

DOI 10.18699/vjgb-24-70

Genotype imputation in human genomic studies

A.A. Berdnikova ^{1, 2}, I.V. Zorkoltseva ¹, Y.A. Tsepilov ¹, E.E. Elgaeva ^{1, 2} 

¹ Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

² Novosibirsk State University, Novosibirsk, Russia

 elizabeth.elgaeva@gmail.com

Abstract. Imputation is a method that supplies missing information about genetic variants that could not be directly genotyped with DNA microarrays or low-coverage sequencing. Imputation plays a critical role in genome-wide association studies (GWAS). It leads to a significant increase in the number of studied variants, which improves the resolution of the method and enhances the comparability of data obtained in different cohorts and/or by using different technologies, which is important for conducting meta-analyses. When performing imputation, genotype information from the study sample, in which only part of the genetic variants are known, is complemented using the standard (reference) sample, which has more complete genotype data (most often the results of whole-genome sequencing). Imputation has become an integral part of human genomic research due to the benefits it provides and the increasing availability of imputation tools and reference sample data. This review focuses on imputation in human genomic research. The first section of the review provides a description of technologies for obtaining information about human genotypes and characteristics of these types of data. The second section describes the imputation methodology, lists the stages of its implementation and the corresponding programs, provides a description of the most popular reference panels and methods for assessing the quality of imputation. The review concludes with examples of the use of imputation in genomic studies of samples from Russia. This review shows the importance of imputation, provides information on how to carry it out, and systematizes the results of its application using Russian samples.

Key words: imputation; genotyping; sequencing; genome-wide association study; human; DNA-microarray.

For citation: Berdnikova A.A., Zorkoltseva I.V., Tsepilov Y.A., Elgaeva E.E. Genotype imputation in human genomic studies. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov Journal of Genetics and Breeding*. 2024;28(6):628-639. DOI 10.18699/vjgb-24-70

Funding. The work was supported by the Russian Science Foundation grant number 22-15-20037 and the Government of Novosibirsk Region.

Acknowledgements. Authors express gratitude to V.S. Fishman for his recommendations for improvement of the article text.

Импутация генотипов в геномных исследованиях человека

A.A. Бердникова ^{1, 2}, И.В. Зоркольева ¹, Я.А. Цепилов ¹, Е.Е. Елгаева ^{1, 2} 

¹ Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

 elizabeth.elgaeva@gmail.com

Аннотация. Импутация – это метод, позволяющий восстанавливать недостающую информацию о генетических вариантах, которые не удалось генотипировать напрямую с помощью ДНК-микрочипов или секвенирования с низким покрытием. Импутация играет важнейшую роль в полногеномном анализе ассоциаций (genome wide associations study, GWAS). Она приводит к существенному увеличению количества изучаемых вариантов, что повышает разрешающую способность метода и увеличивает сопоставимость данных, полученных в разных когортах и/или с помощью разных технологий, что важно при проведении метаанализов. При ее выполнении информацию о генотипах в исследуемой выборке, у которой известна только часть генетических вариантов, дополняют за счет эталонной (референсной) выборки, имеющей более полные данные о генотипах (чаще всего это результаты полногеномного секвенирования). Импутация стала неотъемлемой частью геномных исследований человека благодаря преимуществам, которые она дает, а также увеличению доступности инструментов для импутации и данных референсных выборок. Обзор посвящен импутации в геномных исследованиях человека. В первом разделе приводятся описание технологий получения информации о генотипах человека и характеристика получаемых типов данных. Во втором разделе представлена методология импутации, перечисляются этапы ее проведения и соответствующие программы, дается опи-

сание наиболее популярных референсных панелей и способов оценки качества импутации. В заключении представлены примеры использования импутации в геномных исследованиях выборок из России. Настоящий обзор показывает важность проведения импутации, дает информацию о том, как ее выполнять, и систематизирует результаты ее применения на примере российских выборок.

Ключевые слова: импутация; генотипирование; секвенирование; полногеномный анализ ассоциаций; человек; ДНК-микрочип.

Technologies for obtaining human genotype data and their features

Human genotype data are a key aspect for many genetic studies. There are several technologies developed to read, analyze and interpret genetic information. The most commonly used methods include Sanger sequencing, next generation sequencing (NGS), and DNA microarrays.

Genotyping using DNA microarrays

A DNA microarray (or simply a “microchip” or “chip”, not to be confused with an RNA microarray, which is a different technology) is a small glass or silicon substrate, to which tens of thousands of probes (short single-stranded DNA fragments complementary to certain nucleotide sequences) are attached. These probes are arranged on the chip in such a way that each fragment can be identified by its location (Fig. 1).

During the analysis, fluorescent markers are attached to the studied DNA molecules, which were cut into fragments by restriction endonucleases, and placed on the chip. The target DNA fragments are bound to complementary DNA probes, and all remaining fragments are removed from the chip. Laser beams and computer processing are used to detect the fluorescence of fragments, record the emission (radiation) patterns and subsequently identify the sequences. This method is very fast and allows to simultaneously determine the nucleotide sequence of several DNA fragments (Govindarajan et al., 2012).

An alternative approach to solving the problem of genotyping was implemented by academician A.D. Mirzabekov in domestic developments to create gel microchips (Mirzabekov, 2003). They are a substrate made of glass, plastic or silicone with hemispherical drops of hydrogel fixed on its surface. The distinction of this method is that DNA fragments are immobilized in three-dimensional space, which provides greater

sensitivity and capacity of the microchip. This technology has also found its application in RNA analysis, protein and cell biochips.

There are several strategies for identifying single nucleotide polymorphism (SNP) for microarrays (Fig. 2).

Allele distinction by hybridization (Fig. 2a). The labeled target DNA hybridizes with probes containing a polymorphic site in the center. Correctly paired oligonucleotides are more stable (have a higher melting temperature) compared to duplexes with a non-complementary base. Therefore, after washing the chip under harsh temperature conditions, only correctly paired chains remain on it. It is common to use multiple fragments for each allele to improve the quality of the signal relative to noise (Wang D.G. et al., 1998).

“Golden Gate” analysis by the Illumina company (Fig. 2b). Two allele-specific oligonucleotides, each of which has a 5' end with different universal primers (P1 and P2) (the primers are labeled with a unique fluorophore for subsequent site discrimination), hybridize in solution with genomic DNA. The third oligonucleotide, in addition to the universal primer (P3), has a tail with a “barcode” sequence complementary to the fragment on the chip. The allele-specific primers extended by a polymerase are ligated to a third oligonucleotide, after which the resulting fragments are amplified using the polymerase chain reaction and hybridized onto the chip. The use of multiple barcodes (one for each locus of interest) allows for analysis of several genomic regions at once (Fan et al., 2003).

Arrayed Primer Extension (APEX) (Fig. 2c). Here, the chip contains a DNA fragment, the 5' end of which is fixed to the substrate, and the 3' end finishes with the nucleotide preceding the SNP being detected. Fragments of genomic DNA are hybridized to the chip, while the desired SNP remains unpaired. During the sequencing reaction, the nucleo-

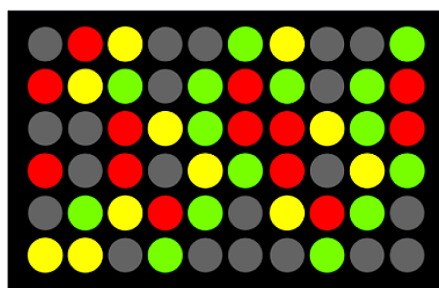
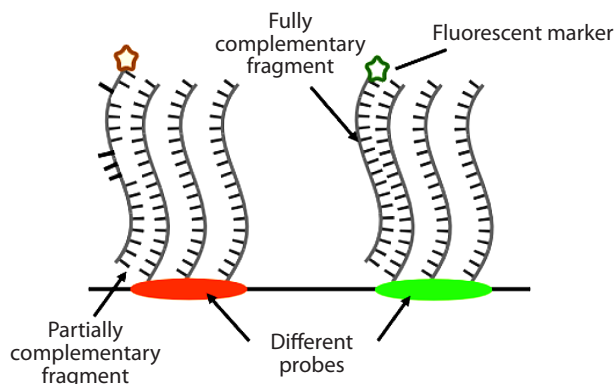


Fig. 1. DNA microarray.

Pseudocolor (red, yellow or green) is determined by the number of molecules bound to the probe and labeled with different dyes. For further explanation of the figure, see the text below.



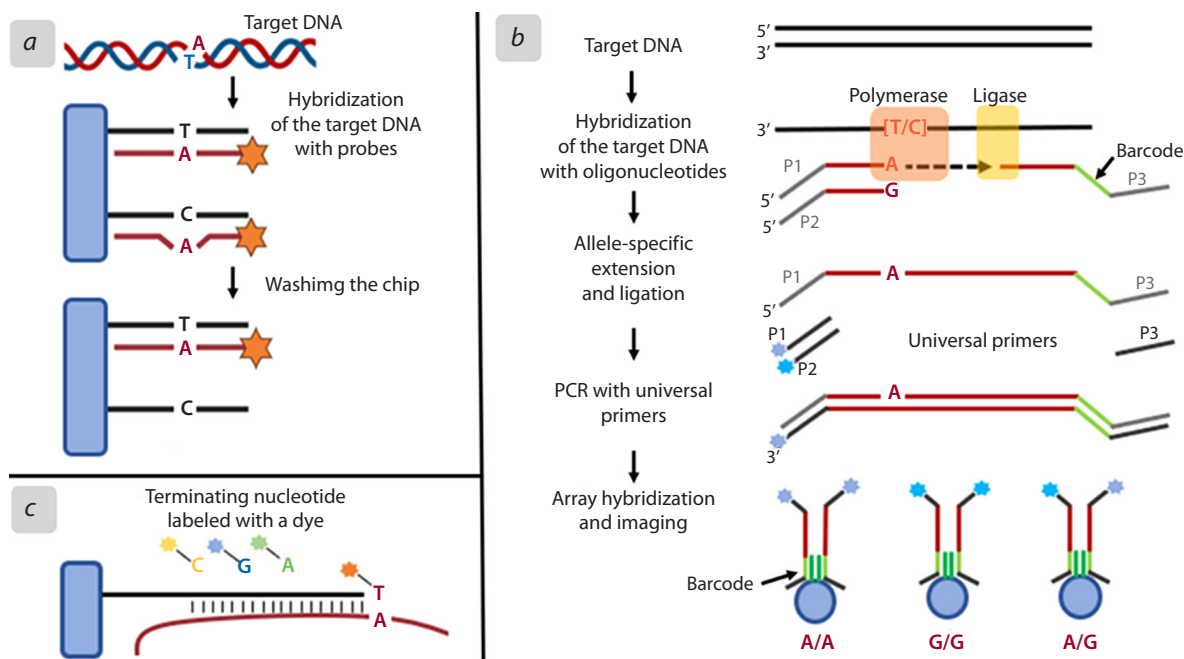


Fig. 2. SNP detection strategies for DNA microarrays.

a – allele-specific hybridization; b – “Golden Gate” analysis by the Illumina company; c – arrayed primer extension.

tide sequence attached to the substrate is extended by one terminating nucleotide labeled with a dye (Kurg et al., 2000). This nucleotide prevents further growth of the DNA chain, and the color of its dye allows you to determine which of the nucleotides (A, T, G or C) is located at the given position.

One of the main advantages of DNA microarrays is their high throughput capability (Hayat, 2002; Brown et al., 2024). The microarray provides the basis for simultaneous genotyping of thousands of different loci and detection of single nucleotide substitutions. Thus, microarrays are used to analyze large samples in order to genotype frequently occurring genetic variants (with a minor allele frequency in the population > 0.01).

However, there are some limitations in interpreting the results. Microarray data are typically binary (indicating the presence or absence of a specific allele), high-throughput (allowing the analysis of thousands or millions of SNPs), and requiring specialized analysis techniques to extract meaningful information. In this case, we are talking about software (for example, GenomeStudio (Illumina Inc., San Diego, CA, CIIIA)), which includes tools for quality control, genotype identification, visualization and data analysis. In addition, microchips can produce both false positive and false negative results. These issues highlight the importance of careful data interpretation and the need to use appropriate statistical methods to control quality and validate results.

Genome sequencing

This chapter describes various sequencing technologies. Around 1976, two methods that could read hundreds of bases in half a day were developed – Sanger and Coulson strand termination and Maxam and Gilbert chemical cleavage

(Maxam, Gilbert, 1977; Sanger et al., 1977). In both methods, the analyzed DNA is placed into four test tubes with different compositions of the reaction mixture for a specific type of nitrogenous base (A, T, G or C). Gilbert’s method uses DNA, radioactively labeled at one end, and a mixture of enzymes that specifically cut it before a certain type of nucleotide. Sanger sequencing, in contrast, involves primers and dideoxynucleotides that stop chain synthesis when radiolabeled dideoxynucleoside triphosphate (ddNTP; different in each tube) is included. Hence, as a result of implementing either method, labeled DNA fragments of different lengths that end with the same base are formed in each tube. Sequences are separated by length using polyacrylamide plate gel electrophoresis (one lane per base type) at single nucleotide resolution. The image obtained on X-ray film after electrophoresis allows researchers to restore the original sequence. The described methods immediately came into use, and by 1987, automated fluorescent Sanger sequencers could read about 1,000 bases per day (Smith et al., 1986; Connell et al., 1987).

In 2005, next generation sequencing (NGS) technologies were first introduced, which are based on two approaches. The first of these is sequencing by hybridization (SBH). The essence of the method is as follows: first, short sections of DNA are fixed on a glass substrate (DNA chip). Then the fragments to be identified are labeled with fluorophore and applied to the chip for hybridization with the fixed areas. Single-stranded DNA is washed away, and the hybridization pattern is read from the color marks and their brightness. An alternative approach in NGS is sequencing by synthesis (SBS) (Shendure et al., 2017).

As a rule, in technologies that use the SBS technique, pre-fragmented sequences are fixed in a flow cell, where cyclic

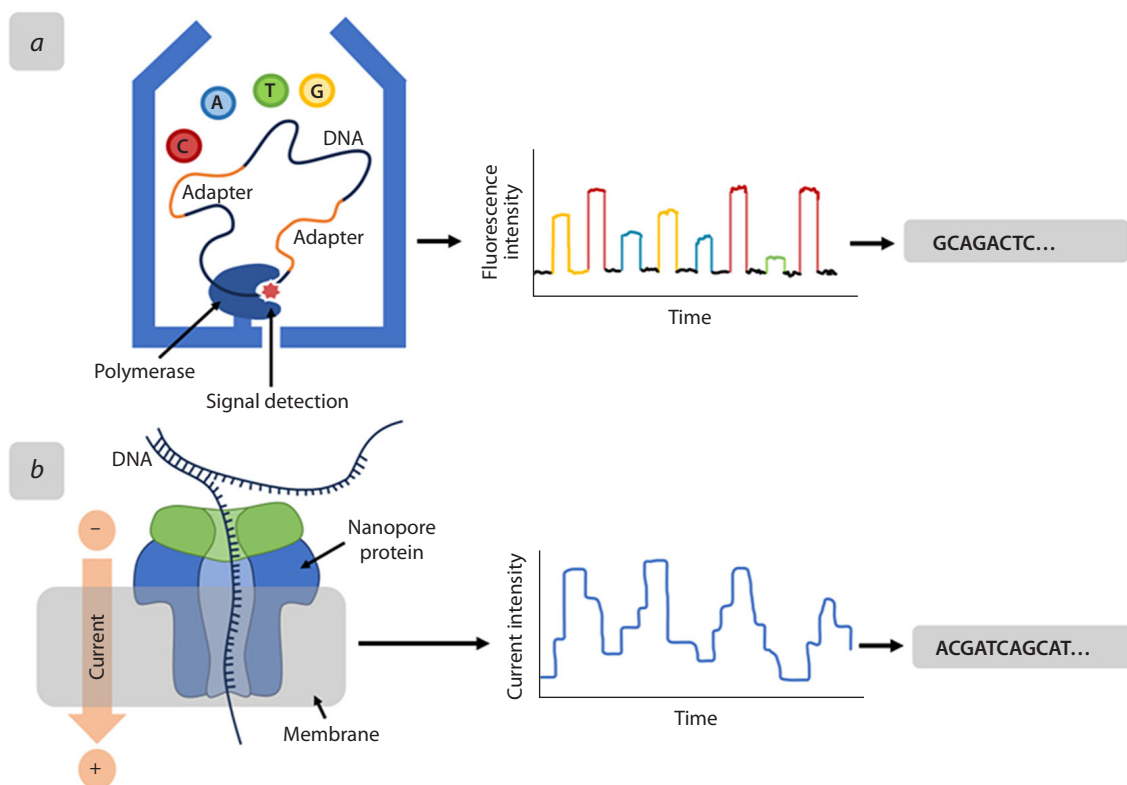


Fig. 3. Third generation sequencing.

a – Pacific Biosciences; *b* – Oxford Nanopore Technology. See the text below for explanation.

synthesis of a new chain occurs. By sequentially adding one of the four deoxynucleotides to the mixture, having removed the previous ones in advance, it is possible to read signals from the cells where the synthesis reaction was successful. Therefore, the output provides information about where which nucleotide is located.

Sequencing technologies with an approach other than NGS were first described in 2008–2009 and named “third generation sequencing” (Check Hayden, 2009). They include two main approaches (Fig. 3).

The first technology, Pacific Biosciences (PacBio) (Rhoads, Au, 2015), is designed to optically monitor DNA synthesis using a polymerase in real time. The structure has a hole less than half the light wavelength that limits fluorescent excitation to a small volume containing only the polymerase and its template (Fig. 3a). With such a device, only fluorescently labeled nucleotides included in the growing DNA strand emit signals of sufficient duration to be read. The error rate in this sequencing method is very high (about 10 %), but the errors are distributed randomly. With long reads and tolerance for high GC content and random errors, PacBio provides *de novo* assemblies of unprecedented quality in terms of accuracy and continuity.

The second major third-generation sequencing technology is Oxford Nanopore (ONT) (Deamer et al., 2016). This technique was first proposed in the 1980s. The special chamber where the sequencing process takes place is filled with an electrolytic solution and divided by a two-layer membrane

with a nanopore (its dimensions are in the nanorange). Once voltage is applied, the electrolyte ions and the DNA molecule begin to move through the pore. Nucleic acid physically interferes with the migration of ions, which leads to fluctuations in current strength, which allows the nucleotide sequence to be determined (Fig. 3b). The main difference from other sequencing technologies is the extreme portability of nanopore devices, which can be as small as a memory stick (USB), as they rely on detecting electronic rather than optical signals.

Comparison of technologies and their application to solve different problems

Most often, Illumina NGS technologies are used for large-scale projects (whole-genome sequencing, transcriptome analysis and epigenetic profiling), but PacBio is more useful for *de novo* assembly, and ONT is more applicable for portable sequencing. The Sanger method is suitable for sequencing short DNA fragments such as individual genes, plasmids or viral genomes.

Also worth mentioning is a sequencing technology competing with Illumina, developed by Complete Genomics and MGI Tech, DNBSEQ-T7 (formerly known as MGISEQ-T7). In DNBSEQ-T7, the clonal amplification process occurs as a rolling circle, i.e., always from the original template, which eliminates the accumulation of DNA polymerase errors (Drmanac et al., 2010). The main advantages of MGI include lower cost compared to Illumina and the ability to process a larger volume of samples in a shorter time. As recent studies

show, the new MGISEQ-2000 sequencer can be used as a full-fledged alternative to Illumina sequencers when conducting whole-genome studies (search for variants, identification of indels), the differences between the two platforms are insignificant (Korostin et al., 2020; Jeon et al., 2021; Feng et al., 2024).

Recently, the effectiveness of using whole-genome sequencing (WGS) for GWAS has been demonstrated (DePristo et al., 2011; Chat et al., 2022). This approach is a promising alternative to genotyping using DNA microarrays, as it allows one to obtain information on a larger proportion of genetic variations, increasing the power of association tests and subsequent fine-mapping analyses (Wang Q.S., Huang, 2022). However, despite the decreasing cost of NGS-based technologies, GWAS mainly use high-throughput and relatively cheap DNA microarrays containing hundreds of thousands to millions of common genetic markers, which make it possible to test almost the entire genome for associations with the trait being studied. SNP genotyping using DNA microarrays can contain up to 5 % errors depending on the manufacturer (Lamy et al., 2006; Yang et al., 2011; Guo et al., 2014). However, existing protocols for quality control of the obtained data can significantly reduce the number of errors (on average by 1.7 %) (Zhao et al., 2018). Thus, microarrays allow fairly accurate genotyping of samples even for species with high heterozygosity (i. e., with greater genetic variation than expected at Hardy–Weinberg equilibrium) (Bourke et al., 2018). Moreover, at the end of 2023, the cost of genotyping a sample on a microchip was an order of magnitude lower than the cost of NGS sequencing, which makes it possible to cover a much larger sample size with the same project budget. Their main disadvantage when conducting GWAS is that they do not allow detection of an association between an SNP and a trait if the genetic variant is not represented on the microarray.

Additional difficulties in using DNA microarrays may arise because the information (such as the location of SNPs on the chromosome) used to design the chip is out of date or differs between manufacturers. The above problems can be solved by imputation of genotyping data (Pasaniuc et al., 2012). This approach allows us to increase the density of coverage for the genetic variants studied (total number of markers) and the proportion of common variants when conducting a meta-analysis (combining data from different studies and/or genotyping platforms) (Li Y. et al., 2009).

A replacement for DNA microarrays could be low-coverage WGS (lcWGS), in which random regions of the genome are sequenced (Chat et al., 2022). Research shows that lcWGS significantly outperforms microarrays in marker density, which also allows for a more thorough assessment of associations with less common variants. Such data also require imputation using haplotypes (e. g., from the 1000 Genomes Project) (Auton et al., 2015). The costs of ultra-low coverage WGS (sequencing depth $\leq 0.5x$) may be comparable to or lower than those of using DNA microarrays, but its potential as an alternative has not yet been sufficiently assessed (Martin et al., 2021).

DNA sequencing and genotyping solve the task of analyzing genetic information in different ways. As such, sequencing allows you to read entire DNA fragments and is therefore applied to identify rare (minor allele frequency $< 0.01\%$) and *de novo* mutations, and is widely used to study the structure of individual genes or genome regions. Genotyping, on the other hand, is a faster and more cost-effective method for analyzing genetic variation, which is particularly useful for large-scale genomic studies involving thousands or even millions of samples. Thus, if the goal of a study is to comprehensively examine the genetic architecture of a trait or disease, sequencing is likely the best approach. However, if the focus of the study is on common genetic variants, or analysis of the population or kinship structure of the sample, then genotyping is often sufficient and more effective (Gresham et al., 2008).

Imputation of genotyping data

Although sequencing the entire genome of hundreds of thousands of people is not yet feasible, significant progress can be made by identifying only a relatively small number of genetic variants in each person. This type of “incomplete” information is still useful because data on any set of SNPs in a group of people allow inferences to be made about many other unobserved variants in the same people. The approach to accomplish this is called imputation.

Methodology

The imputation procedure includes the following stages: quality control of genotyping data, phasing, imputation itself, and at the final step – quality control of imputed genotypes (Fig. 4).

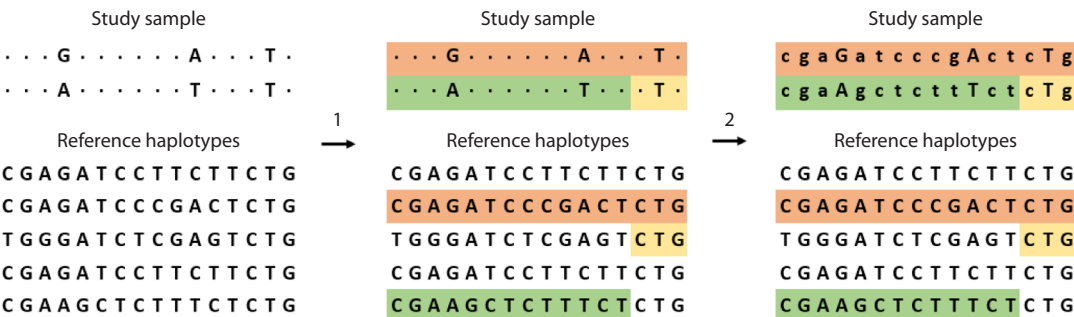


Fig. 4. Imputation of genotyping data.
1 – Phasing; 2 – Imputation itself. See the text below for explanations.

Genetic variants that are located nearby on a chromosome are more likely to be inherited together, which occurs because there are only a few recombinations per chromosome. This principle is called “linkage disequilibrium” (LD). Due to this principle, we observe blocks of haplotypes (haploblocks) – sets of closely spaced genetic variants that were inherited together during evolution.

In imputation, haploblocks are used to identify common short stretches of DNA on chromosomes that individuals in a randomly selected population may have inherited from a common ancestor. By comparing haplotypes in two samples (study and reference) based on a set of common genetic variants, imputation algorithms provide inferences about the genotypes of the studied individuals. Both of these samples must be from the same ethnic group for imputation to produce accurate results (Mills et al., 2020).

Although genotyping data do not contain haplotype information, it can be inferred and reconstructed using stepwise analysis. Phasing is the process of statistically estimating haplotypes. Imputation can be performed on both raw unphased genotyping data and reconstructed mixed haplotypes, although phasing is known to improve imputation accuracy (Anderson et al., 2010). In addition, phasing is often necessary due to the fact that standard imputation algorithms (more about them below) work specifically with haploblocks.

Quality control of genotyping data

An important step in any genomic study is to conduct data quality control. The importance of this step is illustrated by the example of a paper published in Science that was retracted due to insufficient consideration of technical errors in genotyping on an Illumina chip (Marees et al., 2018).

Quality control of DNA microarray genotyping data is divided into two main steps: control at the individual level and control at the marker level. Individual-level control involves removing a sample in the following cases (Anderson et al., 2010):

- there is an observed discrepancy between the phenotype and the genotype (in particular, the phenotypic sex differs from the genetic one);
- the number of heterozygous loci in the genome deviates from the expected value (an overestimation or underestimation of this indicator may indicate sample contamination or inbreeding, respectively);
- the sample contains duplicates, relatives of the first or second degree (similar genotypes will be overrepresented, as a result of which allele frequencies in the population may be displayed unreliably);
- has a different ethnic origin, that is, there is a stratification of the population (the most common approach for identifying such individuals is principal component analysis (PCA) on a kinship matrix).

Data quality control at the level of individual markers also consists of several points that involve the removal of SNPs if:

- minor allele frequency (MAF) < 0.01;
- they are absent from a large part of individuals in the sample;
- they deviate significantly from Hardy–Weinberg equilibrium.



Fig. 5. Visualization of the performance of HMM-based algorithms for four individuals from the reference sample.

Each column is a separate SNP with two alleles (empty and crossed out squares represent different alleles of the same SNP), and each pair of rows represents two copies of DNA (from each parent). Closely related SNPs are grouped by color, and each haplotype is modeled as a mosaic of color combinations (Scheet, Stephens, 2006).

To carry out quality control, a number of publicly available programs are used: PLINK 1.9/PLINK 2 (Purcell et al., 2007; Chang et al., 2015), RICOLI (Lam et al., 2020), SMARTPCA (Price et al., 2006) and FlashPCA (Abraham et al., 2017).

Imputation Tools

Over the past twenty years, several different research groups have developed and published a number of tools for phasing and subsequent imputation, most of which are based on the hidden Markov model (HMM) of Li and Stephens (Li N., Stephens, 2003). This statistical model, first described in 2003, assumes that haplotypes are inherited as haploblocks and that recombination events occur at their boundaries. The model probabilistically reconstructs the studied haplotypes in the form of a mosaic composed of haplotypes from a small reference sample (Fig. 5). It has been shown that methods based on Li and Stephens' HMM are more accurate and efficient (Weale, 2004) than approaches such as Clark's algorithm (Clark, 1990) or the EM algorithm (Expectation-Maximization) (Dempster et al., 1977) (Browning S.R., Browning B.L., 2011). Currently, the most commonly used programs implementing Li and Stephens' HMM are Beagle 5 (Browning B.L. et al., 2021), Eagle2 (Loh et al., 2016) and ShapeIT (Delaneau et al., 2012) for phasing, and also Beagle 5 (Browning B.L. et al., 2018), Impute5 (Rubinacci et al., 2020) and Minimac4 (Das et al., 2016) for imputation. Beagle 5 and ShapeIT2 allow you to perform both of these procedures.

A comparative analysis of current phasing and imputation software showed that, overall, Beagle 5.4 performed slightly better than Impute5 and Minimac4, with a higher concordance rate and high performance even on large data sets (De Marino et al., 2022). However, Minimac4 and Impute5 tend to perform better on rare variants because, unlike Beagle 5.4, which computes clusters of haplotypes and performs calculations based on them, Impute5 and Minimac4 search the entire haplotype space. Minimac4 requires the least amount of memory, but calculations take longer. If memory usage is limited and the loss of accuracy is acceptable, then Minimac4 may be the optimal choice of imputation software.

The above programs can be run from a local server and require reference haplotypes. Nevertheless, most of these large-scale datasets are not publicly available. For this reason, special servers that contain information about different reference panels are most often used for imputation of human data, such as Michigan Imputation Server¹ (Das et al., 2016) and TOPMed Imputation Server² (Das et al., 2016). Researchers can upload their datasets there, configure parameters through the web user interface (select tools, reference panels, etc.), perform phasing and genotype imputation on the server, and download the output files.

As disadvantages of this approach, it is worth noting the need to send your data outside the local server (albeit using secure connection protocols) and possible queues. In addition, users are often limited in the choice of programs or reference panels, and cannot combine multiple panels or integrate their own. However, it is possible to bypass these restrictions, for example, using Docker software (Das et al., 2016), and run imputation on your server. The problem with standalone running is a little more complexity due to manual settings, where the user needs to install additional programs for the pipeline and account for library conflicts.

In Supplementary Material 1³ compares the tools available on the two servers described above.

Reference panels for imputation of human genotyping data

One important issue in genotype imputation is how to select a reference panel that provides high imputation accuracy in the population of interest. As it was shown (Huang, Tseng, 2014), the quality of imputation is affected not only by the size of the panel, but also by the ethnic composition of the reference sample. The most commonly used panels for European populations currently are 1000 Genomes (Sudmant et al., 2015), Haplotype Reference Consortium (HRC) (Haplotype Reference Consortium, 2016) and Trans-Omics for Precision Medicine (TOPMed) (Taliun et al., 2021).

The 1000 Genomes Phase 3 Version 5 reference panel was prepared as part of the 1000 Genomes Project in 2008 (Auton et al., 2015). In total, while using a combination of low-coverage whole-genome sequencing, high-coverage exome sequencing and microarray genotyping, this project was able to characterize 88 million genetic variants (84.7 million SNPs, 3.6 million short insertions/deletions and 60,000 structural variants). This version of the reference panel includes 49 million markers from 2,504 individuals from a mixed population.

The HRC r1.1 2016 reference panel was compiled by the HRC (The Haplotype Reference Consortium) to create a large haplotype reference panel. The HRC panel combines datasets from 20 different studies, most of which were obtained using low-coverage (4–8x) whole-genome sequencing and consist of samples of individuals of predominantly European ancestry. The reference panel consists of 64,976 haplotypes of 32 thousand individuals with 39,235,157 SNPs; it does not contain deletions or insertions.

The TOPMed (The Trans-Omics for Precision Medicine) project was initiated in 2010 with the goal of collecting and analyzing whole-genome sequencing data. As of September 2021, TOPMed has approximately 180 thousand participants, predominantly of non-European origin, from more than 85 different studies. A reference panel was created based on the TOPMed database, which includes 286,068,980 SNPs; 5,815,513 insertions and 16,222,592 deletions in the genotypes of 97,256 individuals. These genetic variants are distributed across 22 autosomes and the X chromosome. TOPMed (Version r2) is the first panel that is based solely on deep whole-genome sequencing data and is significantly superior to previously published alternatives.

Although most genetic studies and reference panels focus on samples of individuals of European ancestry, it is worth noting that there are various projects aimed at studying the genetic diversity of other populations. These include ChinaMAP (10,588 samples and 136.7 million SNPs) (Li L. et al., 2021), NARD (1,779 individuals, 40.6 million SNPs) (Yoo et al., 2019), GAsP (1,739 samples, 1 million autosomal SNPs) (Wall et al., 2019), SG10K (4,810 samples, 89.1 million SNPs) (Wu et al., 2019) for samples of people of Asian descent, AFAM (2,269 samples, 45 million SNPs) (O'Connell et al., 2021) and UGR (4778 samples, 2.2 million markers) (Fatumo et al., 2022) for African Americans. The TOPMed panel can also be used to impute non-European samples of individuals of both African and Asian descent.

The ideal solution when selecting a panel for imputation is to combine data from multiple reference samples to construct a combined reference panel. However, different studies tend to use different quality control and variant filtering strategies, which can make pooling results difficult.

Another major issue is restrictions on shared data use. For example, individual-level genotype information in many reference panels is not publicly available; therefore, it may not be possible to directly combine it with sequencing results from other samples. In this regard, the meta-imputation method was proposed (Yu et al., 2022). Instead of combining reference panels, genotypes are first imputed using multiple reference panels separately and then the imputed results are combined into a consistent data set.

Assessment of imputation quality

The quality of genotyping data imputation can be assessed: 1) using standard imputation quality metrics; 2) empirically (for example, conduct a GWAS on the trait of interest and check the reproducibility of association signals known from the literature, or calculate a polygenic estimate of the trait and compare it with real phenotypes).

Imputation quality metrics can also be divided into two large groups (Stahl et al., 2021): 1) those that assess the quality of imputation without using directly genotyped SNPs and are calculated automatically when running the corresponding imputation software, and 2) those that allow the comparison between imputed SNPs and genotypes and are calculated manually.

Quality metrics in the first group are specific to each individual program. For example, for Minimac4 and Beagle 5, the R^2 indicator is estimated (Marchini, Howie, 2010), which

¹ <https://imputationserver.sph.umich.edu/index.html#pages/home>

² <https://imputation.biodatacatalyst.nhlbi.nih.gov/#pages/home>

³ Supplementary Materials 1–3 are available at:

https://vavilov.elpub.ru/jour/manager/files/Suppl_Berd_Engl_28_6.pdf

is calculated differently for each program, while Impute5 calculates the Info parameter (Marchini, Howie, 2010). Because of their specificity, they are not suitable for comparing the quality of data imputed by different methods. This task is successfully accomplished by metrics from the second group, which include: concordance rate (CR), Imputation Quality Score (IQS) (Lin et al., 2010), Hellinger score (Roshyara et al., 2014), squared Euclidean norm score (SEN) (Roshyara et al., 2014) and others. In practice, standard metrics of the first group are most often used.

While conducting imputation, the posterior probabilities of the genotype are estimated. Thus, for biallelic SNPs in an additive model (where the genotype is coded as 0, 1 and 2, and the reference and alternative allele are 0 and 1, respectively), the estimated probability of individual i to have genotype j at a particular locus is denoted as G_j^i ($j = 0, 1, 2$). This indicator is calculated by appropriate imputation software based on data from the reference and target samples using built-in algorithms (for example, a hidden Markov model, as described above). The dose of the alternative allele is calculated as $D_i = G_1^i + 2G_2^i$.

The R^2 metric is an approximation of the squared correlation between the imputed allele dose and the expected genotype and is calculated as the ratio of the allele dose dispersion and the expected dispersion under Hardy–Weinberg equilibrium

$$\hat{R}_d^2 = \frac{\frac{1}{N} \sum_{i=1}^N (D_i - 2\hat{p})^2}{2\hat{p}(1-\hat{p})}, \quad (1)$$

$$\hat{p} = \sum_{i=1}^N \frac{D_i}{2N},$$

where N is the number of individuals in the sample; D_i is the dose of the imputed allele for the i -th individual; $\hat{p}$ is the allele frequency estimate.

Many modern algorithms (such as Minimac) carry out imputation on pre-phased genotypes, that is, they work with haplotypes. The formula undergoes slight changes, as the set of genotypes is now described as a pool of $2N$ binary encoded alleles

$$\hat{R}_h^2 = \frac{\frac{1}{2N} \sum_{i=1}^{2N} (H_i - \hat{p})^2}{\hat{p}(1-\hat{p})}, \quad (2)$$

$$\hat{p} = \sum_{i=1}^{2N} \frac{H_i}{2N},$$

where H_i is the probability of the imputed allele in the i -th haplotype (varies from 0 to 1 and is estimated by built-in hidden Markov model algorithms); N is the sample size; $\hat{p}$ is the allele frequency estimate. The derivation of the formulas can be found in Supplementary Materials 2 and 3.

When calculating metrics of the second type, part of the information about genotypes in the sample under study is artificially “masked” (removed from the general data set, while maintaining information about these SNPs). Then the resulting gaps are imputed and compared with real genotypes. For instance, CR represents the proportion of correctly calculated SNPs to all SNPs. The Hellinger exponent is a measure of the distance between two genotype probability distributions and is based on the Bhattacharyya coefficient (Bhattacharyya, 1943), which measures the degree of overlap between two distributions. The SEN metric is the scaled Euclidean distance between the true and imputed dose distributions. Both the Hellinger score and the SEN score are calculated for each individual’s distinct SNPs. IQS is based on Cohen’s kappa statistic and allows for random co-occurrence between imputed and real SNPs.

As mentioned at the beginning, in addition to the listed metrics, a polygenic score (PGS) of the trait can be used to control the quality of imputation (Choi et al., 2020). It is a measure of an individual’s genetic risk for a trait, obtained by summing the quantified effect of many common variants (typically with minor allele frequencies $\geq 1\%$) in the genome, each of which may make a small contribution to an individual’s genetic risk for that trait or disease. PGS is typically calculated as a weighted sum of a set of genetic variants, usually SNPs, defined as single base pair variations from a reference genome. The resulting score has a distribution close to normal in the general population, with higher scores indicating higher risk.

In general, the equation for calculating a weighted polygenic risk score for an individual is as follows (Collister et al., 2022):

$$\text{PGS}_i = \sum_j^M \hat{\beta}_j * \text{dosage}_{ij},$$

where M is the number of SNPs in the model; $\hat{\beta}_j$ is an estimate of the effect size of the j -th variant; dosage_{ij} is the genotype encoded 0, 1, 2 for the j -th variant in the genotype of the i -th individual. SNP effect sizes (β) are often obtained from GWAS results.

After calculating the PGS score for a trait, its values are compared with the values of real phenotypes. If there is a significant correlation between these two data sets, we can conclude that the data is of high quality after imputation.

Examples of imputation in genomic studies on Russian samples

Despite the advantages of imputation and phasing described above, there is very little reference to their use in studies of Russian samples. As such, in a 2023 study on depression in a sample of 4,520 individuals from various regions of Russia, imputation was carried out using the HRC and 1000G reference panels using Beagle 5.1 (Pinakhina et al., 2022). Similarly, in a study of the genetic structure of the Western Russian population (sample of 4,145 individuals), the HRC panel was chosen as the panel for imputation; the procedure itself was carried out using Beagle 4.0 and allowed to consider another 10,454,514 imputed genotyped variants in the analysis, in addition to 623,249 genotyped ones (Usoltsev et al., 2023). And in a 2022 study of markers associated with muscle strength and power in 292 Russians (83 of them professional athletes), not only imputation on a 1000G panel, but also phasing using SHAPEIT was carried out (Moreland et al., 2022).

As stated earlier, one of the most important factors for performing high-quality imputation is the correct choice of reference panel. The authors of one work (Kolosov et al., 2022) assessed the reliability of imputation of genotypes of a sample of 230 elderly people from St. Petersburg (501,100 SNP) by such panels as HRC, 1000G, HGDP (Human Genome Diversity Project (Cann et al., 2002) – a reference panel based on 929 people of various ethnic backgrounds). They were able

to increase the total number of variants studied to 37.6, 37.5 and 26.6 million SNPs for each of the panels, respectively, using Beagle 5.1 (the data were pre-phased). In addition, HRC, compared to the other two panels, showed the highest imputation accuracy (IQS and CR metrics).

All of these works use HRC or 1000G as reference panels, but this approach is somewhat outdated and is subject to revision due to the emergence of a larger TOPMed data set, the use of which serves as a kind of gold standard in international studies at the moment. As for the software, various versions of Beagle are used in the reviewed works.

In the mentioned studies on Russian samples, meta-analyses or fine mapping of genes were not carried out; however, as examples from other works show (Barton et al., 2021), thanks to imputation and phasing such analyzes can be done with a significant quality improvement.

Conclusion

From the above, we can conclude that, at the moment, imputation of genotyping data is an integral part of many human genomic studies, in particular GWAS. It provides an increase in the number of SNPs analyzed and makes it possible to combine the results of different studies. Imputation also significantly improves the results of fine mapping, allowing the most accurate identification of specific genetic variants and genes that determine the association of the entire genome region with the trait being studied (Chundru et al., 2019).

It is worth noting that for large-scale studies where sample size and genotyping coverage are important, the combination of DNA microarrays/sequencing with low coverage and further imputation is the most optimal and cheapest data acquisition strategy suitable for most genomic study designs. This combination is used in all major national biobanks, such as UK Biobank (Sudlow et al., 2015), ALIOFs (Ramirez et al., 2022) and others.

Along with the listed advantages, the imputation method has a number of disadvantages and limitations. In particular, reading errors due to low coverage, as well as incorrect selection of parameters for imputation along with an inappropriate reference panel, often lead to low accuracy of the imputed data, which can negatively affect the results of further stages of analysis. It must also be remembered that imputation uses information about haplotypes from the reference sample, so when it becomes outdated, genetic variants that have become frequent in the population relatively recently may be imputed worse (Ali et al., 2022). In addition, a high level of recombination reduces the accuracy of phasing and subsequent imputation of genotypes, and therefore, in some cases, additional recombination analysis is necessary (Weng et al., 2014).

Also, imputation can smooth out genetic differences between individuals in case-control samples (Lau et al., 2024): imputed data may introduce inaccurate genotypes in regions where differences between case and control are expected, and this effect appears regardless of how large and diverse the reference panel is. Finally, when using the method, it is important to remember that what is true for the population as a whole may not always be true for a specific individual.

Currently, there is a wide variety of programs and reference panels for imputation of human genomic data, and, as a conse-

quence, many combinations of them. Due to this, researchers have the opportunity to select the optimal set of imputation tools for the characteristics of the sample and the objectives of a particular study. A review of works on Russian samples showed that the most popular software for imputation is Beagle of various versions, and among reference panels, HRC and 1000G are most often used, which is somewhat different from international practices, where the leader among reference panels is TOPMed.

Greater awareness of the intricacies of imputation and a deliberate approach to the selection of tools will improve the quality of genomic data without increasing the cost of obtaining them, facilitate their integration with the results of other studies, and provide more accurate information about the genetic control of human traits.

References

- Abraham G., Qiu Y., Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics*. 2017; 33(17):2776-2778. DOI 10.1093/bioinformatics/btx299
- Ali A.T., Liebert A., Lau W., Maniatis N., Swallow D.M. The hazards of genotype imputation in chromosomal regions under selection: A case study using the lactase gene region. *Ann. Hum. Genet.* 2022; 86(1):24-33. DOI 10.1111/ahg.12444
- Anderson C.A., Pettersson F.H., Clarke G.M., Cardon L.R., Morris A.P., Zondervan K.T. Data quality control in genetic case-control association studies. *Nat. Protoc.* 2010;5(9):1564-1573. DOI 10.1038/nprot.2010.116
- Auton A., Abecasis G.R., Altshuler D.M., Durbin R.M., Abecasis G.R., Bentley D.R., ... Min Kang H., Korbel J.O., Marchini J.L., McCarthy S., McVean G.A., Abecasis G.R. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. DOI 10.1038/nature15393
- Barton A.R., Sherman M.A., Mukamel R.E., Loh P.-R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat. Genet.* 2021;53(8):1260-1269. DOI 10.1038/s41588-021-00892-1
- Bhattacharyya A. On a measure of divergence between two multinomial populations. *Sankhyā: Ind. J. Stat.* 1946;7(4):401-406
- Bourke P.M., Voorrips R.E., Visser R.G.F., Maliepaard C. Tools for genetic studies in experimental populations of polyploids. *Front. Plant. Sci.* 2018;9:513. DOI 10.3389/fpls.2018.00513
- Brown A., Ampratwum P.O., Ray S.D. Microarray analysis. In: *Encyclopedia of Toxicology*. 4 ed. 2024;6:385-392. DOI 10.1016/B978-0-12-824315-2.00210-4
- Browning B.L., Zhou Y., Browning S.R. A One-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 2018;103(3):338-348. DOI 10.1016/j.ajhg.2018.07.015
- Browning B.L., Tian X., Zhou Y., Browning S.R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* 2021;108(10):1880-1890. DOI 10.1016/j.ajhg.2021.08.005
- Browning S.R., Browning B.L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* 2011;12(10):703-714. DOI 10.1038/nrg3054
- Cann H.M., de Toma C., Cazes L., Legrand M.F., Morel V., Piouffre L., Bodmer J., ... Zhu S., Weber J.L., Greely H.T., Feldman M.W., Thomas G., Dausset J., Cavalli-Sforza L.L. A human genome diversity cell line panel. *Science*. 2002;296(5566):261-262. DOI 10.1126/science.296.5566.261b
- Chang C.C., Chow C.C., Tellier L.C., Vattikuti S., Purcell S.M., Lee J.J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*. 2015;4(1):7. DOI 10.1186/s13742-015-0047-8
- Chat V., Ferguson R., Morales L., Kirchhoff T. Ultra low-coverage whole-genome sequencing as an alternative to genotyping arrays in

- genome-wide association studies. *Front. Genet.* 2022;12:790445. DOI 10.3389/fgene.2021.790445
- Check Hayden E. Genome sequencing: the third generation. *Nature.* 2009;457(7231):768-769. DOI 10.1038/news.2009.86
- Choi S.W., Mak T.S.-H., O'Reilly P.F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* 2020;15(9):2759-2772. DOI 10.1038/s41596-020-0353-1
- Chundru V.K., Marioni R.E., Prendergast J.G.D., Vallerga C.L., Lin T., Beveridge A.J., Gratten J., Hume D.A., Deary I.J., Wray N.R., Visscher P.M., McRae A.F. Examining the impact of imputation errors on fine-mapping using DNA methylation QTL as a model trait. *Genetics.* 2019;212(3):577-586. DOI 10.1534/genetics.118.301861
- Clark A.G. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* 1990;7(2):111-122. DOI 10.1093/oxfordjournals.molbev.a040591
- Collister J.A., Liu X., Clifton L. Calculating polygenic risk scores (PRS) in UK biobank: A practical guide for epidemiologists. *Front. Genet.* 2022;13:818574. DOI 10.3389/fgene.2022.818574
- Connell C., Fung S., Heiner C., Bridgham J., Chakerian V., Heron E., Jones B., Menchen S., Mordan W., Raff M., Recknor M., Smith L.M., Springer J., Woo S., Hunkapiller M. Automated DNA-sequence analysis. *Biotechniques.* 1987;5:342-348
- Das S., Forer L., Schönerr S., Sidore C., Locke A.E., Kwong A., Vrieze S.I., Chew E.Y., Levy S., McGue M., Schlessinger D., Stambolian D., Loh P.-R., Iacono W.G., Swaroop A., Scott L.J., Cucca F., Kronenberg F., Boehnke M., Abecasis G.R., Fuchsberger C. Next-generation genotype imputation service and methods. *Nat. Genet.* 2016;48(10):1284-1287. DOI 10.1038/ng.3656
- De Marino A., Mahmoud A.A., Bose M., Bircan K.O., Terpolovsky A., Bamunusinghe V., Bohn S., Khan U., Novković B., Yazdi P.G. A comparative analysis of current phasing and imputation software. *PLoS One.* 2022;17(10):e0260177. DOI 10.1371/journal.pone.0260177
- Deamer D., Akeson M., Branton D. Three decades of nanopore sequencing. *Nat. Biotechnol.* 2016;34(5):518-524. DOI 10.1038/nbt.3423
- Delaneau O., Marchini J., Zagury J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods.* 2012;9(2):179-181. DOI 10.1038/nmeth.1785
- Dempster A.P., Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Society.* 1977;39(1):1-38. DOI 10.1111/j.2517-6161.1977.tb01600.x
- DePristo M.A., Banks E., Poplin R., Garimella K.V., Maguire J.R., Hartl C., Philippakis A.A., del Angel G., Rivas M.A., Hanna M., McKenna A., Fennell T.J., Kernytzky A.M., Sivachenko A.Y., Cibulskis K., Gabriel S.B., Altshuler D., Daly M.J. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 2011;43(5):491-498. DOI 10.1038/ng.806
- Drmanac R., Sparks A.B., Callow M.J., Halpern A.L., Burns N.L., Kermani B.G., Carnevali P., ... Drmanac S., Oliphant A.R., Bannyai W.C., Martin B., Ballinger D.G., Church G.M., Reid C.A. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science.* 2010;327(5961):78-81. DOI 10.1126/science.1181498
- Fan J.B., Oliphant A., Shen R., Kermani B.G., Garcia F., Gunderson K.L., Hansen M., ... Kruglyak S., Bentley D., Haas J., Rigault P., Zhou L., Stueland J., Chee M.S. Highly parallel SNP genotyping. *Cold Spring Harb. Symp. Quant. Biol.* 2003;68:69-78. DOI 10.1101/sqb.2003.68.69
- Fatumo S., Mugisha J., Soremekun O.S., Kalungi A., Mayanja R., Kintu C., Makanga R., Kakande A., Abaasa A., Asiki G., Kalyesubula R., Newton R., Nyirenda M., Sandhu M.S., Kaleebu P. Uganda genome resource: A rich research database for genomic studies of communicable and non-communicable diseases in Africa. *Cell Genom.* 2022;2(11):100209. DOI 10.1016/j.xgen.2022.100209
- Feng Z., Peng F., Xie F., Liu Y., Zhang H., Ma J., Xing J., Guo X. Comparison of capture-based mtDNA sequencing performance between MGI and illumina sequencing platforms in various sample types. *BMC Genomics.* 2024;25(1):41. DOI 10.1186/s12864-023-09938-6
- Govindarajan R., Duraiyan J., Kaliyappan K., Palanisamy M. Microarray and its applications. *J. Pharm. Bioallied Sci.* 2012;4(6):310. DOI 10.4103/0975-7406.100283
- Gresham D., Dunham M.J., Botstein D. Comparing whole genomes using DNA microarrays. *Nat. Rev. Genet.* 2008;9(4):291-302. DOI 10.1038/nrg2335
- Guo Y., He J., Zhao S., Wu H., Zhong X., Sheng Q., Samuels D.C., Shyr Y., Long J. Illumina human exome genotyping array clustering and quality control. *Nat. Protoc.* 2014;9(11):2643-2662. DOI 10.1038/nprot.2014.174
- Hayat M.A. DNA microarrays technology. In: Handbook of Immunohistochemistry and *in situ* Hybridization of Human Carcinomas. 2002;49-55. DOI 10.1016/S1874-5784(04)80015-1
- Huang G.-H., Tseng Y.-C. Genotype imputation accuracy with different reference panels in admixed populations. *BMC Proc.* 2014;8(S1):S64. DOI 10.1186/1753-6561-8-S1-S64
- Jeon S.A., Park J.L., Park S.-J., Kim J.H., Goh S.-H., Han J.-Y., Kim S.-Y. Comparison between MGI and illumina sequencing platforms for whole genome sequencing. *Genes Genom.* 2021;43(7):713-724. DOI 10.1007/s13258-021-01096-x
- Kolosov N., Rezapova V., Rotar O., Loboda A., Freylikhman O., Melnik O., Sergushichev A., Stevens C., Voortman T., Kostareva A., Konradi A., Daly M.J., Artomov M. Genotype imputation and polygenic score estimation in northwestern Russian population. *PLoS One.* 2022;17(6):e0269434. DOI 10.1371/journal.pone.0269434
- Korostin D., Kulemin N., Naumov V., Belova V., Kwon D., Gorbachev A. Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing. *PLoS One.* 2020;15(3):e0230301. DOI 10.1371/journal.pone.0230301
- Kurg A., Töniss N., Georgiou I., Shumaker J., Tollett J., Metspalu A. Arrayed primer extension: solid-phase four-color DNA resequencing and mutation detection technology. *Genet. Test.* 2000;4(1):1-7. DOI 10.1089/109065700316408
- Lam M., Awasthi S., Watson H.J., Goldstein J., Panagiotaropoulou G., Trubetskoy V., Karlsson R., Frei O., Fan C.-C., De Witte W., Mota N.R., Mullins N., Brügger K., Lee S.H., Wray N.R., Skarabis N., Huang H., Neale B., Daly M.J., Mattheisen M., Walters R., Ripke S. RICOPI: rapid imputation for Consortiums PipeLine. *Bioinformatics.* 2020;36(3):930-933. DOI 10.1093/bioinformatics/btz633
- Lamy P., Andersen C.L., Wikman F.P., Wiuf C. Genotyping and annotation of Affymetrix SNP arrays. *Nucleic Acids Res.* 2006;34(14):e100. DOI 10.1093/nar/gkl475
- Lau W., Ali A., Maude H., Andrew T., Swallow D.M., Maniatis N. The hazards of genotype imputation when mapping disease susceptibility variants. *Genome Biol.* 2024;25(1):7. DOI 10.1186/s13059-023-03140-3
- Li L., Huang P., Sun X., Wang S., Xu M., Liu S., Feng Z., Zhang Q., Wang X., Zheng X., Dai M., Bi Y., Ning G., Cao Y., Wang W. The ChinaMAP reference panel for the accurate genotype imputation in Chinese populations. *Cell Res.* 2021;31(12):1308-1310. DOI 10.1038/s41422-021-00564-z
- Li N., Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics.* 2003;165(4):2213-2233. DOI 10.1093/genetics/165.4.2213
- Li Y., Willer C., Sanna S., Abecasis G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* 2009;10(1):387-406. DOI 10.1146/annurev.genom.9.081307.164242
- Lin P., Hartz S.M., Zhang Z., Saccone S.F., Wang J., Tischfield J.A., Edenberg H.J., Kramer J.R., Goate A.M., Bierut L.J., Rice J.P. A new statistic to evaluate imputation reliability. *PLoS One.* 2010;5(3):e9697. DOI 10.1371/journal.pone.0009697

- Loh P.-R., Danecek P., Palamara P.F., Fuchsberger C., Reshef Y.A., Finucane H.K., Schoenherr S., Forer L., McCarthy S., Abecasis G.R., Durbin R., L Price A. Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.* 2016;48(11):1443-1448. DOI 10.1038/ng.3679
- Marchini J., Howie B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 2010;11(7):499-511. DOI 10.1038/nrg2796
- Marees A.T., de Kluiver H., Stringer S., Vorspan F., Curis E., Marie Claire C., Derks E.M. A tutorial on conducting genome wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* 2018;27(2). DOI 10.1002/mpr.1608
- Martin A.R., Atkinson E.G., Chapman S.B., Stevenson A., Stroud R.E., Abebe T., Akena D., ... Ramesar R., Shiferaw W., Stein D.J., Tefera S., van der Merwe C., Zingela Z. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *Am. J. Hum. Genet.* 2021;108(4):656-668. DOI 10.1016/j.ajhg.2021.03.012
- Maxam A.M., Gilbert W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA.* 1977;74(2):560-564. DOI 10.1073/pnas.74.2.560
- Mills M.C., Barban N., Tropf F.C. An Introduction to Statistical Genetic Data Analysis. Cambridge, MA: MIT Press, 2020
- Mirzabekov A.D. Biochips in the biology and medicine of the XXI century. *Vestnik Rossiyskoy Akademii Nauk = Herald of the Russian Academy of Sciences.* 2003;73(5):412 (in Russian)
- Moreland E., Borisov O.V., Semenova E.A., Larin A.K., Andryushchenko O.N., Andryushchenko L.B., Generozov E.V., Williams A.G., Ahmetov I.I. Polygenic profile of elite strength athletes. *J. Strength. Cond. Res.* 2022;36(9):2509-2514. DOI 10.1519/JSC.0000000000003901
- O'Connell J., Yun T., Moreno M., Li H., Litterman N., Kolesnikov A., Noblin E., ... Wang W., Weldon C.H., Wilton P., Wong C., Auton A., Carroll A., McLean C.Y. A population-specific reference panel for improved genotype imputation in African Americans. *Commun. Biol.* 2021;4(1):1269. DOI 10.1038/s42003-021-02777-9
- Pasaniuc B., Rohland N., McLaren P.J., Garimella K., Zaitlen N., Li H., Gupta N., ... Haas D.W., Liang L., Sunyaev S., Patterson N., de Bakker P.I.W., Reich D., Price A.L. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* 2012;44(6):631-635. DOI 10.1038/ng.2283
- Pinakhina D., Yermakov D., Vergasova E., Kasyanov E., Rukavishnikov G., Rezapova V., Kolosov, ... Plotnikov N., Ilinsky V., Neznanov N., Mazo G., Kibitov A., Rakitko A., Artomov M. GWAS of depression in 4,520 individuals from the Russian population highlights the role of MAGI2 (S-SCAM) in the gut-brain axis. *Front. Genet.* 2022;13:972196. DOI 10.3389/fgene.2022.972196
- Price A.L., Patterson N.J., Plenge R.M., Weinblatt M.E., Shadick N.A., Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 2006;38(8):904-909. DOI 10.1038/ng1847
- Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., Bender D., Maller J., Sklar P., de Bakker P.I.W., Daly M.J., Sham P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 2007;81(3):559-575. DOI 10.1086/519795
- Ramirez A.H., Sulieman L., Schlueter D.J., Halvorson A., Qian J., Ratsimbazafy F., Loperena R., ... Denny J.C., Carroll R.J., Glazer D., Harris P.A., Hripesak G., Philippakis A., Roden D.M.; All of Us research program. The All of Us research program: Data quality, utility, and diversity. *Patterns (N Y).* 2022;3(8):100570. DOI 10.1016/j.patter.2022.100570
- Rhoads A., Au K.F. PacBio Sequencing and its applications. *Genomics Proteomics Bioinformatics.* 2015;13(5):278-289. DOI 10.1016/j.gpb.2015.08.002
- Roshyara N.R., Kirsten H., Horn K., Ahnert P., Scholz M. Impact of pre-imputation SNP-filtering on genotype imputation results. *BMC Genet.* 2014;15(1):88. DOI 10.1186/s12863-014-0088-5
- Rubinacci S., Delaneau O., Marchini J. Genotype imputation using the Positional Burrows Wheeler Transform. *PLoS Genet.* 2020;16(11): e1009049. DOI 10.1371/journal.pgen.1009049
- Sanger F., Nicklen S., Coulson A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA.* 1977;74(12):5463-5467. DOI 10.1073/pnas.74.12.5463
- 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.* 2006;78(4): 629-644. DOI 10.1086/502802
- Shendure J., Balasubramanian S., Church G.M., Gilbert W., Rogers J., Schloss J.A., Waterston R.H. DNA sequencing at 40: past, present and future. *Nature.* 2017;550(7676):345-353. DOI 10.1038/nature24286
- Smith L.M., Sanders J.Z., Kaiser R.J., Hughes P., Dodd C., Connell C.R., Heiner C., Kent S.B.H., Hood L.E. Fluorescence detection in automated DNA sequence analysis. *Nature.* 1986;321(6071): 674-679. DOI 10.1038/321674a0
- Stahl K., Gola D., König I.R. Assessment of imputation quality: comparison of phasing and imputation algorithms in real data. *Front. Genet.* 2021;12:724037. DOI 10.3389/fgene.2021.724037
- Sudlow C., Gallacher J., Allen N., Beral V., Burton P., Danesh J., Downey P., Elliott P., Green J., Landray M., Liu B., Matthews P., Ong G., Pell J., Silman A., Young A., Sprosen T., Peakman T., Collins R. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 2015;12(3):e1001779. DOI 10.1371/journal.pmed.1001779
- Sudmant P.H., Rausch T., Gardner E.J., Handsaker R.E., Abyzov A., Huddleston J., Zhang Y., ... Gerstein M.B., Bashir A., Stegle O., Devine S.E., Lee C., Eichler E.E., Korb J.O. An integrated map of structural variation in 2,504 human genomes. *Nature.* 2015; 526(7571):75-81. DOI 10.1038/nature15394
- Talim D., Harris D.N., Kessler M.D., Carlson J., Szpiech Z.A., Torres R., ... Cupples L.A., Laurie C.C., Jaquish C.E., Hernandez R.D., O'Connor T.D., Abecasis G.R. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature.* 2021; 590(7845):290-299. DOI 10.1038/s41586-021-03205-y
- The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* 2016;48:1279-1283. DOI 10.1038/ng.3643
- Usoltsev D., Kolosov N., Rotar O., Loboda A., Boyarinova M., Moguchaya E., Kolesova E., ... Laiho P., Kostareva A., Konradi A., Shlyakhto E., Palotie A., Daly M.J., Artomov M. Understanding complex trait susceptibilities and ethnical diversity in a sample of 4,145 Russians through analysis of clinical and genetic data. *bioRxiv.* 2023. DOI 10.1101/2023.03.23.534000
- Wall J.D., Stawiski E.W., Ratan A., Kim H.L., Kim C., Gupta R., Suryamohan K., ... Radha V., Mohan V., Majumder P.P., Seshagiri S., Seo J.-S., Schuster S.C., Peterson A.S. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature.* 2019; 576(7785):106-111. DOI 10.1038/s41586-019-1793-z
- Wang D.G., Fan J.-B., Siao C.-J., Berno A., Young P., Sapolsky R., Ghandour G., Perkins N., Winchester E., Spencer J., Kruglyak L., Stein L., Hsie L., Topaloglou T., Hubbell E., Robinson E., Mittmann M., Morris M.S., Shen N., Kilburn D., Rioux J., Nusbaum C., Rozen S., Hudson T.J., Lipshutz R., Chee M., Lander E.S. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science.* 1998;280(5366):1077-1082. DOI 10.1126/science.280.5366.1077
- Wang Q.S., Huang H. Methods for statistical fine-mapping and their applications to auto-immune diseases. *Semin. Immunopathol.* 2022; 44(1):101-113. DOI 10.1007/s00281-021-00902-8
- Weale M.E. A survey of current software for haplotype phase inference. *Hum. Genomics.* 2004;1(2):141. DOI 10.1186/1479-7364-1-2-141
- Weng Z.-Q., Saatchi M., Schnabel R.D., Taylor J.F., Garrick D.J. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. *Gen. Select. Evol.* 2014;46(1):34. DOI 10.1186/1297-9686-46-34

- Wu D., Dou J., Chai X., Bellis C., Wilm A., Shih C.C., ... Wong W.-C., Xie Z., Yeo K.K., Zhang L., Zhai W., Zhao Y. Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell*. 2019;179(3):736-749.e15. DOI 10.1016/j.cell.2019.09.019
- Yang H.-C., Lin H.-C., Kang M., Chen C.-H., Lin C.-W., Li L.-H., Wu J.-Y., Chen Y.-T., Pan W.-H. SAQC: SNP array quality control. *BMC Bioinformatics*. 2011;12(1):100. DOI 10.1186/1471-2105-12-100
- Yoo S.-K., Kim C.-U., Kim H.L., Kim S., Shin J.-Y., Kim N., Yang J.S.W., Lo K.-W., Cho B., Matsuda F., Schuster S.C., Kim C., Kim J.-I., Seo J.-S. NARD: whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants. *Genome Med*. 2019;11(1):64. DOI 10.1186/s13073-019-0677-z
- Yu K., Das S., LeFaive J., Kwong A., Pleinness J., Forer L., Schönherr S., Fuchsberger C., Smith A.V., Abecasis G.R. Meta-imputation: An efficient method to combine genotype data after imputation with multiple reference panels. *Am. J. Hum. Genet*. 2022;109(6):1007-1015. DOI 10.1016/j.ajhg.2022.04.002
- Zhao S., Jing W., Samuels D.C., Sheng Q., Shyr Y., Guo Y. Strategies for processing and quality control of Illumina genotyping arrays. *Brief. Bioinform*. 2018;19(5):765-775. DOI 10.1093/bib/bbx012

Conflict of interest. The authors declare no conflict of interest.

Received December 12, 2023. Revised April 23, 2024. Accepted July 4, 2024.