nature reviews genetics

https://doi.org/10.1038/s41576-025-00895-2

Review article

Advances in haplotype phasing and genotype imputation

Quan Sun[1,2] & Yun Li[3,4]

Abstract

Haplotype phasing – to determine which genetic variants reside on the same chromosome – and genotype imputation – to infer unobserved genotypes – have become indispensable steps to improve genome coverage for genomic analyses such as genome-wide association studies. Several tools exist for haplotype phasing and genotype imputation, all of which have continuously evolved to accommodate the increasing sample sizes of genomic studies and rapidly improving sequencing technologies. To fully leverage these recent advances, researchers must deliberate several practical considerations, including tool choice, quality control filters, data privacy concerns and reference panel choice. Looking ahead, long-read sequencing technologies are poised to bring novel opportunities to this field and drive methodological development.

Sections

[1]Center for Computational and Genomic Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [2]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, USA. [3]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [4]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. ✉e-mail: yunli@med.unc.edu

Introduction

Genotype information is increasingly abundant, as whole-genome sequencing (WGS) and whole exome sequencing costs have continuously decreased and, for studies with large sample sizes but limited funding, array genotyping is an affordable alternative for directly assaying up to ~5 million genetic markers. However, all these data are 'unphased', meaning they do not specify which alleles are inherited together on the same parental chromosome. The process of inferring phased haplotype information from unphased genotypes is called haplotype phasing (or, simply, phasing) (Fig. 1a), which is essential for various genetic analyses, including the calculation of population genetics statistics[1–6], identification of compound heterozygous events[7] and inferences and analyses incorporating local ancestry[8–12]. In addition, phasing is important for estimating missing or unobserved genotypes with the aid of reference panels (Fig. 1b), a process known as genotype imputation (or, simply, imputation)[13–15], which can greatly increase genome coverage and improve the power of genome-wide association studies (GWAS)[16,17]. Therefore, phasing and imputation have become standard practice in current population-scale GWAS with array genotyping data.

Computational methods for phasing and imputation have been proposed since the early 2000s, including PL-EM[18], PHASE[19], fastPHASE[20], Beagle[21], MaCH[22] and IMPUTE2[23]. These early methods for population-based statistical phasing explored different methodologies, including the greedy algorithm[24], expectation-maximization algorithm[18,25], long-range identity-by-descent (IBD)[26], phylogeny tree-based methods[27,28] and coalescent-based hidden Markov model (HMM). HMM coupled with the product of approximate conditionals (PAC) framework was first proposed by Li and Stephens[29,30] and remains predominant in currently used methods. Similarly, most genotype imputation methods have relied (and continue to do so) on PAC-based HMMs where genotypes and haplotypes are modelled conditionally on haplotypes of other individuals.

More recently, owing to the unprecedented increase in the scale of genetic studies in terms of number of individuals and genetic variants, methods have been developed and updated to improve computational efficiency (Fig. 1c). Such methods include Eagle2[31] and SHAPEIT5[32] for phasing; minimac4[13] and IMPUTE5[15] for genotype imputation; and Beagle5 for phasing[33] and genotype imputation[14]. In addition, large-scale imputation reference panels have been constructed for more accurate imputation with the ability to impute rarer variants[34,35]. Web-based imputation servers have also been developed to offer phasing and imputation service to the public without granting users direct access to individual-level genotype data in reference panels[34,36], providing invaluable resources to the community. Lastly, many studies are now focusing on practical considerations of imputation, including post-phasing and post-imputation quality control[37–41], the imputation of disease cohorts[42–44] and different ancestral populations[45–48], and the imputation of variants beyond single-nucleotide polymorphisms (SNPs)[49–52].

In this Review, we discuss recent method developments for haplotype phasing and genotype imputation in unrelated individuals, highlighting various studies to exemplify how these methods facilitate genomic analyses. We also summarize practical considerations for phasing and imputation, especially for post-phasing and post-imputation quality control. We then discuss recent developments enabled by long-read sequencing technologies, as well as promising future directions. Note that we will not discuss phasing and imputation specifically designed for family data, where it is possible to rely on IBD and Mendelian transmission rules compared with linkage disequilibrium-based phasing and imputation in unrelated population samples (the details of which can be found in a previous review[53]). Furthermore, we do not cover human leukocyte antigen imputation and refer interested readers to a previous tutorial[54].

Method developments for haplotype phasing

The HMM framework has been widely adopted and forms the basis of most of the current computational phasing methods, with parameters estimated iteratively using, for example, stochastic expectation-maximization algorithms[55]. More recently, these methods have been updated with computational improvements and to develop rare variant phasing, specifically in the case of Eagle2[31], Beagle5[33] and SHAPEIT5[32], which are state-of-the-art methods and widely used in various genomic analyses[34,35,56,57] (Box 1; see ref. 53 for a comprehensive review of the base frameworks).

Computational improvements

An HMM generally has three important components: hidden states, emission probabilities and transition probabilities. In an HMM for phasing, emission probabilities decide how to select the unobserved haplotypes based on the observed data, and transition probabilities determine how hidden states change along chromosome segments. Hidden states fully determine the underlying unobserved true haplotypes, which are often chosen from a haplotype pool constructed either from external reference haplotypes (phasing with reference) or from already phased haplotypes of the target samples (phasing without reference), thus leading to an important problem of haplotype matching. Given this framework, the computational burden increases quadratically with the number of reference haplotypes and, therefore, computational improvement has been a main focus of modern methods.

The development of the positional Burrows–Wheeler transform (PBWT)[58] was a milestone for modern phasing and imputation methods. It presents a series of algorithms for haplotype data compression and efficient haplotype matching, reducing the computational complexity from quadratic to linear in terms of the number of reference haplotypes. As such, Eagle2, Beagle5 and SHAPEIT5/SHAPEIT4[59] all deploy PBWT, albeit in different manners (Table 1). For instance, Eagle2 uses PBWT to represent the full set of haplotypes, whereas Beagle5 and SHAPEIT5/SHAPEIT4 use PBWT to identify customized reference haplotypes for each target individual.

Phasing of rare variants

As sequencing data from more individuals continue to amass, drastically more rare variants are discovered, far outnumbering common variants. For example, among ~400 million variants detected in the freeze 5 data of the Trans-Omics for Precision Medicine (TOPMed) programme, 97% have a minor allele frequency (MAF) below 1%, with 46% being singletons (minor allele count of 1) that are present in heterozygous form in only one out of the >53,000 individuals[34]. Such an emormous amount of particularly rare variants poses unprecedented challenges for phasing and imputation, which is another focus for modern methods.

Beagle5[33] was the first tool to provide a different approach for phasing rare variants compared with common variants. It adopts a two-stage algorithm where common variants (MAF >0.2%) are phased first to build a haplotype scaffold, and then rare variants are phased borrowing techniques from genotype imputation with phase information converted from imputed allele probabilities. This multistage phasing

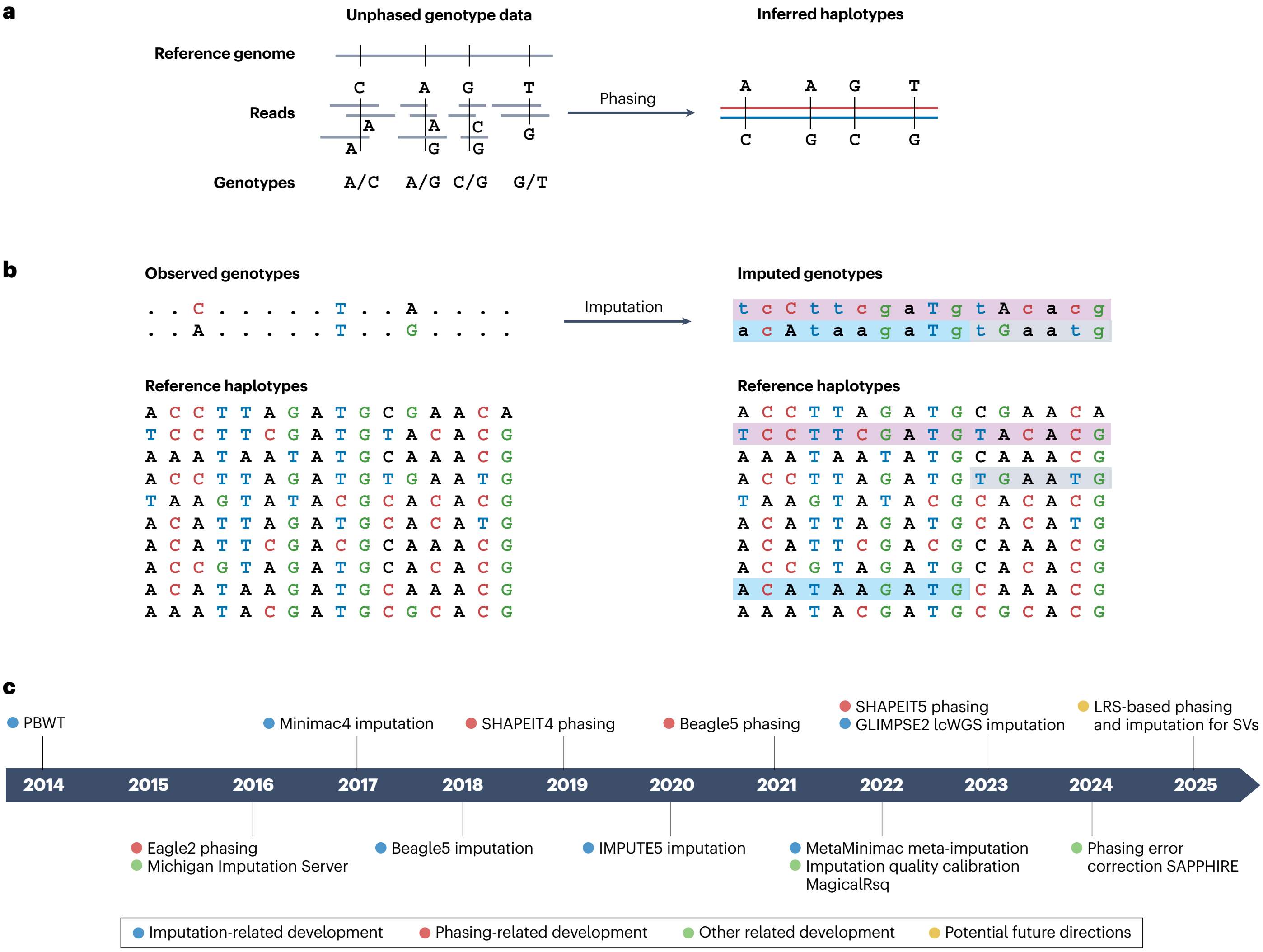
a
Unphased genotype data
Inferred haplotypes
Reference genome
Reads
C A G T
A A C G
A G G
Genotypes
A/C A/G C/G G/T
Phasing
A A G T
C G C G
b
Observed genotypes
. . C . . . . . . T . . A . . . . .
. . A . . . . . . T . . G . . . . .
Imputation
Imputed genotypes
t c C t t c g a T g t A c a c g
a c A t a a g a T g t G a a t g
Reference haplotypes
A C C T T A G A T G C G A A C A
T C C T T C G A T G T A C A C G
A A A T A A T A T G C A A A C G
A C C T T A G A T G T G A A T G
T A A G T A T A C G C A C A C G
A C A T T A G A T G C A C A T G
A C A T T C G A C G C A A A C G
A C C G T A G A T G C A C A C G
A C A T A A G A T G C A A A C G
A A A T A C G A T G C G C A C G
Reference haplotypes
A C C T T A G A T G C G A A C A
T C C T T C G A T G T A C A C G
A A A T A A T A T G C A A A C G
A C C T T A G A T G T G A A T G
T A A G T A T A C G C A C A C G
A C A T T A G A T G C A C A T G
A C A T T C G A C G C A A A C G
A C C G T A G A T G C A C A C G
A C A T A A G A T G C A A A C G
A A A T A C G A T G C G C A C G
c
PBWT
Minimac4 imputation
SHAPEIT4 phasing
Beagle5 phasing
SHAPEIT5 phasing
GLIMPSE2 lcWGS imputation
LRS-based phasing and imputation for SVs
2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025
Eagle2 phasing
Michigan Imputation Server
Beagle5 imputation
IMPUTE5 imputation
MetaMinimac meta-imputation
Imputation quality calibration MagicalRsq
Phasing error correction SAPPHIRE
Imputation-related development
Phasing-related development
Other related development
Potential future directions

Fig. 1 | Conceptual basis and technological development of phasing and imputation. a, A conceptual illustration of phasing. After read alignment with reference genome, we can infer or call genotypes of target individuals, but phase information (that is, information about which alleles are inherited together on the same parental chromosome) is unknown. Phasing is the process to make such inference starting from unphased genotype data. **b**, A conceptual illustration of imputation from array genotype data. Imputation is the process to infer genotypes at untyped markers with the aid of reference panels. Heuristically, it identifies haplotype segments in reference panels that match genotypes at typed markers for imputation of target individuals and then imputes by simply copying over the shared segments. In the right panel (after imputation), imputed genotypes at untyped markers for the target sample are denoted with lower-case letters, with the colour representing the corresponding reference haplotype from which the alleles are copied. **c**, A timeline of recent major developments in phasing and imputation, which begins from the introduction of positional Burrows–Wheeler transform (PBWT), a highly efficient method for haplotype representation that paved the road for more recent phasing and imputation methods focusing on computational improvements. A timeline of earlier evolvement (before 2018) is detailed in ref. 77. lcWGS, low-coverage whole genome sequencing; LRS, long-read sequencing.

strategy was further extended by SHAPEIT5[32], where they adopted a three-stage phasing for common variants (MAF >0.1%), non-singleton rare variants and, specifically, singletons. In brief, common variants are processed with its earlier version SHAPEIT4[59]; rare variants are phased similar to Beagle5; and singleton phasing is achieved through IBD sharing patterns (see Box 1).

Practical considerations for haplotype phasing

Given the recent methodological developments specifically for Eagle2, Beagle5 and SHAPEIT5, an important practical question is which method to use. This decision can be based on the performance of different phasing methods resulting from phasing accuracy and other practical considerations.

Phasing accuracy

Phasing accuracy is commonly quantified on the basis of phasing errors, which are also referred to as switch errors, where two estimated haplotypes are incorrect by being switched versions of the truth. A special case of switch error is called flip errors, in which the switched segment is a single base pair where the two alleles are flipped into the wrong

Box 1 | Frameworks of phasing methods

Eagle2[31] adopts a haplotype copying model similar to previous HMM-based methods, with distinctions in haplotype structure representation and selection of diplotypes, that is, phased haplotype pairs. Most earlier methods approximate the haplotype structure, for example, by merging haplotypes into local clusters[21,167], for computational efficiency. By contrast, Eagle2 represents the full list of haplotypes in a tree structure by using the positional Burrows–Wheeler transform (PBWT)[58]. It further explores diplotypes using a branching-and-pruning beam search, removing unlikely phase paths to limit the search space and achieve computational efficiency. Note that Eagle2 was primarily designed for phasing with reference, but it also implemented phasing without reference, where identity-by-descent (IBD)-based long-range phasing method from its prior work, Eagle1[168], was adopted as an initial step. Compared with Eagle1[24] (which showed decreased accuracy with sample size <50,000), Eagle2 achieved higher accuracy for both small and large haplotype reference panel sizes, as demonstrated by the developers for reference sizes ranging from 15,000 to 100,000 haplotypes.

Beagle5[33] takes the Li and Stephens HMM[29] using a parsimonious state space of individual-specific composite reference haplotypes[14], with a sliding-window approach for memory efficiency. In addition, it implements a two-stage phasing algorithm designed specifically for phasing the increasingly large number of rare variants. The algorithm first phases common variants (minor allele frequency (MAF) >0.2%) with a progressive phasing methodology to build the haplotype scaffold consisting of estimated phase information for common variants, and then infers haplotypes for rare variants borrowing strategies from genotype imputation with phase information converted from imputed allele probabilities.

SHAPEIT5[32], built upon its prior work, SHAPEIT4[59], was designed for phasing rare variants in the context of biobank-scale WGS or whole exome sequencing data. It implements three different phasing modes for three variant types defined by MAF: common variants (MAF >0.1%), rare variants (MAF ≤0.1% and minor allele count >1) and singletons (minor allele count of 1). First, common variants are phased using SHAPEIT4[59], which applies PBWT for computational improvement and offers options to integrate additional layers of information such as reference haplotypes, long-read sequencing data or haplotype scaffold (with phase known for a subset of genotypes). Second, the phased common variants serve as a haplotype scaffold to further phase rare variants, which takes a similar streategy to Beagle5[33], with a sparse data representation where monomorphics (all genotypes being homozygotes of the major allele) are discarded. Finally, SHAPEIT5 phases singletons by leveraging the longest IBD sharing patterns between each target haplotype and the conditioning haplotypes, based on the assumption that singletons reflect recent mutation events and are therefore expected to share the least with other haplotypes. It therefore assigns the minor allele of a singleton to the target haplotype with the shortest shared segment[169].

haplotypes (Fig. 2a). The three major phasing methods, Eagle, Beagle and SHAPEIT, were reported to have similarly high accuracy and low error rates, especially from earlier studies focusing on phasing common variants. For example, phasing accuracy was assessed for Beagle[60], SHAPEIT2[61], SHAPEIT3[62], Eagle2 and two IBD-based methods, SLRP[63] and AlphaPhase[64], showing that SHAPEIT2, SHAPEIT3 and Eagle2 provided the most accurate results in simulated data[65]. Another study compared the phasing accuracies of fastPHASE, Beagle4, IMPUTE2, MaCH, SHAPEIT2, HAPI-UR[66] and Eagle2 and reported that Eagle2, SHAPEIT2 and Beagle4 performed the best alternately in different scenarios, with Eagle2 being the most stable method[37]. Finally, the phasing accuracies of AlphaPhase, Eagle2, SHAPEIT2, SHAPEIT4, Beagle3, Beagle4, Beagle5 and FImpute[67] were compared in another study using cattle pedigree data, which found that either SHAPEIT4 or Beagle5 was the most accurate method under various scenarios[68]. The same study also indicated that most tools achieved high accuracy at short genomic distances (<1 Mb), with minimal differences across these methods[68]. Acknowledging the strengths and weaknesses of these methods, some studies explored the consensus haplotype estimator by aggregating results from different methods and taking the most frequent haplotypes voted across method outputs as the final phased results[37,38]. Such consensus strategies for phasing were shown to achieve more accurate results than each individual method; however, running multiple tools may be computationally intensive, such that practical applications should consider the balance between improved accuracy and higher computational costs.

Although the studies reviewed above provided important insights into phasing accuracy comparison, there is still an insufficient number of benchmarking studies of phasing methods – especially for recent methods focusing on improving computational efficiency for biobank-scale studies (and, thus, for rare variants) – partly owing to the difficulty of obtaining the gold-standard truth of phased haplotypes. Most evaluation studies rely on phased results from family data, but these also have some limitations. For example, regions with low variant density can lead to inaccuracies in family-based phasing[69]; in addition, observed Mendelian inconsistencies due to genotyping errors or de novo mutations may be mistreated as phasing errors[70]. The advantages of long-read sequencing technologies allow for the resolution of longer haplotypes, which may serve as an alternative standard for phasing evaluation.

Owing to their abundance and rareness, phasing error rates for rare variants are substantially higher compared with common variants[32,39]. To improve phasing accuracy for rare variants, a recent post-phasing correction method, SAPPHIRE, has been developed by integrating phasing results and read-based information[39]. It first identifies poorly phased variants based on phasing confidence scores from SHAPEIT5, extracts their associated sequencing reads and then performs rephasing to either validate the current phase or flip the alleles, motivated by the fact that most incorrectly phased rare variants are in the form of flip errors (Fig. 2a). Such correction, based on the SAPPHIRE manuscript, can decrease error rates by up to 17%, particularly for singletons. However, this method requires raw reads data for correction, which may not be readily available and can be rather demanding in storage costs. As noted by the authors, sequencing reads from 200,031 individuals in the UK Biobank require 3.5 petabytes of storage – an amount that is infeasible for most researchers.

Other practical considerations

Beyond accuracy, computational considerations (for example, computional efficiency and hardware requirements such as graphics processing units[71,72]) and study aims (for example, special interest in rare variants and the need to use some specific reference panels such as the four-digit multi-ethnic human leukocyte antigen reference[73]) will also impact practical choices of phasing methods. For example, for phasing ultra-low-coverage WGS (ulcWGS) (that is, sequencing depth ~0.1–1×, which is much shallower than the earlier low-coverage WGS in, for example, the 1000 Genomes project[74]) where genotypes are not determined but represented with genotype likelihood, Beagle5 or SHAPEIT5 are more desirable as they can handle such genotype uncertainty — a feature that Eagle2 does not support. Furthermore, because cloud computing has been widely adopted in the genetics community (for example, the UK Biobank Research Analysis Platform), computational cost is a major consideration. The computational time of Beagle5 and Eagle2 scales linearly with target sample size, whereas SHAPEIT4 claims sublinear scaling by its developers[59]. SHAPEIT5 further improved its computational efficiency over SHAPEIT4 by using a parallelization scheme for the PBWT construction for common variants phasing[32]. Therefore, investigators may choose SHAPEIT5 for large biobank-scale studies with computational efficiency and when studying rare variants is a top priority. Notably, as most methods have implemented a predetermined parameter for the number of conditioning or reference haplotypes of each target haplotype (HMM state size), computational time will generally not be affected by the size of reference panels[31]. Some other considerations include specific variant types of primary interest, including multi-allelic structural variants (SVs) or variants in the major histocompatibility complex region. Currently, only Beagle5 directly supports multi-allelic variants and copy number variations (if encoded correctly). In addition, all the methods may not perform well for phasing long SVs.

As the ultimate goal of most studies is not phasing itself but rather downstream genotype imputation and association, as well as various haplotype-level analyses (Box 2 and Fig. 2b), investigators may prioritize choice of reference panels or ease of use over the actual phasing methods, especially when the focus is on common variants for which most methods perform highly similarly. Imputation servers have precompiled phasing, imputation and reference panels all together in one place, substantially improving accessiblity to general researchers. Therefore, we suggest future method developers to implement their methods on major imputation servers or cloud-based platforms to broaden the impacts.

Methods for genotype imputation

Early imputation methods inferred genotypes at ‘untyped’ (that is, unobserved) markers directly from genotype data. However, computational costs increase quadratically with sample size when modelling

Table 1 | Summary of haplotype phasing and genotype imputation methods

Haplotype phasing			
Method	Eagle2	Beagle5	SHAPEIT5
Algorithm summary	PBWT haplotype representation, beam search for phase paths	Two-stage phasing	Three-stage phasing
Reference panel	Full reference representation	Customized reference for each target sample	Customized reference for each target sample
Specially handle rare variants	No	Yes	Yes, even considering singletons
Allow multi-allelic variants	No, split into biallelic required	Yes	No, split into biallelic required
Allow genotype likelihood	No	No, but Beagle4 does	Yes
Sporadic missing data imputed	Yes	Yes	Yes
Input file format	vcf	vcf	vcf, bcf
Output file format	vcf	vcf	vcf, bcf, xcf
Web availability	Michigan imputation server TOPMed imputation server	UK Biobank research analysis platform	UK Biobank research analysis platform
Genotype imputation			
Method	Minimac4	Beagle5	IMPUTE5
HMM state space	Collapsed reference sequences matched at genotyped positions	A fixed number of composite reference haplotypes	A subset of reference haplotypes represented with PBWT
HMM state space depends on number of reference haplotypes	Yes	No	Yes
Reference panel file format	m3vcf, msav	bref3	imp5
Reference file format relative size	1×	~0.15×	~0.2–0.3×
Output file format	vcf	vcf	vcf, bcf, bgen
Imputation quality metric	Rsq (based on haplotype dosage)	DR2 (based on haplotype dosage)	INFO (based on statistical information of population allele frequency)
Web availability	Michigan imputation server TOPMed imputation server	UK Biobank research analysis platform	UK Biobank research analysis platform

HMM, hidden Markov model; PBWT, positional Burrows–Wheeler transform; TOPMed, Trans-Omics for Precision Medicine.

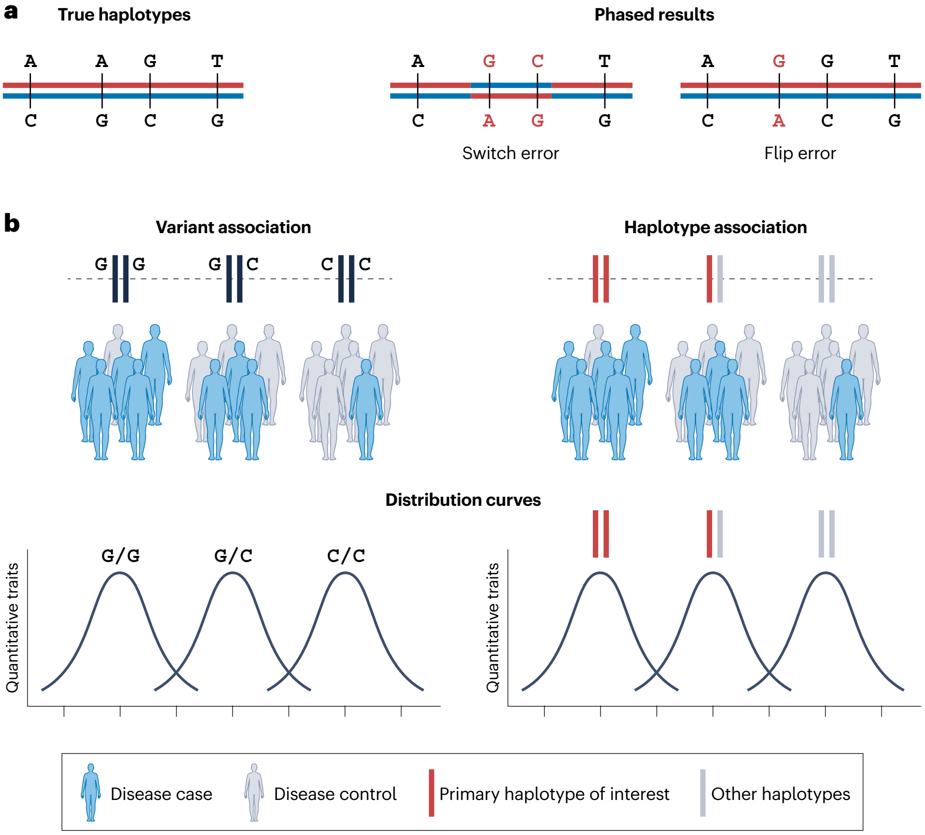

Fig. 2 | Post-phasing considerations and analyses. **a**, Two types of phasing error. Phasing errors can be classified into two categories: switch and flip errors. Switch errors refer to cases where, in certain chromosomal regions, the two haplotypes are incorrectly phased as switched versions of the true haplotypes. These errors are illustrated by regions with mismatched haplotype colours and the affected allele pairs highlighted in red. Flip errors refer to phasing errors at a single base pair where the two alleles are located in the wrong haplotypes. **b**, A comparison between variant association and haplotype association for both binary disease and quantitative trait under an additive model. Variant association focuses on the tests between some phenotype (a binary disease, or quantitative trait as illustrated here) and the genotypes at a specific genome position. Under the additive genetic model, variants are coded as 0, 1 or 2, representing the number of copies of a specific allele. Haplotype association with a primary haplotype of interest (the red haplotype) similarly counts copies of this haplotype and lumps all other haplotypes (grey) as a contrast group. Panel **b** is adapted with permission from ref. 12, Elsevier.

unphased genotype data, prompting the development of 'prephasing' (that is, phasing the target cohort before imputation), which drastically reduces the computational burden (from quadratic to linear complexity) and has become standard practice nowadays. Notably, sporadic missing genotypes in certain individuals are usually imputed in the prephasing step for many phasing methods.

For both genotype- and haplotype-based imputation, heuristically, the imputation task is fulfilled by identifying haplotype segments in reference panels that match imputation target individuals and then imputing the target by simply copying over the shared segments. Several tools exist for imputation from array genotypes (Box 3; see refs. 75–77 for comprehensive reviews of the base frameworks) as well as ulcWGS.

Imputation from array genotypes

Imputation was first developed to handle array genotypes, which range from several hundred thousand to ~5 million markers. Currently, the most widely used methods for genotype imputation from array genotypes are Minimac4[13], Beagle5[14] and IMPUTE5[15], all of which have undergone recent computational improvements to handle increasingly larger reference panels and target cohorts by reducing HMM state space and using new file formats. The Minimac series began as MaCH[22] and was later renamed to Minimac[13,78] when it adopted prephasing before imputation. Compared with its earlier versions, where each individual reference haplotype was treated as a hidden state, Minimac4 improves efficiency by collapsing reference sequences that match at genotyped positions within a chromosome chunk[77], which improves memory efficiency and computational speed. Similarly, Beagle5 reduces computational costs by constructing a fixed number of target-specific composite reference haplotypes through a non-overlapping sliding-window approach to ensure that the personalized composite reference covers the entire window while enabling parallel computing[14]. IMPUTE5 incorporates PBWT for haplotype representation to select a subset of reference haplotypes specifically for each target, reducing computational complexity to linear or even sublinear with respect to target sample sizes[15].

The above three methods also developed their preferred file formats of reference panels for efficient haplotype storage and fast reading (Table 1). Minimac4 takes mvcf format (specifically msav, a new mvcf format)[79], which was first developed in Minimac3 (m3vcf format). This format records only unique representative haplotypes within small genomic segments, following a structure similar to that of vcf files. Beagle5 uses the bref3 format, which similarly partitions chromosomes into chunks and applies different strategies for common and rare variants. This format records only the indices of haplotypes carrying minor alleles for rare variants, while storing common variants as unique allele sequences with pointers linking each haplotype to the corresponding sequence. The bref3 format uses two bytes to store each index compared with only one byte in bref2, which allows for longer chromosome chunks and higher compression of reference files. The imp5 format applied in IMPUTE5 is similar to bref3 in principle, with index recorded for efficient region extraction.

Imputation for ulcWGS

Beyond Minimac4[13], Beagle5[14] and IMPUTE5[15], other methods have also been proposed, including those specifically focusing on imputation from ulcWGS data. As high-coverage WGS is still cost-prohibitive for large-scale samples, ulcWGS, an alternative approach to array genotyping, has been proposed and demonstrated to achieve similar coverage of common variants but better coverage for low-frequency variants compared with array genotyping[80]. More recently, ulcWGS is widely adopted in ancient genome imputation where the coverage is only ~0.1–1× (refs. 81,82). Owing to the low-coverage nature of this type of data, many genomic regions are covered with no or few reads for any given individual, and genotypes are represented in forms of likelihood rather than discrete genotype calls[83]. In these scenarios, standard imputation methods from array data may not be applicable; therefore, methods specifically designed for ulcWGS data have been proposed, including those for human diploid imputation leveraging large reference panels (for example, GeneImp[84], GLIMPSE[83], QUILT[85] and GLIMPSE2[86]) and methods without references (for example, findhap[87], STITCH[88] and magicImpute[89]). Methods in the latter category have been used for imputing genotypes of non-human organisms as well.

Reference-based ulcWGS imputation methods are similar to array imputation in terms of how reference panels are used. Most of the methods also adopt the PAC framework[29] or its variations where conditional probabilities of observed data on reference haplotypes are inferred. Furthermore, GLIMPSE and GLIMPSE2 both apply PBWT (similarly to IMPUTE5) to ensure rapid selection of a subset of most relevant references. However, ulcWGS imputation differs from array imputation in both input format and the core algorithms. Array imputation methods take discrete genotype calls as input and match with reference panels based on these genotyped markers, whereas ulcWGS imputation methods take sequencing reads (as in QUILT) or reads-derived genotype likelihoods (as in GLIMPSE and GLIMPSE2) as input, without requiring deterministic genotypes to start. Therefore, probability-based Gibbs sampling[90] is more widely applied[83,85,91] to account for the higher level of uncertainty in input data. Notably, Beagle4 (but not Beagle5) allows genotype likelihood as input for imputation and therefore can also be easily used for ulcWGS data[92,93].

Practical considerations for imputation

Imputation requires several practical considerations. With increasingly larger reference panels available, it is essential to maintain best practices of pre-imputation quality control, data privacy on imputation servers, choosing a reference panel and assessing imputation quality.

Box 2 | Haplotype-level analyses

Phasing enables haplotype-based associations with phenotypes. In contrast to genome-wide association studies (GWAS), which have been widely applied to screen the whole genome and test the association of each genetic variant individually, haplotype association is often used at a specific locus encompassing multiple variants, especially for loci with multiple distinct signals[170–173]. For example, a study of COVID-19 focused on the chr12q24.13 locus encoding OAS1–OAS3 antiviral proteins[170]. The authors found that the risk of hospitalization was associated with a common *OAS1* haplotype composed of two derived human-specific risk alleles, rs10774671-A and rs1131454-A, which they further confirmed functionally to provide a molecular mechanism explaining COVID-19 severity. These insights would not have been possible without population-based phasing. In addition, haplotypes can help to determine distinct variant effects while accounting for the haplotype background. For example, a study of cystic fibrosis patients identified rs146704092 (p.Val172Met, c.514G>A), a missense variant in *SLC26A9*, to be statistically significantly associated with meconium ileus after conditioning on previously identified associated variants[174]. Noticing that 84 of the 93 p.Val172Met alleles are on a common haplotype that is also known to be associated with meconium ileus, the authors tried to investigate whether this novel variant association could be accounted for by this haplotype. By controlling for the haplotype background — restricting the analysis to individuals with at least one copy of this haplotype — the authors found that the p.Val172Met allele remained associated, with an even greater risk for meconium ileus. Thus, the authors concluded that the association of p.Val172Met cannot be fully explained by its haplotype background, yet again demonstrating the role of haplotypes in phenotypic associations.

Moreover, haplotypes can help explain missing heritabilities[175–177]. A recent study on oculocutaneous albinism (OCA), a rare genetic disorder of pigment production with notably missing heritability, focused on the tyrosinase (TYR) enzyme, as alleles that reduce TYR function are among the most common causes of OCA[175]. Previous GWAS studies suggested that two common variants, rs1042602 and rs1126809, play roles in common pigmentation variation worldwide[178]. Although both variants were associated with reduction in TYR enzyme activities and protein levels[86,179], the full effects were unknown. The authors demonstrated that a disease-causing haplotype, p.[Ser192Tyr; Arg402Gln] (*cis*-YQ), formed by the minor alleles of these two variants, is the most common disease-causing allele (19.1%) for type 1 OCA. Additional haplotypes are also discovered at this locus with potential pathogenic effects on OCA. These results showcase the value of phased haplotypes at the *TYR* locus to comprehensively understand disease-causing alleles for OCA[175], which sets up an example of how phasing can help investigate genetic mechanisms underlying diseases and aid genetic diagnosis.

Haplotype also makes various other downstream analyses possible. For example, local ancestry inference requires phased haplotypes[8,9], and it further enables analyses using local ancestry information[12], for instance, admixture mapping[11,180], GWAS and polygenic risk scores in admixed individuals[10,181–184]. In addition, a recent study demonstrated the potential of leveraging haplotype information for heritability estimation and polygenic risk score development, showing that haplotype-based approaches consistently outperformed genotype-based methods[185]. Future studies could explore the idea of combining genotype and haplotype information to create more accurate and transferable polygenic risk scores to guide clinical research and precision medicine.

Box 3 | Framework of imputation methods

Minimac4[13] adopts the positional Burrows–Wheeler transform (PBWT)[58] for efficient haplotype representation, which allows rapid search and match between target and reference haplotype segments. It collapses reference sequences matched at genotyped positions within chromosome chunks as hidden states, compared with all the individual-level haplotypes in its earlier versions, which greatly improves computational efficiency.

Beagle5[14] constructs a fixed number of target-specific composite reference haplotypes with a sampling-based approach, where each marker window (imputation unit) is divided into consecutive, fixed-length and non-overlapping intervals to ensure the entire marker window will be covered while facilitating parallel computing. It also develops a new file format (bref3) to construct a compact haplotype reference graph, which reduces the computation time required to read in large reference panels. Notably, Beagle5 is the only method that directly supports imputation of multi-allelic variants, without requiring them to be split into multiple biallelic records.

IMPUTE5[15], similar to Minimac4, also applies PBWT[58] for efficient long-range haplotype matching. It selects target-specific hidden states by identifying a subset of best-matching haplotypes that share long identical by state segments with the target. In addition, it also develops a new reference file format (imp5) with index incorporated, which enables quick imputation of a specific genome region. Notably, IMPUTE5, demonstrated by its developers, exhibits sublinear scaling with reference panel sample size, outperforming Minimac4 and Beagle5 in terms of computational speed.

Pre-imputation quality control

Pre-imputation quality control is usually performed to remove samples and variants with, for example, high missing rates, variants with low allele frequency or uncertain strand information. Notably, pre-imputation quality control sometimes also considers the Hardy–Weinberg equilibrium — especially in the early years of implementation — but it may not be appropriate to include in multi-ancestry populations or case–control studies, as it could remove ancestry-specific or disease-associated variants[94–96]. Pre-imputation quality control also includes allele check and strand flip, with caveats for palindromic SNPs (or ambiguous SNPs, with alleles A/T or C/G) as their strand information is not trivial and is usually determined by allele frequency compared with a certain reference panel.

Imputation servers and data privacy

Imputation of human genomes benefits from large-scale reference panels, which are almost impossible to make publicly available owing to privacy concerns and their huge sizes. Given this limitation, researchers have developed imputation servers such as the Michigan Imputation Server, where array genotype data are uploaded for initial quality control, followed by prephasing and imputation with various user-specified controlled-access reference panels that remain secure on the server[36] (Fig. 3a). Imputation servers are exemplary in promoting and accelerating broader usage of reference panels without requiring the researcher to access the individual genotypes within the panels. With open-source code, other imputation servers have also been developed to enable access to specific reference panels, including the TOPMed Imputation Server (for the TOPMed reference panel)[34], NBDC-DDBJ Imputation Server (for Japanese Genotype–Phenotype Archive data)[97] and Taiwan Biobank Imputation Server (for the Taiwan Biobank reference panel)[35], all which provide different ancestry compositions (Table 2).

Some studies have raised privacy concerns for individuals contributing their genotypes to reference panels that are supposedly secure behind the 'walls' of imputation servers, as these reference haplotypes may potentially be reconstructed from artificially designed target haplotypes[98,99]. Recently, a resampling-based approach for sharing reference panels was proposed trying to address the issue[100], but the reduced imputation accuracy in diverse populations has not been comprehensively evaluated. We note that it is not trivial to reconstruct haplotypes in reference panels behind imputation servers, which requires the development of adversarial algorithms[99]; thus, we strongly urge researchers to use imputation servers and reference panels responsibly, without trying to break behind the walls. In parallel, we also encourage developers of imputation servers to impose stronger data protection.

Beyond privacy concerns, current imputation servers have other limitations that warrant future improvements. First, owing to the General Data Protection Regulation restrictions from the European Union (EU), EU-based researchers cannot upload their data to non-EU-based imputation servers. Consequently, the Helmholtz Munich Imputation Server was developed to host genotype data from EU-based researchers, but it has a limited set of reference panels that constrains its use[101]. To overcome these issues, stakeholders should agree to broaden global usage of imputation servers and reference panels. Second, the maximum sample sizes allowed by imputation servers are limited (for example, 110,000 for Michigan Imputation Server and 25,000 for TOPMed Imputation Server). A tool to merge different batches of imputed data was developed as a solution[102], but we anticipate imputation servers will have higher capacities as computational power increases in the future. Third, most — if not all — imputation servers were developed only for genotype-based imputation, without supporting ulcWGS imputation for ancient genomes. Future efforts may consider accommodation of ulcWGS data on imputation servers.

Reference panels

Imputation quality depends highly on reference panels, which may surpass the impact of different imputation methods. When imputation methods were first introduced, researchers could use the single HapMap reference panel[103]. Now, imputation servers have enabled public access to a wide compendium of restricted reference panels, which requires researchers to decide which reference panel to use. Notably, reference panels are often linked to a specific imputation method (Table 2); for example, many imputation servers adapted from the Michigan Imputation Sever typically implement Eagle2 for phasing and Minimac4 for imputation. As such, choice of reference panel is occasionally prioritized over imputation method, as the former greatly impacts imputation quality and there is minimal difference between imputation methods.

Studies have demonstrated that cosmopolitan reference panels consisting of multiple populations or panels better matching genetic ancestry of target individuals perform better than panels based on a single continental ancestry population[75,104,105]. For example, the TOPMed reference panel has demonstrated improved imputation quality and facilitated novel association discoveries in African American and Hispanic or Latino populations[17,106]. Furthermore, many recently constructed population-specific panels have proved beneficial especially in the context of low-frequency and rare variants[107–110]. In addition, case cohorts, where disease-causing allele frequencies differ from general populations, can also benefit from cohort-specific panels at disease-causing genes[43,111]. However, owing to the sample size difference between the general population and cohort-specific panels, the former still perform better for the vast majority of the variants[43]. Therefore, it is valuable to effectively combine imputation results from multiple reference panels, also known as meta-imputation and enabled by MetaMinimac[112].

MetaMinimac[112] combines imputation results from two or more reference panels as a weighted average of the estimated allele counts from each result set (Fig. 3b). Weights are determined by the empirical performance of each panel using leave-one-out imputation at genotyped markers, and are region and individual specific, estimated with an HMM. Results have shown that such meta-imputation can improve imputation quality for non-European populations, especially for low-frequency and rare variants[112–114], consistent with the variant categories that benefit most from population-specific panels. However, one study also reported meta-imputation failed to improve imputation quality for some populations[48], but the authors' evaluation was based only on estimated imputation quality, which may differ dramatically from true imputation quality[40,41]. Looking ahead, more systematic evaluations are needed based on true imputation quality calculated from gold standard genotypes derived from genotyping or sequencing data.

Imputation quality

Although imputation approaches are powerful to expand genome coverage and facilitate various genomic analyses[115–117], they have deficiencies that must be assessed through imputation quality. For example, increasingly large reference panels are composed mostly (>90%) of rare variants[34], which are harder to impute[15,45,118]. To prevent these poorly imputed variants from hindering downstream GWAS and polygenic risk score calculations[119–121], imputation quality must be carefully assessed, and poorly imputed markers are subsequently filtered out. The gold standard approach to evaluate imputation quality of an untyped marker is to calculate the squared Pearson correlation (true R^2) between imputed dosages and the truth across all individuals. It can be calculated at both the genotype and haplotype levels (as in MaCH[22] and Minimac[13,78], respectively), but requiring true genotypes or haplotypes makes true imputation quality unavailable in real applications to impute unknown genotypes. As such, Minimac4 provides the true imputation quality only for genotyped variants, where imputation is performed by masking genotyped variants and treating them as untyped.

Given the limited availability of true imputation quality, various estimated imputation quality metrics have been proposed and usually are directly generated from imputation software. Specifically, Minimac4 reports haplotype-level estimated R^2 (Rsq), defined as the empirical variance of haplotype dosage from imputed data divided by the expected variance given the estimated allele frequency[13,78]. DR2 in Beagle5 takes a similar approach to Rsq in Minimac4 but it explicitly models posterior variance of genotypes given the data[77]. IMPUTE5, instead of using estimated R^2, calculates estimated statistical information (termed INFO) of allele frequency in the imputed and the true genotypes. The above three metrics are proven to be equivalent under Hardy–Weinberg equilibrium[76,77] and are reliable to estimate imputation quality for common variants. However, they are shown to be less accurate for low-frequency and rare variants[40,41,77]; therefore, relying solely on estimated quality metrics to assess imputation accuracy may be inappropriate when comparing imputation strategies or reference panels, especially for rarer variants.

Recent studies have proposed new quality metrics leveraging machine learning techniques, namely MagicalRsq[40] and MagicalRsq-X[41]. MagicalRsq uses imputation summary statistics (Rsq and MAF) along with external population genetics data to train an XGBoost model,

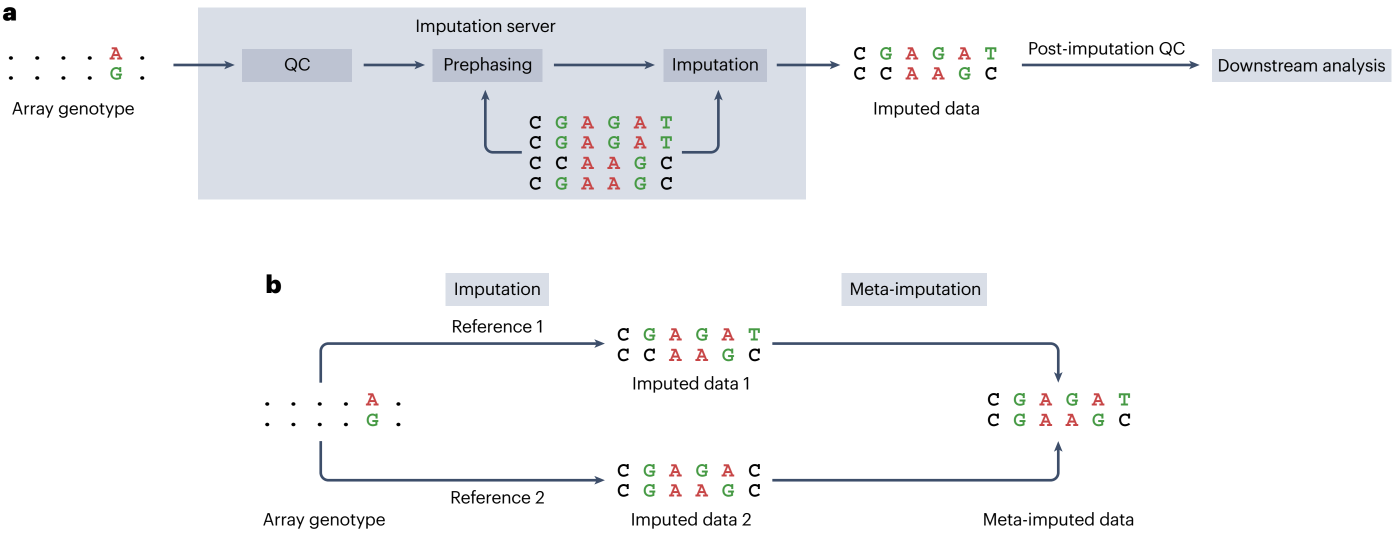

Fig. 3 | Imputation servers, reference panels and meta-imputation. a, An illustration of the imputation workflow starting from array genotype data. Quality control (QC) is first performed, followed by prephasing and imputation. All these steps are integrated in major imputation servers, where users can select reference panels for prephasing and imputation. After obtaining the imputed data, post-imputation QC is performed to remove poorly imputed variants for downstream analyses. **b**, An illustration of two-way meta-imputation. Starting from the same array genotype data, two imputations with two different reference panels are separately performed. Differences manifest between the two results, for example, at the second and the last markers in the illustration. Meta-imputation is performed where the two separately imputed results are combined to generate a consensus or meta-imputed results.

Table 2 | Reference panel characteristics

Reference panel	Number of samples	Number of variants	Population	Continental origins	Available imputation servers
HRC (version r1.1 2016)	32,470	39,635,008	European	EUR	MIS, Sanger, HMIS
1000 G Phase 3 (version 5)	2,504	49,143,605	Multiple	EUR, AFR, NAM, SAM, EAS, SAS	MIS, Sanger, HMIS, TWB, Afrigen-D
TOPMed r3	133,597	445,600,184	Multiple	EUR, AFR, NAM, SAM, EAS, SAS	TOPMed
UK10K	3,781	45,492,035	European	EUR	Sanger
GAsP	1,654	21,494,814	Asian	EAS, SAS	MIS
CAAPA	883	31,163,897	African American	EUR, AFR, NAM	MIS
Taiwan Biobank 1.5k	~1,500	NA	East Asian	EAS	TWB
H3African v6	4,447	130,028,596	Multiple (~50% African)	AFR, EUR	Afrigen-D

AFR, African; EAS, East Asian; EUR, European; HMIS, Helmholtz Munich Imputation Server; NA, not available; NAM, North American; SAM, South American; SAS, South Asian; TWB, Taiwan Biobank Imputation Server.

treating each variant as an observation. It incorporates additional genotypes from a subset of study samples or from variants not used in the imputation process. Such models are shown to be better calibrated compared with the original Rsq, particularly for lower-frequency variants. MagicalRsq-X extends its predecessor by borrowing information from external cohorts and directly applying models pretrained using these external cohorts. However, the calibration would fail if model training cohorts were genetically distant from target cohort. Future studies can further consider incorporating or accounting for genetic dissimilarities when training and applying models to more effectively improve imputation quality calibration.

Phasing and imputation in the long-read era

Methods for phasing and imputation have evolved alongside next-generation sequencing, which is heading towards the long-read era. These technologies produce long reads with an average length of 10 kb (for both Oxford Nanopore Platform and PacBio technologies)[100,122–124] and with the longest read reaching 4 Mb (ref. 125), enabling more reliable read-based phasing (discussed in brief below and more comprehensively in refs. 126–128). Long reads can also better detect SVs that cannot be called effectively with short-read sequencing, making imputation of SVs possible[52].

Read-based phasing

Read-based phasing, or haplotype assembly, addresses the need for phasing directly from sequencing reads. It falls into a different category from the population-based phasing discussed above, where phase is assessed through reads capturing multiple variant sites. For example, Oxford Nanopore Technologies-based workflows phased >99% of heterozygous SNPs in the *APOE* locus, a hotspot for Alzheimer's disease risk haplotypes, using as few as 60 reads[129]. Similarly, Hi-C data integrated with PacBio assemblies via FALCON-Phase[130] extended phase blocks to the scale of Mb in human and cow genomes with 97% accuracy. As another example, long-read sequencing revealed haplotype-specific mutation biases in non-small-cell lung cancer, where chromothripsis-like rearrangements occurred preferentially on one parental chromosome in *EGFR*-mutant tumours[131], demonstrating the use of long-read sequencing in dissecting somatic evolution and allele-specific epigenetic regulation[131].

Read-based phasing can be performed using short-read sequencing data, but they usually provide insufficient information for effective haplotype assembly[126]. An early method HapCUT[132] was based on short-read sequencing data using max-cuts algorithms on fragment metrices to minimize conflicts. The authors updated the method to HapCUT2 to incorporate long-read sequencing data[133]. Another method, WhatsHap[134], applied dynamic programming to optimize the minimum error correction metric, with a parallel implementation called PWhatsHap[135]. More recently, many different haplotype assembly methods have been developed[136–141], including some methods borrowing information from other data types, for example, RNA sequencing[142,143], methylation[144,145], Hi-C[130,146] and single-cell strand-seq[147]. Note that some studies have combined both read-based phasing and genotype-based phasing to further improve phasing accuracy[148,149], but more systematic evaluations are warranted to compare the performance between these two types of phasing method, or to assess how to maximize information provided by reads (for example, read-based phasing versus genotype-based phasing with SAPPHIRE for phasing corrections).

Imputation of SVs

Long-read sequencing technologies also enable the comprehensive study of SVs, which have remained incompletely resolved owing to limitations of short-read sequencing[150]. A recent study performed nanopore long-read sequencing for 1,019 individuals from various populations in the 1000 Genomes Project[150] and focused on the analyses of SVs, providing valuable resources to the community. Another study constructed an imputation reference panel for SVs and imputed SVs for individuals in UK Biobank[52]. They further performed GWAS of SVs, which identified thousands of genome-wide significant associations, including novel signals that were independent of SNP associations. This pioneering study indicates a promising future direction of imputing SVs leveraging long-read sequencing data in diverse populations, which will help uncover biological mechanisms after the landscape of different types of genome variant becomes available. We note that SVs powered by long-read sequencing have also led to novel genome representations using graphics (for example, pangenomes)[151], which have the potential to serve as a standard format of reference panels for phasing and imputation in the future. Complex SVs, including highly polymorphic multi-allelic variants, pose substantial challenges for accurate phasing and imputation; thus, novel methodology and computational tools are needed to facilitate phasing and imputation capable of incorporating long and complex SVs. We also anticipate several other future advances in SV imputation, including meta-imputation combining SV reference panels with existing panels, post-imputation quality evaluation and recalibration, and evaluations of various study design options.

Future perspectives

It is unlikely that completely novel statistical methods will replace the long-established PAC framework for phasing and imputation of short-read sequencing data, but we believe there is potential for algorithmic novelty in the future, especially regarding artificial intelligence, reference panels and long-read sequencing.

Artificial intelligence

Phasing and imputation is poised to benefit from the rapidly evolving artificial intelligence models[152–154]. For example, future studies may attempt to learn haplotype patterns in 'chunks' from large reference panels and apply such chunk-specific models for phasing and imputation in target individuals. Moreover, how artificial intelligence agents[155,156] may help with phasing and genotype imputation, especially some tedious but essential data processing steps, is also an interesting future direction that warrants comprehensive development and evaluation. Methods incorporating various deep learning techniques have been developed, such as accurate data-driven imputation technique (ADDIT)[157], sparse convolutional denoising autoencoder (SCDA)[158], imputation based on bidirectional recurrent neural network (RNN-IMP)[79,159], recurrent neural networks integrating with an additional discriminator network (GRUD)[160] and split-transformer with integrated convolutions for genotype imputation (STICI)[161]. However, these deep learning methods have not been widely evaluated. Impartial and comprehensive evaluations are warranted before conclusions can be drawn for practical utilities. In addition, these methods usually require large amount of data and high computational costs to train reliable models. Future development may focus on integrating deep learning techniques with traditional statistical models to leverage the advantages of both approaches.

Reference panels and diverse variation

We expect that imputation reference panels will continue to grow larger and more diverse, thus computational improvements of imputation methods are still needed. Importantly, some populations are still underrepresented in current reference panels, for example, American Indians[162]. Future efforts are needed to include more comprehensive representations of global populations in reference panels. In addition, reference panels specifically for non-SNP variants are emerging, including copy number variants[49], tandem repeats[50,51] and general SVs[52]. The increasing number of rare variants and the potential of imputing non-SNP variants also present challenges of how to distinguish well-imputed from poorly imputed variants. Our prior study among individuals with cystic fibrosis showed that imputation quality for common SNPs and insertions or deletions (indels) with MAF ≥1% had no substantial differences (relative difference ~2% in terms of true R^2), but rare (MAF <1%) indels had ~35% lower median true R^2 compared with SNPs[43]. We note that these results may be biased towards small indels, and future studies are warranted to evaluate imputation quality for larger and more complex SVs, highlighting the need for better calibrated imputation quality estimate.

Long-read sequencing

With the development of long-read sequencing technologies, many haplotype assembly methods have been proposed. Although long-read sequencing has attractive features compared with short-read sequencing, especially for resolving SVs, the available sample size is still orders of magnitude smaller owing to its high cost. Therefore, future method development may focus on combining short- and long-read sequencing

Glossary

Adversrial algorithms
Algorithms to intentionally 'attack' a computational system or algorithm to assess its robustness.

Emission probabilities
Probabilities of observed data conditional on a particular hidden state in an HMM.

Expectation-maximization
An iterative method to find local maximum likelihood estimate of parameters in a statistical model, where the model depends on unobserved latent variables.

Greedy algorithm
An algorithm that follows the problem-solving heuristic of making the locally optimal choice at each stage.

Haplotype
A physical group of alleles (across multiple genetic variants) on a single chromosome that tends to be inherited together. For example, at two loci with alleles A/a and B/b, there are four possible haplotypes: AB, Ab, aB and ab.

Hardy–Weinberg equilibrium
A theory specifying the relationship between allele and genotype frequencies in a population, attained within one generation under minimal conditions that fit the human population.

Hidden Markov models
(HMM). A statistical model to describe systems where the states of the system are not directly observed (hidden), but their behaviours can be characterized through a sequence of observed data.

Hidden states
Unobserved states in an HMM.

Identity-by-descent
(IBD). DNA segments shared between two or more chromosomes because they were inherited from a common ancestor without recombination.

Linkage disequilibrium
The non-random association of alleles at different loci on a chromosome, that is, correlations among alleles of two genetic variants, usually represented in metrics such as D-prime and R^2.

Minor allele frequency
(MAF). The frequency at which the less common allele occurs in a population.

Probability-based Gibbs sampling
An algorithm for sampling from a specified multivariate probability distribution when it is difficult to directly sampling from the joint distribution. Instead, sampling from the conditional distribution is more practical. It is a type of Markov chain Monte Carlo method.

Transition probabilities
Probabilities of moving from one state to another in an HMM.

reference panels or leveraging long-read sequencing data for error correction in phasing and imputation derived from short-read reference panels. When long-read sequencing costs become feasible for large sample sizes, imputation reference panels based on long-read sequencing data may take place of the current practice, especially given the recent release of a draft human pangenome reference[151]. In addition, compared with short-read data, long-read sequencing better and more comprehensively captures SVs and multi-allelic variants, which are much more complicated than biallelic SNPs. Traditional genotype-based phasing and imputation methods that leverage PAC framework may not be able to handle these more complex types of genetic variant. Novel methodologies and computational pipelines tailored for these data will be in high demand and probably become state of the art in the future.

Long-read sequencing technologies also bring opportunities for more study design options. For example, given a fixed amount of funding, researchers may choose to perform varying mixtures of long-read sequencing, short-read sequencing and array genotyping followed by imputation, within a study cohort. Previous work studied mixed designs, but without long-read sequencing data, and concluded that the best design strategy depends heavily on the study objectives with many aspects to consider[45,80,163,164]. To the best of our knowledge, there are no studies evaluating different study design strategies involving long-read sequencing or long-read sequencing-based imputation, which should be possible in the future with more data becoming available. Important evaluation metrics include the ability to directly capture (either by sequencing or array genotyping) variants of various types (for example, SNPs, simple indels and more complex SVs; biallelic versus multi-allelic variants) at varying MAFs, and the ability to impute variants when not directly measured. Practically, for example, it would be valuable to update tools provided by our previous work, including ABCD[165] (where we estimated power to detect and call genotypes at SNPs across the entire MAF spectrum) and Imputability Database[166] (where we estimated the imputation quality of each variant in the reference panel with different genotyping arrays across various populations).

Conclusions

Since their introduction two decades ago, haplotype phasing and genotype imputation have become standard practices in various genomic analyses. As the methods are still based on the PAC framework, recent method developments for population-level phasing and imputation have focused primarily on computational improvements to accommodate the rapidly growing size of samples and variants. These advances include PBWT for efficient haplotype representation, novel approaches for rare variant phasing, design of new reference file format, and various practical considerations of phasing and imputation. Along with the increasing scale of population genomic datasets and complete genome assemblies from long-read sequencing, the field is rapidly evolving with technological advances presenting various exciting new opportunities for the continued development of phasing and imputation.

Published online: 24 September 2025

References

1. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
2. Fournier, R., Tsangalidou, Z., Reich, D. & Palamara, P. F. Haplotype-based inference of recent effective population size in modern and ancient DNA samples. *Nat. Commun.* **14**, 7945 (2023).
3. Palamara, P. F. & Pe'er, I. Inference of historical migration rates via haplotype sharing. *Bioinformatics* **29**, i180–i188 (2013).
4. Al-Asadi, H., Petkova, D., Stephens, M. & Novembre, J. Estimating recent migration and population-size surfaces. *PLoS Genet.* **15**, e1007908 (2019).
5. Tian, X., Cai, R. & Browning, S. R. Estimating the genome-wide mutation rate from thousands of unrelated individuals. *Am. J. Hum. Genet.* **109**, 2178–2184 (2022).
6. Porubsky, D. et al. Human de novo mutation rates from a four-generation pedigree reference. *Nature* **643**, 427–436 (2025).
7. Lassen, F. H. et al. Exome-wide evidence of compound heterozygous effects across common phenotypes in the UK Biobank. *Cell Genom.* **4**, 100602 (2024).
8. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
9. Browning, S. R., Waples, R. K. & Browning, B. L. Fast, accurate local ancestry inference with FLARE. *Am. J. Hum. Genet.* **110**, 326–335 (2023).
10. Sun, Q. et al. Improving polygenic risk prediction in admixed populations by explicitly modeling ancestral-differential effects via GAUDI. *Nat. Commun.* **15**, 1016 (2024).
11. Horimoto, A. R. V. R. et al. Admixture mapping of chronic kidney disease and risk factors in Hispanic/Latino individuals from Central America country of origin. *Circ. Genom. Precis. Med.* **17**, e004314 (2024).
12. Sun, Q. et al. Opportunities and challenges of local ancestry in genetic association analyses. *Am. J. Hum. Genet.* **112**, 727–740 (2025).
13. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).
This paper introduces the minimac2 imputation method (minimac3 and minimac4 are not associated with any publications).
14. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
This paper introduces the Beagle5 imputation method.
15. Rubinacci, S., Delaneau, O. & Marchini, J. Genotype imputation using the positional burrows wheeler transform. *PLoS Genet.* **16**, e1009049 (2020).
This paper introduces the IMPUTE5 imputation method.
16. Sun, Q. et al. Analyses of biomarker traits in diverse UK Biobank participants identify associations missed by European-centric analysis strategies. *J. Hum. Genet.* **67**, 87–93 (2022).
17. Huerta-Chagoya, A. et al. The power of TOPMed imputation for the discovery of Latino-enriched rare variants associated with type 2 diabetes. *Diabetologia* **66**, 1273–1288 (2023).
18. Qin, Z. S., Niu, T. & Liu, J. S. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **71**, 1242–1247 (2002).
19. Stephens, M. & Scheet, P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76**, 449–462 (2005).
20. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).
21. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
22. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
23. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
24. Carlson, C. S. et al. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120 (2004).
25. Long, J. C., Williams, R. C. & Urbanek, M. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56**, 799–810 (1995).
26. Kong, A. et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).
27. Halperin, E. & Karp, R. M. Perfect phylogeny and haplotype assignment. In *Proc. 8th Annual International Conference on Computational Molecular Biology* 10–19 (ACM, 2004).
28. Halperin, E. & Eskin, E. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics* **20**, 1842–1849 (2004).
29. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
30. Fearnhead, P. & Donnelly, P. Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318 (2001).
31. Loh, P.-R. et al. Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
This paper introduces the Eagle2 phasing method, where PBWT is applied to improve computational efficiency.
32. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat. Genet.* **55**, 1243–1249 (2023).
This paper introduces the SHAPEIT5 phasing method, where singletons are explicitly considered.

33. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
This paper introduces the Beagle5 phasing method (which is different from the Beagle5 imputation publication), where a two-stage phasing strategy is proposed separately for common and rare variants.

34. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
This paper introduces the TOPMed reference panel containing haplotypes from diverse populations, which is more suitable for imputation of global populations compared with previous reference panels, including 1000 Genomes and HRC.

35. Feng, Y.-C. A. et al. Taiwan Biobank: a rich biomedical research database of the Taiwanese population. *Cell Genom.* **2**, 100197 (2022).

36. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
This paper introduces the Michigan imputation server, exemplary in promoting broader usage of reference panels and public servers without accessing individual genotypes contributing to the panels.

37. Al Bkhetan, Z., Zobel, J., Kowalczyk, A., Verspoor, K. & Goudey, B. Exploring effective approaches for haplotype block phasing. *BMC Bioinform.* **20**, 540 (2019).

38. Al Bkhetan, Z., Chana, G., Ramamohanarao, K., Verspoor, K. & Goudey, B. Evaluation of consensus strategies for haplotype phasing. *Brief. Bioinform.* **22**, bbaa280 (2021).

39. Wertenbroek, R., Hofmeister, R. J., Xenarios, I., Thoma, Y. & Delaneau, O. Improving population scale statistical phasing with whole-genome sequencing data. *PLoS Genet.* **20**, e1011092 (2024).
This paper introduces a method to correct phasing errors leveraging raw sequencing.

40. Sun, Q. et al. MagicalRsq: machine-learning-based genotype imputation quality calibration. *Am. J. Hum. Genet.* **109**, 1986–1997 (2022).
This paper introduces a framework to recalculate imputation quality metric for post-imputation quality control, especially for low-frequency and rare variants where the state-of-the-art imputation quality metric (for example, Rsq) performs less well.

41. Sun, Q. et al. MagicalRsq-X: a cross-cohort transferable genotype imputation quality metric. *Am. J. Hum. Genet.* **111**, 990–995 (2024).

42. Aleknonytė-Resch, M., Szymczak, S., Freitag-Wolf, S., Dempfle, A. & Krawczak, M. Genotype imputation in case-only studies of gene-environment interaction: validity and power. *Hum. Genet.* **140**, 1217–1228 (2021).

43. Sun, Q. et al. Leveraging TOPMed imputation server and constructing a cohort-specific imputation reference panel to enhance genotype imputation among cystic fibrosis patients. *HGG Adv.* **3**, 100090 (2022).

44. Lau, W. et al. The hazards of genotype imputation when mapping disease susceptibility variants. *Genome Biol.* **25**, 7 (2024).

45. Liu, E. Y. et al. 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 (2012).

46. Xu, Z. M. et al. Using population-specific add-on polymorphisms to improve genotype imputation in underrepresented populations. *PLoS Comput. Biol.* **18**, e1009628 (2022).

47. Sengupta, D. et al. Performance and accuracy evaluation of reference panels for genotype imputation in sub-Saharan African populations. *Cell Genom.* **3**, 100332 (2023).

48. Cahoon, J. L. et al. Imputation accuracy across global human populations. *Am. J. Hum. Genet.* **111**, 979–989 (2024).

49. Handsaker, R. E. et al. Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).

50. Saini, S., Mitra, I., Mousavi, N., Fotsing, S. F. & Gymrek, M. A reference haplotype panel for genome-wide imputation of short tandem repeats. *Nat. Commun.* **9**, 4397 (2018).

51. Ziaei Jam, H. et al. A deep population reference panel of tandem repeat variation. *Nat. Commun.* **14**, 6711 (2023).

52. Noyvert, B. et al. Imputation of structural variants using a multi-ancestry long-read sequencing panel enables identification of disease associations. *eLife* **14**, RP106115 (2025).
This work performs imputation of SVs using a reference panel based on long-read sequencing data, demonstrating the practical utility of long-read sequencing in the context of imputation, particularly for SVs.

53. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).

54. Sakaue, S. et al. Tutorial: a statistical genetics guide to identifying HLA alleles driving complex disease. *Nat. Protoc.* **18**, 2625–2641 (2023).

55. Tregouet, D. A., Escolano, S., Tiret, L., Mallet, A. & Golmard, J. L. A new algorithm for haplotype-based association analysis: the Stochastic-EM algorithm. *Ann. Hum. Genet.* **68**, 165–177 (2004).

56. Browning, B. L. & Browning, S. R. Statistical phasing of 150,119 sequenced genomes in the UK Biobank. *Am. J. Hum. Genet.* **110**, 161–165 (2023).

57. Sohail, M. et al. Mexican Biobank advances population and medical genomics of diverse ancestries. *Nature* **622**, 775–783 (2023).

58. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).
This paper proposes a series of algorithms for haplotype data compression and efficient haplotype matching, reducing the computational complexity from quadratic to linear in terms of the number of reference haplotypes. It represents a milestone of recent computational development of phasing and imputation methods.

59. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).

60. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).

61. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).

62. O'Connell, J. et al. Haplotype estimation for biobank-scale data sets. *Nat. Genet.* **48**, 817–820 (2016).

63. Palin, K., Campbell, H., Wright, A. F., Wilson, J. F. & Durbin, R. Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genet. Epidemiol.* **35**, 853–860 (2011).

64. Hickey, J. M. et al. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet. Sel. Evol.* **43**, 12 (2011).

65. Herzig, A. F. et al. Strategies for phasing and imputation in a population isolate. *Genet. Epidemiol.* **42**, 201–213 (2018).

66. Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of many thousands of genotyped samples. *Am. J. Hum. Genet.* **91**, 238–251 (2012).

67. O'Connell, J. et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10**, e1004234 (2014).

68. Oget-Ebrad, C. et al. Benchmarking phasing software with a whole-genome sequenced cattle pedigree. *BMC Genom.* **23**, 130 (2022).

69. Choi, Y., Chan, A. P., Kirkness, E., Telenti, A. & Schork, N. J. Comparison of phasing strategies for whole human genomes. *PLoS Genet.* **14**, e1007308 (2018).

70. Lajugie, J. et al. Complete genome phasing of family quartet by combination of genetic, physical and population-based phasing analysis. *PLoS ONE* **8**, e64571 (2013).

71. Chen, G. K., Wang, K., Stram, A. H., Sobel, E. M. & Lange, K. Mendel-GPU: haplotyping and genotype imputation on graphics processing units. *Bioinformatics* **28**, 2979–2980 (2012).

72. Na, J. C., Lee, I., Rhee, J.-K. & Shin, S.-Y. Fast single individual haplotyping method using GPGPU. *Comput. Biol. Med.* **113**, 103421 (2019).

73. Luo, Y. et al. A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nat. Genet.* **53**, 1504–1516 (2021).

74. 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

75. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genom. Hum. Genet.* **10**, 387–406 (2009).

76. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).

77. Das, S., Abecasis, G. R. & Browning, B. L. Genotype imputation from large reference panels. *Annu. Rev. Genom. Hum. Genet.* **19**, 73–96 (2018).

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

79. Kojima, K., Tadaka, S., Okamura, Y. & Kinoshita, K. Two-stage strategy using denoising autoencoders for robust reference-free genotype imputation with missing input genotypes. *J. Hum. Genet.* **69**, 511–518 (2024).

80. Pasaniuc, B. et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* **44**, 631–635 (2012).

81. Hui, R., D'Atanasio, E., Cassidy, L. M., Scheib, C. L. & Kivisild, T. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Sci. Rep.* **10**, 18542 (2020).

82. Sousa da Mota, B. et al. Imputation of ancient human genomes. *Nat. Commun.* **14**, 3660 (2023).

83. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 120–126 (2021).

84. Spiliopoulou, A., Colombo, M., Orchard, P., Agakov, F. & McKeigue, P. GeneImp: fast imputation to large reference panels using genotype likelihoods from ultralow coverage sequencing. *Genetics* **206**, 91–104 (2017).

85. Davies, R. W. et al. Rapid genotype imputation from sequence with reference panels. *Nat. Genet.* **53**, 1104–1111 (2021).

86. Jagirdar, K. et al. Molecular analysis of common polymorphisms within the human tyrosinase locus and genetic association with pigmentation traits. *Pigment. Cell Melanoma Res.* **27**, 552–564 (2014).

87. VanRaden, P. M., Sun, C. & O'Connell, J. R. Fast imputation using medium or low-coverage sequence data. *BMC Genet.* **16**, 82 (2015).

88. Davies, R. W., Flint, J., Myers, S. & Mott, R. Rapid genotype imputation from sequence without reference panels. *Nat. Genet.* **48**, 965–969 (2016).

89. Zheng, C., Boer, M. P. & van Eeuwijk, F. A. Accurate genotype imputation in multiparental populations from low-coverage sequence. *Genetics* **210**, 71–82 (2018).

90. Geman, S. & Geman, D. Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721–741 (1984).

91. Rubinacci, S., Hofmeister, R. J., Sousa da Mota, B. & Delaneau, O. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. *Nat. Genet.* **55**, 1088–1090 (2023).
This paper introduces GLIMPSE2, an imputation method specifically designed for ulcWGS data.

92. Martiniano, R. et al. The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLoS Genet.* **13**, e1006852 (2017).

93. Gamba, C. et al. Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* **5**, 5257 (2014).
94. Royo, J. L. Hardy Weinberg equilibrium disturbances in case–control studies lead to non-conclusive results. *Cell J.* **22**, 572–574 (2021).
95. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 887–893 (2005).
96. Yu, K.-D., Di, G.-H., Fan, L. & Shao, Z.-M. Test of Hardy–Weinberg equilibrium in breast cancer case-control studies: an issue may influence the conclusions. *Breast Cancer Res. Treat.* **117**, 675–677 (2009).
97. Hachiya, T. et al. The NBDC-DDBJ imputation server facilitates the use of controlled access reference panel datasets in Japan. *Hum. Gen. Var.* **9**, 48 (2022).
98. Gürsoy, G., Chielle, E., Brannon, C. M., Maniatakos, M. & Gerstein, M. Privacy-preserving genotype imputation with fully homomorphic encryption. *Cell Syst.* **13**, 173–182.e3 (2022).
99. Mosca, M. J. & Cho, H. Reconstruction of private genomes through reference-based genotype imputation. *Genome Biol.* **24**, 271 (2023).
100. Cavinato, T., Rubinacci, S., Malaspinas, A.-S. & Delaneau, O. A resampling-based approach to share reference panels. *Nat. Comput. Sci.* **4**, 360–366 (2024).
101. Rayner, N. W., Park, Y.-C., Fuchsberger, C., Barysenka, A. & Zeggini, E. Toward GDPR compliance with the Helmholtz Munich genotype imputation server. *Nat. Genet.* **56**, 2580–2581 (2024).
102. Zhu, W. et al. IMMerge: merging imputation data at scale. *Bioinformatics* **39**, btac750 (2023).
103. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
104. Jostins, L., Morley, K. I. & Barrett, J. C. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur. J. Hum. Genet.* **19**, 662–666 (2011).
105. Bai, W.-Y. et al. Genotype imputation and reference panel: a systematic evaluation on haplotype size and diversity. *Brief. Bioinform.* **21**, 1806–1817 (2019).
106. Kowalski, M. H. et al. Use of > 100,000 NHLBI trans-omics for precision medicine (TOPMed) consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.* **15**, e1008500 (2019).
107. Yoo, S.-K. et al. NARD: whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants. *Genome Med.* **11**, 64 (2019).
108. Yu, C. et al. A high-resolution haplotype-resolved reference panel constructed from the China Kadoorie Biobank study. *Nucleic Acids Res.* **51**, 11770–11782 (2023).
109. Cengnata, A. et al. A genotype imputation reference panel specific for native Southeast Asian populations. *NPJ Genom. Med.* **9**, 47 (2024).
110. O'Connell, J. et al. A population-specific reference panel for improved genotype imputation in African Americans. *Commun. Biol.* **4**, 1269 (2021).
111. Panjwani, N. et al. Improving imputation in disease-relevant regions: lessons from cystic fibrosis. *NPJ Genom. Med.* **3**, 8 (2018).
112. Yu, K. et al. Meta-imputation: an efficient method to combine genotype data after imputation with multiple reference panels. *Am. J. Hum. Genet.* **109**, 1007–1015 (2022). **This paper introduces meta-imputation to combine imputed results from multiple reference panels. It is helpful in scenarios where multiple references are suitable, for example, where a small population-specific (or disease cohort) reference panel is available in addition to a large reference panel from general or mismatched populations.**
113. Hwang, M. Y., Choi, N.-H., Won, H. H., Kim, B.-J. & Kim, Y. J. Analyzing the Korean reference genome with meta-imputation increased the imputation accuracy and spectrum of rare variants in the Korean population. *Front. Genet.* **13**, 1008646 (2022).
114. Xu, J. et al. Evaluation of imputation performance of multiple reference panels in a Pakistani population. *HGG Adv.* **6**, 100395 (2025).
115. Quick, C. et al. Sequencing and imputation in GWAS: cost-effective strategies to increase power and genomic coverage across diverse populations. *Genet. Epidemiol.* **44**, 537–549 (2020).
116. Roberts, G. H. L., Santorico, S. A. & Spritz, R. A. Deep genotype imputation captures virtually all heritability of autoimmune vitiligo. *Hum. Mol. Genet.* **29**, 859–863 (2020).
117. Yu, W.-Y. et al. Efficient identification of trait-associated loss-of-function variants in the UK Biobank cohort by exome-sequencing based genotype imputation. *Genet. Epidemiol.* **47**, 121–134 (2023).
118. Si, Y., Vanderwerff, B. & Zöllner, S. Why are rare variants hard to impute? Coalescent models reveal theoretical limits in existing algorithms. *Genetics* **217**, iyab011 (2021).
119. Chen, S.-F. et al. Genotype imputation and variability in polygenic risk score estimation. *Genome Med.* **12**, 100 (2020).
120. Zhang, Z., Xiao, X., Zhou, W., Zhu, D. & Amos, C. I. False positive findings during genome-wide association studies with imputation: influence of allele frequency and imputation accuracy. *Hum. Mol. Genet.* **31**, 146–155 (2021).
121. Appadurai, V. et al. Accuracy of haplotype estimation and whole genome imputation affects complex trait analyses in complex biobanks. *Commun. Biol.* **6**, 101 (2023).
122. Scarano, C. et al. The third-generation sequencing challenge: novel insights for the omic sciences. *Biomolecules* **14**, 568 (2024).
123. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
124. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**, 518–524 (2016).
125. Xu, Y., Luo, H., Wang, Z., Lam, H.-M. & Huang, C. Oxford Nanopore Technology: revolutionizing genomics research in plants. *Trends Plant. Sci.* **27**, 510–511 (2022).
126. Snyder, M. W., Adey, A., Kitzman, J. O. & Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Rev. Genet.* **16**, 344–358 (2015).
127. Garg, S. Computational methods for chromosome-scale haplotype reconstruction. *Genome Biol.* **22**, 101 (2021).
128. Zhang, T. et al. Complex genome assembly based on long-read sequencing. *Brief. Bioinform.* **23**, bbac305 (2022).
129. Maestri, S. et al. A long-read sequencing approach for direct haplotype phasing in clinical settings. *Int. J. Mol. Sci.* **21**, 9177 (2020).
130. Kronenberg, Z. N. et al. Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nat. Commun.* **12**, 1935 (2021).
131. Sakamoto, Y. et al. Phasing analysis of lung cancer genomes using a long read sequencer. *Nat. Commun.* **13**, 3464 (2022).
132. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153–i159 (2008).
133. Bansal, V. Hapcut2: a method for phasing genomes using experimental sequence data. *Methods Mol. Biol.* **2590**, 139–147 (2023).
134. Patterson, M. et al. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *J. Comput. Biol.* **22**, 498–509 (2015).
135. Bracciali, A. et al. PWHATSHAP: efficient haplotyping for future generation sequencing. *BMC Bioinform.* **17**, 342 (2016).
136. Garg, S. et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* **39**, 309–312 (2021).
137. Feng, Z., Clemente, J. C., Wong, B. & Schadt, E. E. Detecting and phasing minor single-nucleotide variants from long-read sequencing data. *Nat. Commun.* **12**, 3032 (2021).
138. Yu, Y., Chen, L., Miao, X. & Li, S. C. SpecHap: a diploid phasing algorithm based on spectral graph theory. *Nucleic Acids Res.* **49**, e114 (2021).
139. Fruzangohar, M., Timmins, W. A., Kravchuk, O. & Taylor, J. HaploMaker: an improved algorithm for rapid haplotype assembly of genomic sequences. *Gigascience* **11**, giac038 (2022).
140. Lin, J.-H., Chen, L.-C., Yu, S.-C. & Huang, Y.-T. LongPhase: an ultra-fast chromosome-scale phasing algorithm for small and large variants. *Bioinformatics* **38**, 1816–1822 (2022).
141. Holt, J. M. et al. HiPhase: jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. *Bioinformatics* **40**, btae042 (2024).
142. Edsgärd, D., Reinius, B. & Sandberg, R. scphaser: haplotype inference using single-cell RNA-seq data. *Bioinformatics* **32**, 3038–3040 (2016).
143. Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y. & Lappalainen, T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat. Commun.* **7**, 12817 (2016).
144. Akbari, V. & Jones, S. J. M. Phasing DNA methylation. *Methods Mol. Biol.* **2590**, 219–235 (2023).
145. Fu, Y. et al. MethPhaser: methylation-based long-read haplotype phasing of human genomes. *Nat. Commun.* **15**, 5327 (2024).
146. Ouchi, S., Kajitani, R. & Itoh, T. GreenHill: a de novo chromosome-level scaffolding and phasing tool using Hi-C. *Genome Biol.* **24**, 162 (2023).
147. Henglin, M. et al. Graphasing: phasing diploid genome assembly graphs with single-cell strand sequencing. *Genome Biol.* **25**, 265 (2024).
148. Yang, W.-Y. et al. Leveraging reads that span multiple single nucleotide polymorphisms for haplotype inference from sequencing data. *Bioinformatics* **29**, 2245–2252 (2013).
149. Bansal, V. Integrating read-based and population-based phasing for dense and accurate haplotyping of individual genomes. *Bioinformatics* **35**, i242–i248 (2019).
150. Schloissnig, S. et al. Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature* **644**, 442–452 (2025).
151. Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
152. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
153. Dalla-Torre, H. et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* **22**, 287–297 (2025).
154. Consens, M. E. et al. Transformers and genome language models. *Nat. Mach. Intell.* **7**, 346–362 (2025).
155. Durante, Z. et al. Agent AI: surveying the horizons of multimodal interaction. Preprint at https://doi.org/10.48550/arXiv.2401.03568 (2024).
156. Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N. & Narayanan, A. AI agents that matter. Preprint at https://doi.org/10.48550/arXiv.2407.01502 (2024).
157. Choudhury, O., Chakrabarty, A. & Emrich, S. J. Highly accurate and efficient data-driven methods for genotype imputation. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **16**, 1107–1116 (2019).
158. Chen, J. & Shi, X. Sparse convolutional denoising autoencoders for genotype imputation. *Genes* **10**, 652 (2019).
159. Kojima, K. et al. A genotype imputation method for de-identified haplotype reference information by using recurrent neural network. *PLoS Comput. Biol.* **16**, e1008207 (2020).
160. Chi Duong, V. et al. A rapid and reference-free imputation method for low-cost genotyping platforms. *Sci. Rep.* **13**, 23083 (2023).
161. Mowlaei, M. E. et al. STICI: split-transformer with integrated convolutions for genotype imputation. *Nat. Commun.* **16**, 1218 (2025).

162. Sun, Q. et al. Polygenic scores of cardiometabolic risk factors in american indian adults. *JAMA Netw. Open* **8**, e250535 (2025).
163. Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. R. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* **21**, 940–951 (2011).
164. Zöllner, S. Sampling strategies for rare variant tests in case-control studies. *Eur. J. Hum. Genet.* **20**, 1085–1091 (2012).
165. Kang, J. et al. AbCD: arbitrary coverage design for sequencing-based genetic studies. *Bioinformatics* **29**, 799–801 (2013).
166. Duan, Q., Liu, E. Y., Croteau-Chonka, D. C., Mohlke, K. L. & Li, Y. A comprehensive SNP and indel imputability database. *Bioinformatics* **29**, 528–531 (2013).
167. Browning, B. L. & Browning, S. R. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet. Epidemiol.* **31**, 365–375 (2007).
168. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
169. Platt, A., Pivirotto, A., Knoblauch, J. & Hey, J. An estimator of first coalescent time reveals selection on young variants and large heterogeneity in rare allele ages among human populations. *PLoS Genet.* **15**, e1008340 (2019).
170. Banday, A. R. et al. Genetic regulation of OAS1 nonsense-mediated decay underlies association with COVID-19 hospitalization in patients of European and African ancestries. *Nat. Genet.* **54**, 1103–1116 (2022).
171. Michalek, D. A. et al. A multi-ancestry genome-wide association study in type 1 diabetes. *Hum. Mol. Genet.* **33**, 958–968 (2024).
172. Lucas, E. R. et al. Genome-wide association studies reveal novel loci associated with pyrethroid and organophosphate resistance in Anopheles gambiae and Anopheles coluzzii. *Nat. Commun.* **14**, 4946 (2023).
173. Bråten, L. S., Ingelman-Sundberg, M., Jukic, M. M., Molden, E. & Kringen, M. K. Impact of the novel CYP2C:TG haplotype and CYP2B6 variants on sertraline exposure in a large patient population. *Clin. Transl. Sci.* **15**, 2135–2145 (2022).
174. Aksit, M. A. et al. Pleiotropic modifiers of age-related diabetes and neonatal intestinal obstruction in cystic fibrosis. *Am. J. Hum. Genet.* **109**, 1894–1908 (2022).
175. Loftus, S. K. et al. Haplotype-based analysis resolves missing heritability in oculocutaneous albinism type 1B. *Am. J. Hum. Genet.* **110**, 1123–1137 (2023). **This paper sets up an example of how phasing or haplotype-level analyses can help better understand disease-causing alleles, elucidate genetic mechanisms underlying diseases and aid genetic diagnosis.**
176. Khankhanian, P., Gourraud, P.-A., Lizee, A. & Goodin, D. S. Haplotype-based approach to known MS-associated regions increases the amount of explained risk. *J. Med. Genet.* **52**, 587–594 (2015).
177. Albiñana, C. et al. Genetic correlates of vitamin D-binding protein and 25-hydroxyvitamin D in neonatal dried blood spots. *Nat. Commun.* **14**, 852 (2023).
178. Sollis, E. et al. The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
179. Lin, S. et al. Evidence that the Ser192Tyr/Arg402Gln in *cis* tyrosinase gene haplotype is a disease-causing allele in oculocutaneous albinism type 1B (OCA1B). *NPJ Genom. Med.* **7**, 2 (2022).
180. Shriner, D. Overview of admixture mapping. *Curr. Protoc.* **3**, e677 (2023).
181. Duan, Q. et al. A robust and powerful two-step testing procedure for local ancestry adjusted allelic association analysis in admixed populations. *Genet. Epidemiol.* **42**, 288–302 (2018).
182. Atkinson, E. G. et al. Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat. Genet.* **53**, 195–204 (2021).
183. Hou, K. et al. Admix-kit: an integrated toolkit and pipeline for genetic analyses of admixed populations. *Bioinformatics* **40**, btae148 (2024).
184. Hou, K. et al. Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nat. Genet.* **55**, 549–558 (2023).
185. Meisner, J., Benros, M. E. & Rasmussen, S. Leveraging haplotype information in heritability estimation and polygenic prediction. *Nat. Commun.* **16**, 126 (2025).

Author contributions

Q.S. researched data for the article. All authors contributed substantially to discussion of the content. Q.S. wrote the article. All authors reviewed and/or edited the manuscript before submission.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Genetics* thanks Christian Fuchsberger and Daniel Lawson, who co-reviewed with Eva König, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.