ANNUAL REVIEWS

Annual Review of Genomics and Human Genetics

Genotype Imputation from Large Reference Panels

Sayantan Das,[1] Gonçalo R. Abecasis,[1] and Brian L. Browning[2]

[1]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029, USA; email: sayantan@umich.edu, goncalo@umich.edu

[2]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, Washington 98195-7720, USA; email: browning@uw.edu

Annu. Rev. Genom. Hum. Genet. 2018. 19:73–96

First published as a Review in Advance on May 23, 2018

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

https://doi.org/10.1146/annurev-genom-083117-021602

Keywords

imputation, genotype imputation, genome-wide association study, GWAS

Abstract

Genotype imputation has become a standard tool in genome-wide association studies because it enables researchers to inexpensively approximate whole-genome sequence data from genome-wide single-nucleotide polymorphism array data. Genotype imputation increases statistical power, facilitates fine mapping of causal variants, and plays a key role in meta-analyses of genome-wide association studies. Only variants that were previously observed in a reference panel of sequenced individuals can be imputed. However, the rapid increase in the number of deeply sequenced individuals will soon make it possible to assemble enormous reference panels that greatly increase the number of imputable variants. In this review, we present an overview of genotype imputation and describe the computational techniques that make it possible to impute genotypes from reference panels with millions of individuals.

INTRODUCTION

The field of human genetics has made great strides since R.A. Fisher first explored the genetic architecture of quantitative traits a century ago (29). In 2005, the advent of single-nucleotide polymorphism (SNP) genotyping arrays made it possible to conduct the first genome-wide association study (GWAS) (52). Although this study had a very small sample (~100 macular degeneration cases and ~50 controls), it was the first in a series of studies that definitively implicated the alternate complement pathway in macular degeneration. Just two years later, the Wellcome Trust Case Control Consortium published a landmark GWAS, the largest of its time, which analyzed ~17,000 cases and controls for seven common diseases (92). Along with replicating many previously implicated genetic loci, the study revealed multiple new risk loci for several diseases, including Crohn's disease and rheumatoid arthritis.

The success of these early GWASs led to an explosion of interest in the field. Numerous GWASs have systematically evaluated the contributions of genetic factors to various complex diseases (68) along with quantitative traits such as human height (53), body mass index (60), and cholesterol (33). These studies revealed unexpected pathways in disease etiology, such as the importance of complement factor genes in macular degeneration (52), the role of the central nervous system in obesity susceptibility (60), and the function of genes in the autophagy pathway in Crohn's disease (7). They have also provided evidence for previously suspected molecular mechanisms [e.g., the role of *IL-23* signaling in psoriasis (70) and the role of *APOE* in Alzheimer's disease (15)] and are expected to enable the development of new drugs and treatment strategies (75, 97). Although GWASs are now commonplace, they have greatly changed human genetics in the last 10 years by providing a new, systematic method that can provide deeper insights into disease biology.

One limitation of SNP genotyping arrays is that they assay only a small fraction of human genetic variation. The variants assayed on SNP arrays are chosen based on the linkage disequilibrium structure of the human genome. The International HapMap Project facilitated the design of the first arrays to be used for GWASs (44–46). Without imputation, GWASs that test variants on a commercial genotyping array must rely on pairwise linkage disequilibrium between an assayed SNP and a causal variant to detect association between the assayed SNP and trait. However, rare variants, which are more often associated with dramatic functional consequences, tend to have low levels of pairwise linkage disequilibrium with common variants on SNP genotyping arrays (32), which makes it difficult to detect signals of association from rare variants. However, given sufficient coverage, whole-genome sequencing can detect the rarest of mutations with very high accuracy (5). Although next-generation technologies have significantly reduced the cost of sequencing a genome, it remains prohibitively expensive to whole-genome sequence the millions of samples that are included in genetic studies each year (34).

A more cost-efficient way of genotyping rare variants is to impute them (58). Rare-variant genotypes that are not directly assayed on GWAS arrays can be reconstructed by comparing each sample to a reference panel of sequenced genomes. This method of estimating genotypes or genotype probabilities at markers that have not been directly genotyped is known as genotype imputation. The first two GWASs to use genotype imputation were a study of type 2 diabetes in Finnish samples (85) and the Wellcome Trust Case Control Consortium study (92). In the type 2 diabetes study, imputation helped the researchers identify and replicate multiple risk variants and compare their results with those of two other studies that used different genotyping arrays. Since then, imputation has been a key step in the analysis of human genetic studies—accelerating fine-mapping efforts, aiding the combination of results across studies (meta-analysis), and increasing the power of gene mapping analyses (66). Some examples of the benefits of genotype imputation are given below.

Fine Mapping

Imputation provides a higher-resolution view of a genetic region by adding more variants, thereby increasing the chances of identifying a causal variant. For example, a study on blood triglyceride levels that aimed to fine map the *GCKR* gene found that the strongest signal came from a missense variant that was imputed and later confirmed by direct genotyping (76). Similarly, a recent fine-mapping study on type 2 diabetes used imputation to enhance the discovery and increase the SNP resolution of causal type 2 diabetes risk alleles (63).

Meta-Analysis

Imputation also helps in meta-analysis by facilitating the combination of results across studies. Different studies often use different genotyping arrays containing different sets of variants. For instance, only 20% of the SNPs included on the Affymetrix 6.0 SNP array are included on the Illumina 660K array. Genotype imputation can generate a common set of variants that can be analyzed across all the studies to boost power. The first two examples of an imputation-based meta-analysis date from early 2008. In both cases, researchers were able to combine studies conducted using different arrays and identify new association signals that could not be discerned in any of the original studies individually (93, 94). Since then, this approach has been successful in discovering associated loci for many different traits, including type 1 and type 2 diabetes (16, 23), Parkinson's disease (14), coronary artery disease (73), different types of cancer (4, 51, 84), height (95), body mass index (60, 86), and lipid levels (33).

Increasing the Power of Association Studies

Another benefit of imputation lies in increasing the power to detect an association signal. When SNPs are genotyped in only a portion of the samples, imputation can increase the effective sample size by filling in the missing genotypes. This was demonstrated in a study on triglycerides and cholesterol, where a common variant in a known risk gene (*LDLR*) was missed when only genotyped SNPs were analyzed but was then identified following imputation (93). This was because the genotyping chip used to assay most of the samples did not contain the common variant or any variant strongly correlated with the common variant. Some simulation studies have shown that imputation can increase power by up to 10% when compared with testing only genotyped SNPs (87), while others have predicted more modest gains (6, 37). The differences in these estimates can be attributed to differences in experimental design, including filtering thresholds and the different genotyping arrays that were used (87).

Other Benefits

Imputed data have also been used to test for pleiotropic effects by imputing a known risk variant into multiple disease studies. As a case in point, Hoffmann et al. (38) found evidence suggesting a pleiotropic effect from a *HOXB13* mutation across multiple cancers. Imputation has been used to estimate other types of genetic variations, such as copy number variants (36) and classical HLA alleles (24, 49, 55, 98).

GENOTYPE IMPUTATION METHODS: MAJOR MILESTONES

Estimation of missing data is a ubiquitous problem in statistics, and human genetic studies are no exception. However the advent of GWASs ushered in a new era, with a new type of imputation

problem that traditional methods were ill equipped to solve (96). The primary reason is the extremely high rate of missing data: Commercial GWAS arrays genotype <1% of the known genetic variants, and the remaining >99% of the genetic variation is missing data that needs to be imputed. The second reason is that common statistical imputation techniques, such as linear regression, regression trees, and *k*-nearest neighbors, do not model key characteristics of genetic data (linkage patterns, recombination hot spots, mutations, genotyping errors, etc.). These challenges necessitated the development of statistical methods and computational tools created explicitly for genotype imputation in GWASs (see **Figure 1**).

The basic intuition behind genotype imputation is as follows: Any two individuals, even if apparently unrelated, can share short stretches of chromosome derived from a distant common ancestor. Consequently, once a study sample is genotyped on a commercial array (with mostly missing data), the observed genotypes can be used to identify DNA segments shared between the study sample and a reference panel of sequenced genomes (with no missing data). In this way, a study haplotype can be represented as a mosaic of short segments of related haplotypes found in the reference panel, enabling one to impute the sites that were not genotyped (see **Figure 2**). Points where the reference haplotype template changes represent historical recombination events. Points where the observed target allele differs from the template allele represent historical mutation events, gene conversion events, genotype error, or even erroneously assigned matches. Since a study haplotype can be represented by many possible mosaics of reference haplotype segments, a probabilistic framework is needed to summarize information from all possible mosaics into imputed alleles.

The Li and Stephens Model

Although multiple research groups have developed numerous genotype imputation methods over the last decade, the basic framework behind most of them is fundamentally the same and is known as the Li and Stephens model. It was first described in 2003 (56) to allow haplotype estimation methods to handle large stretches of chromosome, where individual haplotypes are all unique but are expected to share contiguous, mosaic stretches with other haplotypes in the sample. A modified version of this approach was implemented in fastPHASE by Scheet & Stephens (83) to enable genotype phasing of larger samples. The framework uses a hidden Markov model (HMM) (80) to describe the data, where the observed genotypes of unknown phase in a study sample represent the observed data of the HMM, while an underlying and unobserved set of phased genotypes represent the hidden states of the HMM. The HMM framework was immediately beneficial and provided notable and substantial improvements in the quality of inferred haplotypes compared with previous approaches (65).

The Li and Stephens model state space can be visualized as a two-dimensional grid of HMM states, with rows corresponding to reference haplotypes and columns corresponding to markers in the reference panel. In **Figure 2**, each allele on each reference haplotype corresponds to an HMM state. Each study sample haplotype is assumed to trace an unobserved path through the grid, proceeding left to right from the first reference marker to the last reference marker. This path is equivalent to a mosaic of templates. A new segment in the mosaic begins when the path switches reference haplotypes (rows) between one marker and the next. The probability of a template switch between markers is determined by the HMM transition probabilities and is closely related to the population recombination rate. The probability that an observed allele differs from the template allele is determined by the HMM emission probabilities. Given an observed haplotype with missing alleles, the probability of each possible path through the HMM states can be calculated. A path probability is penalized (i.e., decreased) each time the path switches reference haplotypes (via the

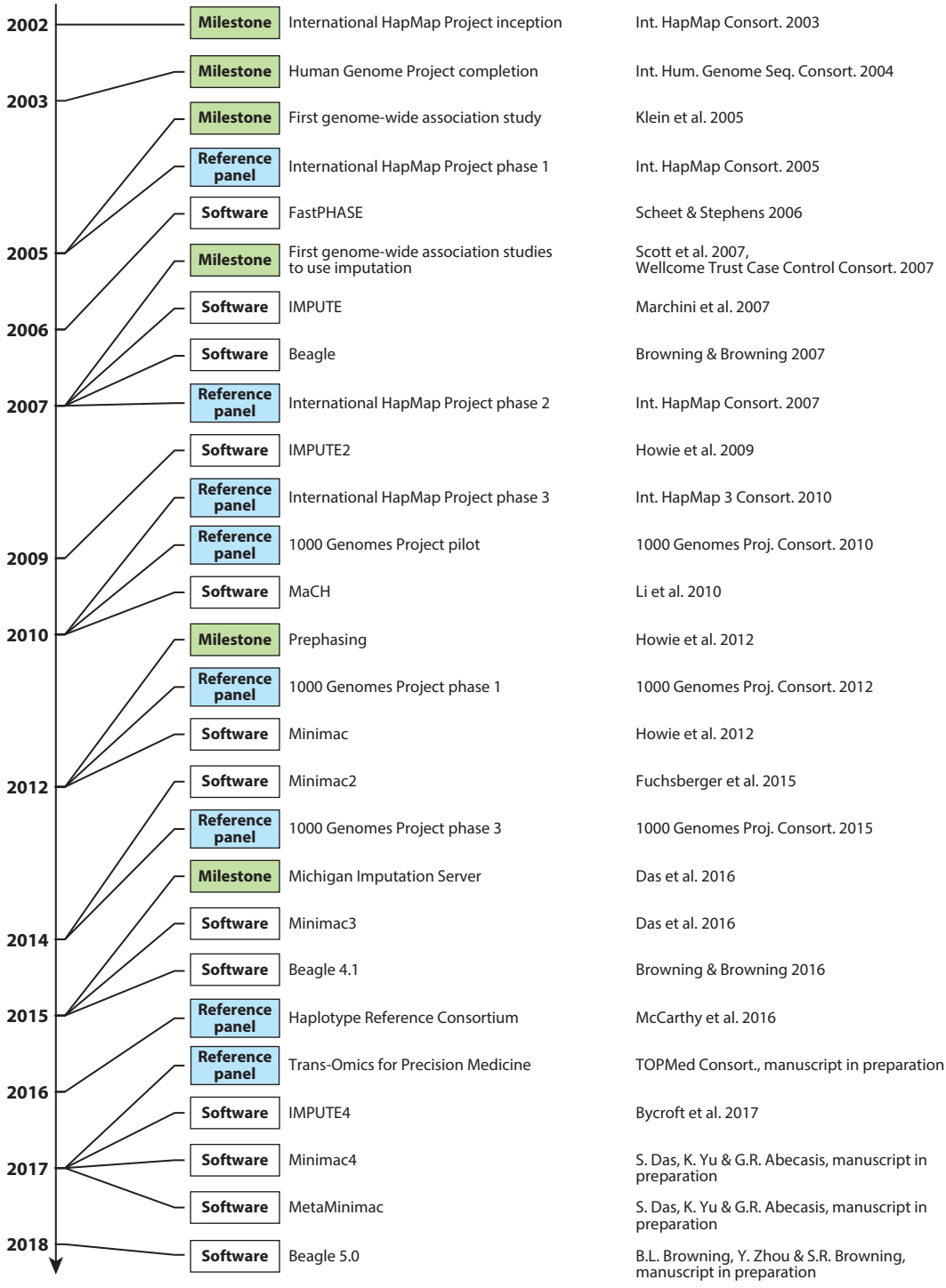

Figure 1

A brief time line summarizing the major developments in genotype imputation. Each major development has been categorized as a milestone (*green*), a reference panel (*blue*), or software (*white*).

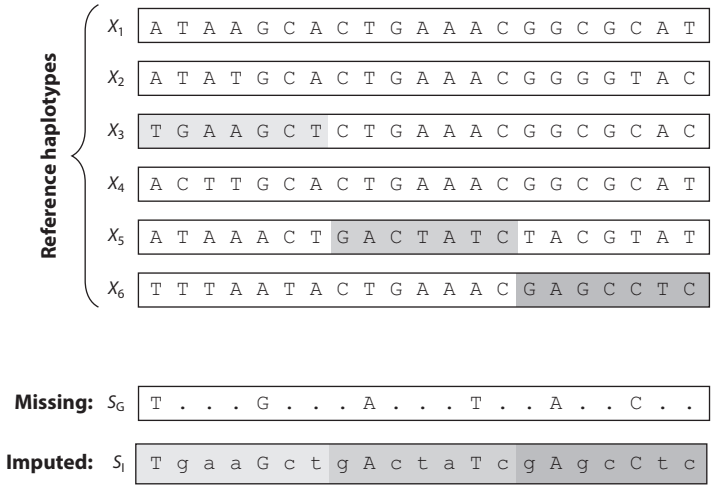

Figure 2

An illustration of genotype imputation, showing the process of imputation for a study haplotype (S_G) genotyped at 6 markers using a reference panel of sequenced haplotypes at 21 markers. The alleles in S_G are used to match short segments from the reference panel. For example, in the first genomic segment, the alleles T and G imply that the corresponding segment might have been copied from haplotype X_3. In the second segment, the alleles A and T imply that haplotype X_5 might have been copied. Proceeding similarly, the study haplotype can be represented as a mosaic of DNA segments from haplotypes X_3, X_5, and X_6. Consequently, the missing sites can be imputed to obtain the final imputed haplotype, S_I.

HMM transition probabilities) and each time the reference allele on the path differs from the observed allele (via the HMM emission probabilities). The probability that the (unobserved) path of a target haplotype goes through a particular HMM state (the state probability) can be calculated efficiently with the HMM forward–backward algorithm (80). Imputed allele probabilities at a marker are obtained from the state probabilities. The probability that the target haplotype carries a particular allele is the sum of the state probabilities corresponding to reference haplotypes that carry the allele.

Although most contemporary imputation tools employ an HMM framework, they differ in how they define the state space and the parameters of the HMM. **Table 1** summarizes the major imputation tools in the last decade. While fastPHASE, MaCH, and IMPUTE were quite similar, the first Beagle imputation algorithm was different because it did not employ the usual transition and emission functions, and the haplotype model was constructed from both reference and study samples as opposed to only reference samples (11). However, the second Beagle imputation algorithm (introduced in version 4.1) uses the Li and Stephens model and is similar to the other tools (9).

While the term genotype imputation typically refers to imputing variants not directly assayed in a GWAS, it has also been used to refer to inferring genotypes from genotype likelihoods that are estimated from SNP array or low-coverage sequence data. This inference of genotypes from genotype likelihoods, which is also called genotype refinement, is outside the scope of this review. However, HMMs are also used by several methods in this setting in tools such as Beagle, MaCH, SNPTools, and Stitch (10, 20, 57, 91).

Other imputation engines have employed direct methods of haplotype matching that do not employ the HMM framework. For example, PBWT uses the positional Burrows–Wheeler transformation to find set-maximal matches at each marker of the genotyped sample, which are in turn used to assign alleles at the missing markers (27). While such methods are highly

Table 1 Genotype imputation tools that employ a hidden Markov model (HMM)

Tool	Year	Description of state space	Computational complexity	HMM parameter functions
FastPHASE	2006	All genotype configurations from a fixed number of localized haplotype clusters	Maximization-step linear in number of haplotypes, quadratic in number of clusters	Depends on recombination and mutation rates; parameters are fit using an expectation–maximization algorithm
IMPUTE	2007	All genotype configurations from all reference haplotypes	Quadratic in number of haplotypes	Depends on a fine-scale recombination map that is fixed and provided internally by the program
Beagle	2007	All genotype configurations from a variable number of localized haplotype clusters	Quadratic in number of haplotypes	Empirical model with no explicit parameter functions
IMPUTE2	2009	All reference haplotypes	Phasing quadratic in number of haplotypes, imputation linear in number of haplotypes	Same as IMPUTE
MaCH	2010	All genotype configurations from all reference haplotypes	Quadratic in number of haplotypes	Depends on recombination rate, mutation rate, and genotyping error; parameters are fit using a Markov chain Monte Carlo or expectation–maximization algorithm
Minimac and Minimac2	2012	All reference haplotypes	Linear in number of haplotypes	Same as MaCH
Minimac3	2016	All unique allele sequences observed in reference data in a small genomic segment	Linear in number of haplotypes	Same as MaCH, but parameter estimates are precalculated and fixed
Beagle 4.1	2016	All reference haplotypes at genotyped markers	Linear in number of haplotypes	Depends on recombination rates and error rates, which are precalculated and fixed
Minimac4	2017	Collapsed allele sequences from reference data that match at genotyped positions in small genomic segments	Linear in number of haplotypes	Same as Minimac3
IMPUTE4[a]	2017	All possible reference haplotypes	Linear in number of haplotypes	Same as IMPUTE2
Beagle 5.0	2018	A user-specified number of reference haplotypes	Linear in number of haplotypes	Same as Beagle 4.1

This table describes the typical state space and parameter functions used to model the Li and Stephens framework. Minimac and IMPUTE2 were the first tools to use the prephasing approach. Minimac3 and Beagle 4.1 exploit local haplotype redundancy to reduce the size of the state space and hence the computational burden.

[a]IMPUTE4 uses the same HMM as IMPUTE2; however, to reduce memory usage and increase speed, it uses compact binary data structures and takes advantage of high correlations between inferred copying states in the HMM to reduce computation.

computationally efficient, in current panels, their imputation accuracy is reduced because the methods do not integrate over all possible mosaic configurations but instead use the longest haplotype matches flanking each location to impute each genotype. Additionally, some other genotype imputation methods [e.g., PLINK (79), SNPMStat (59), TUNA (72), and UNPHASED (26)] use SNP-tagging approaches to carry out imputation. Although these methods are simpler and can

be very fast, they do not utilize information from the entire chromosome to perform imputation and hence generally provide less accurate allele estimates.

Prephasing

Genotype imputation is a highly computationally intensive process because of the probabilistic framework and a high rate of missing data. One of the major milestones to reduce the computational burden in the Li and Stephens framework was the introduction of prephasing. This idea involves a two-step imputation process: the initial step of prephasing (i.e., haplotype estimation) of the GWAS genotypes and a subsequent step of imputation into the estimated study haplotypes (41). These separate steps benefited researchers in several ways. First, the decomposed haplotypes could be reused for imputation from different reference panels, allowing researchers to conveniently explore the trade-offs of different imputation strategies. Second, separating the phasing (or haplotype estimation step) from imputation allowed researchers to quickly benefit from separate advances in phasing and imputation technology, which no longer needed to be tightly integrated. Third, it reduced the complexity of the imputation step from quadratic to linear in the number of reference haplotypes, because prephasing allowed matches to be found by comparison against phased haplotypes rather than against all pairs of haplotypes. Although splitting up the process does marginally reduce the imputation accuracy in some populations, such as African Americans (41), the ability to use much larger reference panels with prephasing makes it possible to attain greater imputation accuracy. The current versions of Minimac, IMPUTE, and Beagle all employ this prephasing approach.

Public Reference Panels

Over the years, the quality of genotype imputation has benefited greatly from improved genotyping technologies (e.g., high-density genotyping arrays) and more efficient analytical methods (e.g., prephasing) but most notably from the increase of genetic information in publicly available data sets. Examples of such data sets include those from the International HapMap Project (45–47), the 1000 Genomes Project (1000G) (1), the UK10K Project (43), the Haplotype Reference Consortium (HRC) (69), and the Trans-Omics for Precision Medicine (TOPMed) program (71). The development of next-generation sequencing technologies has led to a rapid increase in the sizes of data sets used as reference panels for genotype imputation. For example, while the first release of the International HapMap Project panel had 269 individuals (45), the subsequent 1000G phase 3 panel had 2,504 individuals with low-coverage sequence data (3), and upcoming reference panels from the TOPMed program are expected to soon include more than 100,000 deeply sequenced samples. **Table 2** summarizes the major public reference panels.

For association studies, the immediate benefits of a larger panel include a more detailed catalog of genetic variants, which increases the chance of imputing a causal variant, and better imputation accuracy, which improves the power of downstream association analyses, especially for rare variants (58). Panels derived from disease-focused sequencing efforts often have restrictions that limit direct access to the underlying haplotypes (e.g., HRC and TOPMed), making their use as a broadly distributed imputation resource challenging. These challenges led to the evolution of web imputation servers, another major milestone in the field of genotype imputation.

Web Imputation Servers

Imputation servers enable users to upload GWAS data to a remote server through secure file transfer protocols. The server then carries out imputation along with automated quality control,

Table 2 The most commonly used public reference panels to date

Reference panel	Number of reference samples	Number of sites (autosomes + X chromosome)	Average sequencing coverage	Ancestry distribution	Publicly available	Indels available	Reference
International HapMap Project phase 3	1,011	1.4 million	NA[a]	Multiethnic	Yes	No	47
1000G phase 1	1,092	28.9 million	2–6×	Multiethnic	Yes	Yes	1
1000G phase 3	2,504	81.7 million	7× genomes, 65× exomes	Multiethnic	Yes	Yes	3
UK10K Project	3,781	42.0 million	7× genomes, 80× exomes	European	Yes	Yes	89
HRC	32,470	40.4 million	4–8×[b]	Predominantly European[c]	Partially[d]	No	69
TOPMed	60,039	239.7 million	30×	Multiethnic	Partially[e]	Yes	71

Abbreviations: 1000G, 1000 Genomes Project; HRC, Haplotype Reference Consortium; indel, insertion or deletion; NA, not applicable; TOPMed, Trans-Omics for Precision Medicine.

[a]The International HapMap Project phase 3 data were genotyped on the Illumina Human1M and Affymetrix 6.0 SNP arrays.

[b]The HRC panel was obtained by combining sequencing data across many low-coverage (4–8×) and a few high-coverage sequencing studies.

[c]The only non-European samples in the HRC panel are through the 1000G reference panel (which was a contributing study).

[d]Most of the HRC samples (~27,000) are available for download through controlled access from the European Genome-Phenome Archive.

[e]Some of the TOPMed samples (~18,000) are available for download through controlled access from the Database of Genotypes and Phenotypes (dbGaP).

such as checking strand orientation, allele labels, file integrity, minor allele frequency distribution, and per-sample missingness (see **Figure 3**). After the analysis is complete, users receive a notification and download link to the imputed data, which are encrypted with a onetime password. A major benefit of imputation servers is that they allow researchers to spend more time on analyzing and interpreting their data instead of learning about imputation tools and data preprocessing. In addition, they provide a uniform platform for comparing and consolidating results across studies, thereby aiding collaborative efforts. Finally, for reference panels derived from disease-focused sequencing studies or from other studies that also have data-sharing restrictions, keeping reference panel data behind a firewall, where it can be used for low-risk analyses without exposing individual-level data, greatly increases the number of users who can benefit from these panels. Imputation servers have also motivated the development of remote web servers for other genomic analyses, such as estimation of principal components [Locating Ancestry from Sequence Reads (LASER)] (88) and phenome-wide analysis of participants in electronic health record–based studies [such as those in the Michigan Genomics Initiative (MGI) PheWeb (**http://pheweb.sph.umich.edu**)]. Currently, the University of Michigan and the Wellcome Sanger Institute host web imputation servers, using the underlying imputation engines Minimac3 and PBWT, respectively (19, 27). The Michigan server has imputed more than 18 million genomes for ~3,000 registered users, while the Sanger server has imputed ~9 million genomes and has ~550 users. The current throughput of the Michigan server, which consists of 23 multiprocessor computers (~600 cores in total), is ~7 million genomes per month using the International HapMap Project phase 2 panel and ~900,000 genomes per month using the HRC panel [estimates include time to prephase using Eagle (61)].

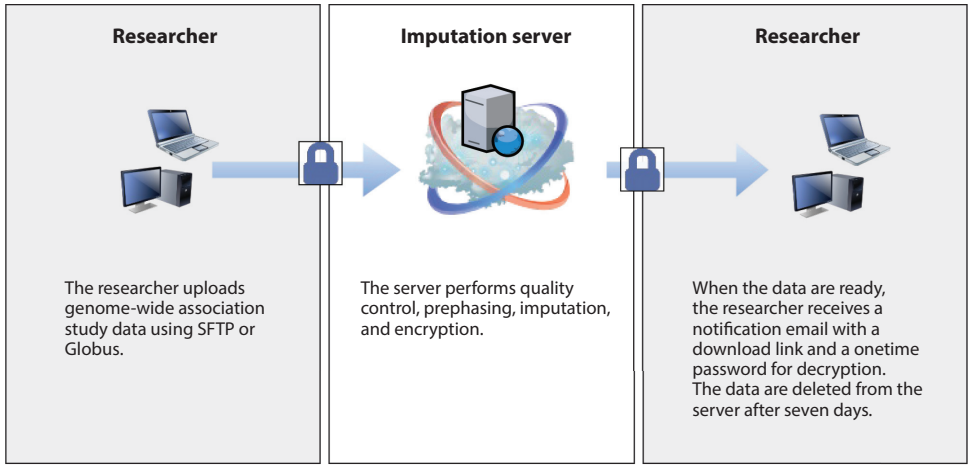

Figure 3

An outline of the pipeline in the Michigan imputation server.

COMPUTATIONAL METHODS FOR LARGE REFERENCE PANELS

In the last 15 years, the cost of DNA sequencing has decreased by five orders of magnitude (35). The super-exponential decrease in the cost of genome sequencing is widely known. Less well known is the even more rapid decrease in the cost of genotype imputation since 2009. Timing results in 2009 for one of the first imputation engines (IMPUTE 0.5.0) showed an imputation time of 0.43 s per genotype per target sample when imputing from 1,200 reference samples (8). Timing results in 2015 for Minimac3 and Beagle 4.1 showed imputation times of 6.4×10^{-6} and 7.5×10^{-6} s per genotype per sample, respectively, when imputing from 2,452 reference samples from the 1000G phase 3 panel (9). During this time, these imputation methods also increased the number of reference samples that could be included in an imputation analysis by two to three orders of magnitude. In this section, we discuss computational methods for reducing computation time and memory requirements. The methods described here can be combined to produce a large cumulative effect.

Use of a Custom Subset of the Reference Panel

One of the first strategies for reducing computation time was to use only a subset of the available reference haplotypes. The idea is to select a small subset of reference haplotypes that appear to be closely related to the sample of interest in a genomic region and to impute genotypes in the target haplotype in that region using the small custom reference panel instead of the full reference panel. This technique was first implemented in IMPUTE2 (40, 42). If the custom reference panel for a region contains the reference haplotypes that are most closely related to the target haplotype and if the algorithm for selecting the custom reference panel is sufficiently fast, genotype imputation using the custom reference panel can be much faster than imputation from the full reference panel and have similar accuracy (40, 42). The first method for selecting closely related haplotypes was based on Hamming distance (40), but more recent approaches select closely related haplotypes based on long identity-by-state segments (43). Typically, a different custom subset of the reference panel is selected for each observed haplotype and each region of interest.

Prephasing Target Data

The first genotype imputation methods assumed unphased target genotype data and internally estimated the target genotype phase during genotype imputation (8, 57, 67). The first step away from this paradigm was the realization that, instead of imputing unphased genotypes, one could impute alleles directly onto the estimated target haplotypes (40). Once methods began performing haploid imputation, the next step was to separate the genotype phasing and imputation steps and to require the input target data to be phased (41).

Unlike most optimizations, which trade increased algorithmic and software complexity for reduced run time, the use of prephased target data simplifies the algorithms and software. This reduced complexity makes it easier to discover and implement additional computational optimizations.

Specialized Input Formats for Reference Data

One of the challenges of imputation from large reference panels is reading and storing the reference genotype data. A genotype is stored in 4 bytes in the Variant Call Format (VCF) (17). If there are 1 million reference samples, VCF storage requirements can exceed 1 TB per megabase of chromosome. Consequently, genotype data for large reference panels are typically stored on hard disk drives in compressed form. General-purpose compression algorithms, such as gzip (22), greatly reduce the reference panel file size but do not address the problem of storing reference genotype data in memory during analysis. In addition, the time required for decompressing large reference panels compressed with gzip is substantial and can exceed the time required for imputation. These limitations of general-purpose compression algorithms have motivated the development of specialized compression formats for reference genotype data that achieve high data compression or permit fast retrieval of individual genotypes from the compressed representation so that reference haplotypes can be stored in compressed form in memory when imputing genotypes.

One of the first specialized compression formats to be developed for reference data was based on the Burrows–Wheeler transform (27). Two additional compression formats were developed specifically for genotype imputation. One of these formats (M3VCF) (19) exploits local redundancy among haplotypes. Rather than storing all haplotypes in a chromosome segment, only the unique allele sequences are stored, along with a map from the reference haplotypes to the allele sequence carried by each haplotype. A second compression format, called binary reference format (bref) (9), employs this same general strategy for markers with a high nonmajor allele frequency but uses an alternate compression scheme for markers with a low nonmajor allele frequency, which are the bulk of markers in large reference panels. For a low-frequency marker, bref stores a list of reference haplotypes that carry the minor allele. If the marker has more than two alleles, it stores a separate list for each nonmajor allele. The allele on a given haplotype can be found by searching the lists of haplotypes. If the haplotype is found in a list, the haplotype carries the corresponding nonmajor allele. If the haplotype is not found in any list, the haplotype carries the major allele. For low-frequency markers, bref increases the time required to query the allele carried by a given haplotype, but the query time is not prohibitive if the lists of haplotypes are sorted in increasing order and a binary search algorithm is used. For reference panels with more than 100,000 samples, the use of the bref format reduces computation time by more than 30% and the use of the M3VCF format reduces computation time by more than 90% compared with the use of the VCF format (9).

Clustering of Identical Reference Haplotype Segments

In short segments of the genome, the same allele sequence can be carried by many haplotypes in the reference panel. HMM states that correspond to identical reference haplotypes in a genomic

segment will have the same emission probabilities at each marker in the segment. Consequently, in a short region, the identical haplotypes can be clustered together, and the HMM forward–backward algorithm calculations can traverse the relatively small number of unique reference allele sequences rather than all reference haplotypes. This speeds up HMM calculations within small regions with no loss in accuracy. However, the full set of reference haplotypes must still be traversed when performing the HMM forward and backward algorithm calculations across region boundaries. Minimac3 implements a version of local clustering that is based exclusively on the reference panel, which allows the local clustering to be precomputed for a reference panel and used in all subsequent imputation analyses (19). This precomputed clustering also provides the basis for the Minimac3 M3VCF format (19). Beagle 4.1 performs local clustering on the fly during imputation because the short regions used for clustering are defined by the genotyped markers in the target samples (9).

Imputation via Linear Interpolation

Although it is most natural to calculate HMM state probabilities in one forward pass and one backward pass through the markers in the reference panel, it is also possible to perform these calculations in two stages. In a two-stage approach, the HMM forward–backward algorithm is performed first using only the markers that are genotyped in the target samples. In the second stage, the forward–backward algorithm is performed in each chromosome interval bounded by two adjacent genotyped markers or by a chromosome boundary and the adjacent genotyped marker. For the Li and Stephens model, the one-stage and two-stage approaches produce identical HMM state probabilities.

The two-stage approach to HMM calculations permits two optimizations based on linear interpolation. The first optimization uses linear interpolation in the second stage instead of the forward–backward algorithm to calculate HMM state probabilities at imputed markers. After the first stage, HMM state probabilities are calculated for genotyped markers, and HMM state probabilities at imputed markers are estimated by linear interpolation on genetic distance. Over short genetic distances, linear interpolation generally provides an accurate approximation of the HMM state probability (9).

A second optimization arises from the observation that interpolated HMM state probabilities in an interval are bounded by the HMM state probabilities at the bounding genotyped markers. If the state probabilities at the bounding genotyped markers are sufficiently small for a reference haplotype, the interpolated HMM state probabilities in the interval can be approximated by 0, and the linear interpolation step can be skipped altogether.

These optimizations based on linear interpolation can also be combined with clustering of identical haplotype segments. In particular, one can cluster reference haplotypes that have identical allele sequences between the two genotyped markers before performing linear interpolation.

Reducing Memory Requirements

Imputation from large reference panels must be performed within the constraints imposed by the available computer memory. A standard approach to reducing memory requirements for genome-wide imputation is to divide the genome into small overlapping genomic windows and perform imputation in each window separately. Although one could compensate for increasing reference panel size by decreasing the length of the analysis window, there are limits to this strategy. If the window is too short, imputation accuracy will suffer if information from outside the window is ignored. This is most evident when imputing rare variants from large reference panels. Individuals

who share a rare variant will typically share a long haplotype around the variant, and a major benefit of large panels is their ability to facilitate identification of these long shared haplotypes. If the genomic window is too short, the shared long haplotype containing the rare variant will be truncated by the window boundary, and imputation accuracy will decrease.

Another approach to reducing memory requirements is multithreading, which allows multiple CPU cores to share a single copy of the reference and target genotype data. This provides some reduction in memory use, but memory requirements generally will still increase linearly with the number of computational threads because each thread must allocate memory for its probability calculations.

Fortunately, all of the above-described optimizations that reduce compute time have the added benefit of reducing memory requirements. This makes it possible to use relatively long analysis windows (say, >10 cM in length) when imputing from large reference panels.

MEASURING IMPUTATION ACCURACY

Most imputation methods estimate a probability distribution for the allele carried by each haplotype at each imputed marker. Posterior genotype probabilities can be derived from the posterior allele probabilities under the assumption of Hardy–Weinberg equilibrium (40). One of the most common uses of imputed genotype data is to test each imputed marker for association with a trait. Standard regression-based approaches for genetic association studies, including generalized linear models and linear mixed models, extend naturally to imputed data by replacing the allele dose of the observed genotype with the expected allele dose of the imputed genotype, which is the sum of the posterior allele probabilities for each haplotype (66). In this context, it is helpful to assess accuracy for imputed markers so that poorly imputed markers may be excluded prior to association testing.

The most interpretable measures of imputation accuracy are based on the correlation between the imputed and true dose of an allele. One correlation-based measure is the r^2 measure (41), which is the squared correlation between the true and estimated dose of an allele across all imputed samples. Note that the MaCH r^2 measure (57) is slightly different from the Minimac r^2 measure (which is also reported by Beagle) since the former is based on correlation between true and estimated diploid dosages across all samples instead of haploid dosages across all haplotypes (they are equivalent under assumptions of Hardy–Weinberg equilibrium). For each haplotype, the true allele dose is 0 or 1, and the estimated allele dose is the posterior allele probability. A formula for the Minimac r^2 measure is derived in the sidebar titled Estimating r^2.

The r^2 measure has two attractive features: It can be estimated from posterior allele probabilities without knowledge of the true allele dose on each chromosome if the allele probabilities are well calibrated (see the sidebar titled Estimating r^2), and it has a useful interpretation in terms of sample size and statistical power when testing a binary trait. This relationship is as follows: If $r^2 = r_0^2$ for an imputed marker, the power of an allelic test with N samples and imputed alleles is approximately equal to the power of an allelic test with $r_0^2 N$ samples and true alleles. Thus, the r^2 measure can be interpreted as the effective reduction in sample size when testing imputed alleles rather than the true alleles for association with a binary trait. Pritchard & Przeworski (78) gave a derivation of this result in the context of two correlated markers. This result generalizes to the correlation between the estimated allele dose and the true allele dose (18). It is common to exclude poorly imputed markers from downstream analysis by requiring the r^2 measure for an imputed variant to exceed some threshold. Thresholds of 0.3 or larger are commonly used. An r^2 threshold of 0.3 means that one is willing to accept an effective reduction in sample size of approximately two-thirds when performing an allelic test with imputed alleles. Since r^2 is a correlation, it is defined in terms of the variance of the true allele dose, and thus if it is correctly

ESTIMATING r^2

One attractive feature of r^2, the squared correlation between true and imputed allele dose, is that it can be estimated from posterior allele probabilities without knowing the true allele on each chromosome. Here, we derive an estimate of r^2 in terms of the posterior allele probabilities.

Let X be 1 if a chromosome carries the allele of interest and be 0 otherwise, and let Z be the estimated posterior allele probability that $X = 1$. Then r^2 is defined to be the squared correlation of X and Z. We say that the posterior allele probabilities are correctly calibrated if $E[X \mid Z] = Z$. If the posterior allele probabilities are correctly calibrated, we can use the law of total expectation and the fact that $X^2 = X$ to obtain

$$\begin{aligned} E[X^2] &= E[X] = E[E[X \mid Z]] = E[Z] \\ \mathrm{Var}(X) &= E[X^2] - E[X]^2 \\ &= E[Z] - E[Z]^2 \end{aligned}$$

and

$$\begin{aligned} \mathrm{Cov}(X, Z) &= E[XZ] - E[X]E[Z] \\ &= E[E[XZ \mid Z]] - E[E[X \mid Z]]\, E[Z] \\ &= E[Z^2] - E[Z]E[Z] \\ &= \mathrm{Var}(Z). \end{aligned}$$

Consequently,

$$\begin{aligned} r^2 &= \frac{(\mathrm{Cov}(X, Z))^2}{\mathrm{Var}(X)\mathrm{Var}(Z)} \\ &= \frac{\mathrm{Var}(Z)}{\mathrm{Var}(X)} \\ &= \frac{E[Z^2] - E[Z]^2}{E[Z] - E[Z]^2} \end{aligned}$$

If there are n imputed chromosomes and z_i is the estimated reference allele probability in the ith haplotype, one can estimate $E[Z^k]$ as $E[Z^k] \approx (1/n) \sum z_i^k$ and r^2 as

$$r^2 \approx \frac{n \sum z_i^2 - (\sum z_i)^2}{n \sum z_i - (\sum z_i)^2}.$$

estimated, its interpretation does not depend on allele frequency. However, the estimate of r^2 becomes noisier (i.e., larger standard error) (see **Figure 4**). Consequently, one could consider applying a frequency-dependent r^2 threshold for marker filtering. One limitation of r^2 is that there must be enough imputed samples so that it can be accurately estimated. Another is that it is defined in terms of a particular allele. If a marker has only two alleles, the estimated r^2 will be the same for both alleles. If a marker has more than two alleles, one could combine the alternate alleles into a single composite allele or estimate r^2 separately for each allele.

A second correlation-based measure of imputation accuracy that is closely related to r^2 is allelic R^2 (8). The allelic R^2 measure differs from r^2 in that it estimates the correlation between the true allele dose and the most probable (i.e., best guess) allele dose instead of the estimated allele dose. Thus, the r^2 measure is more closely aligned than allelic R^2 to the power of downstream analyses

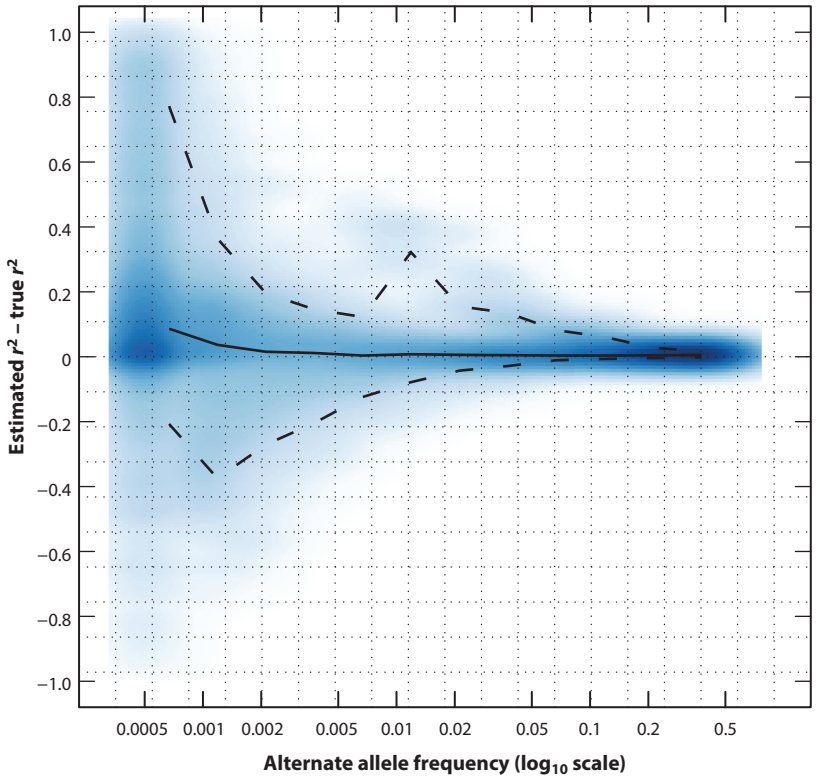

Figure 4

Plot of noise in estimated imputation r^2 as a function of nonreference allele frequency when imputing 10 European-ancestry genome-wide association study samples from the Trans-Omics for Precision Medicine reference panel ($n = 18{,}000$). The x axis is the alternate allele frequency on a $\log_{10}$ scale. The y axis is the difference between the estimated imputation accuracy (from Minimac) and the true imputation r^2. The blue shaded area shows the smoothed scatter plot. The solid line is the mean difference at each frequency, and the upper and lower dashed lines are the 5% and 95% quantiles, respectively, for the difference at each frequency.

that use estimated allele dose instead of directly observed genotypes. Allelic R^2 more closely mirrors the loss in power one might expect when using best-guess genotypes in analyses instead of estimated allele doses. One particular limitation of allelic R^2 is that it cannot be computed when the most probable target allele is the same for all target haplotypes.

A third commonly used measure of imputation accuracy that is not directly based on correlation is IMPUTE's info measure, which is an estimate of the ratio of statistical information about the population allele frequency in the imputed genotypes and in the true genotypes (66). However, it can be shown that if the Hardy–Weinberg equilibrium holds, then IMPUTE's info measure is equal to the Minimac r^2 measure (where the genotype probabilities to calculate IMPUTE's info measure are estimated under Hardy–Weinberg equilibrium).

FACTORS AFFECTING IMPUTATION ACCURACY

Multiple factors can affect imputation accuracy:

- *Size of reference panel*: Increasing the size of the reference panel generally increases imputation accuracy, especially for rarer variants, provided that the level of genetic similarity between

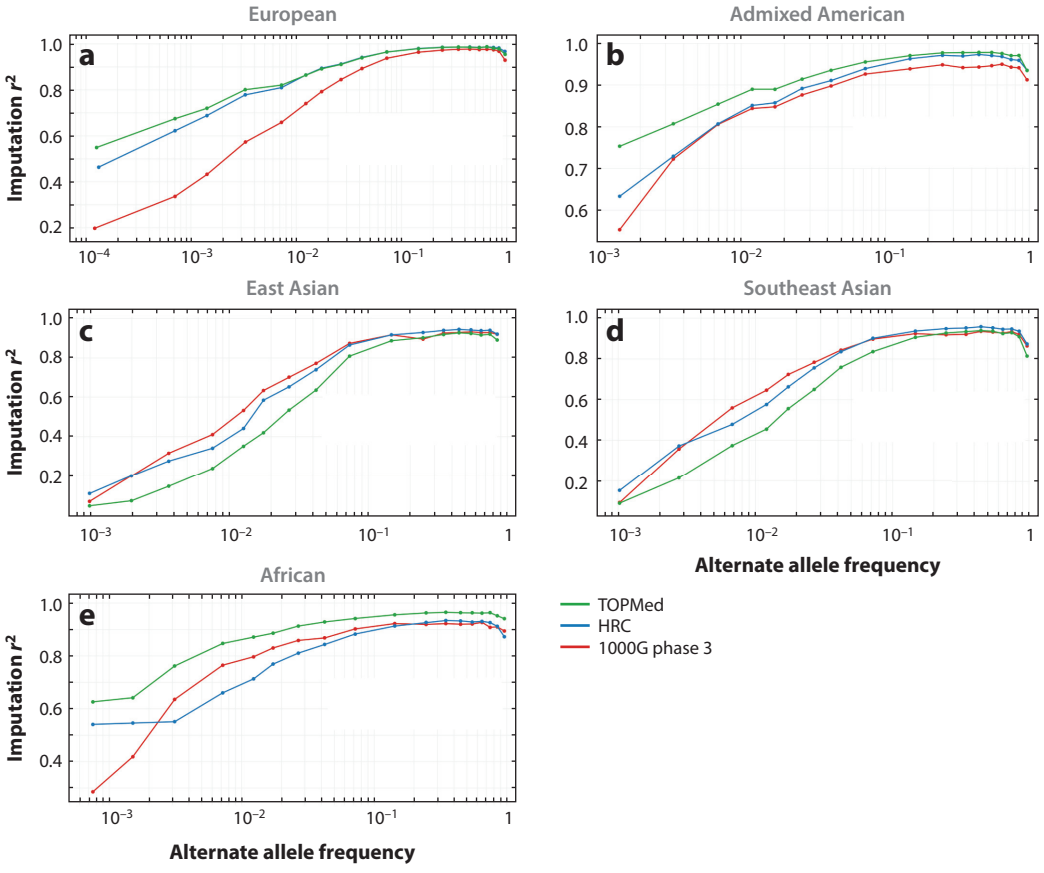

Figure 5

Imputation accuracy for five ancestries: (*a*) European, (*b*) admixed American, (*c*) East Asian, (*d*) Southeast Asian, and (*e*) African. We extracted 10 samples from each of these ancestries from the 1000 Genomes Project (1000G) phase 3 data, masked all variants except those on the Illumina 1M chip, and imputed them using the Trans-Omics for Precision Medicine (TOPMed) (with 18,000 samples), Haplotype Reference Consortium (HRC), and 1000G phase 3 (after removing overlaps) reference panels. The aggregate r^2 (measuring the imputation accuracy) is plotted as a function of the alternate allele frequency.

the reference panel and study samples is maintained. A larger panel provides a larger set of template haplotypes to match against, which improves the imputation accuracy. For example, for imputing into European samples, the HRC panel, which has 33,000 samples, is a better a choice than the 1000G phase 3 panel, which has 2,500 samples (see **Figure 5*a***). However, expanding the reference panel to include samples with little genetic similarity to the target panel or less accurately estimated haplotypes may decrease imputation accuracy. This is evident from **Figure 5*b*–*e***, where the HRC panel, despite being a superset of the 1000G phase 3 panel, decreases the imputation accuracy in non-European samples.

- *Density of genotyping array*: A denser genotyping array increases the number of sites to match with, thereby improving the chances of finding shared haplotype segments. For example, the Illumina Human Omni5Exome array (~5 million variants) is a better choice for imputation than the Human Omni2.5Exome array (~2.5 million variants). The Human Omni5Exome array increases the average imputation accuracy (r^2) by 0.1 for variants with a minor allele

frequency of approximately 0.01 when imputing into European samples using the HRC reference panel.

- *Minor allele frequency of variant being imputed (in the reference panel)*: Rare variants are harder to impute than common variants because rare variants will generally be observed only a few times in the reference panel. This makes it harder to establish the haplotype background for these variants and reduces the set of template haplotypes available for matching. Overall, this is expected to lower imputation accuracy. The plots of imputation accuracy (**Figure 5**) demonstrate this feature, as we see that the imputation accuracy always increases with the minor allele frequency (controlling all other factors).
- *Haplotype accuracy in reference and study samples*: Genotype imputation depends on finding haplotype segments that are shared between reference haplotypes and a target haplotype. Haplotype phase error tends to break up shared haplotype segments. Increases in reference panel size over time should lead to lower haplotype phase error rates (62, 74). Large, accurately phased reference panels will also make it possible to improve haplotype phase estimates in the study samples (61).
- *Sequencing coverage of reference panel*: The sequencing coverage directly correlates with the accuracy of the genotypes in the reference panel, which in turn correlates with the accuracy of the inferred haplotypes. For example, in **Figure 5*a***, the pilot TOPMed panel gives considerably higher accuracy than the HRC panel, although the latter is twice as large. This is most likely because the TOPMed panel was generated with high-resolution sequencing (an average of 40× coverage), whereas the HRC panel was generated by combining low-coverage sequencing data across different studies.

Imputation accuracy depends strongly on the number of copies of the allele in the reference panel. In a study with simulated UK European data, mean imputation accuracy for a fixed minor allele count remained relatively constant as the reference panel size increased from 50,000 to 200,000 samples (9). However, this stable relationship between minor allele count and imputation accuracy may not hold in other contexts. For example, as reference panels grow in size, more alleles will arise from recurrent mutation. Recurrent mutation introduces the same allele on different haplotypic backgrounds. For a recurrent mutation, the number of copies of the allele that share the same haplotypic background as the target haplotype is more relevant to imputation accuracy than the minor allele count in the reference panel. Similarly, genotype imputation methods cannot impute de novo variants. If a de novo variant is also present in the reference panel owing to recurrent mutation, it generally will not be on the same haplotypic background in both the reference panel and the target sample.

IMPUTATION FROM MULTIPLE REFERENCE PANELS

The most common strategy for imputation is to use a single publicly available data set (such as the 1000G or HRC data set), which works quite well for studies with samples of European ancestry. However, studies on non-European populations or relatively homogeneous European subpopulations often benefit more from using custom reference panels of samples with greater genetic similarity. For instance, in a study on samples in Sardinia, an island in Italy with a genetically isolated population, a custom reference of Sardinian samples provided better imputation accuracy than the 1000G and other large European panels (77). This result has also been replicated in other studies on homogeneous European subpopulations (21) and non-European populations, such as African Americans (25). Additionally, disease-specific studies often benefit from a hybrid approach that combines a custom reference panel with an existing public reference panel. This approach is particularly beneficial for disease studies because it significantly increases the accuracy of imputing

a causal variant (since the hybrid panel is enriched with more copies of rare alleles). One example highlighting this feature is a recent study on prostate cancer that was eventually able to impute the rare *HOXB13* G84E variant only after using a hybrid imputation approach that combined the 1000G data with an enriched set of cases carrying the mutation (38), even though previous studies had suggested that it might not be possible to impute this variant (82). The hybrid method is also useful in studies with homogeneous subpopulations, given that it enriches the panel with genetically similar samples while retaining the benefits of large public panels.

The hybrid approach currently requires studies to merge multiple reference panels, and the streamlining of this process is still an active field of research. One proposed approach has been to combine the reference panels by first treating them as reference panels for each other and then cross-imputing the missing variants (40). Although this approach enables one to use all variants found in any panel, the performance of this method has not been fully evaluated. One study found this approach to be neither helpful nor harmful for large, population-specific panels (43). At present, reference panels do not generally include monomorphic markers. The merging of reference panels would be facilitated if reference panels included auxiliary data identifying the markers that are monomorphic and the observed allele at each monomorphic marker.

Some studies have repeated the imputation process for each reference panel. For example, a study repeatedly imputed against multiple reference panels to reveal a missense variant in the *MYH6* gene (c.2161C>T) associated with high risk of sick sinus syndrome (39). The ideal solution with multiple reference panels would be to call the variants in all reference panels jointly from their respective sequence alignment files for all the samples. The HRC panel was generated by jointly calling variants from the respective sequence alignment files from 20 contributing studies. However, variant calling is highly computationally intensive and may not be a feasible or practical solution for merging large reference panels. In the era of imputation servers, a question still to be answered is whether future methods will be able to combine imputation results from multiple servers (each using its own private panel) into a better set of posterior genotype probabilities for each sample than could be derived using a simple panel.

FUTURE DIRECTIONS

Genotype imputation is now a standard part of the human geneticist's toolbox. When combined with a large set of sequenced genomes, genotype imputation extends the resolution of genotyping arrays, allowing many variants that are not directly genotyped (particularly including rare variants and non-SNP variants) to be studied at scale. We expect that this capability will accelerate studies of rare variants, shortening the time between when whole-genome and whole-exome sequencing experiments can discover a variant and when scientists can explore its effect by examining downstream phenotypic consequences in large numbers of carriers.

As reference panels and human genetic studies continue to grow in scale, we expect continued research in imputation methods. There are several interesting challenges, including the potential for general-purpose machine learning methods—which are evolving rapidly, with the emergence of deep convolution neural networks and other highly efficient and effective computational techniques—to eventually compete with or displace the current methods based on HMM and positional Burrows–Wheeler transformation, which have been crafted specifically to model features and properties of human genetic data.

Imputation of human genetic variation uses detailed observation of a reference set of samples to enable understanding of unobserved genotypes in samples where only a scaffold of markers is measured. Imputation experiments are increasingly being attempted in other areas of biology and human genetics, and it will be interesting to see whether any of these methods and approaches

become as useful and ubiquitous as genotype imputation is in GWASs. Notable examples include ongoing attempts to impute gene expression levels and other types of genomic variation using GWAS array data (31, 90, 99) in order to search for associations between expression levels, protein levels, methylation patterns (81), and other genomic states and human diseases and traits characterized in large sets of genotyped individuals. Another notable example is the imputation of epigenomic states in detailed assessments of cellular function and biology (28). Here, a large set of epigenomic assays for expression levels, methylation patterns, transcription factor binding, and other information are carried out in a small set of cell lines; in parallel, a subset of these experiments are carried out in additional cell lines, forming a scaffold that can be used to predict results of missing assays.

DISCUSSION

In the last decade, the cost of sequencing a human genome has dropped from $10,000,000 to $1,000 (35). At the same time, continued improvements in genotype imputation methods have enabled genotype imputation to remain an attractive low-cost alternative to sequencing. The decrease in sequencing cost enhances the advantages of genotype imputation because large-scale human sequencing will make it possible to assemble reference panels of hundreds of thousands (and eventually millions) of individuals. These enormous reference panels will make it possible to accurately impute variants with very low population frequency. In turn, imputation will provide an accessible strategy for studying the variants discovered by sequencing in tens of millions of genotyped samples.

Imputation methods are now mature and highly efficient, which permits genotype imputation to be offered as a free web-based service (19). Some related areas that require further development include methods for merging reference panels and methods for merging multiple sets of imputed data when an individual has been imputed from multiple, possibly overlapping reference panels. In addition, the ability of imputation methods to impute non-single-nucleotide variation and structural variation will need to be continually reassessed as sequencing technology improves the number and genotype accuracy of these variants in reference panels. Current methods are more than four orders of magnitude faster than the first imputation methods based on the Li and Stephens haplotype mosaic model. These methods have evolved through successive iterations and refinements, and we anticipate that further iterations and refinements will be needed as the scale of human genetic data sets continues to increase.

Genotype imputation methodology has outstripped advances in reference panels (9), and there is a clear need for large reference panels that can take full advantage of the capabilities of current imputation methods. One obstacle to creating large reference panels is the computational challenge of jointly calling genotypes in large samples of sequenced individuals. This is an active area of methodological research, with recent methods leveraging cloud computing to address the computational demands (50). A second challenge is the scarcity of sequenced samples that are consented for general research use. In October 2017, only 1,245 individuals in the Database of Genotypes and Phenotypes (dbGaP) (64) compilation of individual-level genomic data for general research use (accession phs000688.v1.p1) had whole-genome sequence data. One approach to including samples with restrictive sample consents is to include restrictions on the use of reference panels. The sequence data for most of the individuals in the HRC reference panel are available via application to the European Genome-Phenome Archive (accession EGAD00001002729) (54). The HRC reference panel may be used for genotype phasing and imputation but not for any other purpose. Phasing and imputation servers provide an alternative way to perform phasing and imputation from large reference panels that have restrictive sample consents (19).

Beyond simply increasing reference sample size, there is also a need for reference panels to better represent human genetic variation (3) through the inclusion of deeply sequenced individuals from diverse populations. Large, diverse reference panels will ensure that the benefits of genotype imputation are available to all.

DISCLOSURE STATEMENT

G.R.A. is a member of the scientific advisory boards for 23andMe and Regeneron Pharmaceuticals and a consultant for Merck.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number R01HG008359 to B.L.B. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

LITERATURE CITED

1. 1000 Genomes Proj. Consort. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–73
2. 1000 Genomes Proj. Consort. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65
3. 1000 Genomes Proj. Consort. 2015. A global reference for human genetic variation. *Nature* 526:68–74
4. Al Olama AA, Kote-Jarai Z, Berndt SI, Conti DV, Schumacher F, et al. 2014. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat. Genet.* 46:1103–9
5. Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, et al. 2015. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* 6:10001
6. Anderson CA, Pettersson FH, Barrett JC, Zhuang JJ, Ragoussis J, et al. 2008. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am. J. Hum. Genet.* 83:112–19
7. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* 40:955–62
8. 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.* 84:210–23
9. Browning BL, Browning SR. 2016. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* 98:116–26
10. Browning BL, Yu Z. 2009. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am. J. Hum. Genet.* 85:847–61
11. Browning SR. 2008. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet.* 124:439–50
12. Browning SR, Browning BL. 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.* 81:1084–97
13. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, et al. 2017. Genome-wide genetic data on ~500,000 UK Biobank participants. bioRxiv 166298. **https://doi.org/10.1101/166298**
14. Chang D, Nalls MA, Hallgrimsdottir IB, Hunkapiller J, van der Brug M, et al. 2017. A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* 49:1511–16
15. Coon KD, Myers AJ, Craig DW, Webster JA, Pearson JV, et al. 2007. A high-density whole-genome association study reveals that *APOE* is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J. Clin. Psychiatry* 68:613–18

16. Cooper JD, Smyth DJ, Smiles AM, Plagnol V, Walker NM, et al. 2008. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat. Genet.* 40:1399–401
17. Danecek P, Auton A, Abecasis GR, Albers CA, Banks E, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156–58
18. Das S. 2017. *Next generation of genotype imputation methods*. PhD Thesis, Dep. Biostat., Univ. Mich., Ann Arbor
19. Das S, Forer L, Schonherr S, Sidore C, Locke AE, et al. 2016. Next-generation genotype imputation service and methods. *Nat. Genet.* 48:1284–87
20. Davies RW, Flint J, Myers S, Mott R. 2016. Rapid genotype imputation from sequence without reference panels. *Nat. Genet.* 48:965–69
21. Deelen P, Menelaou A, van Leeuwen EM, Kanterakis A, van Dijk F, et al. 2014. Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands.' *Eur. J. Hum. Genet*. 22:1321–26
22. Deutsch LP. 1996. *GZIP file format specification version 4.3*. RFC 1952, Netw. Work. Group, Internet Eng. Task Force, Fremont, CA. **https://tools.ietf.org/search/rfc1952**
23. Diabetes Genet. Replication Meta-Anal. (DIAGRAM) Consort., Asian Genet. Epidemiol. Netw. Type 2 Diabetes (AGEN-T2D) Consort., South Asian Type 2 Diabetes (SAT2D) Consort., Mex. Am. Type 2 Diabetes (MAT2D) Consort., Type 2 Diabetes Genet. Explor. Next-Gener. Seq. Multi-Ethnic Samples (T2D-GENES) Consort., et al. 2014. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* 46:234–44
24. Dilthey A, Leslie S, Moutsianas L, Shen J, Cox C, et al. 2013. Multi-population classical HLA type imputation. *PLOS Comput. Biol.* 9:e1002877
25. Duan Q, Liu EY, Auer PL, Zhang G, Lange EM, et al. 2013. Imputation of coding variants in African Americans: better performance using data from the exome sequencing project. *Bioinformatics* 29:2744–49
26. Dudbridge F. 2008. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum. Hered.* 66:87–98
27. Durbin R. 2014. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* 30:1266–72
28. Ernst J, Kellis M. 2015. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.* 33:364–76
29. Fisher RA. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Philos. Trans. R. Soc. Edinb.* 52:399–433
30. Fuchsberger C, Abecasis GR, Hinds DA. 2015. minimac2: faster genotype imputation. *Bioinformatics* 31:782–84
31. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, et al. 2015. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47:1091–98
32. Gibson G. 2012. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* 13:135–45
33. Glob. Lipids Genet. Consort. 2013. Discovery and refinement of loci associated with lipid levels. *Nat. Genet*. 45:1274–83
34. Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17:333–51
35. Green ED, Rubin EM, Olson MV. 2017. The future of DNA sequencing. *Nature* 550:179–81
36. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, et al. 2015. Large multiallelic copy number variations in humans. *Nat. Genet.* 47:296–303
37. Hao K, Chudin E, McElwee J, Schadt EE. 2009. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genet*. 10:27
38. Hoffmann TJ, Sakoda LC, Shen L, Jorgenson E, Habel LA, et al. 2015. Imputation of the rare *HOXB13* G84E mutation and cancer risk in a large population-based cohort. *PLOS Genet*. 11:e1004930
39. Holm H, Gudbjartsson DF, Sulem P, Masson G, Helgadottir HT, et al. 2011. A rare variant in *MYH6* is associated with high risk of sick sinus syndrome. *Nat. Genet.* 43:316–20
40. Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genet*. 5:e1000529

41. Howie BN, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* 44:955–59
42. Howie BN, Marchini J, Stephens M. 2011. Genotype imputation with thousands of genomes. *G3* 1:457–70
43. Huang J, Howie B, McCarthy S, Memari Y, Walter K, et al. 2015. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* 6:457–70
44. Int. HapMap Consort. 2003. The International HapMap Consortium. *Nature* 426:789–96
45. Int. HapMap Consort. 2005. A haplotype map of the human genome. *Nature* 437:1299–320
46. Int. HapMap Consort. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–61
47. Int. HapMap 3 Consort. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58
48. Int. Hum. Genome Seq. Consort. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931–45
49. Jia X, Han B, Onengut-Gumuscu S, Chen WM, Concannon PJ, et al. 2013. Imputing amino acid polymorphisms in human leukocyte antigens. *PLOS ONE* 8:e64683
50. Jun G, Wing MK, Abecasis GR, Kang HM. 2015. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* 25:918–25
51. Klein AP, Wolpin BM, Risch HA, Stolzenberg-Solomon RZ, Mocci E, et al. 2018. Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. *Nat. Commun.* 9:556
52. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* 308:385–89
53. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467:832–38
54. Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, et al. 2015. The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.* 47:692–95
55. Leslie S, Donnelly P, McVean G. 2008. A statistical method for predicting classical HLA alleles from SNP data. *Am. J. Hum. Genet.* 82:48–56
56. Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165:2213–33
57. 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.* 34:816–34
58. Li Y, Willer CJ, Sanna S, Abecasis GR. 2009. Genotype imputation. *Annu. Rev. Genom. Hum. Genet.* 10:387–406
59. Lin DY, Hu Y, Huang BE. 2008. Simple and efficient analysis of disease association with missing genotype data. *Am. J. Hum. Genet.* 82:444–52
60. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, et al. 2015. Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518:197–206
61. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, et al. 2016. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* 48:1443–48
62. Loh P-R, Palamara PF, Price AL. 2016. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* 48:811–16
63. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW. 2018. Fine-mapping of an expanded set of type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. bioRxiv 245506. **https://doi.org/10.1101/245506**
64. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, et al. 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* 39:1181–86
65. Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, et al. 2006. A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* 78:437–50
66. Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11:499–511
67. 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.* 39:906–13

68. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9:356–69
69. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, et al. 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* 48:1279–83
70. Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, et al. 2009. Genome-wide scan reveals association of psoriasis with IL-23 and NF-κB pathways. *Nat. Genet.* 41:199–204
71. Natl. Heart Lung Blood Inst. 2018. *Trans-Omics for Precision Medicine (TOPMed) program*. **https://www.nhlbi.nih.gov/science/trans-omics-precision-medicine-topmed-program**
72. Nicolae DL. 2006. Testing untyped alleles (TUNA)—applications to genome-wide association studies. *Genet. Epidemiol.* 30:718–27
73. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, et al. 2015. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* 47:1121–30
74. O'Connell J, Sharp K, Shrine N, Wain L, Hall I, et al. 2016. Haplotype estimation for biobank-scale data sets. *Nat. Genet.* 48:817–20
75. Okada Y, Wu D, Trynka G, Raj T, Terao C, et al. 2014. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506:376–81
76. Orho-Melander M, Melander O, Guiducci C, Perez-Martinez P, Corella D, et al. 2008. Common missense variant in the glucokinase regulatory protein gene is associated with increased plasma triglyceride and C-reactive protein but lower fasting glucose concentrations. *Diabetes* 57:3112–21
77. Pistis G, Porcu E, Vrieze SI, Sidore C, Steri M, et al. 2015. Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur. J. Hum. Genet.* 23:975–83
78. Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69:1–14
79. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81:559–75
80. Rabiner LR. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77:257–86
81. Rawlik K, Rowlatt A, Tenesa A. 2016. Imputation of DNA methylation levels in the brain implicates a risk factor for Parkinson's disease. *Genetics* 204:771–81
82. Saunders EJ, Dadaev T, Leongamornlert DA, Jugurnauth-Little S, Tymrakiewicz M, et al. 2014. Fine-mapping the *HOXB* region detects common variants tagging a rare coding allele: evidence for synthetic association in prostate cancer. *PLOS Genet.* 10:e1004129
83. Scheet P, 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.* 78:629–44
84. Schully SD, Yu W, McCallum V, Benedicto CB, Dong LM, et al. 2011. Cancer GAMAdb: database of cancer genetic associations from meta-analyses and genome-wide association studies. *Eur. J. Hum. Genet.* 19:928–30
85. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, et al. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316:1341–45
86. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, et al. 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* 42:937–48
87. Spencer CC, Su Z, Donnelly P, Marchini J. 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLOS Genet.* 5:e1000477
88. Taliun D, Chothani SP, Schönherr S, Forer L, Boehnke M, et al. 2017. LASER server: ancestry tracing with genotypes or sequence reads. *Bioinformatics* 33:2056–58
89. UK10K Consort. 2015. The UK10K project identifies rare variants in health and disease. *Nature* 526:82–90
90. Wang J, Gamazon ER, Pierce BL, Stranger BE, Im HK, et al. 2016. Imputing gene expression in uncollected tissues within and beyond GTEx. *Am. J. Hum. Genet.* 98:697–708
91. Wang Y, Lu J, Yu J, Gibbs RA, Yu F. 2013. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Res.* 23:833–42

92. Wellcome Trust Case Control Consort. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–78
93. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. 2008. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.* 40:161–69
94. Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, et al. 2009. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* 41:25–34
95. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, et al. 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* 46:1173–86
96. Yu Z, Schaid DJ. 2007. Methods to impute missing genotypes for population data. *Hum. Genet.* 122:495–504
97. Zhang J, Jiang K, Lv L, Wang H, Shen Z, et al. 2015. Use of genome-wide association studies for cancer research and drug repositioning. *PLOS ONE* 10:e0116477
98. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, et al. 2014. HIBAG—HLA genotype imputation with attribute bagging. *Pharmacogenom. J.* 14:192–200
99. Zhou W, Han L, Altman RB. 2016. Imputing gene expression to maximize platform compatibility. *Bioinformatics* 33:522–28