ARTICLE

GIGI: An Approach to Effective Imputation of Dense Genotypes on Large Pedigrees

Charles Y.K. Cheung,[1] Elizabeth A. Thompson,[2] and Ellen M. Wijsman[1,3,4,*]

Recent emergence of the common-disease-rare-variant hypothesis has renewed interest in the use of large pedigrees for identifying rare causal variants. Genotyping with modern sequencing platforms is increasingly common in the search for such variants but remains expensive and often is limited to only a few subjects per pedigree. In population-based samples, genotype imputation is widely used so that additional genotyping is not needed. We now introduce an analogous approach that enables computationally efficient imputation in large pedigrees. Our approach samples inheritance vectors (IVs) from a Markov Chain Monte Carlo sampler by conditioning on genotypes from a sparse set of framework markers. Missing genotypes are probabilistically inferred from these IVs along with observed dense genotypes that are available on a subset of subjects. We implemented our approach in the Genotype Imputation Given Inheritance (GIGI) program and evaluated the approach on both simulated and real large pedigrees. With a real pedigree, we also compared imputed results obtained from this approach with those from the population-based imputation program BEAGLE. We demonstrated that our pedigree-based approach imputes many alleles with high accuracy. It is much more accurate for calling rare alleles than is population-based imputation and does not require an outside reference sample. We also evaluated the effect of varying other parameters, including the marker type and density of the framework panel, threshold for calling genotypes, and population allele frequencies. By leveraging information from existing genotypes already assayed on large pedigrees, our approach can facilitate cost-effective use of sequence data in the pursuit of rare causal variants.

Introduction

Strategies used for identifying the genetic basis of human disease have evolved considerably over the past few decades. Pedigrees have been central to the discovery of genes relevant to simple Mendelian traits, leading to the identification of nearly 4,500 such genes by the end of 2011.[1] More recently, genome-wide association studies (GWASs) of large population-based samples have been used to search for variants influencing complex traits based on the common-disease-common-variant hypothesis.[2] However, although GWASs have yielded many candidate loci,[3] common variants now appear to explain only a small percentage of heritability.[4] Empirical evidence[5–9] also suggests that most complex diseases are likely be explained by rare variants. This hypothesis is leading to a resurgence in the use of large pedigrees, because the analysis of sequence data collected in large pedigrees is a particularly efficient design for identifying rare variants that affect disease risk.[10,11]

Methods now exist that overcome many earlier computational challenges for large pedigrees. Although exact computation is not feasible for large pedigrees with even a moderate number of markers,[12] Markov Chain Monte Carlo (MCMC)-based methods enable feasible and accurate analyses of large pedigrees with many markers on large pedigrees.[13–16] Recent advances continue to improve MCMC methodology[17,18] and have been implemented in (for example) the MORGAN package.[18]

Although essential to the identification of causal variants, generation of very dense genotypes from platforms that include next-generation sequencing technologies is both expensive and challenging. First, the total cost of producing dense genotypes on many subjects remains expensive, especially for sequence data. Nevertheless, it can be important to carry out a deep and comprehensive analysis of all variants in a region of interest in order to reach a conclusion about a causal locus.[19,20] Second, it is not always possible to produce genotypes on all subjects because of the quality and quantity of available DNA. This issue is particularly acute in the case of high-throughput sequencing. Together, these two potential issues can inhibit optimal analyses. One solution is to genotype a subset of individuals and carry out genotype imputation to infer missing genotypes on unobserved subjects. Genotype imputation is a cost-effective approach to leverage existing genotype data, which is often available on many subjects, with new dense genotypes collected on just a few subjects.

Multiple population-based and pedigree-based genotype imputation methods exist. Genotype imputation, as a general example of imputation,[21] typically infers missing data by borrowing information from correlated observations. Imputation in population-based samples leverages information from the correlation among dense markers due to linkage disequilibrium (LD) observed in outside reference samples of unrelated individuals.[22–27] In contrast, imputation in pedigrees uses the correlation of genotypes among relatives derived from sharing of genomic segments identical by descent (IBD) within pedigrees. For small pedigrees, Burdick and colleagues developed an imputation

[1]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; [2]Department of Statistics, University of Washington, Seattle, WA 98195, USA; [3]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA 98195, USA; [4]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
*Correspondence: wijsman@u.washington.edu
http://dx.doi.org/10.1016/j.ajhg.2013.02.011.

method and applied it to imputation of dense genotypes.[28,29] Genotype imputation was demonstrated in the most recent generation of a pedigree by using dense marker data available in the oldest generations, with sparse markers available in all generations. Rule-based long-range phasing methods, which detect long strings of nonconflicting homozygous genotypes to identify shared haplotypes between relatives, have also been developed.[30,31]

These existing genotype imputation methods have major limitations for use in large pedigrees. Population-based methods cannot impute genotypes on relatives who are completely unobserved for marker genotypes when used in the context of ignored pedigree structure. In addition, although high imputation accuracy can often be achieved when sufficient numbers of subjects in a reference panel are available,[32] imputation of rare variants is particularly difficult.[33,34] Existing pedigree-based methods also have major limitations. Burdick's method cannot handle large pedigrees with many markers because of computational constraints. Although existing rule-based methods[30,31] can handle large pedigrees, they are ad hoc, require high-quality dense genotype data on subjects for whom we want to impute data, and do not account for recombination events.

Here, we present a computationally efficient approach for imputing dense genotypes in large pedigrees, which is implemented in the program GIGI (Genotype Imputation Given Inheritance). Our MCMC-based approach uses a sparse set of markers typed on most subjects plus dense markers typed on a few subjects. By analyzing both simulated and real data, we demonstrate that our approach can impute many alleles accurately on many subjects in large pedigrees, even including some relatives who are completely unobserved for genotypes. In addition, we evaluate parameters that affect imputation quality, and we demonstrate that GIGI is substantially more accurate for imputing rare alleles in a large pedigree than the state-of-the-art population-based approach in BEAGLE.[26,32]

Material and Methods

Overview

To clarify what follows, we first present some terminology. We define a framework marker panel as a relatively sparse set of markers that are used jointly for inferring inheritance states[35] along a chromosome of interest in a particular pedigree. The framework marker panel could consist of markers of any types, including short tandem repeats (STRs) and SNPs, or a combination of these or other types of markers. These framework markers are assumed to be in linkage equilibrium (LE) and ideally should be genotyped on a large fraction of subjects in a pedigree. We define a dense marker panel as additional markers with missing genotypes on some subjects that we want to impute. For example, these dense markers could be genotypes obtained from sequence data or from a dense SNP panel and could be typed on fewer and even different subjects than the framework panel. Our goal is to impute genotypes of dense markers on the unobserved subjects. Here the imputation relies on correlation resulting from inheritance in the pedigree. The inheritance of shared segments of chromosome is represented by inheritance vectors (IVs).[35] The imputation approach consists of four steps. (1) Sample IVs at the positions of the framework markers conditional on the observed genotypes at the framework markers. (2) Sample IVs at the positions of the dense markers conditional on the IVs sampled at the positions of the framework markers and the meiotic map. (3) Estimate the probability distribution for each unobserved genotype at the dense marker positions conditional on all observed dense genotypes, known or estimated allele frequencies for the dense markers, and position-specific IVs corresponding to the dense markers. (4) Call genotypes via the estimated probabilities and user-specified thresholds.

Details

Sampling IVs at the Positions of Framework Markers

We use the program gl_auto[18] in MORGAN to infer IVs at the positions of framework markers. This program infers IVs by using observed genotypes of the framework markers (G_F^{ob}) and population allele frequencies, in a manner similar to other pedigree-based linkage analysis methods.[13,36–38] The program samples IVs from probabilities obtained by either exact or MCMC-based computation.[13,17,39] In the exact sampling approach, gl_auto uses the Lander-Green algorithm[35] to compute the multipoint likelihood $P(G_F^{ob})$. After computing the likelihood, gl_auto performs Monte Carlo sampling of IVs[40] with the Baum-Welch algorithm.[41] However, the Lander-Green algorithm restricts computation to use in only small pedigrees. To handle large pedigrees, gl_auto uses a hybrid MCMC sampler based on both the Elston-Stewart[42] and Lander-Green algorithms, with components of this likelihood stored for subsequent efficient Monte Carlo sampling of IVs.[17] Evaluation of an older version of this hybrid sampler suggests that it outperforms SimWalk2,[15] a widely used MCMC-based linkage analysis program, in terms of accuracy and computational speed for the use of dense markers typed on pedigrees.[16] Results have also shown that the current sampler in MORGAN performs even better than this older sampler.[17] We sample a set of IVs at the positions of the framework markers.

Sampling IVs at the Positions of Dense Markers

Let $S_{\cdot v}$ denote the inheritance vector at the position of a dense marker v. $S_{\cdot v} = (S_{1v}, \ldots, S_{mv})$ is composed of a collection of segregation indicators S_{iv}, for $i = 1, \ldots, m$, in a pedigree with m meiosis.[35,36] The dense marker v is flanked on the left by a framework marker j and on the right by a framework marker $j + 1$. Analogously, let $S_{\cdot j}$ denote the IVs at position j and S_{ij} denotes its segregation indicator for $i = 1, \ldots, m$.

We seek to sample from $P(S_{\cdot v} = \cdot | G_F^{ob})$ the probability distribution of IVs at the position of dense marker v conditional on the observed framework markers. Because IVs are highly correlated across nearby positions, we infer IVs sampled at the position of dense marker v by using the IVs sampled at the positions of the framework markers. Because the use of a moderate number of framework markers generally extracts much of the information about the IVs in a pedigree,[16,43] IVs sampled at dense positions should already be well inferred when they are conditioned on IVs sampled at framework positions. Given the Haldane Map function,[44] IVs sampled at the positions of the closest flanking framework markers contain all information for the inference of IVs at the position of dense marker v,[45] i.e., $P(S_{\cdot v} = s_{\cdot v} | S_{\cdot F} = s_{\cdot F}) = P(S_{\cdot v} = s_{\cdot v} | S_{\cdot j} = s_{\cdot j}, S_{\cdot j+1} = s_{\cdot j+1})$, where $S_{\cdot F}$ is the IVs at framework

markers and $s_{\cdot v}$ and $s_{\cdot F}$ are configurations of IVs. Conditional on the IVs of nearby flanking framework positions that are highly correlated with the yet-to-be-sampled IVs at dense positions, a set of IVs is sampled marginally for each dense marker v. Since the m meioses in a pedigree are independent, sampling $S_{\cdot v}$ corresponds to sampling each S_{iv} independently. Under the Haldane map function, $P(S_{iv}=s_{iv}, S_{i\,j+1}=s_{i\,j+1}|S_{ij}=s_{ij}) = P(S_{iv}=s_{iv}|S_{ij}=s_{ij})P(S_{i\,j+1}=s_{i\,j+1}|S_{iv}=s_{iv})$, where s_{iv} specifies whether the chromosome is inherited maternally or paternally at position v. Because $P(S_{iv}=s_{iv}|S_{ij}=s_{ij}, S_{i\,j+1}=s_{i\,j+1}) = P(S_{iv}=s_{iv}, S_{i\,j+1}=s_{i\,j+1}|S_{ij}=s_{ij})/P(S_{i\,j+1}=s_{i\,j+1}|S_{ij}=s_{ij})$, it is straightforward to sample S_{iv} conditional on the IVs at the flanking markers.

Each term is calculated easily given the Haldane map function. At position v, one IV ($S^k_{\cdot v}$) is sampled from the jointly sampled IVs obtained at positions j and $j+1$. We repeat this process to sample a total of n such IVs. This set of $S^k_{\cdot v}$, for $k = 1,\ldots, n$, provides an estimate of the probability $P(S_{\cdot v}=s_{\cdot v}|G^{ob}_F)$, because

$$\begin{aligned} P\left(S_{\cdot v}=s_{\cdot v}\,|\,G^{ob}_F\right) &= \sum_{s_{\cdot F}} P\left(S_{\cdot F}=s_{\cdot F}\,|\,G^{ob}_F\right) P\left(S_{\cdot v}=s_{\cdot v}\,|\,S_{\cdot F}=s_{\cdot F}, G^{ob}_F\right) \\ &= \sum_{s_{\cdot F}} P\left(S_{\cdot F}=s_{\cdot F}\,|\,G^{ob}_F\right) P\left(S_{\cdot v}=s_{\cdot v}\,|\,S_{\cdot F}=s_{\cdot F}\right) \\ &= \sum_{s_{\cdot F}} P\left(S_{\cdot F}=s_{\cdot F}\,|\,G^{ob}_F\right) P\left(S_{\cdot v}=s_{\cdot v}\,|\,S_{\cdot j}=s_i, S_{\cdot j+1}=s_{\cdot j+1}\right). \end{aligned}$$

Then $\widehat{P}(S_{\cdot v}=s_{\cdot v}|G^{ob}_F) = \frac{1}{n}\sum_{k=1}^{n} P(S_{\cdot v}=S^k_{\cdot v}|S_{\cdot j}=S^k_{\cdot j}, S_{\cdot j+1}=S^k_{\cdot j+1})$ is the required estimate, because $S_{\cdot F}$ is realized from $P(S_{\cdot F}=s_{\cdot F}|G^{ob}_F)$.

Imputing Dense Genotypes

We estimate the probability distribution of the missing genotype of subject i of dense marker v (G_{iv}), conditional on the observed genotypes of all framework markers (G^{ob}_F), the observed genotypes (G^{ob}_v) of dense marker v, and the allele frequencies of dense marker v. For each genotype configuration g, our estimator is based on the calculation:

$$\begin{aligned} P\left(G_{iv}=g\,|\,G^{ob}_F, G^{ob}_v\right) &= \sum_s P\left(G_{iv}=g\,|\,S_{\cdot v}=s, G^{ob}_F, G^{ob}_v\right) P\left(S_{\cdot v}=s\,|G^{ob}_F, G^{ob}_v\right) \\ &\cong \sum_s P\left(G_{iv}=g\,|\,S_{\cdot v}=s, G^{ob}_F, G^{ob}_v\right) P\left(S_{\cdot v}=s\,|\,G^{ob}_F\right) \end{aligned} \quad \text{(Equation 1)}$$

$$\cong \sum_s P\left(G_{iv}=g\,|\,S_{\cdot v}=s, G^{ob}_v\right) P\left(S_{\cdot v}=s\,|\,G^{ob}_F\right). \quad \text{(Equation 2)}$$

Equation 1 is an exact equality if dense marker v is one of the framework markers: i.e., $G^{ob}_v \subseteq G^{ob}_F$. In general, Equation 1 is a good approximation when $P(S_{\cdot v}=s|G^{ob}_F, G^{ob}_v) \cong P(S_{\cdot v}=s|G^{ob}_F)$, which says that the inference of IVs at the position of dense marker v is not influenced much by the addition of the genotypes of dense marker v, given that we already observe the genotypes of the framework markers. Equation 2 is a good approximation when $P(G_{iv}=g|S_{\cdot v}=s, G^{ob}_F, G^{ob}_v) \cong P(G_{iv}=g|S_{\cdot v}=s, G^{ob}_v)$. Indeed, this approximation is an exact equality if the framework markers are in linkage equilibrium with dense marker v, as is assumed in the Lander-Green algorithm.[35] See Appendix A for further discussion.

$P(G_{iv}=g|S_{\cdot v}=s, G^{ob}_v)$ in Equation 2 is calculated by

$$P\left(G_{iv}=g\,|\,S_{\cdot v}=s, G^{ob}_v\right) = \frac{P\left(G_{iv}=g, G^{ob}_v\,|\,S_{\cdot v}=s\right)}{\sum_k P\left(G_{iv}=k, G^{ob}_v\,|\,S_{\cdot v}=s\right)}. \quad \text{(Equation 3)}$$

Each term in Equation 3 can be computed efficiently.[14,36] The second term of Equation 2 is estimated by the sampled IVs at position v. Because IVs are sampled conditionally on G^{ob}_F, Equation 4 provides a Monte Carlo estimator for imputation of G_{iv}:

$$\widehat{P}\left(G_{iv}=g\,|\,G^{ob}_F, G^{ob}_v\right) = \frac{1}{n}\sum_{k=1}^{n} P\left(G_{iv}=g\,|\,S^k_{\cdot v}, G^{ob}_v\right), \quad \text{(Equation 4)}$$

where $S^k_{\cdot v}$ is the IVs sampled at iteration k, for $k = 1,\ldots, n$. Equation 4 assumes that all $S^k_{\cdot v}$ are consistent with the observed genotypes of dense marker v. For practical purposes, we propose a modified estimator, Equation 5, that is based only on the sampled IVs that are consistent with the observed genotypes of marker v:

$$\widehat{P}\left(G_{iv}=g\,|\,G^{ob}_F, G^{ob}_v\right) = \frac{1}{n^*}\sum_{k=1}^{n} P\left(G_{iv}=g\,|\,S^k_{\cdot v}, G^{ob}_v\right), \quad \text{(Equation 5)}$$

where $n^* = \sum_{k=1}^{n} I(P(G^{ob}_v|S^k_{\cdot v}) > 0)$. $I(P(G^{ob}_v|S^k_{\cdot v}) > 0)$ is an indicator that $S^k_{\cdot v}$ is consistent with G^{ob}_v. Thus, n^* is the number of sampled IVs that are consistent with the observed genotypes at dense marker v. A more thorough discussion of the estimators is presented in Appendix B.

Calling Genotypes

Although we can leave the imputed results as estimated probabilities, we can also call genotypes. By using a confidence-based genotype-calling approach, we call both alleles if $\widehat{P}(G_{iv}=g|G^{ob}_F, G^{ob}_v) > t_1$, where t_1 is a user-defined threshold. In allele calling, we first use genotype calling. If we cannot call the complete genotype, we call one of the two alleles if $\widehat{P}(G_{iv}=a/.|G^{ob}_F, G^{ob}_v) > t_2$, where $a/.$ denotes that the genotype contains an a allele. Although this second threshold t_2 can be arbitrary, we set $t_2 = t_1 + (1-t_1/2)$. A reason for this choice is that for a diallelic marker, the algorithm will select the more likely allele when the estimated probability of the heterozygous configuration is equal to t_1. Besides the confidence-based genotype-calling approach, we can alternatively call the most-probable genotype. In this approach, a genotype call is always made.

Evaluating Imputation Performance

Measuring Quality

We used three metrics to evaluate imputation quality. Call rate measures the percentage of alleles called, accuracy measures the percentage of alleles called correctly among the alleles called, and consistency measures the percentage of IVs that are consistent with the observed genotypes at a marker locus. In real data, these metrics were calculated by averaging over all marker loci and across all subjects. In simulated data, these metrics were further averaged over all simulation replicates. In addition, we summarized the call rate by subject.

Simulated Data

We simulated data on a 5-generation pedigree of 52 subjects (Figure 1A). Although this pedigree is beyond the limit of exact computational methods for multipoint computation, the use of gl_auto's MCMC option enabled computation on this large pedigree. We used simulated descent patterns from a previous study[16] to obtain genotypes in nonfounders after simulating genotypes in founders. We analyzed several replicates with different descent patterns. Results from the first ten replicates gave consistent interpretation and were therefore deemed a sufficient sample size.

We simulated both framework and dense markers on a chromosome of 100 cM. We simulated two types of framework markers: diallelic and 4-allelic. The diallelic markers with uniform allele frequencies spaced uniformly at one marker per 0.5 cM represented a SNP linkage panel. The 4-allelic markers with uniform allele frequencies spaced uniformly at one marker per 4 cM represented a STR marker panel. The 4-allelic markers represent what might exist in a region where there has been some follow-up genotyping. To

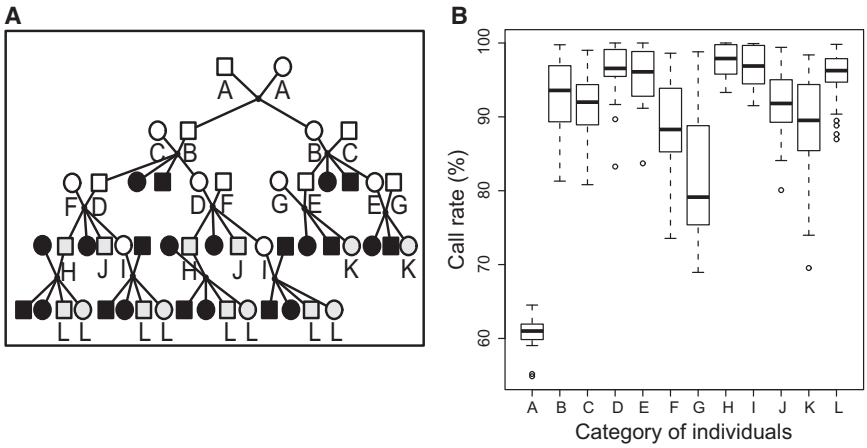

Figure 1. Call Rate across Classes of Subjects in the Simulated Pedigree of 52 Subjects

(A) Different designs of subjects observed for genotypes are indicated by different shading schemes: all subjects (shaded or not shaded); many subjects (any shaded); and few subjects (black shaded). Classes of subjects are indicated by letters.

(B) We used the $S_f M_m^8$ framework panel from simulation. Classes of subjects are as in (A).

examine a density of STR markers that is more commonly available in legacy samples in initial genome scans, we also thinned the STRs to a density of one marker per 8 cM. To test the effect of framework marker density and observed data patterns, we also thinned the framework SNP markers and varied the number of observed subjects, as described below. In addition to the framework markers, we simulated 25,000 uniformly spaced diallelic dense markers at a density of one marker per 0.004 cM with population allele frequencies simulated from the uniform [0, 1] distribution. These markers approximate markers from SNP-chip or variants that might be available from high-throughput sequencing. After generating the complete marker data set by simulating the founder alleles and gene-dropping through the descent patterns as described above, we retained dense SNPs in only 22 subjects, as indicated by the unlabeled subjects in Figure 1A. We imputed genotypes on the 30 labeled subjects.

Analysis of Simulated Data

We carried out our analysis under seven different designs that were organized into three different types of framework marker panels (Table S1 available online). The first set of designs consisted of only SNPs (S), where all (52) subjects (S_a), many (36) subjects (S_m), or few (22) subjects (S_f) were observed for genotypes in the framework panels (Figure 1A). The design S_a is unrealistic but provides a benchmark for optimal inference of IVs, whereas other designs capture more realistic situations where only subjects from the more recent generations are available. The second set of designs consisted of only STRs (M) in the framework panels typed on many subjects at a spacing of either one STR per 4 cM (M_m^4) or one STR per 8 cM (M_m^8). Finally, the third set of designs consisted of both SNPs and STRs in the framework panel, where few (22) subjects typed for SNPs were combined with many (36) subjects typed for STRs at a spacing of either one STR per 4 cM ($S_f M_m^4$) or one STR per 8 cM ($S_f M_m^8$). These hybrid panels capture the situation where STR markers observed on many subjects from older studies are combined with newly collected dense markers observed on fewer subjects.

We evaluated our approach's ability to impute rare alleles. By stratifying on SNPs with minor allele frequencies (MAFs) less than either 0.05 or 0.01, we computed call rate and accuracy of imputation specifically for the heterozygotes. In addition, we computed call rate and accuracy of imputation of doubletons. A doubleton is defined here as the closest pair of SNPs in an individual, such that the two SNPs were within 1 cM and the MAF < 0.05. This analysis was restricted to SNPs with <0.05 because of the low number of doubletons at lower MAF under our simulation conditions. We used the default ($S_f M_m^8$) framework panel in this analysis.

We also evaluated four other conditions that might affect the imputation quality. First, we varied the density of the framework markers, because the density of markers might affect the inference of IVs.[16] For this purpose, we thinned the original 0.5 cM spaced framework SNPs to obtain 1 and 2 cM spaced SNPs. Second, we varied the call threshold on accuracy and call rate by varying call thresholds ranging from $t_1 = 0.5$ to 0.999999 (~1) while fixing t_2 midway between t_1 and 1. We refer to the case where the call threshold was ~1 as practically deterministic. Unless otherwise stated, the default $t_1 = 0.8$ and $t_2 = 0.9$ call thresholds were used. Third, we investigated the effect of MAF on imputation accuracy. By using the true MAF used to simulate the data, we evaluated accuracy by binning markers into MAF bins of size 0.01. Finally, we investigated the effect of the distance of dense markers from the closest framework markers under a STR-only framework panel (M_m^8), again by binning dense markers by the distance from their closest framework markers.

Analysis of Real Data

Analysis of a real data set allowed evaluation of our approach in data that contains complexities not captured well in simulated data. These include, but are not limited to, variable marker informativeness, potential misspecification of the genetic map and allele frequencies, undetected genotyping errors, and LD between markers. We used a 5-generation pedigree of 95 members,[46,47] with some branches from the original pedigree omitted because they contained neither sparse nor dense marker data. Of the subjects retained, the average sibship sizes in the second, third, fourth, and fifth generations were 3, 3.8, 2, and 1.4. This pedigree also included one large sibship of size 9. We focused on imputing SNPs in a ~50 cM interval defined as a region of interest for a cardiovascular trait.[47] Our original data set contained 60 subjects observed for 323 SNPs and 64 subjects observed for 21 STRs in the lowest four generations. Most of these SNPs were tightly linked with a few adjacent SNPs.

We performed an analysis that resembled the situation where we had legacy genome scan marker data and just collected new denser markers on a few subjects (Table S2). We retained SNP genotypes on 13 subjects scattered throughout different branches of the pedigree, and we masked SNP genotypes on the other 47 subjects. To infer IVs, we used a framework panel composed of 21 STRs typed on 64 subjects and 29 SNPs typed on 13 subjects. These 29 SNPs were chosen because they have high MAFs, so incorporating them into the framework panel should improve the inference of IVs.[48] The dense genotypes consisted of 294 SNPs typed on the same 13 subjects typed for SNPs in the framework panel. Finally, we imputed missing genotypes on those 294 SNP markers on all other 82 subjects, including both subjects with masked genotypes

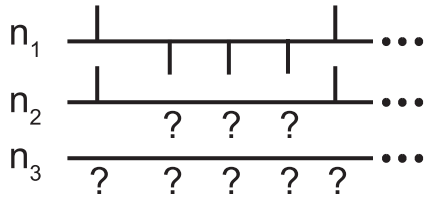

Figure 2. Different Subjects Have Different Levels of Genotypes
Some subjects (n_1 of them) had observed genotypes for both framework markers (top ticks) and dense markers (bottom ticks); n_2 of the subjects had observed genotypes for framework markers but had missing genotypes (symbol ?) for dense markers; n_3 of the subjects were completely unobserved for both framework and dense markers.

(47) and subjects with no observed SNP genotypes (35). The masked genotypes from the 47 subjects allowed us to evaluate the accuracy of imputation. The meiotic map for the SNP markers was obtained by linear interpolating from Haldane map position of STR markers[49] with sequence positions of both STRs and SNPs. The population allele frequencies of the SNP markers were estimated by Loki v.246[13] with this pedigree along with three other large pedigrees with similar European ethnicity.

Comparison with BEAGLE

We compared GIGI to BEAGLE, a state-of-the-art population-based genotype imputation approach.[26,32] BEAGLE can be used by ignoring the pedigree structure. Because BEAGLE uses information from population-level LD while not incorporating the pedigree structure, we sought to understand how the use of different sources of information could affect genotype imputation in a pedigree. We used the same real-data pedigree and region of interest. We computed the accuracy and call rate over all genotypes, as well as separately, over rare SNPs. Here, we defined a rare allele to be the minor allele of a SNP with minor allele frequency less than 0.05. The accuracy of imputing rare alleles is especially important because the primary motivation for using such pedigrees could be to identify rare variants that affect disease risk or phenotypic variation.

We evaluated GIGI and BEAGLE v.3.3 (Table S3) with the real data set. Because BEAGLE does not output genotype probabilities if we use both STRs and SNPs, we modified the previous analysis to perform a SNP-only analysis (Figure 2: $n_1 = 13$, $n_2 = 47$, $n_3 = 35$). Similar to the previous analysis, the same 13 subjects were given the complete genotype data. In BEAGLE's terminology, these genotype data were the outside reference samples used to infer haplotypes of dense markers. In the other 47 subjects, we kept genotypes of 35 approximately evenly spaced SNPs and masked genotypes of the remaining 288 SNPs. Under this setup (design FW), we compared GIGI and BEAGLE based on the imputed results of the masked SNPs on the 47 subjects.

We also evaluated BEAGLE under other designs (Table S3). When markers are tightly linked and when sample size of the outside reference is large, BEAGLE is more likely to perform well. In Design L1, we supplied BEAGLE with more markers by using a leave-one-out analysis where we imputed one SNP at a time, based on all other SNPs. In this leave-one-out analysis, genotypes of each SNP in the 47 subjects were omitted sequentially and were subsequently imputed back. In design FWO, we added genotypes from 202 subjects to the outside reference panel (Figure 2: $n_1 = 13 + 202$, $n_2 = 47$, $n_3 = 35$). These 202 subjects were derived from three other pedigrees of similar ethnic background and who were typed on the same SNP platform. In design L1O, we supplied both dense markers and additional outside reference samples in a leave-one-out analysis. In each design, we called genotypes by using both the most probable genotype-calling and the threshold-based approaches. To evaluate the performance of imputation of rare alleles, we performed a subgroup analysis in heterozyogous genotypes, each containing a rare allele.

Results

Simulated Data

Data Patterns and Framework Marker Panels

High call rates were obtained in most subjects from multiple branches in the simulated large pedigree (Figure 1B). In the $S_f M_m^8$ framework panel, subjects descended from the central pedigree, who tended to share more alleles with relatives, had higher call rates than married-in spouses (96.1% versus 88.7% for group D versus F; 95.4% versus 81.2% for E versus G). In addition, high call rates were observed in subjects from the bottom generation who had multiple relatives typed for dense markers but were not themselves typed for dense markers (95.8% in group L). Also, high call rates were observed even in some subjects who were not typed for either sparse framework markers or dense markers (>95% in groups D, E, and I).

The call rate depended much more on the number of subjects typed than on the density of framework markers. Among different framework panels considered (Table 1), the design where only a few subjects were typed for framework SNPs (S_f) gave the lowest call rate (78.8%). Regardless of the type of panel, having more subjects typed for the framework panel increased the call rate to 89.1%–92.1% for S_m and all STR panels. Genotyping the majority of subjects for the framework panel (92.1% for S_m) is nearly as beneficial as genotyping all subjects (93.5% for S_a). In contrast, altering marker density did not strongly influence the call rate. Doubling the density of STR markers increased the call rate only slightly (89.1% versus 90.7% for M_m^8 versus M_m^4). Similarly, increasing density by adding SNP markers on a few subjects to an existing STR panel only slightly improved the call rate, when the STR panel was either sparse (89.1% versus 90.9% for M_m^8 versus $S_f M_m^8$) or dense (90.7% versus 91.5% for M_m^4 versus $S_f M_m^4$).

When we called a genotype, it was highly accurate across all conditions considered (Table 1). Among SNP-only panels, accuracy was the lowest in the S_f design (98.7%). Typing more subjects for SNPs (S_m) increased the accuracy only slightly (99.2%). Doubling the density of STR markers also only slightly increased the accuracy (98.6% versus 99.2% for M_m^8 versus M_m^4). In addition, the 4 cM spaced STR panel typed on many subjects (M_m^4) was similar in accuracy to the denser but diallelic SNP panel typed on many subjects (S_m). Unlike call rate, accuracy did not improve from increasing density by adding SNP markers on a few subjects to an existing STR panel, whether the STR panel was sparse (98.6% for M_m^8 and $S_f M_m^8$) or dense (99.2% for M_m^4 and $S_f M_m^4$).

Both the call rate and accuracy increased only slightly when the density of the SNP framework panel increased

Table 1. The Effect of Different Framework Panels on Imputation Quality and Different Designs Evaluated with Simulated Data

Panel: Quality Metric (%)	SNPs Only			STRs Only		SNPs and STRs	
	S_a	S_m	S_f	M_m^4	M_m^8	$S_f M_m^4$	$S_f M_m^8$
Called	93.5 (91.7, 95.8)[a]	92.1 (90.1, 94.7)	78.8 (77.8, 79.6)	90.7 (87.2, 93.1)	89.1 (86.3, 90.2)	91.5 (89.7, 93.1)	90.9 (89.5, 92.8)
Accuracy	99.6 (99.4, 99.7)	99.2 (97.8, 99.6)	98.7 (97.0, 99.5)	99.2 (98.9, 99.4)	98.6 (98.1, 99.0)	99.2 (98.6, 99.6)	98.6 (98.0, 99.4)
Consistency	93.6	92.1	90.2	71.0	54.2	92.4	91.6

[a]Range across ten runs: (low, high).

(Table 2). Among SNP-only panels typed on many subjects, the call rate increased slightly when doubling the SNP density from one marker per 2 cM (90.5%) to one marker per 1 cM (91.7%) and again when doubling density from one marker per 1 cM to one marker per 0.5 cM (92.1%). Similarly, although accuracy increased slightly when doubling SNP density from one marker per 2 cM (98.9%) to one marker per 1 cM (99.2%), it did not further increase when doubling from one marker per 1 cM to one marker per 0.5 cM. Both gains, however, were modest, because call rate and accuracy were high even at the 2 cM density. Overall, these marginal increases in both call rate and accuracy were consistent with the previous results from increasing the STR marker density (Table 1).

Unlike call rate and accuracy, consistency depended strongly on the density of framework markers (Table 2). All panels that contained the 0.5 cM spaced SNPs had high consistency (>90.2%) (Table 1). However, consistency decreased as the density of framework markers decreased. As the marker spacing in SNP-only panels decreased from 0.5 to 2 cM, consistency decreased from 92.1% to 69.5% (Table 2). Similarly, as the marker spacing in STR-only panels decreased from 4 to 8 cM, consistency also decreased from 71.0% to 54.2% (Table 1). Even though the 8 cM spaced STR panel (M_m^8) had the lowest consistency (54.2%), call rate and accuracy were still high.

GIGI called rare alleles with high accuracy (Table S4). For SNPs with MAF < 0.05, under the default call threshold GIGI imputed 88.1% of the heterozygous genotypes with a high accuracy of 89.6%. Lowering the MAF by defining rare alleles as MAF < 0.01 slightly increased the call rate and accuracy, yielding 88.5% of the heterozygous genotypes called with an accuracy of 90.0%. In addition, GIGI called doubletons at MAF < 0.05 with imputation quality similar to that of calling single heterozygotes, achieving a call rate of 83.2% and accuracy of 86.6%. With the "practically deterministic threshold," GIGI called rare heterozygous, either as singletons or doubletons (as defined in the Material and Methods), with an accuracy of more than 99.7% for SNPs with MAF < 0.01 or 0.05, although the call rate diminished to about 30% because at such a stringent threshold, often it was not possible to call both alleles in a genotype.

Table 2. The Effect of Different Marker Density on Imputation Quality Evaluated with the S_m Panel in Simulated Data

Quality Metric (%)	Spacing of SNPs in Framework Panel (cM) Typed on S_m		
	0.5	1	2
Called	92.1 (90.1, 94.7)[a]	91.7 (90.3, 93.8)	90.5 (88.6, 93.5)
Accuracy	99.2 (97.8, 99.6)	99.2 (97.8, 99.5)	98.9 (98.1, 99,4)
Consistency	92.1	84.5	69.5

[a]Range across ten runs: (low, high).

Other Parameters

Call thresholds affected both call rate and accuracy but in different directions. The use of a more stringent call threshold decreased the call rate (Figure 3A). For instance, under the design $S_f M_m^8$ (Figure 3A), the call rate decreased from 95.8% to 81.0% as the call threshold increased from $t_1 = 0.6$ to $t_1 = 0.99$. In contrast, the use of a more stringent threshold increased the accuracy: accuracy increased from 97.8% to 99.9% as the call threshold increased from $t_1 = 0.6$ to $t_1 = 0.99$. However, the change in accuracy was less dramatic than that of the call rate, because accuracy was already high at a liberal call threshold (97.8% for $t_1 = 0.6$). In this particular simulation, a reasonable balance between call rate and accuracy was achieved at the call threshold of $t_1 = 0.8$.

The MAF of dense markers also affected quality metrics. At the default $t_1 = 0.8$ call threshold, the call rate decreased as the MAF increased (Figure 4). Also, there was a sudden drop in the call rate at MAF = 0.2 (Figure 4A). Besides call rate, accuracy decreased as the MAF increased from 0 to 0.2 but was approximately constant near 99.1% for frequencies above 0.2. We also called alleles by using the "practically deterministic" threshold (Figure 4B). Similar to the call rate with $t_1 = 0.8$, the call rate with the practically deterministic threshold decreased when the MAF increased. In contrast, the imputation accuracy was almost perfect regardless of the MAF.

The confidence-based call threshold directly determines when the call algorithm relies heavily on population allele frequencies. Because allele frequencies had no impact on calling genotypes when we used the practically deterministic threshold (Figure 4B), calls were made only when forced by very tight constraints between the sampled IVs and observed genotypes. As we relaxed the call threshold, additional calls were made with input from population

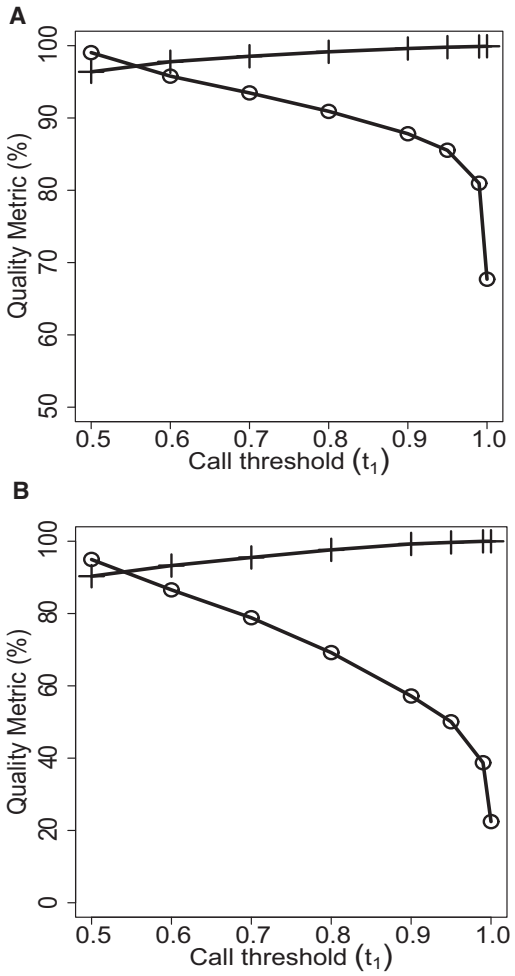

Figure 3. Call Rate and Accuracy as a Function of Call Threshold in Simulated and Real Pedigree
Call rate is indicated by circle and accuracy is indicated by a plus sign.
(A) Analysis of simulated data: we used the $S_f M_m^8$ framework panel.
(B) Analysis of real data: see text for the description of the analysis.

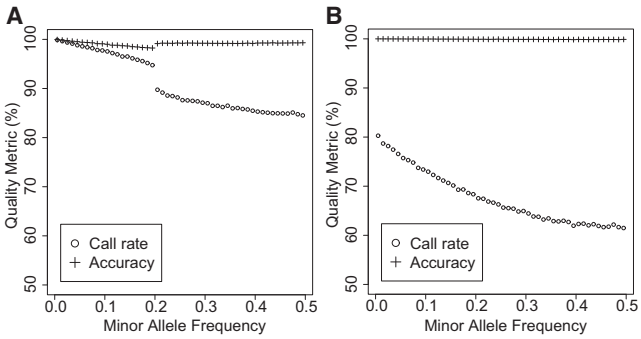

Figure 4. Call Rate and Accuracy as a Function of the True Minor Allele Frequency
We used the $S_f M_m^8$ framework panel from simulation. Different call thresholds were used: (A) $t_1 = 0.8$, $t_2 = 0.9$ and (B) practically deterministic ($t_1 \cong 1.0$).

allele frequencies. These additional calls include alleles transmitted from unobserved founder chromosomes that can be called only by using population allele frequencies, so the use of a call threshold that exceeds 0.5 will call those alleles as the major alleles. Hence, when we used a call threshold of $t_1 = 0.8$, unobserved alleles from SNPs with MAF < 0.2 were called as the major allele whereas SNPs with MAF ≥ 0.2 were not called because their major allele frequencies then fell below the call threshold (Figures 4A and S1). As expected, the MAF at which the call rate suddenly changed was proportional to $1 - t_1$ (Figures S2 and S3), even though the calls were made from the same underlying estimated genotype probabilities.

The distance between dense genotypes and their respective nearest framework markers affected consistency much more than the call rate and accuracy (Figure 5). Under the M_m^8 panel, consistency decreased substantially as dense genotypes were farther from the nearest framework markers, e.g., from ~63% to ~45% as the map distance increased from 0 to 4 cM. In contrast, the accuracy and call rate did not greatly drop as the map distance increased, even though decreasing trends were observed.

Real Data

High imputation accuracy and call rate were also obtained on the real data. In the real pedigree, the method called 68% of alleles among the 47 subjects that could be validated and achieved an accuracy of 97.6% via the default threshold (Figure 3B). Relaxation of the call threshold to $t_1 = 0.6$ increased the call rate to 85% but with a decline to 93% in the accuracy. Similar to the simulated data, allele call rate was inversely related to the population allele frequency.

Comparison with BEAGLE

GIGI called rare heterozygous genotypes with substantially higher accuracy than did BEAGLE (Table 3). Under design FW and with the most probable genotype calling, GIGI called these genotypes with an accuracy of 64.4%, in contrast to BEAGLE, which achieved an accuracy of only 4.6%. Increasing the number of dense markers and providing more subjects in the reference panel (designs L1 and L1O) improved BEAGLE's accuracy in calling rare heterozygous genotypes (up to 26.4%), but the accuracy was still much lower than that for GIGI. In addition, GIGI called 46.2% more rare genotypes for relatives who were completely untyped. These genotypes were not called by BEAGLE because BEAGLE did not impute genotypes on completely unobserved subjects. With the confidence-based calling with the default threshold (Table 3), the same trends were observed.

We also compared the overall genotype accuracy and genotype call rate in GIGI and BEAGLE (Table 3). Under the design FW and with the most probable genotype calling, GIGI called genotypes with higher accuracy than BEAGLE (79.7% versus 70.2%). However, the availability of outside reference (FWO) or dense framework marker panel (L1) improved both accuracy and call rate in BEAGLE. In particular, the joint use of dense framework SNPs and outside references (L1O) improved the imputation accuracy of

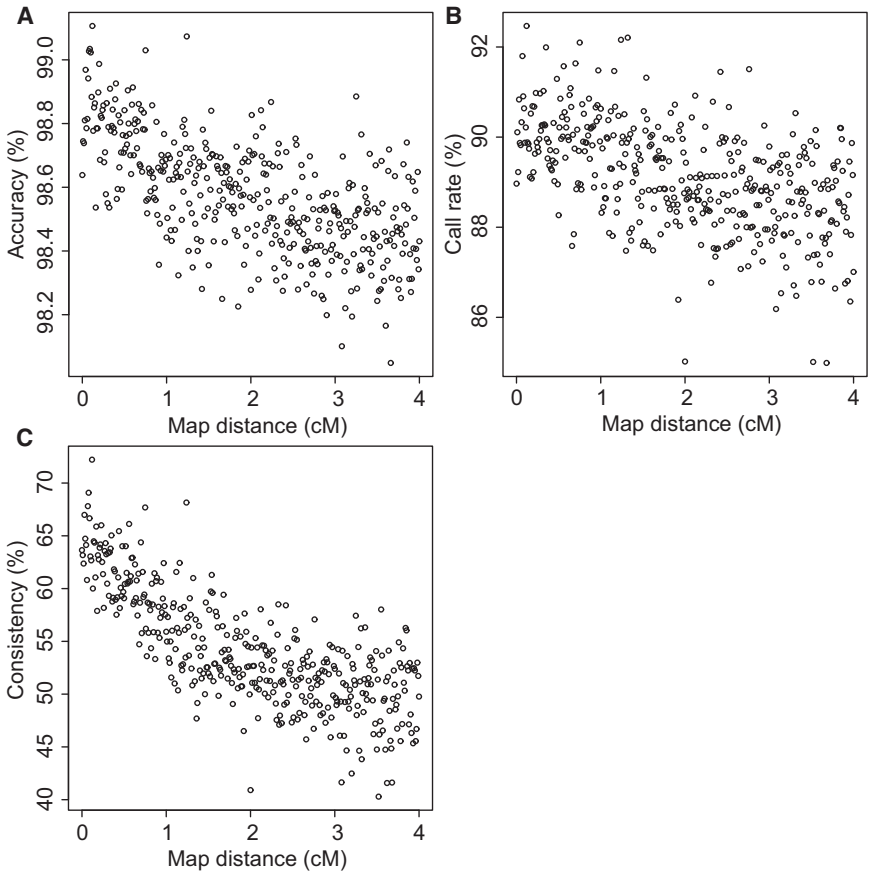

Figure 5. Impact of Distance from the Nearest Framework Marker
We used the M_m^8 framework panel from simulation. We measured the (A) accuracy, (B) call rate, and (C) consistency.

BEAGLE substantially (95.4%). The accuracy of imputing genotypes via GIGI could increase substantially to 96.4% when using the default threshold. However, the tradeoff was that a considerable fraction of genotypes was not called (52.3%).

Discussion

We have introduced an approach for carrying out genotype imputation in potentially large pedigrees. By harnessing existing computational tools that combine exact computation with MCMC-based sampling, our imputation approach can be used in pedigrees that range from small to very large with a range of possible missing data that could include founders. Our results demonstrate that imputed genotypes, including those with rare alleles, can be accurate and obtained with a high call rate. Results from analysis of the simulated data suggest that the number of subjects genotyped for a framework panel has a higher influence on the quality of imputed genotypes than does marker density. Also, results from the analysis of real data show that our genotype imputation approach has higher accuracy in imputing rare alleles than does the population-based approach as implemented in the state-of-the-art BEAGLE.

Our imputation approach can efficiently incorporate new collections of very dense markers, including high-throughput sequencing data, into studies involving existing genome scan data. Results obtained here suggest that an existing framework panel does not need to have high density to infer IVs needed for genotype imputation. This is consistent with both theory[50] and past results[16,43] that show diminishing gains in pedigrees in determining inheritance at a particular position with increasing marker density. Our results suggest that markers from existing genome scans can be leveraged to allow genotype imputation of dense markers on many individuals when these existing marker genotypes are coupled with dense markers typed on some subjects. As demonstrated in our real pedigree analysis, genotypes of some informative dense markers, such as those with high minor allele frequencies typed on multiple individuals, can also be included in the framework marker panel.

Results from the comparison between GIGI and BEAGLE have several implications. First, GIGI provides much higher accuracy in calling rare alleles than does BEAGLE. This is an expected outcome, because explicitly modeling the transmission of genomic segments via the pedigree structure allows rare alleles on such segments to be reliably called. In contrast, BEAGLE is not accurate in calling rare alleles, a result that agrees with other studies.[33,34] When rare alleles are segregating in only one pedigree, increasing the reference sample size is unlikely to help impute such rare alleles. Such pedigree-specific rare or ultra-rare alleles could be typical, especially for causal disease alleles, as is suggested by the very large number of alleles known for some disease loci.[9] Second, conditioning on the pedigree structure together with marker data in the pedigree allows imputation of genotypes in relatives who are unobserved for any genotypes. Both of these considerations are important when the motivation is to identify rare causal variants in pedigrees. Third, BEAGLE excels in imputing common variants when both dense markers and an adequate number of reference samples are present. Under these conditions, BEAGLE imputes common alleles quite accurately and calls them with high confidence; however, the availability of dense framework markers typed on many subjects is not always guaranteed in pedigree studies. A potential future direction would be to integrate the use

Table 3. Comparison between GIGI and BEAGLE under Various Designs of the Real Data

			Program				
			GIGI	BEAGLE			
			Framework	Framework		Leave-One-Out	
Call Choice	Group[a]	Metric[b]	FW[c]	FW[c]	FWO[c]	L1[c]	L1O[c]
Most probable	rare	A	64.4	4.6	4.6	26.4	14.9
	overall	A	79.7	70.2	73.3	82.5	95.4
Threshold	rare	A	69.0	5.3	4.6	15.7	18.2
		C	82.0	86.5	100	57.3	76.4
	overall	A	96.4	88.8	91.5	95.1	98.0
		C	47.7	47.1	43.8	54.1	93.6

[a]Rare = among the 87 heterozygous genotypes that each contain a rare allele.
[b]A, accuracy (%); C, call rate (%). Under the most probable genotype calling approach, C = 100% and therefore has been omitted.
[c]Design. FW, framework. Refer to Table S3 for the description of the designs.

of information from population LD that BEAGLE uses into GIGI to improve the call rate for common variants. We also acknowledge that BEAGLE has potential for improvement in use for genotype imputation. For instance, BEAGLE has an option to detect pairwise IBD segments[51,52] that could potentially be used for this purpose.

The success of genotype imputation requires reliable inference of IVs. Ultimately, this requires the use of informative framework markers. The methods implemented in MORGAN sample IVs from the appropriate conditional distribution,[18] giving accurate results for computations on large pedigrees,[16] and here we showed how imputation quality is affected by the density, type, and number of subjects observed for the framework markers. In practice, users of these imputation methods will need to determine whether they have sufficiently informative framework markers for their own data, possibly by means of an existing information measure.[48] In real data, genotyping errors could also affect the reliability of IVs, so genotyping errors should also first be cleaned. This topic is beyond the scope of this current paper but will be addressed in the future.

Two notable features of our approach allow efficient imputation in large pedigrees. First, our approach separates the inference of IVs from imputation of dense genotypes. One advantage of this strategy is that it circumvents the linkage equilibrium assumption between markers that is needed for application of the Lander-Green algorithm. This is an advantage because the estimated probability of IVs could be incorrect if the linkage equilibrium assumption is violated, which could lead to an increase in false-positive linkage signals.[53,54] Another advantage is computational efficiency, which is achieved because IVs needed to be sampled only once via sparse framework markers. This approach is in contrast to the computationally intensive approach used in MERLIN to incorporate LD through the use of haplotype blocks.[37] Second, our approach uses a state-of-the-art MCMC sampler for analysis of large pedigrees. This allows us to sample IVs to enable analyses that are otherwise computationally intractable on large pedigrees. Computation is relatively rapid, given a sample of IVs: on an Intel L5420 Xeon 2.50 GHz processor, GIGI used 26 min to impute genotypes for 25,000 dense markers, given 1,000 sampled IVs on a 52 member pedigree. In this example, gl_auto required 3.5 hr for 30,000 Monte Carlo iterations for multipoint computation on 213 framework markers that span 100 cM. Therefore, if the 1,000 sampled IVs had not previously been obtained, in this example imputation on the entire largest human chromosome would require ~12 hr of computation, because computation time is approximately linear in the number of markers. Parallelization of computation involving both chromosomes and pedigrees can, of course, keep throughput computation time relatively low.

Genetic analyses can be performed with imputed dense genotypes to identify variants that affect traits. In large pedigrees, it might be fruitful to limit the initial search space to regions where there is positive evidence for linkage with the trait, because only here is there sufficient joint segregation of trait and markers to provide strong confidence in any implicated variants. In these regions, we can then search for causal variants with different approaches. One approach is to perform a measured genotype approach on imputed SNPs, treating them as covariates to adjust out a linkage signal[55] in, for example, a variance component analysis. Another approach is to perform a family-based association test that is suitable for small[56] or large[57] pedigrees. Yet another approach is to perform exploratory analyses via simple filters to correlate disease status with rare variants. Because many types of analysis require genotypes on many subjects, the use of imputed genotypes will enable these types of analyses. In any case, where imputation is used, the most significant results should be checked with direct genotyping, just as is standard for population-based studies.[58]

Our genotype imputation approach, as implemented in GIGI, can facilitate cost-effective genetic analyses,

including but not limited to the identification of rare causal variants in complex traits. Because rare alleles affecting traits can be enriched in pedigrees, the use of large pedigrees is an efficient design to detect signals that are statistically significant. Such pedigrees are emerging as an important class of data used to identify rare causal variants. Statistical analyses of such large pedigrees via imputed dense genotypes could benefit from increased power. Other potential extensions to our approach include inferring haplotypes of dense markers, providing an option for multiple imputation, and providing guidance in selecting which subjects to genotype for dense variants.

Appendix A: Approximation in the Inference of IVs

To achieve computational efficiency, we use the approximation that $P(S_{\cdot v} = s|G_F^{ob}, G_v^{ob}) \cong P(S_{\cdot v} = s|G_F^{ob})$. Making this approximation allows us to sample IVs by using only framework markers. This approximation states that the knowledge of G_v^{ob} does not dramatically influence the inference of $S_{\cdot v}$ given that G_F^{ob} is already observed. This approximation is reasonable because meiotic events in a chromosome do not occur frequently and the use of a moderately sparse set of markers can often extract much of the information of the IVs in a pedigree.[16,43] Also see Appendix B.

Appendix B: Convergence Property of the Estimators

Under the assumption that dense marker v is in linkage equilibrium with the framework markers, the estimator $\tilde{P}(G_{iv}|G_F^{ob}, G_v^{ob}) = \sum_{k=1}^{n} P(G_{iv}|S_{\cdot v}^k, G_v^{ob}) P(G_v^{ob}|S_{\cdot v}^k) / \sum_{k=1}^{n} P(G_v^{ob}|S_{\cdot v}^k)$ converges to $P(G_{iv}|G_F^{ob}, G_v^{ob})$. The modified estimator $\widehat{P}(G_{iv}|G_F^{ob}, G_v^{ob}) = \frac{1}{n^*} \sum_{k=1}^{n} P(G_{iv}|S_{\cdot v}^k, G_v^{ob})$, where $n^* = \sum_{k=1}^{n} I(P(G_v^{ob}|S_{\cdot v}^k) > 0)$, converges to a quantity like $P(G_{iv}|G_F^{ob}, G_v^{ob})$ that replaces the emission probability $P(G_v^{ob}|s)$ by the emission function $I(P(G_v^{ob}|s) > 0)$ at position v. In test data sets, the estimates from the two estimators are often quite similar.

Proof of Convergence

To see this, let $h(G_v^{ob}, s)$ be the *generic* emission function at the position v, conditional on the IVs s at the position v. Most commonly, $h(G_v^{ob}, s) = P(G_v^{ob}|s)$.

We show equality between $P(S_{\cdot v} = s|G_F^{ob}, G_v^{ob})$ and $\tilde{p} = P(S_{\cdot v} = s|G_F^{ob}) h(G_v^{ob}, s) / \sum_w P(S_{\cdot v} = w|G_F^{ob}) h(G_v^{ob}, w)$ when $h(G_v^{ob}, s) = P(G_v^{ob}|s)$. For brevity, we omit the inclusion of allele frequencies into the equation below. We assume that the dense marker v is not in linkage disequilibrium with the framework markers, which are indexed from 1 to M.

Define $\alpha_j(s) = P(G_1^{ob}, \ldots, G_j^{ob}, S_{\cdot j} = s)$ and $\beta_j(s) = P(G_{j+1}^{ob}, \ldots, G_M^{ob}|S_{\cdot j} = s)$.

$$
\begin{aligned}
\tilde{p} &= \frac{P(S_{\cdot v} = s \mid G_F^{ob}) h(G_v^{ob}, s)}{\sum_w P(S_{\cdot v} = w \mid G_F^{ob}) h(G_v^{ob}, w)} = \frac{P(S_{\cdot v} = s, G_F^{ob}) h(G_v^{ob}, s)}{\sum_w P(S_{\cdot v} = w, G_F^{ob}) h(G_v^{ob}, w)} \\
&= \frac{\sum_x \sum_y P(S_{\cdot v} = s \mid S_{\cdot j} = x, S_{\cdot j+1} = y, G_F^{ob}) P(S_{\cdot j} = x, S_{\cdot j+1} = y, G_F^{ob}) h(G_v^{ob}, s)}{\sum_w \sum_x \sum_y P(S_{\cdot v} = w \mid S_{\cdot j} = x, S_{\cdot j+1} = y, G_F^{ob}) P(S_{\cdot j} = x, S_{\cdot j+1} = y, G_F^{ob}) h(G_v^{ob}, w)} \\
&= \frac{\sum_x \sum_y P(S_{\cdot v} = s \mid S_{\cdot j} = x, S_{\cdot j+1} = y) P(S_{\cdot j} = x, S_{\cdot j+1} = y, G_F^{ob}) h(G_v^{ob}, s)}{\sum_w \sum_x \sum_y P(S_{\cdot v} = w \mid S_{\cdot j} = x, S_{\cdot j+1} = y) P(S_{\cdot j} = x, S_{\cdot j+1} = y, G_F^{ob}) h(G_v^{ob}, w)} \\
&= \frac{\sum_x \sum_y P(S_{\cdot v} = s \mid S_{\cdot j} = x, S_{\cdot j+1} = y) \alpha_j(x) P(S_{\cdot j+1} = y \mid S_{\cdot j} = x) P\left(G_{j+1}^{ob} \mid S_{\cdot j+1} = y\right) \beta_{j+1}(y) h(G_v^{ob}, s)}{\sum_w \sum_x \sum_y P(S_{\cdot v} = w \mid S_{\cdot j} = x, S_{\cdot j+1} = y) \alpha_j(x) P(S_{\cdot j+1} = y \mid S_{\cdot j} = x) P\left(G_{j+1}^{ob} \mid S_{\cdot j+1} = y\right) \beta_{j+1}(y) h(G_v^{ob}, w)} \\
&= \frac{\sum_x \sum_y \alpha_j(x) P(S_{\cdot v} = s \mid S_{\cdot j} = x) h(G_v^{ob}, s) P(S_{\cdot j+1} = y \mid S_{\cdot v} = s) P\left(G_{j+1}^{ob} \mid S_{\cdot j+1} = y\right) \beta_{j+1}(y)}{\sum_w \sum_x \sum_y \alpha_j(x) P(S_{\cdot v} = w \mid S_{\cdot j} = x) h(G_v^{ob}, w) P(S_{\cdot j+1} = y \mid S_{\cdot v} = w) P\left(G_{j+1}^{ob} \mid S_{\cdot j+1} = y\right) \beta_{j+1}(y)} = \frac{P(S_{\cdot v} = s, G_F^{ob}, G_v^{ob})}{\sum_w P(S_{\cdot v} = w, G_F^{ob}, G_v^{ob})} \\
&= P(S_{\cdot v} = s \mid G_F^{ob}, G_v^{ob}),
\end{aligned}
$$

which holds because

$$
\begin{aligned}
&P(S_{\cdot v} = s \mid S_{\cdot j} = x, S_{\cdot j+1} = y) P(S_{\cdot j+1} = y \mid S_{\cdot j} = x) \\
&\quad = P(S_{\cdot v} = s \mid S_{\cdot j} = x) P(S_{\cdot j+1} = y \mid S_{\cdot v} = s)
\end{aligned}
$$

by the property of the Haldane Map function.

If $h(G_v^{ob}, s) = P(G_v^{ob}|s)$, this equation becomes the usual calculation of $P(S_{\cdot v} = s|G_F^{ob}, G_v^{ob})$. This result tells us that the proper way to update the probability distribution of $S_{\cdot v}$ after adding genotypes of the dense marker v is to reweight the top and bottom by the emission probability of the dense marker v. Alternatively, we can define

$$h(G_v^{ob}, s) = P(G_v^{ob} \text{ is compatible with } s) = \begin{cases} 1 \text{ if } P(G_v^{ob} \mid s) > 0 \\ 0 \text{ if } P(G_v^{ob} \mid s) = 0 \end{cases} = I(P(G_v^{ob} \mid s) > 0).$$

This emission function uses only the deterministic information of G_v^{ob} and does not depend on the allele frequency of the dense marker v. Therefore, we do not have to worry about making the unrealistic assumption that the tightly linked markers are independent of each other.

Now, we calculate $P(G_{iv}|G_F^{ob}, G_v^{ob})$.

$$P(G_{iv} \mid G_F^{ob}, G_v^{ob}) = \sum_s P(G_{iv} \mid S_{\cdot v} = s, G_F^{ob}, G_v^{ob}) P(S_{\cdot v} = s \mid G_F^{ob}, G_v^{ob}) = \sum_s P(G_{iv} \mid S_{\cdot v} = s, G_v^{ob}) P(S_{\cdot v} = s \mid G_F^{ob}, G_v^{ob}) = \sum_s P(G_{iv} \mid S_{\cdot v} = s, G_v^{ob}) \frac{P(S_{\cdot v} = s \mid G_F^{ob}) P(G_v^{ob} \mid s)}{\sum_w P(S_{\cdot v} = w \mid G_F^{ob}) P(G_v^{ob} \mid w)}.$$

A natural estimator of $P(G_{iv}|G_F^{ob}, G_v^{ob})$ is to plug in $\widehat{P}(S_{\cdot v} = s|G_F^{ob})$ for $P(S_{\cdot v} = s|G_F^{ob})$.

$\widehat{P}(S_{\cdot v} = s|G_F^{ob}) = \frac{1}{n}\sum_{k=1}^{n} I(S_{\cdot v}^k = s|G_F^{ob})$ is an empirical estimator of $P(S_{\cdot v} = s|G_F^{ob})$ using the realized MCMC samples $S_{\cdot v}^1, \ldots, S_{\cdot v}^n$. We propose the estimator

$$\begin{aligned} \tilde{P}(G_{iv} \mid G_F^{ob}, G_v^{ob}) &= \sum_s P(G_{iv} \mid S_{\cdot v} = s, G_v^{ob}) \times \frac{\widehat{P}(S_{\cdot v} = s \mid G_F^{ob}) P(G_v^{ob} \mid s)}{\sum_w \widehat{P}(S_{\cdot v} = w \mid G_F^{ob}) P(G_v^{ob} \mid w)} \\ &= \frac{\sum_s P(G_{iv} \mid S_{\cdot v} = s, G_v^{ob}) \frac{1}{n} \sum_{k=1}^{n} I(S_{\cdot v}^k = s \mid G_F^{ob}) P(G_v^{ob} \mid s)}{\sum_w \frac{1}{n} \sum_{k=1}^{n} I(S_{\cdot v}^k = w \mid G_F^{ob}) P(G_v^{ob} \mid w)} \\ &= \frac{\sum_{k=1}^{n} P(G_{iv} \mid S_{\cdot v}^k, G_v^{ob}) P(G_v^{ob} \mid S_{\cdot v}^k)}{\sum_{k=1}^{n} P(G_v^{ob} \mid S_{\cdot v}^k)} \\ &\xrightarrow{p} P(G_{iv} \mid G_F^{ob}, G_v^{ob}). \end{aligned}$$

Alternatively, we replaced $P(G_v^{ob}|s)$ by $I(P(G_v^{ob}|s) > 0)$. We used this estimator in Equation 5.

$$\begin{aligned} \widehat{P}(G_{iv} \mid G_F^{ob}, G_v^{ob}) &= \frac{\sum_{k=1}^{n} P(G_{iv} \mid S_{\cdot v}^k, G_v^{ob}) I(P(G_v^{ob} \mid S_{\cdot v}^k) > 0)}{\sum_{k=1}^{n} I(P(G_v^{ob} \mid S_{\cdot v}^k) > 0)} \\ &= \frac{1}{n^*} \sum_{k=1}^{n} P(G_{iv} \mid S_{\cdot v}^k, G_v^{ob}) I(P(G_v^{ob} \mid S_{\cdot v}^k) > 0) \\ &= \frac{1}{n^*} \sum_{k=1}^{n} P(G_{iv} \mid S_{\cdot v}^k, G_v^{ob}) \end{aligned}$$

Supplemental Data

Supplemental Data include three figure and four tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

This research was supported by funding from the National Institutes of Health grants R37GM046255, P01HL030086, P50AG05136, R01MH094293, and R01MH092367.

Received: December 10, 2012
Revised: January 15, 2013
Accepted: February 27, 2013
Published: April 4, 2013

Web Resources

The URLs for data presented herein are as follows:

BEAGLE, http://faculty.washington.edu/browning/beagle/beagle.html
GIGI, http://faculty.washington.edu/wijsman/software.shtml
Loki, http://faculty.washington.edu/wijsman/software.shtml
MORGAN, http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml

References

1. Amberger, J., Bocchini, C., and Hamosh, A. (2011). A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). Hum. Mutat. *32*, 564–567.
2. Collins, F.S., Guyer, M.S., and Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. Science *278*, 1580–1581.
3. Manolio, T.A., Brooks, L.D., and Collins, F.S. (2008). A HapMap harvest of insights into the genetics of common disease. J. Clin. Invest. *118*, 1590–1605.
4. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
5. Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science *305*, 869–872.
6. Bodmer, W., and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. Nat. Genet. *40*, 695–701.
7. Gorlov, I.P., Gorlova, O.Y., Frazier, M.L., Spitz, M.R., and Amos, C.I. (2011). Evolutionary evidence of the effect of rare variants on disease etiology. Clin. Genet. *79*, 199–206.
8. Sanna, S., Li, B.S., Mulas, A., Sidore, C., Kang, H.M., Jackson, A.U., Piras, M.G., Usala, G., Maninchedda, G., Sassu, A., et al. (2011). Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. PLoS Genet. *7*, e1002198.
9. Leigh, S.E.A., Foster, A.H., Whittall, R.A., Hubbart, C.S., and Humphries, S.E. (2008). Update and analysis of the University College London low density lipoprotein receptor familial hypercholesterolemia database. Ann. Hum. Genet. *72*, 485–498.
10. Ott, J., Kamatani, Y., and Lathrop, M. (2011). Family-based designs for genome-wide association studies. Nat. Rev. Genet. *12*, 465–474.
11. Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. *11*, 415–425.

12. Elston, R.C. (1997). 1996 William Allan Award Address. Algorithms and inferences: the challenge of multifactorial diseases. Am. J. Hum. Genet. *60*, 255–262.
13. Heath, S.C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. *61*, 748–760.
14. Sobel, E., and Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am. J. Hum. Genet. *58*, 1323–1337.
15. Sobel, E., Papp, J.C., and Lange, K. (2002). Detection and integration of genotyping errors in statistical genetics. Am. J. Hum. Genet. *70*, 496–508.
16. Wijsman, E.M., Rothstein, J.H., and Thompson, E.A. (2006). Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo provides practical approaches for genome scans on general pedigrees. Am. J. Hum. Genet. *79*, 846–858.
17. Tong, L., and Thompson, E. (2008). Multilocus lod scores in large pedigrees: combination of exact and approximate calculations. Hum. Hered. *65*, 142–153.
18. Thompson, E. (2011). The structure of genetic linkage data: from LIPED to 1M SNPs. Hum. Hered. *71*, 86–96.
19. Musunuru, K., Pirruccello, J.P., Do, R., Peloso, G.M., Guiducci, C., Sougnez, C., Garimella, K.V., Fisher, S., Abreu, J., Barry, A.J., et al. (2010). Exome sequencing, *ANGPTL3* mutations, and familial combined hypolipidemia. N. Engl. J. Med. *363*, 2220–2227.
20. Rosenthal, E.A., Ronald, J., Rothstein, J., Rajagopalan, R., Ranchalis, J., Wolfbauer, G., Albers, J.J., Brunzell, J.D., Motulsky, A.G., Rieder, M.J., et al. (2011). Linkage and association of phospholipid transfer protein activity to *LASS4*. J. Lipid Res. *52*, 1837–1846.
21. Little, R.J., and Rubin, D.B. (1987). Statistical Analysis with Missing Data (New York: J. Wiley & Sons).
22. Stephens, M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. *68*, 978–989.
23. Scheet, P., and Stephens, M. (2006). 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*, 629–644.
24. Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. *34*, 816–834.
25. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.
26. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.
27. Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.
28. Burdick, J.T., Chen, W.M., Abecasis, G.R., and Cheung, V.G. (2006). In silico method for inferring genotypes in pedigrees. Nat. Genet. *38*, 1002–1004.
29. Chen, W.M., and Abecasis, G.R. (2006). Estimating the power of variance component linkage analysis in large pedigrees. Genet. Epidemiol. *30*, 471–484.
30. Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., Rafnar, T., et al. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. Nat. Genet. *40*, 1068–1075.
31. Daetwyler, H.D., Wiggans, G.R., Hayes, B.J., Woolliams, J.A., and Goddard, M.E. (2011). Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics *189*, 317–327.
32. Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. *84*, 210–223.
33. Krithika, S., Valladares-Salgado, A., Peralta, J., Escobedo-de La Peña, J., Kumate-Rodríguez, J., Cruz, M., and Parra, E.J. (2012). Evaluation of the imputation performance of the program IMPUTE in an admixed sample from Mexico City using several model designs. BMC Med. Genomics *5*, 12.
34. Li, L., Li, Y., Browning, S.R., Browning, B.L., Slater, A.J., Kong, X.Y., Aponte, J.L., Mooser, V.E., Chissoe, S.L., Whittaker, J.C., et al. (2011). Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. PLoS ONE *6*, e24945.
35. Lander, E.S., and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA *84*, 2363–2367.
36. Kruglyak, L., Daly, M.J., Reeve-Daly, M.P., and Lander, E.S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. Am. J. Hum. Genet. *58*, 1347–1363.
37. Abecasis, G.R., Cherny, S.S., Cookson, W.O., and Cardon, L.R. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet. *30*, 97–101.
38. Lathrop, G.M., Lalouel, J.M., Julier, C., and Ott, J. (1984). Strategies for multilocus linkage analysis in humans. Proc. Natl. Acad. Sci. USA *81*, 3443–3446.
39. Thompson, E.A., and Heath, S.C. (1999). Estimation of conditional multilocus gene identity among relatives. Statistics in Molecular Biology and Genetics: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology, *33*, 95–113.
40. Ploughman, L.M., and Boehnke, M. (1989). Estimating the power of a proposed linkage study for a complex genetic trait. Am. J. Hum. Genet. *44*, 543–551.
41. Baum, L.E. (1972). An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. Inequalities *3*, 1–8.
42. Elston, R.C., and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. Hum. Hered. *21*, 523–542.
43. Wilcox, M.A., Pugh, E.W., Zhang, H.P., Zhong, X.Y., Levinson, D.F., Kennedy, G.C., and Wijsman, E.M. (2005). Comparison of single-nucleotide polymorphisms and microsatellite markers for linkage analysis in the COGA and simulated data sets for Genetic Analysis Workshop 14: Presentation Groups 1, 2, and 3. Genet. Epidemiol. *29* (*Suppl 1*), S7–S28.
44. Haldane, J.B.S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. J. Genet. *8*, 299–309.
45. Sobel, E., Sengul, H., and Weeks, D.E. (2001). Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. Hum. Hered. *52*, 121–131.

46. Wijsman, E.M., Brunzell, J.D., Jarvik, G.P., Austin, M.A., Motulsky, A.G., and Deeb, S.S. (1998). Evidence against linkage of familial combined hyperlipidemia to the Apolipoprotein AI-CIII-AIV gene complex. Arterioscler. Thromb. Vasc. Biol. *18*, 215–226.
47. Wijsman, E.M., Rothstein, J.H., Igo, R.P., Jr., Brunzell, J.D., Motulsky, A.G., and Jarvik, G.P. (2010). Linkage and association analyses identify a candidate region for APOB level on chromosome 4q32.3 in FCHL families. Hum. Genet. *127*, 705–719.
48. Kruglyak, L. (1997). The use of a genetic map of biallelic markers in linkage studies. Nat. Genet. *17*, 21–24.
49. Matise, T.C., Chen, F., Chen, W.W., De La Vega, F.M., Hansen, M., He, C.S., Hyland, F.C.L., Kennedy, G.C., Kong, X.Y., Murray, S.S., et al. (2007). A second-generation combined linkage physical map of the human genome. Genome Res. *17*, 1783–1786.
50. Boehnke, M. (1994). Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. Am. J. Hum. Genet. *55*, 379–390.
51. Browning, S.R., and Browning, B.L. (2010). High-resolution detection of identity by descent in unrelated individuals. Am. J. Hum. Genet. *86*, 526–539.
52. Browning, B.L., and Browning, S.R. (2011). A fast, powerful method for detecting identity by descent. Am. J. Hum. Genet. *88*, 173–182.
53. Huang, Q.Q., Shete, S., and Amos, C.I. (2004). Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. Am. J. Hum. Genet. *75*, 1106–1112.
54. Schaid, D.J., McDonnell, S.K., Wang, L., Cunningham, J.M., and Thibodeau, S.N. (2002). Caution on pedigree haplotype inference with software that assumes linkage equilibrium. Am. J. Hum. Genet. *71*, 992–995.
55. Almasy, L., and Blangero, J. (2004). Exploring positional candidate genes: linkage conditional on measured genotype. Behav. Genet. *34*, 173–177.
56. Lunetta, K.L., Faraone, S.V., Biederman, J., and Laird, N.M. (2000). Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. Am. J. Hum. Genet. *66*, 605–614.
57. Bourgain, C., Hoffjan, S., Nicolae, R., Newman, D., Steiner, L., Walker, K., Reynolds, R., Ober, C., and McPeek, M.S. (2003). Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. Am. J. Hum. Genet. *73*, 612–626.
58. Thomas, D.C., Casey, G., Conti, D.V., Haile, R.W., Lewinger, J.P., and Stram, D.O. (2009). Methodological issues in multistage genome-wide association studies. Stat. Sci. *24*, 414–429.