Accurate Imputation of Rare and Common Variants in a Founder Population From a Small Number of Sequenced Individuals

Lawrence H. Uricchio,[1] Jessica X. Chong,[1] Kevin D. Ross,[1] Carole Ober,[1*†] and Dan L. Nicolae[1,2,3*†]

[1]*Department of Human Genetics, University of Chicago, Chicago Illinois*
[2]*Department of Medicine, University of Chicago, Chicago Illinois*
[3]*Department of Statistics, University of Chicago, Chicago Illinois*

Advances in DNA sequencing technologies have greatly facilitated the discovery of rare genetic variants in the human genome, many of which may contribute to common disease risk. However, evaluating their individual or even collective effects on disease risk requires very large sample sizes, which involves study designs that are often prohibitively expensive. We present an alternative approach for determining genotypes in large numbers of individuals for all variants discovered in the sequence of relatively few individuals. Specifically, we developed a new imputation algorithm that utilizes whole-exome sequencing data from 25 members of the South Dakota Hutterite population, and genome-wide single nucleotide polymorphism (SNP) genotypes from >1,400 individuals from the same founder population. The algorithm relies on identity-by-descent sharing of phased haplotypes, a different strategy than the linkage disequilibrium methods found in most imputation algorithms. We imputed genotypes discovered in the sequence data to on average ∼77% of chromosomes among the 1,400 individuals. Median R^2 between imputed and directly genotyped data was >0.99. As expected, many variants that are vanishingly rare in European populations have risen to larger frequencies in the founder population and would be amenable to single-SNP analyses. *Genet. Epidemiol.* 36:312–319, 2012.

Key words: Exome sequencing; association study; IBD calculation; complex pedigrees; imputation

Supporting Information is available in the online issue at wileyonlinelibrary.com.
†These authors contributed equally to this work.
Open source software will be made available to interested parties by e-mail request.
*Correspondence to: Dan L. Nicolae, 5734 S. University Avenue, Eckhart 127, Chicago, IL 60637. E-mail: nicolae@galton.uchicago.edu / Carole Ober, 920 E 58th Street, CLSC 425, Chicago, IL 60637. E-mail: c-ober@genetics.uchicago.edu
Received 5 October 2011; Revised 4 January 2012; Accepted 9 January 2012
Published online 28 March 2012 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.21623

INTRODUCTION

Most common diseases are largely heritable, yet much of the genetic variation that contributes to disease risk remains unknown [Eichler et al., 2010]. Although genome-wide association studies (GWAS) have successfully uncovered many common human genomic variants that impact disease risk, it is now clear that new approaches must be developed to find the remaining risk-associated variation. Under some common models of disease risk, rare genetic variation plays a significant role [Clarke and Cooper, 2010; Manolio et al., 2009]. Moreover, due to the rapidly falling DNA sequencing prices, the rate of discovery of rare variants in the human genome has recently accelerated. However, it can be difficult to test the contribution of individual rare variants to disease risk, even in association studies of thousands of individuals. In order to gain sufficient statistical power for these tests, the variants must have large effect sizes or they must be grouped together, often in nonobvious ways [Oexle, 2010].

Founder populations offer an alternative to association studies of unrelated subjects or families [Lander and Schork, 1994; Ober and Cox, 1998; Peltonen et al., 2000]. In a founder population, each individual's genome is composed of fragments of a subset of the genetic material introduced by a relatively small number of founder individuals. Although the overall amount of variation in a founder population is small compared to nonfounder populations, some variants that are rare in the general population (and potentially deleterious) occur at higher frequencies in founder populations due to the effects of random genetic drift [Lowe et al., 2009]. It is now possible in principle to find such variants by DNA sequencing, but despite the falling sequencing prices, deep sequencing at the whole-exome (or whole-genome) level in large numbers of subjects remains prohibitively expensive for many academic research groups. Imputation of genotypes for sequenced variants discovered in a small number of individuals to a larger sample is an appealing alternative. Indeed, imputation of sequenced variants was recently used to find a causal variant for sick sinus syndrome in the Icelandic population [Holm et al., 2011].

Imputation of genotypes in both related and unrelated individuals has been performed successfully for many years (see [Li et al., 2009]). Imputation algorithms in unrelated individuals typically rely on genotype training data from a large number of unrelated individuals [Li et al. 2010]. Family-based imputation approaches

identify regions of identity-by-descent (IBD) sharing, and some linkage packages have built-in support for imputation (e.g., MERLIN [Abecasis et al., 2002]), but there are limitations on maximum family size and minimum relatedness of individuals that can be analyzed by these approaches. Imputation algorithms for unrelated individuals take advantage of the correlation structure (linkage disequilibrium, LD) among SNPs in the same region, structure that is inferred from a reference panel [Browning and Browning, 2009; Li et al., 2010; Marchini et al., 2007; Servin and Stephens, 2007].

Founder populations offer advantages for imputation studies. First, founder populations have less genetic variation than nonfounder populations, so less input data (e.g., fewer exome-sequenced individuals) are needed to impute genotypes to many individuals. Second, many founder populations have maintained detailed genealogical records (e.g., [Steinberg et al., 1967]), which can be used to calculate accurate kinship coefficients that confer additional information about the global likelihood of IBD between all pairs of individuals (e.g., [Abney, 2009]). Dense genotype information provided by Affymetrix or Illumina arrays can be used in concert with kinship coefficients to calculate local IBD-sharing relationships, which can be leveraged for the purposes of imputation.

We developed an imputation algorithm that takes advantage of the relatedness among members of the Hutterite population to assign genotypes for sequenced variants discovered in only 25 Hutterites to more than 1,400 Hutterites with genome-wide SNP genotypes. Our method first determines phase, and then locally computes IBD near variants discovered in the exome sequences, and finally imputes genotypes to the remaining individuals. We show that we can phase and impute genotypes discovered through exome sequencing to members of this founder population with high accuracy and low missing information. We show that a large proportion of the exonic variants that are not present in dbSNP 129 and, therefore, likely to be rare in the general population, occur at larger frequencies in the Hutterites. This approach can be used for subsequent GWAS or for the global detection of IBD. These algorithms are ideally suited to founder populations, particularly those with large, complex pedigrees.

MATERIALS AND METHODS

HUTTERITES

The Hutterites are a religious isolate that originated in central Europe in the 16th century; after a series of migrations and population bottlenecks, they settled in the United States in the 1870s, at which time their population size increased exponentially [Hostetler, 1974]. Today there are >40,000 Hutterites living on communal farms in the northern plains of the United States and western Canada, all of whom descended from <90 founders [Martin, 1970]. The Hutterites living in South Dakota, the subjects of our genetic studies, are descendants of just 64 of the 90 founders [Ober et al., 2001]. The imputation studies presented here focused on 1,414 South Dakota Hutterites who are related to each other in a 3,672 person pedigree. We used Affymetrix arrays (500k, 5.0 and 6.0) to measure genotypes at the loci of 246,057 common bi-allelic SNPs in these Hutterites as previously described [Ober et al., 2008; Ober et al., 2009]. This set of common bi-allelic SNPs is referred to as "GWAS SNPs" herein.

Twenty-five Hutterites were selected from among the 1,414 genotyped individuals for whole-exome sequencing to maximize the number of genotyped descendents and minimize the pairwise relatedness between the 25 subjects. Whole-exome sequencing was performed as previously described [Caliskan et al., 2011].

PHASING

Our phasing algorithm draws on ideas from conventional phasing approaches, and the long-range phasing algorithm described in Kong et al. [2008], in which surrogate parents (i.e., relatives of the proband through either the mother or the father) are used to determine phase. Surrogates are relatives who are identical-by-state (IBS) with the proband at 1,000 consecutive SNPs, which implies IBD. Phase information in the surrogates is then used to phase the proband. Kong et al. report that this method provides nearly complete phase information among genotyped individuals in the Icelandic population.

Although a 1,000 SNP IBS window provides strong evidence of IBD, this method is not sensitive to regions of IBD composed of fewer SNPs. Furthermore, in an inbred pedigree, classification of surrogates as relatives of only the mother or only the father is often not possible as both parents can be related to the proposed surrogate. This classification is especially difficult in the absence of parental genotype information. To circumvent this requirement, we developed a simple local phasing tool that first exhaustively uses nuclear family information to phase, and then uses information in more distant relatives to probabilistically calculate phase. Importantly, this approach does not require the classification of surrogate parents as maternal or paternal relatives.

Our phasing algorithm is executed simultaneously on all genotyped members of the Hutterite pedigree, and each small chromosomal region (100 SNPs herein) is phased independently. SNPs cannot be phased by Mendelian transmission within a trio when both parents and the proband are heterozygous or when the proband is heterozygous and one or both parents have missing genotype information. However, in cases where a nuclear family has two genotyped children, the proportion of SNPs at which both parents and all children are heterozygous reduces by a factor of 2 compared to a family with only one genotyped child. As more members of a family (such as other children, grandparents, grandchildren) are genotyped, the number of instances where all individuals are heterozygous at a particular SNP becomes very small. This is particularly useful for phasing in the Hutterites, in whom the mean number of genotyped sibs is approximately four and grandparents or grandchildren are available in many cases. We exploit these features of the pedigree to phase families along direct lines of descent.

We first exhaustively apply single-SNP Mendelian logic to phase as many SNPs as possible. After this step is completed, we use phase information of SNPs proximal to an unphased SNP of interest to compare partially phased haplotypes among family members. If a recombination has not occurred within a nuclear family near this locus, then one or more pairs of family members will share a haplotype at this locus IBD. We search for pairs of individuals displaying local IBD sharing within the family, and we pass

any phase information that is complete in one individual to the other individual.

In practice, the method described above cannot be applied near sites of recombination, which can be detected in nuclear families when comparing parental (partially phased) haplotypes to the haplotypes of their offspring. A recombination is detected when contiguous SNPs that are heterozygous in the parent (and hence differentiate the parent's two chromosomes) support the inheritance of one of the parent's chromosomes proximal to a given SNP, and the inheritance of the other of the parent's chromosomes distal to the SNP in the offspring. When a recombination is detected between a parent and offspring, we do not use the partial haplotype-matching technique described above. Instead, we use the other parent's phase information (when available) to phase the proband. This is a practical solution because it is rare for a recombination event to occur in both parents in the same small region of the genome.

For genotyped Hutterites who lack genotype information for any parents or children, or who lack phase information at 5% or more SNPs after application of the above methods, the algorithm then searches for surrogate parents who share a haplotype with the proband in a manner similar to that used by Kong et al. If the proband and a phased relative share a haplotype IBD with high confidence (discussed below), we use the phase information in the relative to phase the proband.

IBD-SHARING CALCULATION

IBD-sharing likelihoods can be calculated from either phased or unphased genotype data, but phased data are preferred because breakpoints in IBD sharing are more easily determined from phased data. While this is not critical when comparing long regions of IBD (1,000 SNPs, for example), breakpoint uncertainty can greatly change the relative length of short IBD tracts. It is also simpler to distinguish between IBD = 2 and IBD = 1 segments with phased data, which is critical for inbred populations where the IBD = 2 configuration is possible between even distant relatives.

We perform a simple calculation that approximates the probability of IBD for two IBS haplotypes of any length. We first prune for LD in the region using a greedy strategy that leaves only SNPs with pairwise $r^2 < 0.3$ for all pairs. We then treat all remaining SNPs as probabilistically independent. We compare each remaining nonphased chromosomal segment to phase information from relatives who are >98% phased in that region by the previous methods. We first search for uninterrupted regions of IBS between pairs of individuals. If such a region is discovered, we approximate the probability of IBD in the following manner:

$$P(IBD = 1 \mid IBS) = \frac{P(IBS \mid IBD=1)P(IBD=1)}{P(IBS \mid IBD=0)P(IBD=0)+P(IBS \mid IBD=1)P(IBD=1)}, \quad (1)$$

where $P(IBD = 1)$ and $P(IBD = 0)$ are given by the kinship coefficient K and $(1 - K)$, respectively. We calculate kinship coefficients from the Hutterite pedigree [Abney, 2009]. $P(IBS \mid IBD = 1)$ is 1 (assuming low genotyping error and mutation rate). $P(IBS \mid IBD = 0)$ is given by the product of the (assumed independent) allele frequencies in the region of the locus of interest. Allele frequencies are estimated using the set of 1,414 genotyped Hutterites. This calculation is similar in form for unphased data (but the denominator must be summed over all three possible states, $IBD = 0$, $IBD = 1$, $IBD = 2$) [Purcell et al., 2007].

Our calculation is an approximation to the true probability because there is some remaining LD after pruning the SNPs in the region to be independent, and because we use estimated allele frequencies. We require that this approximation exceed a threshold in order to assign local IBD relationships. This threshold is a tuning parameter in the phasing algorithm. We set this threshold to 0.95, which is low compared to default values of analogous parameters of programs that focus on nonfounder populations, such as Beagle's fastIBD method [Browning and Browning, 2011]. This threshold would likely result in an extremely high false positive rate if our method was naively applied to a nonfounder population (or a founder population with more than the 64 Hutterite founders). However, because the Hutterites in our studies are descended from 64 founders, there are at most 128 different possible chromosomes at any one locus. Many of the founder haplotypes were eliminated by random processes, especially early in the history of the population following a bottleneck. Therefore, it is likely that even most short IBS tracts are IBD in the Hutterites.

IMPUTATION

Imputation is performed on a SNP by SNP basis for each SNP present in the exome sequences. Each of the 25 exome-sequenced individuals' genotypes at each SNP is filtered by phred [Ewing and Green, 1998] based quality score (>90) and depth of coverage (≥16×); passing calls are then used as input to the imputation algorithm.

Our imputation algorithm requires phasing of the sequenced variants to place each allele onto a local haplotype. Once an allele is placed into the context of a local haplotype, it is possible to compare phased haplotypes directly between relatives. Therefore, we first phase the 25 exome-sequenced individuals, who were also genotyped for the GWAS SNPs. We assign each homozygous individual's allele to local haplotypes defined by the genotypes of nearby GWAS SNPs for that individual. Genotypes of heterozygous individuals in the sequencing sample are phased if they share one local haplotype IBD with a phased individual. The approximate probability of pairwise local IBD is again calculated with the equation given above. If there is a high likelihood of IBD between an unphased, sequenced individual and a phased, sequenced individual, we then phase the heterozygous individual at this SNP of interest.

After phasing as many of the sequenced individuals as possible, we repeat the pairwise IBD calculation with other phased (nonsequenced) individuals in the Hutterite pedigree. We prioritize this process by first searching for an IBD match between the subject of interest and the sequenced/phased individual with whom the subject has the highest kinship coefficient. If a high probability of IBD is found both proximal and distal to the SNP of interest, we are able to assign a haploid genotype to the nonsequenced individual. As a result, some individuals will have a haploid imputed genotype if an IBD match is found for only a single chromosome.

GENE-DROPPING SIMULATIONS

We used gene-dropping simulations to estimate the proportion of each founder's genome expected to be represented in the 25 exome-sequenced and the larger sample of genotyped Hutterites. Each founder in the Hutterite pedigree was seeded with a unique genotype at each of 1,000 independently segregating loci. Each unique allele was passed to the descendants of each founder according to the rules of Mendelian inheritance. This model accurately captures the expected proportion of genetic material inherited by each Hutterite from each founder individual under a neutral model, but does not model recombination, selection, or other genetic processes. At the conclusion of the simulations, we measured the fraction of the alleles remaining in the 1,414 individuals genotyped for the GWAS SNPs that were also present among the 25 exome-sequenced individuals.

RESULTS

ACCURACY AND COMPLETENESS OF IMPUTED GENOTYPES

We first assessed the accuracy and completeness of our imputation algorithm by masking the genotypes of 2,000 GWAS SNPs and phasing blind to these genotypes. These SNPs were selected to match the minor allele frequency spectrum of the variants in the 25 exome sequences (Fig. S1). We then used the genotypes at these SNPs for the 25 individuals in our exome sequencing study as input to our imputation algorithm and imputed genotypes to the remaining individuals with phased GWAS SNP data. A summary of the haploid call rates is presented in Figure 1A. The median haploid call rate per SNP was 77.2% and the median diploid call rate was 60.9%. We also measured the proportion of genotypes that were phased among all sites of a given allele frequency, as this will greatly influence the completeness of imputation. The proportion of genotypes phased varies between 86% and 99% and generally decreases with increasing minor allele count (Fig. S3). At singleton SNPs, 68% of heterozygous individuals were phased.

To measure the accuracy of the algorithm, we computed the squared Pearson correlation, R^2 [Browning and Browning, 2009], between GWAS genotypes and imputed genotypes for each of the 2,000 SNPs. The median R^2 between imputed genotypes and GWAS genotypes was greater than 0.99 for each minor allele frequency bin, but the mean for singleton SNPs was 0.751. This reflects the sensitivity of the method to the accuracy of the phasing of the single-heterozygous individual at singleton SNPs (Fig. 1B). The call rate was not strongly correlated with concordance rate (Pearson's correlation coefficient of 0.08).

As we have a complete and accurate 13-generation pedigree of Hutterites in our studies, it is possible to apply simple Mendelian rules to estimate the percentage of original genetic material from each founder that is present among the 1,414 genotyped individuals. Furthermore, it is possible to estimate the proportion of this variation that is present in the 25 exome-sequenced individuals. Using pedigree-based simulations of Mendelian inheritance, we found that there is a 79% chance that any given founder variant that remains among the 1,414 individuals (i.e., has not been eliminated by drift) is also present among the 25 exome-sequenced individuals. This indicates that with perfect genotyping and imputation, we would be able to impute to an average of 79% of the chromosomes among the 1,414 genotyped Hutterites. The median and mean haploid call rates for the 2,000 test SNPs were 77.2% and 70.8%, respectively, indicating that our imputation algorithm provides near-optimal completeness of information for our reference panel.

EXOME SEQUENCING AND IMPUTATION OF SEQUENCED VARIANTS

The 25 exome sequences revealed 12,134 novel variants and 35,596 previously annotated (dbSNP build 129)

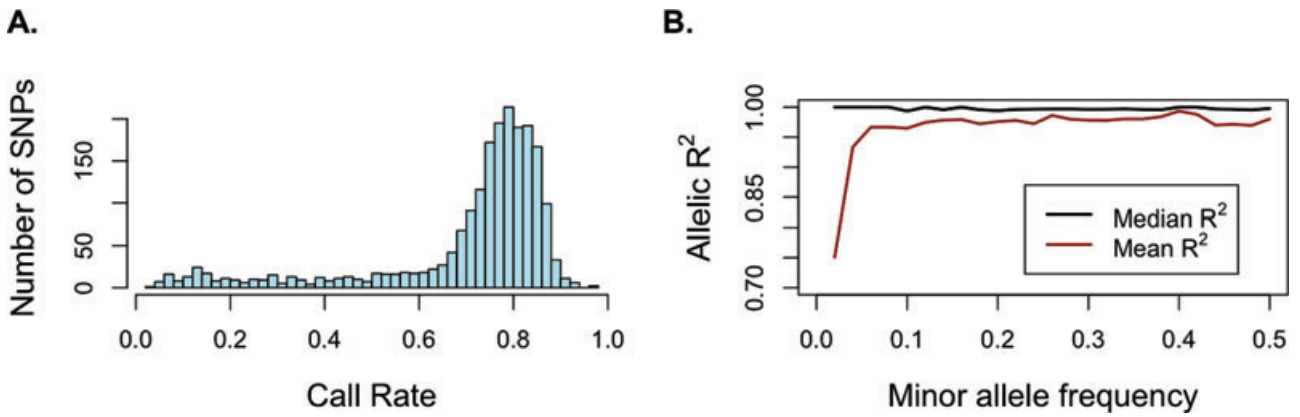

Fig. 1. Accuracy and completeness of the imputation algorithm. The genotypes of 2,000 GWAS SNPs that matched the site-frequency spectrum of the exome sequencing data were masked in the GWAS SNP dataset. We then performed our phasing algorithm in the absence of these genotypes. Using this phase data and the genotypes of the same 25 Hutterites for whom whole-exome sequences are available, we imputed genotypes for the 2,000 masked SNPs to the rest of the 1,414 person sample. (A) The median haploid call rate per SNP was 77%, where call rate was calculated as N/2,828 (N = number of total genotypes imputed; 2,828 is the number of chromosomes among the 1,414 GWAS SNP genotyped individuals). (B) Median and mean R^2 for SNPs of varying allele frequencies. R^2 is the Pearson's correlation coefficient between imputed genotypes and GWAS genotypes. Allele frequency bins are each 0.02 wide.

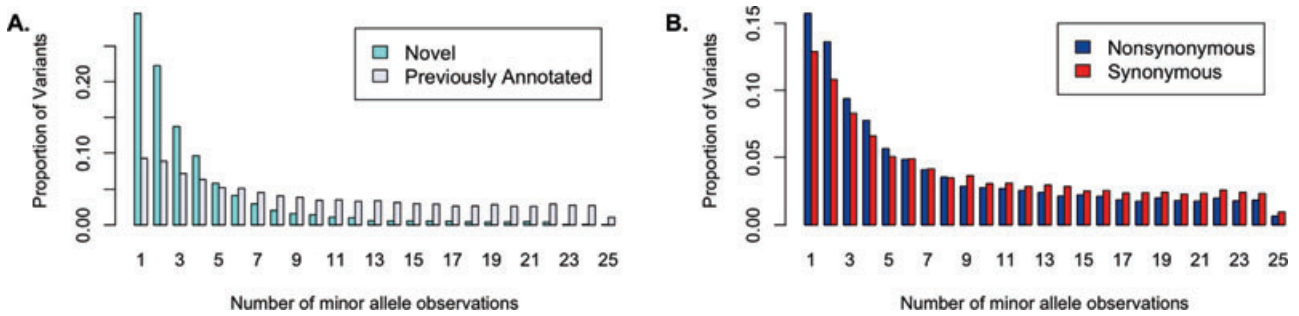

Fig. 2. Site-frequency spectra of (A) novel vs. previously annotated SNPs and (B) nonsynonymous vs. synonymous SNPs that were present in the exome sequencing study. Both nonsynonymous and novel variants are shifted to lower frequencies relative to synonymous/previously annotated variants. Variants are included in the figure only if a diploid genotype call could be made in all 25 individuals.

variants after quality control. The site-frequency spectra (SFS) of minor allele counts of both novel and annotated variants and the SFS of nonsynonymous and synonymous mutations are shown in Figure 2. Not surprisingly, the distributions of novel and nonsynonymous alleles are shifted to lower frequencies compared to annotated and synonymous mutations, respectively. However, 70.4% of novel variants are present at least twice in the 25 sequenced exomes, corresponding to a sample frequency at or above 4%. Because these variants are not present in dbSNP 129, their allele frequency among all European-derived samples is likely to be much lower.

Using genotype calls at these sequenced variants as input to our imputation algorithm, we phased and imputed genotypes to the remaining individuals among the 1,414 genotyped for the GWAS SNPs. We performed a leave-one-out cross validation to verify that these imputed genotypes had similar accuracy and completeness profiles to the data presented in Figure 1. The accuracy as measured by Pearson's R^2 was similar to the results presented in Figure 1, although the completeness diminished slightly (median haploid call rate of 72%, Fig. S2). This reduction in median call rate was expected because of the reduction in sample size for the sequenced panel, and because the average kinship between pairs of exome-sequenced individuals is lower than the Hutterite population average. Finally, we selected three imputed SNPs that were candidates in other ongoing studies for further study. Two of these SNPs are novel SNPs discovered in the exome sequences. We independently genotyped these three SNPs with Taqman® assays (Applied Biosystems, Carlsbad, CA). Concordance between Taqman genotypes and imputed genotypes was above 99% for each SNP (Table SI).

LD BETWEEN IMPUTED AND GWAS VARIANTS

We measured the r^2 value between each of the GWAS SNPs and each of the imputed exome variants using every available imputed diploid genotype. For each imputed SNP, we computed the maximum r^2 value with any GWAS SNP. We report a cumulative distribution of maximum r^2 values for imputed SNPs in Figure 3. Forty percent of the imputed SNPs have a maximum r^2 less than 0.798, and 25% of the SNPs have a maximum r^2 of less than 0.610. Furthermore, 40% of novel SNPs have a maximum r^2 of less than 0.642 and 25% of novel SNPs have a maximum r^2 of less than 0.491. As a consequence, the evidence of association at the many of the imputed SNPs may not be efficiently captured by single-SNP analyses of the GWAS SNPs.

IMPUTATION WITH ADDITIONAL SEQUENCING SUBJECTS

Lastly, we assessed the effect of larger samples of sequenced individuals on the accuracy and completeness of our imputation algorithm using the same set of 2,000 test SNPs discussed previously. In general, selecting the group of individuals that will maximize the call rate of our imputation algorithm is a difficult problem, so individuals were selected as sequencing candidates with a greedy strategy that provides good coverage. First, we selected the person with the highest average kinship to the other 1,413 individuals in our study. We subsequently selected individuals who had the highest average kinship to the remaining, unselected individuals, but who have kinship not exceeding 0.1 with any other individual already selected for sequencing. We composed two groups of individuals using this strategy, with sample sizes of 48 and 96 individuals, the group of 48 being a subset of the group of 96. The median haploid call rate for imputed SNPs in the larger sample of 1,414 individuals was 88.1% when 48 individuals were sequenced, while the median haploid call rate was 94.1% when 96 individuals were sequenced (Fig. 4). The median diploid call rates improved to 78.9% and 89.5% for the 48 and 96 individual groups, respectively. The mean R^2 value improved at low allele frequency relative to Figure 1 (mean R^2 of 0.865 for 48 individuals, 0.875 for 96 individuals).

DISCUSSION

In this paper, we described a new method to impute genotypes for variants discovered in a subset of individuals in a founder population to other members of the population. We demonstrated the high accuracy of the imputation algorithm across the full range of allele frequencies, and then assigned imputed genotypes at each variant site to a large proportion of the chromosomes in the population. As predicted, many variants that are rare in European populations occur at estimated frequencies of >4% in the Hutterites. Although we were motivated by the hypothesis that rare

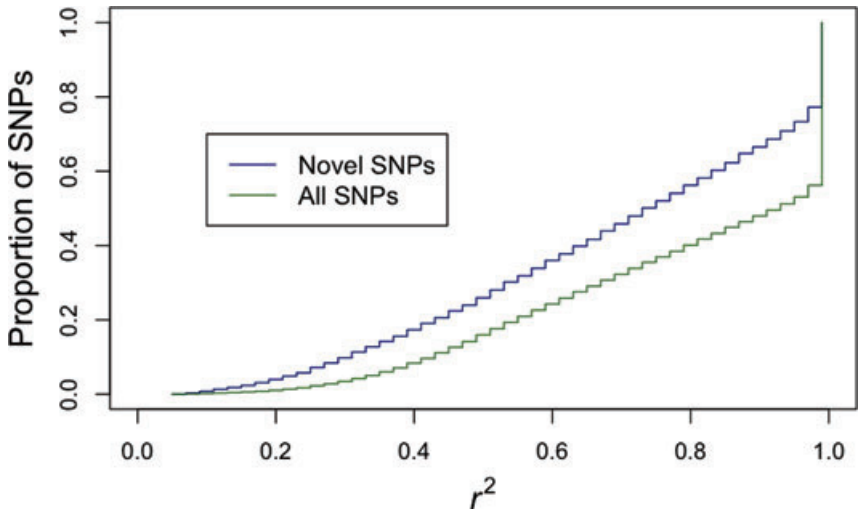

Fig. 3. Linkage disequilibrium between GWAS SNPs and imputed SNPs. We measured pairwise LD between imputed and GWAS SNPs. r^2 values were computed between each GWAS SNP and each imputed SNP on each chromosome. Here, we plot the cumulative distribution of maximum r^2 values discovered for each imputed SNP with the GWAS SNPs. Overall, 40% of imputed SNPs have a maximum r^2 of less than 0.798 and 25% have a maximum r^2 of less than 0.610, while 40% of novel SNPs have a maximum r^2 of less than 0.642 and 25% have a maximum r^2 of less than 0.491.

variation may contribute to risk for common diseases, our algorithm is equally well suited to imputing and ultimately investigating associations with common variants.

Our imputation method was designed to impute genotypes in a founder population for future GWAS of common diseases and quantitative phenotypes. As founder populations have less genetic diversity than nonfounder populations, with a preferential loss of rare alleles, it was previously unknown whether there was a significant amount of variation in the Hutterites that is not tagged by the GWAS SNPs. In fact, 25% of the novel variants discovered by exome sequencing have no corresponding GWAS SNP with an r^2 value of greater than 0.491. If any of these novel, sequenced variants with low maximum r^2 are disease- or phenotype-associated SNPs, our power to detect these associations is greater with our method relative to a study using only GWAS SNPs.

By sequencing only 25 Hutterite exomes, we were able to impute variants to 77% of chromosomes among our larger dataset of 1,414 individuals. Sequencing a larger number of Hutterites will not only enhance our ability to discover genetic variation that is present in the Hutterite pedigree, but will also allow us to impute genotypes to a greater proportion of the sample. For example, imputation of 2,000 GWAS SNPs that matched the allele frequency spectrum of the exome data from a sample of 96 Hutterites garnered a median haploid call rate of approximately 94% and diploid call rate of 89.5% (Fig. 4).

An ultimate consideration is the effect of the imputation coverage and accuracy on downstream association tests. Because of the high accuracy we observe in our studies (similar in magnitude to that of direct genotyping), we believe that this could result in only a minor decrease in power compared to directly genotyping the same subjects. Coverage could have a higher impact, as nonrandom missing genotypes can lead to an inflation in type 1 error. This bias can occur when the missing data patterns are correlated with both the phenotype (subjects closer related to the sequenced individuals have a different phenotype distribution than subjects further related) and with specific alleles (such as the minor allele). In our studies, we address this issue by testing for association between missing genotypes (coded as 0/1) and phenotype. The results of these tests can be used to filter SNPs out of the analysis in the same way as it is done with deviations from Hardy-Weinberg equilibrium. Note that a similar strategy is used in classical GWAS—a significant difference in missing genotypes in cases and controls could flag SNPs during the QC process. Thus, overall, we believe that biases introduced by imputation can be assessed and directly addressed by filtering imputed SNPs based on call rates or nonrandom missing genotypes, as we do for directly genotyped SNPs.

Our approach to generating imputed genotypes is dependent on phasing and accurate calculation of local IBD-sharing relationships. In our study, we followed large scale Affymetrix genotyping in a founder population with sequencing of a subset of these individuals. The quality of phasing depends greatly on the pedigree (larger nuclear families are easier to phase) and on the IBD-sharing information (that allows the use of "surrogate parents"). The IBD inference depends on the density and informativeness of the GWAS SNPs. The number of sequenced individuals affects the coverage of the pedigree (the imputed SNP call rates), but it has little influence on the accuracy of imputation. This is different than the approaches to imputation in nonfounder populations that commonly rely on LD structure that is inferred from a reference panel [Browning and Browning, 2009; Li et al., 2010; Marchini et al., 2007; Stephens and Donnelly, 2003]. LD-based approaches are best suited to cases where fine scale recombinations can be accurately estimated from the reference data. The accuracy of the imputation for these algorithms depends greatly on the quality of the reference panel (the number of subjects and the density and informativeness of the SNPs in the reference panel), and on the local LD

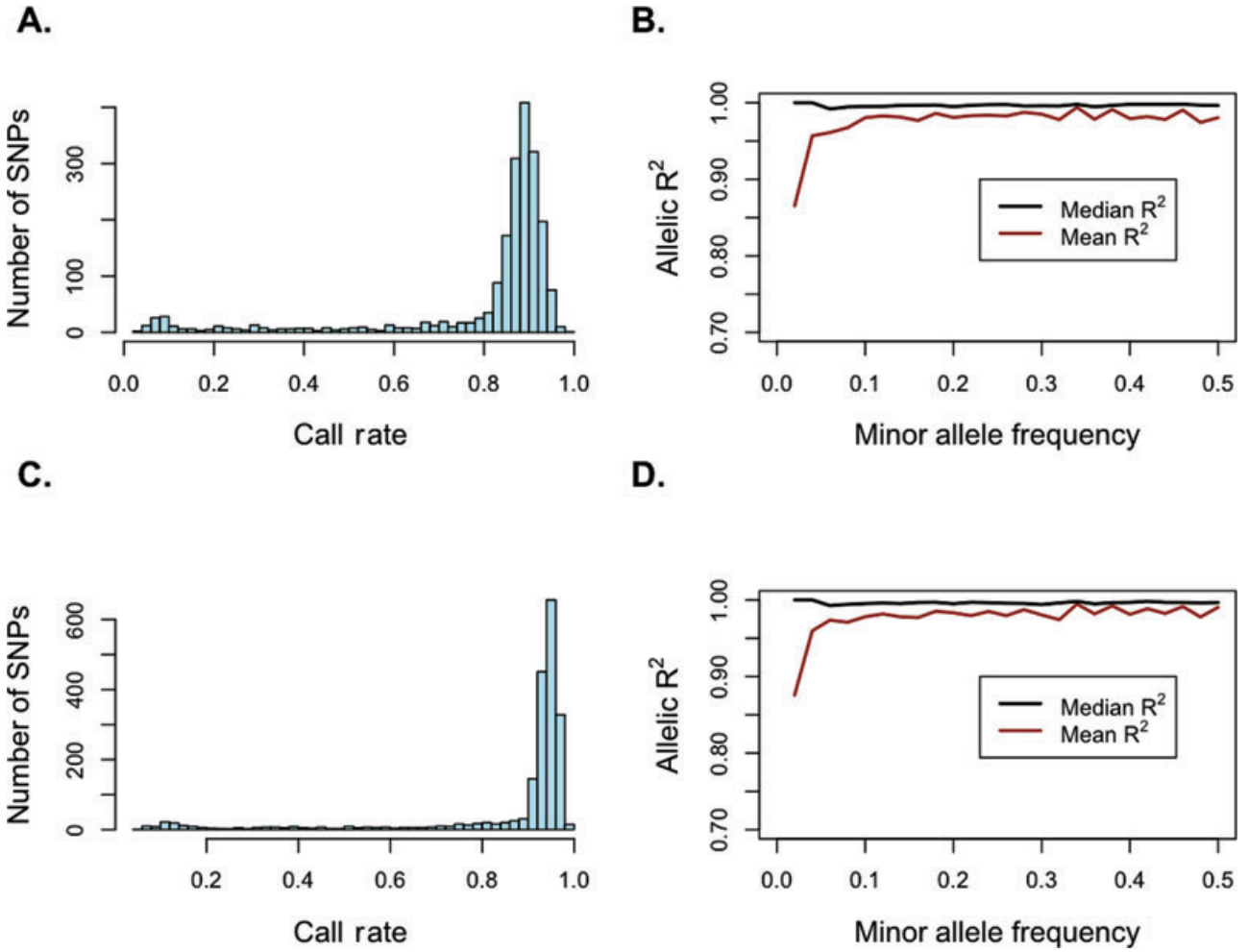

Fig. 4. Accuracy and completeness with additional subjects. We imputed genotypes for the same 2,000 SNPs shown in Figure 1, using the genotypes at these SNPs for sets of 48 individuals (A and B) and 96 individuals (C and D) as input to our imputation algorithm. Median haploid call rate improved to 94.1% for the group of 96 and 88.1% for the group of 48, whereas median diploid call rates improved to 89.5% and 78.9%, respectively. Accuracy, as reflected by the R^2 value, is improved at low frequency relative to Figure 1 because it becomes easier to phase low-frequency alleles as the number of reference individuals increases. Allele frequency bins are each 0.02 wide.

structure. Thus, even though the quality of the set of SNPs (number and frequency spectrum) and of the reference panel influence the performance for both groups of imputation methods, the two strategies take advantage of different characteristics of the data. This suggests that, in the context of a founder population, a combination of these strategies can be constructed to further improve coverage and imputation accuracy.

The genotypes generated by this study and future whole-genome sequencing/imputation studies will be included in GWAS of common diseases and disease-associated quantitative phenotypes in the Hutterites, which will allow direct assessment of the effects of exonic (coding) variants vs. noncoding variants on disease risk, including the subset of variants that are rare in the general population but relatively common in the Hutterites. The phased haplotypes may shed further light on recombination events in the Hutterites [Coop et al., 2008] and facilitate studies of the origin of Mendelian disease mutations in the Hutterites [Chong et al., 2011].

ACKNOWLEDGMENTS

We thank Gaixin Du for computational assistance. We thank our anonymous reviewers for their suggestions. This work was supported by NIH grants HL101651, HL085197, and HD21244. Exome sequencing studies were performed at the NHLBI-funded Exome Sequencing Service at the Broad Institute. The authors state that no competing interests exist.

REFERENCES

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30(1):97–101.

Abney M. 2009. A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. Bioinformatics 25(12):1561–1563.

Browning BL, Browning SR. 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(2):210–223.

Browning BL, Browning SR. 2011. A fast, powerful method for detecting identity by descent. Am J Hum Genet 88(2):173–182.

Caliskan M, Chong JX, Uricchio L, Anderson R, Chen P, Sougnez C, Garimella K, Gabriel SB, dePristo MA, Shakir K, Matern D, Das S, Waggoner D, Nicolae DL, Ober, C. 2011. Exome sequencing reveals a novel mutation for autosomal recessive non-syndromic mental retardation in the TECR gene on chromosome 19p13. Hum Mol Genet 20(7):1285–1289.

Chong JX, Oktay AA, Dai Z, Swoboda KJ, Prior TW, Ober C. 2011. A common spinal muscular atrophy deletion mutation is present on a single founder haplotype in the US Hutterites. Eur J Hum Genet 19:1045–1051.

Clarke AJ, Cooper DN. 2010. GWAS: heritability missing in action? Eur J Hum Genet 18(8):859–861.

Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. 2008. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. Science 319(5868):1395–1398.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11(6):446–450.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8(3):186–194.

Holm H, Gudbjartsson DF, Sulem P, Masson G, Helgadottir HT, Zanon C, Magnusson OT, Helgason A, Saemundsdottir J, Gylfason A, Stefansdottir H, Gretarsdottir S, Matthiasson SE, Thorgeirsson GM, Jonasdottir A, Sigurdsson A, Stefansson H, Werge T, Rafnar T, Kiemeney LA, Parvez B, Muhammad R, Roden DM, Darbar D, Thorleifsson G, Walters GB, Kong A, Thorsteinsdottir U, Arnar DO, Stefansson K. 2011. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nat Genet 43(4):316–320.

Hostetler JA. 1974. Hutterite society. Baltimore: Johns Hopkins University Press.

Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, Olason PI, Ingason A, Steinberg S, Rafnar T, Sulem P, Mouy M, Jonsson F, Thorsteinsdottir U, Gudbjartsson DF, Stefansson H, Stefansson K. 2008. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet 40(9):1068–1075.

Lander ES, Schork NJ. 1994. Genetic dissection of complex traits. Science 265(5181):2037–2048.

Li Y, Willer C, Sanna S, Abecasis G. 2009. Genotype imputation. Annu Rev Genomics Hum Genet 10:387–406.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34(8):816–834.

Lowe JK, Maller JB, Pe'er I, Neale BM, Salit J, Kenny EE, Shea JL, Burkhardt R, Smith JG, Ji W, Noel M, Foo JN, Blundell ML, Skilling V, Garcia L, Sullivan ML, Lee HE, Labek A, Ferdowsian H, Auerbach SB, Lifton RP, Newton-Cheh C, Breslow JL, Stoffel M, Daly MJ, Altshuler DM, Friedman JM. 2009. Genome-wide association studies in an isolated founder population from the Pacific Island of Kosrae. PLoS Genet 5(2):e1000365.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747–753.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39(7):906–913.

Martin AO. 1970. The founder effect in a human isolate: evolutionary implications. Am J Phys Anthropol 32(3):351–367.

Ober C, Abney M, McPeek MS. 2001. The genetic dissection of complex traits in a founder population. Am J Hum Genet 69(5):1068–1079.

Ober C, Cox NJ. 1998. The genetics of asthma. Mapping genes for complex traits in founder populations. Clin Exp Allergy 28 (Suppl 1):101–105; discussion 108–110.

Ober C, Nord AS, Thompson EE, Pan L, Tan Z, Cusanovich D, Sun Y, Nicolae R, Edelstein C, Schneider DH, Billstrand C, Pfaffinger D, Phillips N, Anderson RL, Philips B, Rajagopalan R, Hatsukami TS, Rieder MJ, Heagerty PJ, Nickerson DA, Abney M, Marcovina S, Jarvik GP, Scanu AM, Nicolae DL. 2009. Genome-wide association study of plasma lipoprotein(a) levels identifies multiple genes on chromosome 6q. J Lipid Res 50(5):798–806.

Ober C, Tan Z, Sun Y, Possick JD, Pan L, Nicolae R, Radford S, Parry RR, Heinzmann A, Deichmann KA and others. 2008. Effect of variation in CHI3L1 on serum YKL-40 level, risk of asthma, and lung function. N Engl J Med 358(16):1682–1691.

Oexle K. 2010. A remark on rare variants. J Hum Genet 55(4):219–226.

Peltonen L, Palotie A, Lange K. 2000. Use of population isolates for mapping complex traits. Nat Rev Genet 1(3):182–190.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575.

Servin B, Stephens M. 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3(7):e114.

Steinberg AG BH, Kurczynski TW, Martin AO, Kurczynski EM 1967. Genetic studies on an inbred human isolate. In: Crow JF NJ, editor. Human genetics. Baltimore: The Johns Hopkins Press. p 267–289.

Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73(5):1162–1169.