisposing to autoimmune and inflammatory diseases are not associated with the risk of type 2 diabetes
Sanders	Am J Psychiatry	2008/01	No significant association of 14 candidate genes with schizophrenia in a large European ancestry sample: implications for psychiatric genetics
Sanna	Nature Genetics	2008/01	Common variants in the GDF5-UQCC region are associated with variation in human height
Scott	Science	2007/04	A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants
Scott	PNAS	2009/04	Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry
Willer	Nature Genetics	2008/01	Newly identified loci that influence lipid concentrations and risk of coronary artery disease
Willer	Nature Genetics	2008/12	Six new loci associated with body mass index highlight a neuronal influence on body weight regulation
Zeggini	Nature Genetics	2008/03	Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes

References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97-101 (2002).

Andridge RR and Little RA. A review of hot deck imputation for survey nonresponse. *submitted*.

Barrett JC and Cardon LR. Evaluating coverage of genome-wide association studies. *Nat Genet* **38**, 659-662 (2006).

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ; NIDDK IBD Genetics Consortium, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, and Daly MJ. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40**, 955-962 (2008).

Baum LE. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* **3**, 1-8 (1972).

Bentley D.R. Whole-genome re-sequencing. *Curr Opin Genet Dev* **16**, 545-552 (2006).

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).

Botstein D, and Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease future approaches for complex diseases. *Nat Genet* **33 Suppl**, 228-237 (2003).

Browning SR, and Browning BL. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *Am J Hum Genet* **81**, 1084-1097 (2007).

Browning BL, and Browning SR. A Unified Approach to Genotype Imputation and

Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *Am J Hum Genet* **84**, 210-223 (2009).

Burdick JT, Chen WM, Abecasis GR and Cheung VG. In silico method for inferring genotypes in pedigrees. Nat Genet 38, 1002-1004 (2006).

Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L and Nickerson DA. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet* **33**, 518-521 (2003).

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L and Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* **74**, 106-120 (2004).

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA and Pritchard JK. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38**, 1251-1260 (2006).

Daly MJ, Rioux JD, Schaffner SE, Hudson TJ and Lander ES. Highresolution haplotype structure in the human genome. *Nature Genetics* **29**, 229-232 (2001).

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ and Altshuler D. Efficiency and power in genetic association studies. *Nat Genet* **37**, 1217-1223 (2005).

Dewan A, Klein RJ and Hoh J. Linkage disequilibrium mapping for complex disease genes. *Methods Mol Biol* **376**, 85-107 (2007).

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Boström K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Råstam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjögren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D and Purcell S. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331-1336 (2007).

Eiberg H, Mohr J and Nielsen LS. Linkage relationships of human coagulation factor XIIIB. *Cytogenet Cell Genet* **37**, 463 (1984).

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I. Identification of a variant associated with adult-type hypolactasia. *Nat Genet* **30**, 233-237 (2002).

Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, Domingo R Jr, Ellis MC, Fullan A, Hinton LM, Jones NL, Kimmel BE, Kronmal GS, Lauer P, Lee VK, Loeb DB, Mapa FA, McClelland E, Meyer NC, Mintier GA, Moeller N, Moore T, Morikang E, Prass CE, Quintana L, Starnes SM, Schatzman RC, Brunke KJ, Drayna DT, Risch NJ, Bacon BR and Wolff RK. A novel MHC-class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat Genet* **13**, 399-408 (1996).

Gaulton KJ, Willer CJ, Li Y, Scott LJ, Conneely KN, Jackson AU, Duren WL, Chines PS, Narisu N, Bonnycastle LL, Luo J, Tong M, Sprau AG, Pugh EW, Doheny KF, Valle TT, Abecasis GR, Tuomilehto J, Bergman RN, Collins FS, Michael Boehnke M and Mohlke KL. Comprehensive association study of type 2 diabetes and related quantitative traits with 222 candidate genes. *Diabetes* **57**, 3136-3144 (2008).

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, and Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* **106**, 9362-9367 (2009).

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA and Cox DR. Whole-Genome patterns of common DNA variation in three human populations. *Science* **307**, 1072-1079 (2005).

Hirschhorn JN and Daly MJ. Genome-Wide Association Studies for Common Diseases and Complex Traits. *Nat Rev Genet* **6**, 95-108 (2005).

Hoh J and Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet* **4**, 701-709 (2003).

Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, Scheet P. Genotype-Imputation Accuracy across Worldwide Human Populations. *Am J Hum Genet* **84**, 235-50 (2009).

Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, Cooper GM, Roos C, Voight BF, Havulinna AS, Wahlstrand B, Hedner T, Corella D, Tai ES, Ordovas JM, Berglund G, Vartiainen E, Jousilahti P, Hedblad B, Taskinen MR, Newton-Cheh C, Salomaa V, Peltonen L, Groop L, Altshuler DM, and Orho-Melander M. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* **40**, 189–97 (2008).

Kruglyak L, Daly MJ, Reeve-Daly MP and Lander ES. Parametric and nonparametric linkage analysis: A unified multipoint approach. *Am J Hum Genet* **58**, 1347-1363 (1996).

Kruglyak L and Nickerson DA. Variation is the spice of life. *Nature Genet* **27**, 234-236 (2001).

Lander ES and Green P. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA* **84**, 2363-7 (1987).

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, et al; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

Lander ES and Schork NJ. Genetic dissection of complex traits. *Science* **265**, 2037-2048 (1994).

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, and Venter JC. The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254 (2008).

Li B, and Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**, 311-321 (2008).

Li M, Atmaca-Sonmez P, Othman M, Branham KE, Khanna R, Wade MS, Li Y, Liang L, Zareparsi S, Swaroop A and Abecasis GR. CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nat Genet* **38**, 1049–1054 (2006).

Li N and Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213-2233 (2003).

Li Y, Willer CJ, Sanna S, and Abecasis GR. Genotype imputation. *Annu Rev Genomics Hum Genet* In Press (2009).

Lin S, Chakravarti A and Cutler DJ. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat Genet* **36**, 1181-1188 (2004).

Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, Inouye M, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet* **40**, 768-775 (2008).

Madsen BE, and Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5**, e1000384 (2009).

Marchini J, Culter D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR and Donnelly P. A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet* **78**, 437-450 (2006).

Markianos K, Daly MJ and Kruglyak L. Efficient multipoint linkage analysis through reduction of inheritance space. *Am J Hum Genet* **68**, 963-977 (2001).

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, and Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**, 356-369 (2008).

Mott R and Flint J. Simultaneous detection and fine mapping of quantitative trait loci in mice using heterogeneous stocks. *Genetics* **160**, 1609-18 (2002).

Mott R, Talbot CJ, Turri MG, Collins AC and Flint J. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc Natl Acad Sci USA* **97**, 12649-12654 (2000).

Nicolae DL. Testing untyped alleles (TUNA)-applications to genome-wide association studies. *Genet Epidemiol* **30**, 718-727 (2006).

Pe'er I., de Bakker PI, Maller J, Yelensky R, Altshuler D and Daly MJ. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet* **38**, 663-667 (2006).

Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford University Press (2003).

Pritchard JK, and Przeworski M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69**, 1-14 (2001).

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ and Sham PC. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* **81**, 559-575 (2007).

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

Rubin DB. Multiple Imputation in Sample Surveys - a Phenomenological Bayesian Approach to Nonresponse. in American Statistical Association Proceedings of the Survey Research Methods Section, pp. 20–34 (1978).

Saaristo T, Peltonen M, Lindström J, Saarikoski L, Sundvall J, Eriksson JG and T Jaakko. Cross-sectional evaluation of the Finnish Diabetes Risk Score: a tool to identify undetected type 2 diabetes, abnormal glucose tolerance and metabolic syndrome. *Diabetes Vasc Dis Res* **2**, 67-72 (2005).

Sanger F, Nicklen S, and Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**,5463–67 (1977).

Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, Bonnycastle LL, Shen H, Timpson N, Lettre G, Usala G, Chines PS, Stringham HM, Scott LJ, Dei M, Lai S, Albai G, Crisponi L, Naitza S, Doheny KF, Pugh EW, Ben-Shlomo Y, Ebrahim S, Lawlor DA, Bergman RN, Watanabe RM, Uda M, Tuomilehto J, Coresh J, Hirschhorn JN, Shuldiner AR, Schlessinger D, Collins FS, Davey Smith G, Boerwinkle E, Cao A, Boehnke M, Abecasis GR, and Mohlke KL. Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat Genet* **40**, 198-203 (2008).

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* **15**, 1576-1583 (2005).

Scheet P and Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**, 629-44 (2006).

Scott JL, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS and Boehnke M. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341-1345 (2007).

Scott LJ, Muglia P, Kong XQ, Guan W, Flickinger M, Upmanyu R, Tozzi F, Li JZ, Burmeister M, Absher D, Thompson RC, Francks C, Meng F, Antoniades A, Southwick AM, Schatzberg AF, Bunney WE, Barchas JD, Jones EG, Day R, Matthews K, McGuffin P, Strauss JS, Kennedy JL, Middleton L, Roses AD, Watson SJ, Vincent JB, Myers RM, Farmer AE, Akil H, Burns DK, Boehnke M. Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. *Proc Natl Acad Sci USA* **106**, 7501-7506 (2009).

Silander K, Scott LJ, Valle TT, Mohlke KL, Stringham HM, Wiles KR, Duren WL, Doheny KF, Pugh EW, Chines P, Narisu N, White PP, Fingerlin TE, Jackson AU, Li C, Ghosh S, Magnuson VL, Colby K, Erdos MR, Hill JE, Hollstein P, Humphreys KM, Kasad RA, Lambert J, Lazaridis KN, Lin G, Morales-Mena A, Patzkowski K, Pfahl C, Porter R, Rha D, Segal L, Suh YD, Tovar J, Unni A, Welch C, Douglas JD, Epstein MP, Hauser ER, Hagopian W, Buchanan TA, Watanabe RM, Bergman RN, Tuomilehto J, Collins FS and Boehnke M. A Large Set of Finnish Affected Sibling Pair Families With Type 2 Diabetes Suggests Susceptibility Loci on Chromosomes 6, 11, and 14. *Diabetes* **53**, 821-829 (2004).

Skol AD, Scott LJ, Abecasis GR and Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* **38**, 209-213 (2006).

Stephens M and Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* **76**, 449-462 (2005).

Stephens M., Smith NJ and Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**, 978-989 (2001).

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

Valle T, Tuomilehto J, Bergman RN, Ghosh S, Hauser ER, Eriksson J, Nylund SJ, Kohtamaki K, Toivanen L, Vidgren G, Tuomilehto-Wolf E, Ehnholm C, Blaschak J, Langefeld CD, Watanabe RM, Magnuson V, Ally DS, Hagopian WA, Ross E, Buchanan TA, Collins FS and Boehnke M. Mapping genes for NIDDM. Design of the Finland-United States Investigation of NIDDM Genetics (FUSION) Study. *Diabetes Care* **21**, 949-958 (1998).

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, *et al*. The sequence of the human genome. *Science* **291**, 1304-1351 (2001).

Visscher PM, Hill WG, and Wray NR. Heritability in the genomics era–concepts and misconceptions. *Nat Rev Genet* **9**,255–266 (2008).

Wainwright B Scambler P, Farrall M, Schwartz M, and Williamson R. Linkage between the cystic fibrosis locus and markers on chromosome 7q. *Cytogenet Cell Genet* **41**, 191-192 (1986).

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–76 (2008)

Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, Strait J, Duren WL, Maschio A, Busonero F, Mulas A, Albai G, Swift AJ, Morken MA, Narisu N, Bennett D, Parish S, Shen H, Galan P, Meneton P, Hercberg S, Zelenika D, Chen WM, Li Y, Scott LJ, Scheet PA, Sundvall J, Watanabe RM, Nagaraja R, Ebrahim S, Lawlor DA, Ben-Shlomo Y, Davey-Smith G, Shuldiner AR, Collins R, Bergman RN, Uda M, Tuomilehto J, Cao A, Collins FS, Lakatta E, Lathrop GM, Boehnke M, Schlessinger D, Mohlke KL and Abecasis GR. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40**, 161-169 (2008).

Willer CJ, Scott LJ, Bonnycastle LL, Jackson AU, Chines P, Pruim R, Bark CW, Tsai YY, Pugh EW, Doheny KF, Kinnunen L, Mohlke KL, Valle TT, Bergman RN, Tuomilehto J, Collins FS and Boehnke M. Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genet Epidemiol* **30**, 180-190 (2006).

Zaitlen N, Kang HM, Eskin E and Halperin E. Leveraging the HapMap correlation structure in association studies. *Am J Hum Genet* **80**, 683-691 (2007).

Zeggini E, Scott LJ, Saxena R, Voight BF, for the Diabetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Meta-analysis of genome-wide association data and large-scale replication identifies several additional susceptibility loci for type 2 diabetes. *Nat Genet* **40**, 638-645 (2008).

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS; Wellcome Trust Case Control Consortium (WTCCC), McCarthy MI and Hattersley AT. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336-1341 (2007).