Genotype Imputation in Genome-Wide Association Studies

Eleonora Porcu,[1,2,3] Serena Sanna,[3] Christian Fuchsberger,[1] and Lars G. Fritsche[1]

[1]Department of Biostatistics, Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, Michigan
[2]Dipartimento di Scienze Biomediche, Università di Sassari, Sassari, Italy
[3]Istituto di Ricerca Genetica e Biomedica (IRGB), CNR, Monserrato, Cagliari, Italy

ABSTRACT

Imputation is an in silico method that can increase the power of association studies by inferring missing genotypes, harmonizing data sets for meta-analyses, and increasing the overall number of markers available for association testing. This unit provides an introductory overview of the imputation method and describes a two-step imputation approach that consists of the phasing of the study genotypes and the imputation of reference panel genotypes into the study haplotypes. Detailed steps for data preparation and quality control illustrate how to run the computationally intensive two-step imputation with the high-density reference panels of the 1000 Genomes Project, which currently integrates more than 39 million variants. Additionally, the influence of reference panel selection, input marker density, and imputation settings on imputation quality are demonstrated with a simulated data set to give insight into crucial points of successful genotype imputation. *Curr. Protoc. Hum. Genet.* 78:1.25.1-1.25.14. © 2013 by John Wiley & Sons, Inc.

Keywords: genome-wide association studies • imputation • linkage disequilibrium • inference • imputation • 1000 Genomes Project • HapMap Project • rare variants • genotyping arrays

INTRODUCTION

Since the first successful genome-wide association study (GWAS; Klein et al., 2005), the GWAS approach has been applied to many complex traits and diseases, leading to the identification of more than 5,500 variants significantly associated with at least one of more than 200 complex phenotypes (*http://www.genome.gov/gwastudies*). The approach assesses the genomes of several hundred or thousand individuals by using high-density genotyping arrays (see *UNIT 2.9*) to interrogate from several hundred thousand to millions of genetic markers (e.g., most commonly in platforms from Illumina and Affymetrix). Such markers were chosen to be representative of the most common genomic variation in human populations but only represent a small fraction of the more than 50 million common and rare variants that have been so far discovered (*http://www.ncbi.nlm.nih.gov/projects/SNP*). As a consequence, with only few exceptions, the identified associated variants that were discovered in these initial GWAS reports do not represent the putative causal variants, but variants that are associated indirectly, i.e., due to their linkage disequilibrium (LD) to the causal variants. An increase in the density of genotyped markers will increase the likelihood of refining the association signal and pinpoint the causal variants.

The analysis of whole-genome sequencing data of large studies would represent a superior solution because it would allow a comprehensive testing of all, even the rare, genetic variants. Generating and analyzing whole-genome sequencing of a thousand individuals is still not yet feasible for small to medium size laboratories (Metzker, 2010; Wetterstrand, 2013), but genotype imputation, a statistical framework, provides an efficient strategy for inferring and assessing in silico sequence data. Such an approach, combined with the information from two large international consortial efforts that aim to systematically and comprehensively catalog human variations of different populations, the HapMap Project and the 1000 Genomes Project (International HapMap 3 Consortium, 2010; 1000 Genomes Project Consortium, 2012), represents a

Current Protocols in Human Genetics 1.25.1-1.25.14, July 2013
Published online July 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471142905.hg0125s78

transitional solution. These projects, by genotyping or sequencing individuals from different ancestry groups, allow the determination of genomic variations and LD structures within ethnicities and represent a reference set for inferring untyped markers by looking at haplotype similarities of individuals under study.

IMPUTATION METHODS: OVERVIEW

Genotype imputation consists of inferring untyped markers in a study sample by using the LD structure among markers assessed in an external reference panel for which a much denser genetic map is available (Fig. 1.25.1). Typically, the study sample is genotyped with a commercial genotyping platform for hundreds of thousands to millions of single nucleotide polymorphisms (SNPs) located across the entire genome.

The HapMap Consortium database (International HapMap 3 Consortium, 2010) has commonly served as the reference panel for most of the GWAS published to date, but its use is now being replaced by the larger and more comprehensive set of individuals characterized within the 1000 Genomes Project (1KG; 1000 Genomes Project Consortium, 2012). Indeed, while the HapMap set characterized 270 individuals with genotyping arrays for ~3 million markers, the 1KG reference set has been generated from whole-genome sequencing of 1,092 individuals (181 samples from Admixed American, 246 from African, 286 from East Asian, and 379 from European ancestry groups), leading to the discovery of ~39.7 million bi-allelic variants; ~1.4 million markers are short indels and large deletions, and the rest are SNPs. Imputation performed with this much denser data set will yield a higher resolution of the genome for detection of association signals, thus increasing the power of the existing GWAS to identify novel variants beyond what was found after imputation with the HapMap data set and to pinpoint the causal variants at known associated loci (Huang et al., 2009).

Currently, there are several programs able to perform genotype imputation (e.g., PLINK, BEAGLE, MaCH+minimac, fastPHASE, and IMPUTE2), each implementing different algorithms and with different limitations and accuracy (Scheet and Stephens, 2006; Purcell et al., 2007; Browning and Browning, 2009; Howie et al., 2009, 2012; Li et al., 2010). PLINK and BEAGLE are

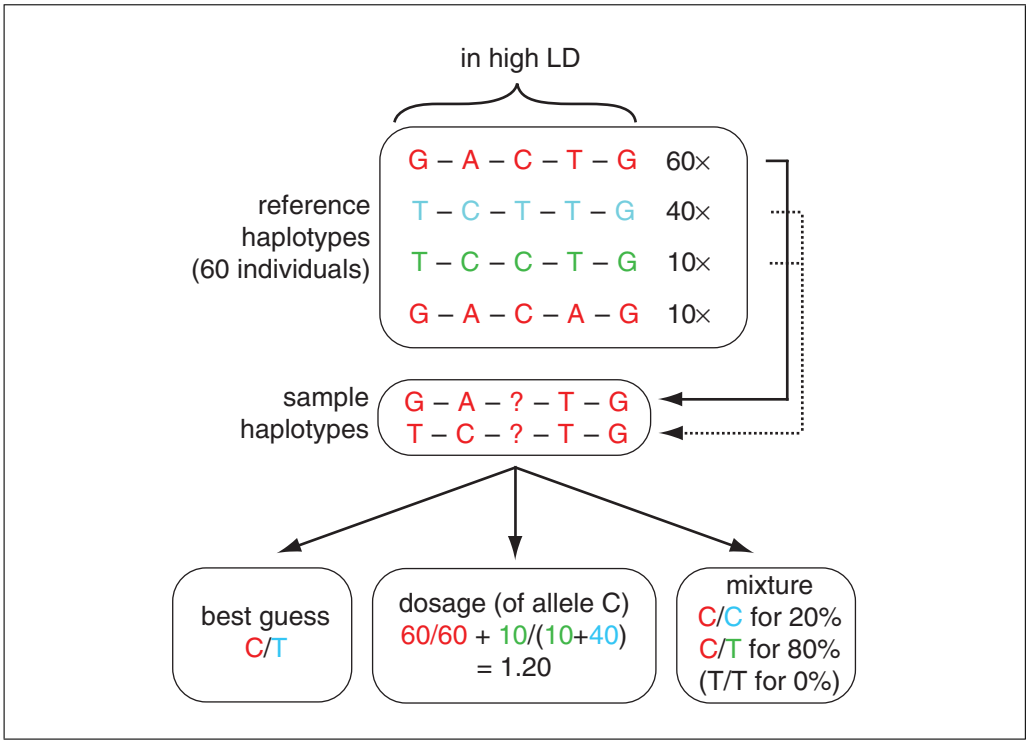

Figure 1.25.1 Imputation of a missing sample genotype using a reference haplotype panel. The most common reference haplotype can be unambiguously assigned to the upper sample haplotype, while two reference haplotypes come into consideration for the lower sample haplotype. Three methods to describe the imputed genotype are shown: best guess, dosage, and mixture. Only the dosage and mixture methods take into account the uncertainty present.

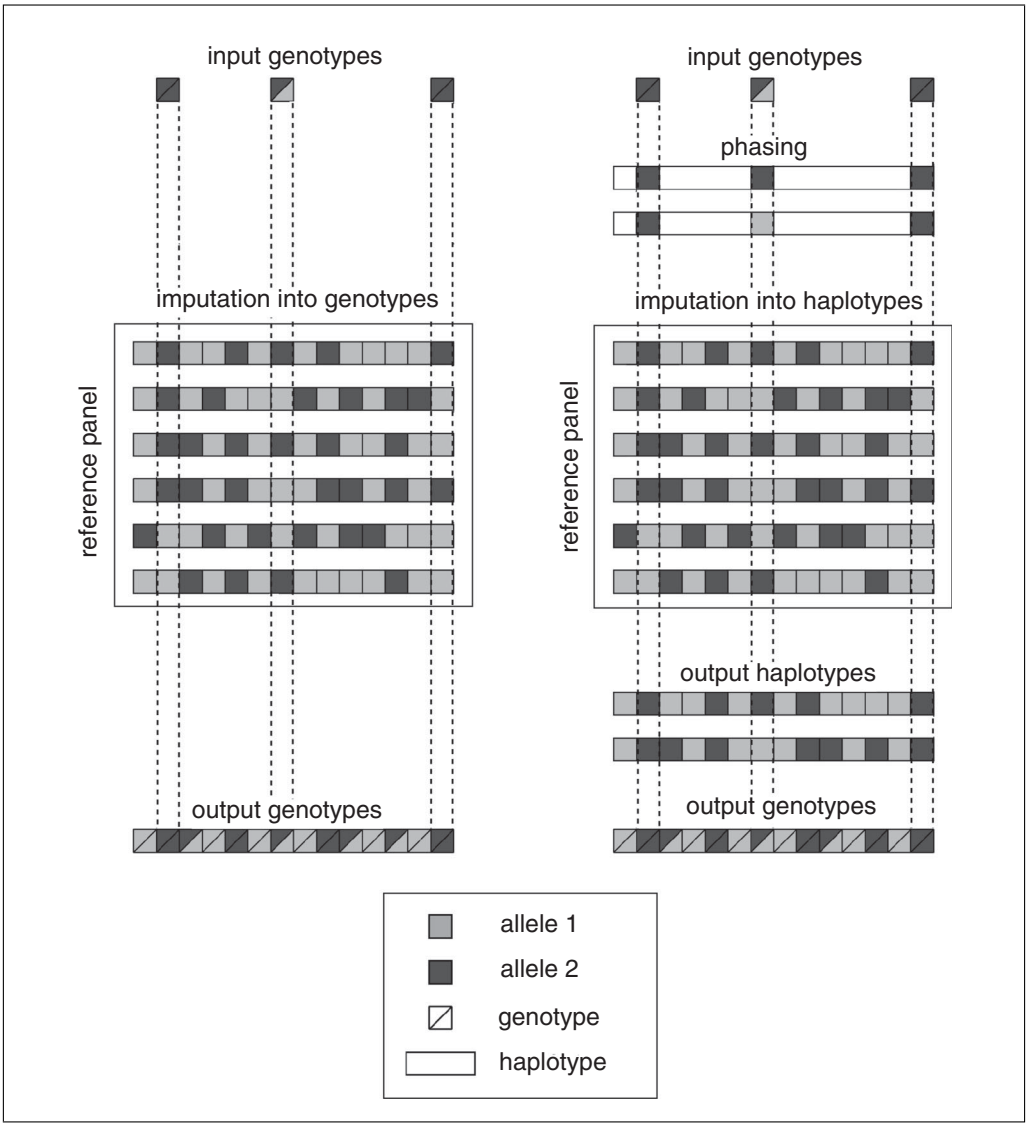

Figure 1.25.2 Comparison of the single-step imputation (left) and two-step imputation (right) approaches. Although the input and reference panels are identical for both approaches, the main difference is speed. Identification of the reference haplotype combination(s) that can best explain the input genotypes (single-step imputation) takes longer than identification of the two reference haplotypes that most likely represent the input haplotypes (phased genotypes; two-step imputation). The relative positions of genotyped variants are indicated by dashed lines.

computationally more efficient because they focus on genotypes for a relatively small number of neighboring markers when imputing each missing genotype. IMPUTE2, MaCH, and fastPHASE are computationally more intensive but provide a better estimate of missing genotypes because they take into account all available markers when imputing each missing genotype. This strategy improves imputation accuracy, particularly for rare variants.

To balance the increasing computational burden necessary for large, dense reference panels such as 1KG, the developers of IMPUTE2 and MaCH+minimac each introduced the two-step imputation process, which consists of an initial prephasing (i.e., haplotype estimation) of the GWAS genotypes and a subsequent imputation of reference panel markers into the estimated study haplotypes (Fig. 1.25.2).

This separation of the computationally intensive phasing from the data-intensive imputation can substantially reduce computation time, compared with the one-step imputation, which simultaneously estimates missing genotypes and considers the uncertainty of the phase of SNPs. For some populations the imputation quality might be slightly lower in the two-step approach, compared with their specific one-step algorithms (e.g., for African

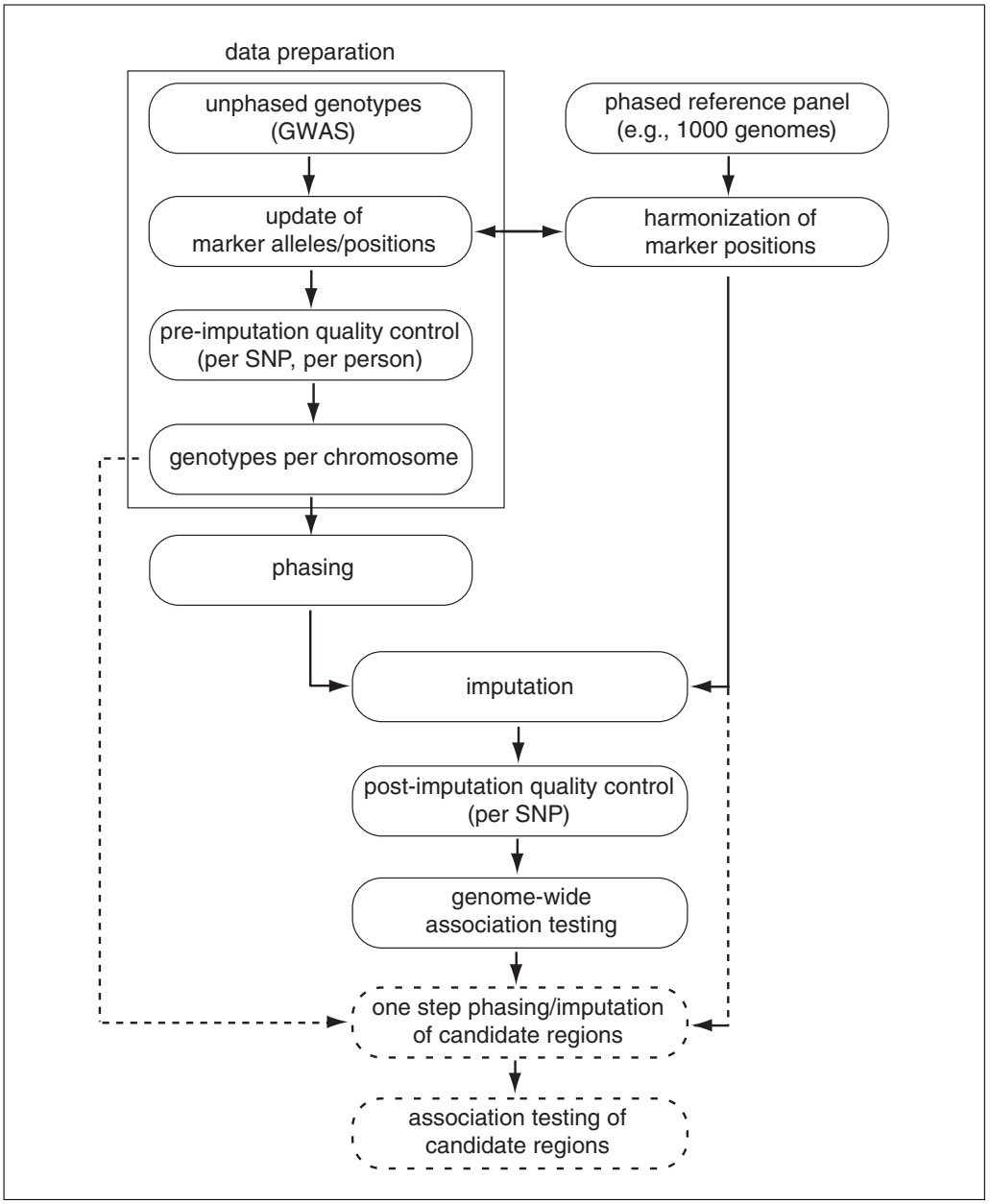

Figure 1.25.3 Workflow for the individual steps required to successfully perform a two-step imputation. An optional one-step imputation of candidate regions is indicated by the boxes with dashed lines.

American populations). However, the time saving usually outweighs the decrease in imputation quality (Fig. 1.25.2). Moreover, it is still possible to specifically and more accurately impute the genotypes of target regions in a single step once such targets are found.

The computational burden of the prephasing step increases quadratically (using IMPUTE2 or MaCH) with the number of study haplotypes, while it increases linearly with the number of reference panel haplotypes in the imputation step. Thus, this two-step approach is well suited for the latest and still growing reference panels that feature denser marker sets and larger sample sizes. Moreover, once imputation is done for a sample study and an imputation with updated reference panels is desired (e.g., from HapMap or 1KG projects or to an updated release), the time-consuming phasing can be skipped, and the phased study genotypes at hand can be directly used for the imputation.

The two-step imputation procedure suitable for IMPUTE2 and MaCH+minimac provides both the accuracy and power to impute common, as well as rare markers of large studies and the latest reference panels (Howie et al., 2012). Here the focus will be

on the general practice for genotype imputation using a two-step imputation on the basis of MaCH+minimac. Because of their comparable approaches, the introduced MaCH+minimac imputation is largely transferable to IMPUTE2. Detailed documentation and example scripts for both platforms can be found in the "1000 Genomes Imputation Cookbook" of MaCH+minimac and IMPUTE2 (see Internet Resources). A schematic workflow is represented in Figure 1.25.3, and a detailed overview is presented below.

DATA PREPARATION

Preparing the data for phasing and imputation is a crucial step of the analysis. Inaccurately cleaned genotype data can lead to false-positive and false-negative associations. A good starting point is to format the genetic data in the pedigree file format (pedfile). This file has a row for each individual, where the first five columns contain the pedigree information (IDs of family, individual, father, mother, and sex), whereas the other columns are usually genotypes. As a side note, because the phasing procedure ignores relatedness between individuals, father and mother IDs could also be set to 0 for all samples. In addition, all individuals with available genotypes should be included in the pedfile, regardless of their phenotypic assessment. The pedfile is always associated with a file (datfile or mapfile) containing the description of the columns following the pedigree information of the pedfile; in this case, it will contain the list of genetic markers.

When creating a pedfile, it is important to align the genotypes to the same strand as the reference panel chosen for the imputation; for example, the latest 1KG panel release is all mapped to forward strand of NCBI build 37/hg19. MaCH+minimac offers an autoflip option to identify possible allele strand flips by looking at label match and frequencies; however, this option is unable to identify flips for very common ambiguous variants (A/T or C/G). As most genotyping platforms generally avoided such strand-ambiguous alleles in their designs, the negative impact of such flipped ambiguous alleles will be minimal when using the autoflip option. Alternatively, overall exclusion of ambiguous variants might help to prevent errors of flipped alleles, although it should be carefully considered because of the consequent reduction of marker density.

Once the files are formatted, several software tools can be used (e.g., PEDSTATS, Merlin, PLINK, or PedCheck) to perform quality control (QC; O'Connell and Weeks, 1998; Abecasis and Wigginton, 2005; Wigginton and Abecasis, 2005; Purcell et al., 2007; see *UNIT 1.19*). While it has been shown (Southam et al., 2011) that the imputation accuracy does not appear to be substantially affected by a GWAS QC step, this observation is valid only for common variants and may not be generalized to the imputation of low frequency (1% to 5% minor allele frequency; abbreviated MAF) and rare variants (<1% MAF).

The best practice procedures of GWAS QC are described in detail in *UNIT 1.19* (Turner et al., 2011). Usually, before phasing and imputation, QC filters are applied to the samples and to the markers. Filters on samples typically include removing samples with low call rates, or that are duplicated, and inconsistencies of genotypes on X and Y chromosomes with reported gender. Careful inspection of batch effects and presence of population stratification should also be taken into account. Filters on markers include removing SNPs with low call rates, very low frequency, Mendelian inconsistencies, and deviation from Hardy-Weinberg equilibrium (HWE).

In general, QC checks are platform- and study-specific, and the most appropriate filters have to be identified for each sample data individually. Additional filters may be necessary. For example, for a case-control study it would be suitable to remove the SNPs with high call-rate differences between cases and controls or to restrict the check on deviation from HWE only in the control sample; in fact, a deviation from HWE in the case sample could be a signal of association (see *UNIT 1.18*). Similarly, the corresponding thresholds are platform- and study-specific; therefore, performing general statistical checks to validate the QC procedure, e.g., quantile–quantile (Q–Q) analyses and evaluation of inflation of association statistics, is recommended. Depending on the observed indication for population stratification, adjusting the output χ^2 and p values by genomic controls or by principal component analysis might be required. Because these indicators might be distorted in the imputed data set, they should be determined beforehand with only genotyped markers to evaluate appropriate QC filters.

Using unified names or naming schemes for identical markers is recommended (e.g., rs429608, rs116503776, chr6:31930462, and 6:31930462 represent the same SNP) because it enables matching of markers between the data sets by name. Numeric marker

annotations ("<chromosome>:<position>") are sometimes used by public sets of reference haplotypes or by resequencing analyses whose detected variants do not match to the National Center for Biotechnology Information (NCBI) single nucleotide polymorphism database (dbSNP; *http://www.ncbi.nlm.nih.gov/SNP*) or have not yet received a reference ID (mostly insertions and deletions). One solution is to convert all marker names in the GWAS data to their numeric annotations. Alternatively, minimac's alias file, which contains all marker name pairs that might have been labeled differently between the data sets, can be used for name harmonization in the imputation step. In general, markers that are duplicated or map to multiple positions in the genome should be removed from all data sets.

Finally, the recoded and formatted data files need to be sorted in ascending chromosomal order and split by chromosome, because phasing will be performed one chromosome at a time. Because of the pseudoautosomal regions PAR1 (chrX:60,001-2,699,520/chrY: 10,001-2,649,520)/PAR2 (chrX:154,931,044-155,260,560/chrY:59,034,050-59,363,566) and the hemizygosity of nonpseudoautosomal regions of male individuals, the preparation of X-chromosomal data (if present) requires additional care. Usually, the recommendation is to phase females and males separately, as well as to split the X chromosome into nonpseudoautosomal and pseudoautosomal regions.

STEP 1: PREPHASING

Before running imputation, the genotypes of the GWAS individuals will be phased, i.e., their most likely haplotypes will be estimated. If genotypes of unrelated individuals are phased, the estimated haplotypes are less likely to represent the true allelic configurations of the corresponding chromatids, but rather represent the best-guess mosaic of both chromosomes that underwent numerous chromosome crossovers. These crossovers represent assumed recombination events that disrupted regions of linkage disequilibrium in the common history of the study individuals. Consequently, the lengths of these chromosomal stretches are smaller in outbred than in founder populations, and for this reason, in founder populations the haplotype estimation may be more accurate.

To phase genotypes for a group of individuals, MaCH, which uses a hidden Markov chain, can be run with the option `–phase`. Two key parameters, rounds and states, can lead to more accurate results if appropriately set. The rounds parameter represents the number of iterations that the Markov sampler uses for haplotyping (at least 20 rounds recommended). The states parameter represents the number of sampled haplotypes in each iteration (at least 200 states recommended). Larger numbers lead to better quality but can require much more computing time and memory usage. The computational cost of the prephasing step increases quadratically with the number of states and linearly with the number of the rounds. If substantial computing resources are available, 600 or even 800 states should be considered to obtain optimal results.

Genotypes of males on the X and Y chromosomes at nonpseudoautosomal regions are already phased because of their hemizygosity and just need to be manually converted into the haplotype format (see Internet Resources for the "Minimac: 1000 Genomes Imputation Cookbook").

Besides large genome association analysis tool sets, there are small helper tools or example shell codes available that can conveniently generate some additionally required input files (e.g., `*.map file into *.dat and *.snp files`) and can efficiently split chromosomes into overlapping chunks or offer quick manipulations of huge files. For a more memory-efficient and parallel processing, we recommend ChunkChromosome, a freely available tool (*http://genome.sph.umich.edu/wiki/ChunkChromosome*; Table 1.25.2) that automatically splits each chromosome into overlapping chunks, allowing imputation of chromosomes to be run in multiple, simultaneous, lower memory–demanding jobs. Because overlapping markers are selected between the chunks, Chunk Chromosome circumvents decreasing imputation accuracy at the chunk borders.

STEP 2: IMPUTATION

In this step, the reference panel genotypes are imputed into the phased genotypes of the GWAS sample. The imputation into haplotypes takes significantly less time than the imputation into genotypes (one-step imputation; see Fig. 1.25.2). However, the whole process is still time consuming and largely depends on the size of the available computing cluster. As with phasing, performing the imputation in chunks is recommended, as this will be more memory efficient and faster. Furthermore, if the ChunkChromosome tool is used, it interfaces with minimac to ensure SNPs

that overlap between chunks are only imputed once.

There are three ways to describe the imputed genotypes: best-guess genotype/haplotype, allele dosages, and genotype probabilities (Fig. 1.25.1). The standard output files of minimac contain, for each individual and for each marker, the allele dosages of the reference allele (Al1). The dosage is the expected number of copies of Al1 and is a real number between 0 and 2. Al1 is an arbitrary allele, and typically, it is the first allele read in the reference haplotypes. The specific allele Al1 for each marker is stored in the info file generated by minimac.

If the option `–probs` is selected, minimac will also output a file that contains two columns with the probabilities for the homozygous and heterozygous states of allele 1, respectively. This option is turned off by default because the resulting file may generate several additional gigabytes of data.

The output file size depends on the marker density of the reference panel, number of study individuals, and output format, but in general, imputed GWAS data becomes very large, especially when using the 1KG reference panel. Directly compressing the output file using the option `–gzip` is recommended. The compressed dosage and probability format files for 1,000 individuals with 8 million markers are ~10 GB and ~15 GB, respectively, whereas these numbers will be four to five times as large when using a more recent 1000 Genomes Project reference panel of ~40 million markers (2012-03-14 release).

If desired, it is also possible to output the best-guess haplotypes, which represent the alleles with the highest posterior probability of each chromosome. However, treating these estimated genotypes as true genotypes in subsequent analyses will lead to misleading results, especially for poorly imputable rare variants that are likely to be called homozygous for the common allele. It is important to remember that the imputation is not perfect, so use of the dosages (or probabilities) to account for the uncertainty of genotypes is highly recommended.

Reported run times for the imputation step using the European ancestry group from the 1000 Genomes Project (379 individuals; 37.4 million SNPs) as reference was ~24 CPU min per study individual (Howie et al., 2012), when using IMPUTE2. For minimac, a rough estimate for the imputation run time is ~1 hr to impute 1 million markers in 1,000 individuals using a reference panel with 100 haplotypes. For example, imputation of 40 million markers in 1,000 individuals using the European ancestry group would take ~300 hr or 12 to 13 days on a modern single-core machine. These estimates are approximate because they may change in reality, given cluster settings (e.g., disk access speed or memory conflict with parallel jobs on the same core; see the minimac reference in Internet Resources for the corresponding formula).

MEASURING IMPUTATION QUALITY

The imputation quality is commonly measured with a parameter called Rsq, i.e., the estimate of the squared correlation between imputed and true genotypes or, in other words, the ratio of the variances of imputed and true allele counts. This parameter is calculated and stored in the info file. Nonmonomorphic variants with Rsq>0.3 are usually considered as successfully imputed variants. Most of currently published GWAS and meta-analysis papers that used the MaCH software for imputation have chosen this threshold to discard poorly imputed (common) variants (Scott et al., 2007; de Bakker et al., 2008; Sanna et al., 2008). However, one should be aware that most if not all of these studies focused on common variants, and a more stringent threshold might be required for rare variants. In fact, some studies have chosen to use more stringent thresholds, e.g., Rsq>0.50 (Meschia et al., 2011) or variable Rsq thresholds, depending on allele frequency (Liu et al., 2012).

There are several factors that affect imputation quality, and in this section, we focus on three factors: choice of reference panel, quality of input genotypes or haplotypes, and number of genotypes in input. To evaluate the impact of these aspects, we simulated haplotypes of 1,000 unrelated individuals of European ancestry at SNPs present in the OmniExpress array on chromosome 20 (17,250 SNPs) with the software HAPGEN (Su et al., 2011) and then performed several runs of prephasing (MaCH) and imputation (minimac) using different settings.

To assess and compare imputation accuracy, we looked at mean Rsq values and the proportion, as well as the total number, of variants that were considered successfully imputed (Rsq>0.3). In this context, it should be noted that the accuracy of predicted Rsq values is, in general, high for common variants, but decreases with frequency, thereby limiting the

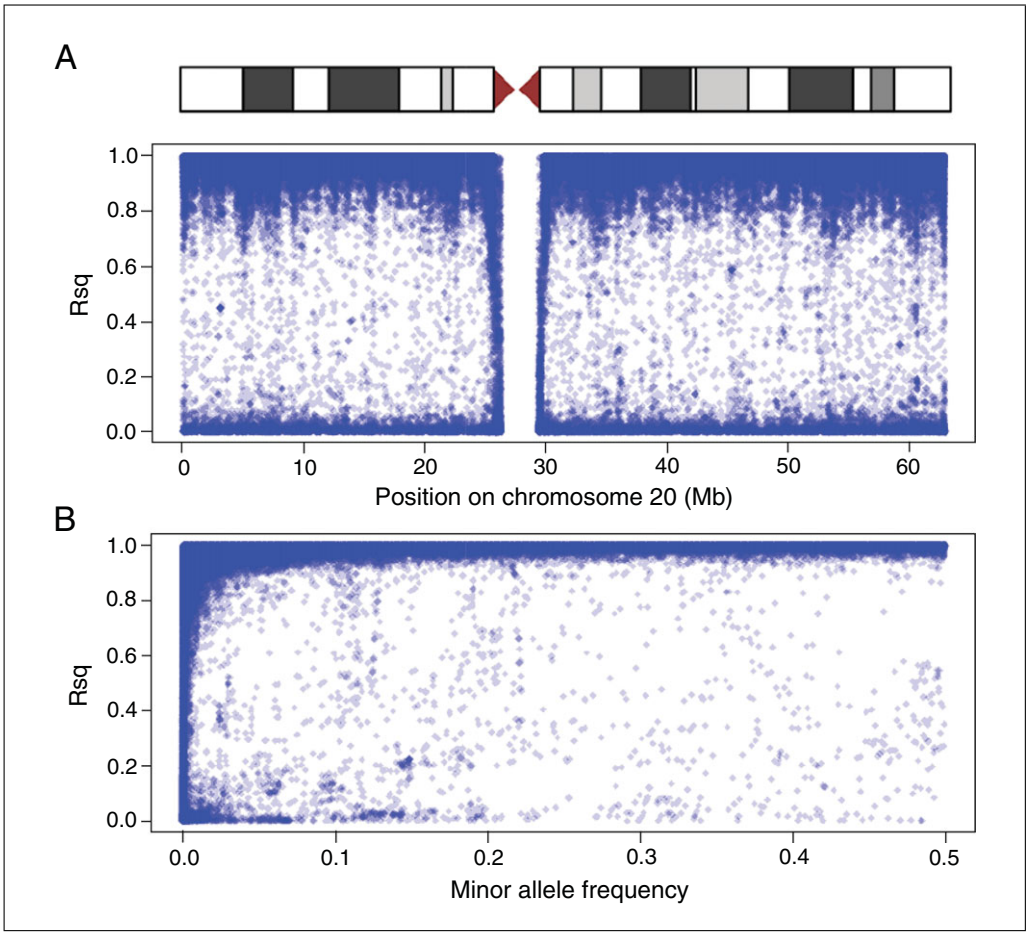

Figure 1.25.4 Imputation accuracy for imputed SNPs on chromosome 20. Imputation quality was measured by MaCH's estimated Rsq obtained from a 1000G-EUR panel imputation of a sample of 1000 unrelated European individuals. (**A**) Imputation accuracy plotted versus physical position. The corresponding chromosomal ideogram is shown above the plot. (**B**) Imputation accuracy plotted versus minor allele frequency.

applicability of such general filters, especially for rare variants.

Choosing the Best Reference Panel

The genotype imputation technique uses LD patterns to infer untyped markers; thus, the ideal reference panel includes individuals selected from the same population as the study samples. We phased the simulated genotypes using 25 rounds and 300 states and performed imputation using the total reference panel of the 1KG project (ALL, n=1,092), as well as its subsets stratified by European (EUR, n=379), East Asian (ASN, n=286), African (AFR, n=246), and Admixed American (AMR, n=181) ancestry groups. More details about the 1KG samples can be found on the 1000 Genomes Project Web page (see Internet Resources).

Because imputation quality is homogeneous over the chromosome, except for the region around the centromere (Fig. 1.25.4A), but not over the allele frequency spectrum (Fig. 1.25.4B), we grouped variants in different MAF bins. As expected, considering that we simulated haplotypes of European ancestry, the 1KG-EUR panel provided the highest imputation quality among all ethnicity-specific panels, with remarkable differences for low-frequency and rare variants (MAF<5%; Fig. 1.25.5 and Table 1.25.1). Although the overall imputation accuracy was very comparable between the ALL and the EUR panel imputation (Fig. 1.25.5), the imputation with the ALL panel resulted in slightly more successfully imputed variants than with the EUR panel (332,415 and 330,835 variants, respectively; Table 1.25.1). Notably, this marginal gain of variants was achieved by an approximately three-fold increased computational burden of the imputation step. A possible explanation for this observation is that these additional variants were relatively rare or not present in the EUR panel, but more

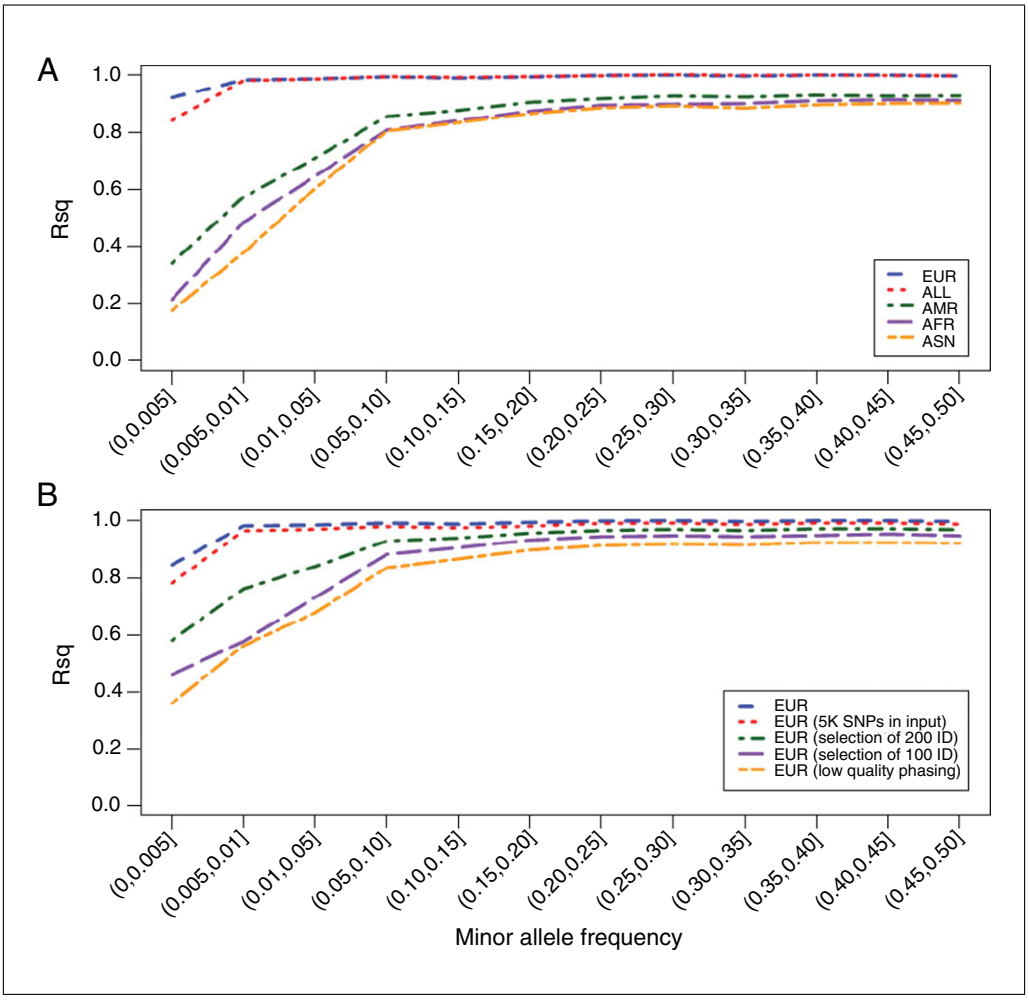

Figure 1.25.5 Factors influencing the imputation quality. All imputations were performed on a sample of European ancestry and were based on 17,250 genotyped SNPs phased with 25 rounds and 300 states, unless otherwise stated. The total 1KG reference panel of 1,092 individuals (ALL, *n*=1,092) was partially stratified by European (EUR, *n*=379), East Asian (ASN, *n*=286), African (AFR, *n*=246), and Admixed American (AMR, *n*=181) ancestry groups. (**A**) Mean Rsq values of successfully imputed SNPs versus minor allele frequency bins. Mean Rsq is calculated on SNP that were present in all panels (159,438 SNPs). (**B**) Mean Rsq values of successfully imputed SNPs versus minor allele frequency bins for different imputations with the EUR reference panel (*n*=379) using the reduced marker density (5,750 SNPs in study sample), reduced EUR reference panel size (*n*=200 and *n*=100, respectively), or low-quality phasing (1 round per 100 states). Details can also be found in Table 1.25.1.

frequent in other ancestry groups whose additional haplotype information increased imputation quality.

Clearly, genetic similarity between study samples and reference populations leads to better quality. However, if the ancestry of the study sample is unknown or is a known mixture of different ancestries, there are, in general, three options to be considered: (1) use a reference panel that has not been subdivided in ancestry groups (e.g., 1KG-ALL); (2) perform imputation of a defined genomic region on each available reference panel, and then select the population that offers the highest imputation quality; or (3) use a study-specific reference panel generated from a subset of the study population by resequencing or genotyping it on a high-density SNP array. The last approach can be particularly efficient if applied to isolated populations (Kong et al., 2008). However, one should carefully balance the pros (availability of population-specific variants, same haplotypic background) and the cons (additional costs and time), as well as evaluate how many samples and what coverage is actually needed to improve the performance of an imputation over the current 1KG panel.

In addition to genetic similarity, larger reference panels may lead to more accurate

Table 1.25.1 Comparison of Average Imputation Accuracies and Variant Numbers Obtained from Different Phasing and/or Imputation Settings

Reference panel	Reference panel (*n*)	Mean Rsq[a]			Number of variants[a]		
		Rare MAF≤1%	Less common 1%<MAF≤5%	Common MAF>5%	Rare MAF ≤ 1%	Less common 1%<MAF≤5%	Common MAF>5%
Phasing: 25 round 300 states; 17,250 genotyped SNPs							
ALL	1,092	0.962 (0.496)	0.988 (0.976)	0.995 (0.990)	140,538 (277,177)	61,954 (62,769)	129,923 (130,573)
AFR	246	0.580 (0.095)	0.662 (0.484)	0.870 (0.845)	31,741 (333,144)	49,903 (77,287)	130,606 (135,415)
AMR	181	0.646 (0.162)	0.747 (0.602)	0.911 (0.889)	41,256 (225,439)	50,034 (66,352)	127,027 (130,929)
ASN	286	0.625 (0.069)	0.727 (0.426)	0.881 (0.846)	7,843 (132,212)	20,379 (40,311)	119,217 (125,350)
EUR	379	0.963 (0.869)	0.987 (0.979)	0.995 (0.989)	139,119 (154,594)	61,931 (62,415)	129,785 (130,645)
EUR	200	0.731 (0.632)	0.841 (0.834)	0.955 (0.949)	84,044 (99,776)	63,410 (63,967)	130,315 (131,292)
EUR	100	0.660 (0.509)	0.762 (0.725)	0.929 (0.919)	43,412 (61,486)	59,649 (63,876)	130,181 (131,693)
Phasing: 25 round 300 states; 5,750 genotyped SNPs							
EUR	379	0.939 (0.814)	0.973 (0.964)	0.988 (0.979)	137,955 (159,780)	62,660 (63,285)	139,632 (140,961)
Phasing: 1 round 100 states; 17,250 genotyped SNPs							
EUR	379	0.572 (0.392)	0.706 (0.671)	0.894 (0.885)	99,063 (165,152)	57,643 (62,019)	129,389 (131,106)

[a] Applied quality filter was Rsq > 0.3; values for the unfiltered sets are given in parentheses. Abbreviations: ALL, all individuals from the current 1000 Genomes release; AFR, African; AMR, Admixed American; ASN, East Asian; EUR, European; MAF, minor allele frequency.

imputation by increasing the chance to find perfect matches for the study haplotypes and possibly allow the imputation of rare, haplotype-specific variants. For example, if we select only 100 or 200 haplotypes from the European reference panel, the quality decreases, especially at low frequencies (Fig. 1.25.5B). Thus, we recommend using the most updated reference panel release, which will likely contain not only more individuals, but also more variants.

Quality of Input Genotypes/Haplotypes

As mentioned in the previous section, performing the standard battery of QC filters before phasing, as well as running phasing with the recommended parameters, is highly recommended. To assess the impact of inaccurate haplotyping, we re-ran phasing with only 1 round and 100 states, followed by imputation with the 1KG-EUR reference panel, and compared the results with the ones of the previous setting of 25 rounds and 300 states (Table 1.25.1).

Whereas lower imputation quality was observed in all frequency bins, the accuracy most severely dropped for less frequent and rare variants. Although the number of successfully imputed SNPs (286,095) was higher compared with any imputation results of genetically more distant populations (≤218,317), their Rsq mean values were comparable (Fig. 1.25.5 and Table 1.25.1).

Number of Genotypes in Input

Imputation quality usually improves with the number of SNPs genotyped in the study by providing a refined definition of study haplotypes and by increasing the chance to discriminate between multiple similar reference haplotypes. In our example, we performed imputation by selecting 5,750 of the total set of 17,250 genotyped markers, and we again

evaluated the resulting imputation accuracy. To preserve a similar coverage over the chromosome, we selected one out of every three markers. Although the total number of successfully imputed variants were slightly higher in the sparser data set compared with the denser data set (340,247 and 332,415, respectively), their mean Rsq values were lower, especially for the rare variants (Rsq=0.939 and Rsq=0.963, respectively; Fig. 1.25.5 and Table 1.25.1).

The results obtained for the EUR panel show that if the reference panel is genetically close to the study sample, it is possible to have a good coverage of well imputed variants even if we use a sparser set of genotypes. Coverage is thus important, but it has only a moderate impact with respect to the choice of the reference panel and the accuracy of prephasing. However, lower quality is expected when the markers are reduced but are not homogenous along the chromosome, e.g., for custom arrays where SNPs are located in genes of interest (Voight et al., 2012).

ASSOCIATION TESTING

After postimputation quality control and filtering, the imputed genotypes can be tested for trait associations. Rounding the obtained allele dosages to integers (i.e., selecting the most likely or best-guess genotype; Fig. 1.25.1) would enable the application of the whole spectrum of classic genetic association tests. However, the obtained results, especially of variants with lower imputation quality and lower allele frequencies, should be treated with extreme care. Instead, the recommendation is to account for the genotype uncertainty by analyzing the obtained allele dosages or genotype probabilities (Fig. 1.25.1) as continuous variables in a regression model.

Table 1.25.2 Software Programs that Aid in Data Management for Imputation

Tool or software (Web page)	Typical usage
PLINK (*http://pngu.mgh.harvard.edu/~purcell/plink*)	Whole-genome association analysis.
Merlin (*http://www.sph.umich.edu/csg/abecasis/merlin/index.html*)	Whole-genome association analysis.
PEDSTATS (*http://www.sph.umich.edu/csg/abecasis/PedStats/index.html*)	Quick validation and summary of any pair of pedigree (`.ped`) and data (`.dat`) files.
PedCheck (*http://www.genomeutwin.org/member/cores/stat/linkage/pedcheck.html*)	Quality control of pedigree files, detects marker typing incompatibilities in pedigree data.
liftOver (*http://genome.ucsc.edu/cgi-bin/hgLiftOver*)	Conversion of positions between different genome assemblies. (Download the liftOver executable and the appropriate chain file.)
ChunkChromosome (*http://genome.sph.umich.edu/wiki/ChunkChromosome*)	Splitting chromosomes into overlapping chunks.
GTOOL (*http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html*)	Generation of subsets of genotype data, conversion of different formats of genotype data (e.g., pedigree file to IMPUTE format and vice versa), merging of genotype data sets, orientation of genotype data according to a strand file.
VCFtools (*http://vcftools.sourceforge.net*)	Reduction of the complete reference panel to certain individuals, e.g., from a certain ancestry group (e.g., AMR, AFR, ASN, and EUR) or extraction of the haplotypes of a candidate region for a targeted one step imputation.
UNIX command-line utilities	Handling and modification of very large files, e.g., awk, cut, sed, or sort.
SNP and indel imputability (*http://www.unc.edu/~yunmli/1000G-imp*)	Estimation of the ability to impute a candidate SNP or the variants of a given region of interests based on the standard SNP set of several genotyping arrays.
prob2plink (*http://www.sph.umich.edu/csg/yli/prob2plink.V001.tgz*)	Conversion of MaCH prob+info output into PLINK dosage file.

Most of the imputation-based analysis software tools can handle allele dosages or genotype probabilities, but they often require a specific input format for genotype probabilities and the marker information. The common imputation platforms are accompanied by analyses tools that can directly use the corresponding and compressed output files. For the MaCH+minimac these are MACH2QTL (quantitative traits) or MACH2DAT (discrete traits; see Internet Resources). The use of alternative imputation software might offer more flexibility, but it often comes with the requirement to unpack and/or reformat the generally huge output files. Instructions on how to load imputation data to the different software tools can usually be found on the developers' Web pages, while a quick online search might reveal powerful conversion tools (Table 1.25.2).

Finally, in case interesting candidate regions are identified and a more accurate phasing and imputation of these regions is intended, it might be helpful to repeat the imputation of these target regions in a single step directly into the reference panel genotypes and use more rounds and more states (Fig. 1.25.3). Although the resulting gain in imputation quality and thus power compared with the two-step imputation might be small (Howie et al., 2012), it could be crucial for borderline association signals.

CONCLUSIONS

Genotype imputation has become a fundamental tool for genetic analyses. At no cost, it improves the power of GWAS by increasing the resolution of each single study and by allowing direct comparison of findings. After its initial use for imputation of HapMap data, the approach rapidly changed to handle large amounts of data generated with the advent of next-generation sequencing technologies. We can envisage future modifications that increase accuracy at rare and structural variants and further reduce computational time. Here, together with a baseline tutorial, we have described the main factors affecting imputation quality and their impacts on the full allele frequency spectrum. It is clear that the choice of the reference panel plays a fundamental role in the process; thus, we expect that future releases of the 1000 Genomes Project, with more individuals from diverse populations, as well as the study of specific reference panels, will enhance the discovery and dissection of genetic components of complex diseases.

LITERATURE CITED

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56-65.

Abecasis, G.R. and Wigginton, J.E. 2005. Handling marker-marker linkage disequilibrium: Pedigree analysis with clustered markers. *Am. J. Hum. Genet.* 77:754-767.

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.

de Bakker, P.I., Ferreira, M.A., Jia, X., Neale, B.M., Raychaudhuri, S., and Voight, B.F. 2008. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* 17:R122-R128.

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.

Howie, B.N., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* 44:955-959.

Huang, L., Li, Y., Singleton, A.B., Hardy, J.A., Abecasis, G., Rosenberg, N.A., and Scheet, P. 2009. Genotype-imputation accuracy across worldwide human populations. *Am. J. Hum. Genet.* 84:235-250.

International HapMap 3 Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52-58.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C., and Hoh, J. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* 308:385-389.

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

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.

Liu, E.Y., Buyske, S., Aragaki, A.K., Peters, U., Boerwinkle, E., Carlson, C., Carty, C., Crawford, D.C., Haessler, J., Hindorff, L.A., Marchand, L.L., Manolio, T.A., Matise, T., Wang, W., Kooperberg, C., North, K.E., and L,i, Y. 2012. Genotype imputation of Metabochip SNPs using a study-specific reference panel of ∼4,000 haplotypes in African Americans from the Women's Health Initiative. *Genet. Epidemiol.* 36:107-117.

Meschia, J.F., Nalls, M., Matarin, M., Brott, T.G., Brown, R.D. Jr., Hardy, J., Kissela, B., Rich, S.S., Singleton, A., Hernandez, D., Ferrucci, L., Pearce, K., Keller, M., and Worrall, B.B. 2011. Siblings With Ischemic Stroke Study Investigators. Siblings with ischemic stroke study: Results of a genome-wide scan for stroke loci. *Stroke.* 42:2726-2732.

Metzker, M.L. 2010. Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11:31-46.

O'Connell, J.R. and Weeks, D.E. 1998. PedCheck: A program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.* 63:259-266.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81:559-575.

Sanna, S., Jackson, A.U., Nagaraja, R., Willer, C.J., Chen, W.M., Bonnycastle, L.L., Shen, H., Timpson, N., Lettre, G., Usala, G., Chines, P.S., Stringham, H.M., Scott, L.J., Dei, M., Lai, S., Albai, G., Crisponi, L., Naitza, S., Doheny, K.F., Pugh, E.W., Ben-Shlomo, Y., Ebrahim, S., Lawlor, D.A., Bergman, R.N., Watanabe, R.M., Uda, M., Tuomilehto, J., Coresh, J., Hirschhorn, J.N., Shuldiner, A.R., Schlessinger, D., Collins, F.S., Davey Smith, G., Boerwinkle, E., Cao, A., Boehnke, M., Abecasis, G.R., and Mohlke, K.L. 2008. Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat. Genet.* 40:198-203.

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.

Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U., Prokunina-Olsson, L., Ding, C.J., Swift, A.J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X.Y., Conneely, K.N., Riebow, N.L., Sprau, A.G., Tong, M., White, P.P., Hetrick, K.N., Barnhart, M.W., Bark, C.W., Goldstein, J.L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T.A., Watanabe, R.M., Valle, T.T., Kinnunen, L., Abecasis, G.R., Pugh, E.W., Doheny, K.F., Bergman, R.N., Tuomilehto, J., Collins, F.S., and Boehnke, M. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316:1341-1345.

Southam, L., Panoutsopoulou, K., Rayner, N.W., Chapman, K., Durrant, C., Ferreira, T., Arden, N., Carr, A., Deloukas, P., Doherty, M., Loughlin, J., McCaskie, A., Ollier, W.E., Ralston, S., Spector, T.D., Valdes, A.M., Wallis, G.A., Wilkinson, J.M., arcOGEN Consortium, Marchini, J., and Zeggini, E. 2011. The effect of genome-wide association scan quality control on imputation outcome for common variants. *Eur. J. Hum. Genet.* 19:610-614.

Su, Z., Marchini, J., and Donnelly, P. 2011. HAPGEN2: Simulation of multiple disease SNPs. *Bioinformatics* 27:2304-2305.

Turner, S., Armstrong, L.L., Bradford, Y., Carlson, C.S., Crawford, D.C., Crenshaw, A.T., de Andrade, M., Doheny, K.F., Haines, J.L., Hayes, G., Jarvik, G., Jiang, L., Kullo, I.J., Li, R., Ling, H., Manolio, T.A., Matsumoto, M., McCarty, C.A., McDavid, A.N., Mirel, D.B., Paschall, J.E., Pugh, E.W., Rasmussen, L.V., Wilke, R.A., Zuvich, R.L., and Ritchie, M.D. 2011. Quality control procedures for genome-wide association studies. *Curr. Protoc. Hum. Genet.* 68:1.19.1-1.19.18.

Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J., Frayling, T.M., Heid, I.M., Jackson, A.U., Johnson, T., Kilpelainen, T.O., Lindgren, C.M., Morris, A.P., Prokopenko, I., Randall, J.C., Saxena, R., Soranzo, N., Speliotes, E.K., Teslovich, T.M., Wheeler, E., Maguire, J., Parkin, M., Potter, S., Rayner, N.W., Robertson, N., Stirrups, K., Winckler, W., Sanna, S., Mulas, A., Nagaraja, R., Cucca, F., Barroso, I., Deloukas, P., Loos, R.J., Kathiresan, S., Munroe, P.B., Newton-Cheh, C., Pfeufer, A., Samani, N.J., Schunkert, H., Hirschhorn, J.N., Altshuler, D., McCarthy, M.I., Abecasis, G.R., and Boehnke, M. 2012. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* 8:e1002793.

Wetterstrand, K. 2013. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). *www.genome.gov/sequencingcosts*.

Wigginton, J.E. and Abecasis, G.R. 2005. PEDSTATS: Descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics* 21:3445-3447.

INTERNET RESOURCES

http://www.sph.umich.edu/csg/abecasis/MaCH/tour

Tutorial for the MACH 1.0 program for carrying out genotype imputation.

http://genome.sph.umich.edu/wiki/MaCH_FAQ

Frequently asked questions about the MaCH program.

http://genome.sph.umich.edu/wiki/Minimac

Using the minimac program to carry out genotype imputation.

http://genome.sph.umich.edu/wiki/Minimac:_1000_Genomes_Imputation_Cookbook

The 1000 Genomes Imputation Cookbook contains detailed documentation and example scripts for the MaCH+minimac platform.

http://genome.sph.umich.edu/wiki/IMPUTE2:_1000_Genomes_Imputation_Cookbook

The 1000 Genomes Imputation Cookbook contains detailed documentation and example scripts for the IMPUTE2 platform.

http://www.1000genomes.org
The 1000 Genomes Project Web site.

http://hapmap.ncbi.nlm.nih.gov
The HapMap Project Web site.

http://www.unc.edu/~yunmli/software.html
Web site for Li Group Software.

https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html
HAPGEN software for simulating haplotypes.