INTEGRATED REFERENCE PANEL OF MACACA *MULATTA HAPLOTYPES* FOR GENOTYPE IMPUTATION IN POPULATIONS OF MIXED ORIGIN

Sharko FS, Prokopenko AV, Izotova AO, Toshchakov SV ✉

National Research Centre "Kurchatov Institute", Moscow, Russia

Genotype imputation makes it possible to considerably reduce the cost of whole-genome data acquisition preserving the information value, but its accuracy depends directly on the availability of a high-quality phased reference panel of haplotypes. The existing panels for *Macaca mulatta* cover mostly the Indian species lineage and do not reflect the genetic structure of laboratory populations of mixed origin. This retrospective bioinformatics study aimed to produce and validate a phased reference panel based on the consolidation of publicly available whole-genome datasets from 618 *M. mulatta* individuals of both geographic lineages processed in accordance with the common standardized protocol. The resulting panel includes 33,457,491 SNP markers. Validation on samples of Indian and Chinese origin with ultra-low coverage (0.1x–0.9x) confirmed high imputation accuracy, which was highest for the Indian lineage dominant in the panel. Testing on 10 independent samples with the 0.6x coverage showed a high average imputation reliability score (INFO = 0.898). The created resource opens the prospects for cost-effective genotyping of large *M. mulatta* cohorts and provides the basis for whole-genome association and population genetic studies in domestic and foreign primate research centers.

Keywords: genotype imputation, reference panel, *Macaca mulatta*, whole-genome sequencing, GLIMPSE2, SHAPEIT5, bioinformatics

Funding: the study was supported by the Ministry of Science and Higher Education of the Russian Federation (Federal Scientific and Technical Program for the Development of Genetic Technologies for 2019–2030), agreement No. 075-15-2025-491 dated 30.05.2025.

Acknowledgements: the study was conducted using the equipment of the shared use center "Complex for Modeling and Processing Data from Mega-Class Research Facilities", National Research Centre "Kurchatov Institute", http://ckp.nrcki.ru/.

Author contribution: Toshchakov SV — research design, manuscript writing; Sharko FS — bioinformatics analysis, manuscript draft; Prokopenko AV, Izotova AO — manuscript editing.

Compliance with ethical standards: the studies were approved by the Ethics Committee of the National Research Centre "Kurchatov Institute" (protocol №2-2пр dated June 30, 2025).

✉ **Correspondence should be addressed:** Stepan V. Toshchakov
Academician Kurchatov pl, 1, Moscow, 123098, Russia; stepan.toshchakov@gmail.com

Received: 14.07.2026 **Accepted:** 14.08.2026 **Published online:** 21.08.2026

DOI: 10.24075/brsmu.2026.045

ИНТЕГРИРОВАННАЯ РЕФЕРЕНСНАЯ ПАНЕЛЬ ГАПЛОТИПОВ *MACACA MULATTA* ДЛЯ ИМПУТАЦИИ ГЕНОТИПОВ В ПОПУЛЯЦИЯХ СМЕШАННОГО ПРОИСХОЖДЕНИЯ

Ф. С. Шарко, А. В. Прокопенко, А. О. Изотова, С. В. Тощаков ✉

Национальный исследовательский центр «Курчатовский институт», Москва, Россия

Импутация генотипов позволяет значительно снизить затраты на получение полногеномных данных при сохранении информативности, однако ее точность напрямую зависит от наличия качественной фазированной референсной панели гаплотипов. Существующие панели для *Macaca mulatta* охватывают преимущественно индийскую линию вида и не отражают генетическую структуру лабораторных популяций смешанного происхождения. Целью ретроспективного биоинформатического исследования было создать и отвалидировать фазированную референсную панель на основе консолидации 618 общедоступных полногеномных данных *M. mulatta* обеих географических линий, обработанных по единому стандартизированному протоколу. Итоговая панель включает 33 457 491 SNP-маркер. Валидация на образцах индийского и китайского происхождения при сверхнизком покрытии (0,1×–0,9×) подтвердила высокую точность импутации, максимальную для доминирующей в панели индийской линии. Апробация на 10 независимых образцах при покрытии 0,6x показала высокий средний показатель достоверности импутации (INFO = 0,898). Созданный ресурс открывает возможности для экономически эффективного генотипирования больших когорт *M. mulatta* и служит основой для проведения полногеномных ассоциативных и популяционно-генетических исследований в отечественных и зарубежных приматологических центрах.

Ключевые слова: импутация генотипов, референсная панель, *Macaca mulatta*, полногеномное секвенирование, GLIMPSE2, SHAPEIT5, биоинформатика

Финансирование: работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации (Федеральная научно-техническая программа развития генетических технологий на 2019–2030 г.), соглашение №075-15-2025-491 от 30.05.2025.

Благодарности: работа выполнена с использованием оборудования центра коллективного пользования «Комплекс моделирования и обработки данных исследовательских установок мега-класса» НИЦ «Курчатовский институт», http://ckp.nrcki.ru/.

Вклад авторов: С. В. Тощаков — дизайн работы, написание рукописи; Ф. С. Шарко — биоинформатический анализ, подготовка черновика; А. В. Прокопенко, А. О. Изотова — редактирование рукописи.

Соблюдение этических стандартов: исследования одобрены этическим комитетом НИЦ «Курчатовский институт» (протокол №2-2пр от 30 июня 2025 г.).

✉ **Для корреспонденции:** Степан Владимирович Тощаков
пл. Академика Курчатова, д. 1, г. Москва, 123098, Россия; stepan.toshchakov@gmail.com

Статья получена: 14.07.2026 **Статья принята к печати:** 14.08.2026 **Опубликована онлайн:** 21.08.2026

DOI: 10.24075/vrgmu.2026.045

The rhesus macaque (*Macaca mulatta*) is one of the model organisms most often used in biomedical research. In the area of infectious diseases this species is used as a model for studying human diseases of various etiologies, including HIV/AIDS, tuberculosis, hepatitis and influenza [1]. Furthermore, due to the high phylogenetic similarity of the macaque's and human nervous systems, macaques are widely used in neurobiology, specifically when modeling neurodegenerative processes and Alzheimer's disease [2], a number of rare congenital disorders [3], as well as for preclinical assessment of neuroactive medicinal products [4]. The wide use of this species in experimental biomedicine determines the great interest in studying the genetic basis of the *M. mulatta* phenotypic variation, particularly differences in the susceptibility to diseases and the response to therapy [5].

Genome-wide association studies (GWAS) are the key instrument for solving such problems, but such studies involving large cohorts are limited by the high cost of whole genome sequencing [6]. When the budget is limited, approaches that make it possible to reduce the cost of obtaining genomic data without the decrease in their information content are of particular importance. The low-coverage sequencing with subsequent genotype imputation is an effective alternative to deep sequencing. Such an approach makes it possible to reconstruct missing genotypes based on the known haplotypes from the standard (reference) panel, significantly reducing the cost of analysis compared to the standard coverage (from 20x) and allowing for the population studies on the sets of thousands of samples [7]. Availability of the reference panel that is representative, phased, and genetically close to the target sample is the key factor determining the imputation effectiveness [8].

A number of genomic resources focused on the cataloging and annotation of known genetic variants have been created for *M. mulatta*. Of those the largest one is mGAP representing a catalogue of the annotated variants obtained as a result of sequencing 2.5 thousand animals [5]. The RhesusBase database including 46.1 million polymorphic sites [9] and the MACSNVdb resource for interspecies comparisons within the genus *Macaca* (74.5 million SNV) [10] should be also noted. These resources solve the problems of cataloging, annotating, and searching for known variants, but do not provide a phased haplotype reference panel necessary for the reconstruction of whole-genome genotypes in new, previously unsequenced samples.

In current practice, phased reference panels for genotype imputation represent a basic instrument used in large genomic projects related with human studies and studies involving livestock [11, 12]. Extensive *Macaca mulatta* whole-genome sequencing data sets are available in international repositories. In 2025, the first specialized reference panel for this species was published, providing high genotype imputation accuracy even with the 0.5× coverage [13]. However, this resource is primarily confined to the Indian Macaca mulatta lineage. The genetic structure of artificially created closed populations, whose genesis is associated with the heterogeneous ancestry of the source stock, the fragmentary nature of archive data on introduction, and many years of panmixia, remains beyond the scope of this analysis [14, 15]. For populations of mixed origin, the optimal approach involves employing an integrated panel that incorporates the major geographic lineages of the species, thereby combining population specificity with versatility.

This study aimed to describe the development and validation of a phased reference panel for *M. mulatta* genotype imputation. The panel was generated using the results of whole-genome sequencing of 618 samples of Indian and Chinese origin obtained from open sources, processed in accordance with a single standardized protocol employing contemporary phasing and imputation algorithms. When assessing imputation accuracy using independent test samples from both lineages, high R^2 and concordance values were obtained within the coverage depth range of 0.1x–0.9x, with the expected advantage of the Indian lineage dominant in the panel. The resource developed provides the basis for genome-wide association and population genetic studies involving rhesus macaques in both domestic and international primatology centers, where the genetic structure of populations may be heterogeneous and apriori unknown [16, 17].

METHODS

Collection and systematization of publicly available whole-genome sequencing data

A systematic search for *M. mulatta* whole-genome sequencing data was conducted in the NCBI SRA [18], EA [19], and NGDC [20] international repositories. Data were selected based on the following criteria: confirmed fact of belonging to the species *M. mulatta*, sequencing depth of at least 20x, as well as availability of verified meta-data on the animal's geographic origin (India or China) and sex. A total of 1808 samples were identified during the primary search. After filtering based on coverage values and meta-data validation the number of candidates was 982. The animals defined in metadata as interspecies hybrids (*M. mulatta* × *M. fascicularis*) and samples of undefined geographic origin were excluded from further analysis.

In the final phase of sampling a stratification procedure was applied aimed at achieving the composition balanced based on sex within each population group and minimizing potential systematic distortions due to uneven presentation of data from different studies. The resulting data set included 618 unique animals from three projects: PRJNA251548 (BCM, $n = 21$), PRJNA382404 (OHSU, $n = 397$), and CRA014717 (KIZ, $n = 200$). Among all samples 418 were of Indian origin and 200 were of Chinese origin; there were 410 females and 208 males. Information about the geographic origin and sex of each sample is provided in the dataset-description.xlsx file in the Zenodo open repository: https://doi.org/10.5281/zenodo.21296424.

Bioinformatics processing and SNP identification

The common standardized bioinformatics processing pipeline that ensured data consistency at all stages was developed and used for all 618 samples. The raw reads were filtered by quality using fastp v0.23.2 [21] with the following parameters: -w 12 -z 7 -V -g --poly_g_min_len 5 -x --poly_x_min_len 10 -5 -3 -M 30 -n 1 -e 20 -l 50 --c. The filtered reads were mapped to the M. mulatta reference genome (assembly Mmul_10, GCF_003339765.1) using Bowtie2 v2.3.5.1 [22] with default settings. SNPs were identified using bcftools v1.15 [23] (mpileup and call commands) with the following key parameters: --redo-BAQ --min-BQ 30 --per-sample-mF --annotate DP,AD --multiallelic-caller --variants-only. In the next phase, insertions and deletions were deleted from the resulting VCF files using vcftools v0.1.16 (--remove-indels parameters) to leave only single nucleotide substitutions (SNPs) for further analysis. The resulting VCF file was subjected to further strict filtering using vcftools v0.1.16 [24] based on the following criteria: genotype quality (GQ) ≥ 30, read depth (DP) 15–180, percentage of missing genotypes per site ≤ 0.02, minimal minor allele frequency (MAF) ≥ 0.01.

Reference panel phasing

The SHAPEIT5 v5.1.0 software package [25] developed for large sets of WG data and characterized by high accuracy of haplotype phase reconstruction when dealing with large samples was used to reconstruct the chromosomal phases of haplotypes of the filtered data set. Stage-by-stage phasing with the genome division into independent blocks was applied to increase the computing efficiency and reduce the RAM requirements. At first, the SHAPEIT5_phase_common module was used for processing in the 20 cM-long chromosome blocks; independent phasing with subsequent quality assessment was performed for each block. To ensure the complete chromosome assembly integrity and eliminate artifacts arising at block boundaries, further phasing was performed in overlapping regions at the junctions (overlap of 2 cM) followed by merging the phases into a single chromosome sequence. As a result, fully phased panel was produced that contained two haplotypes (maternal and paternal) for each of 618 animals, which was essential for subsequent genotype imputation.

Genotype imputation and panel validation

To assess the accuracy of the reference panel produced, two samples of known geographic origin were excluded from the panel: M00016 (Indian population) and CRR1023808 (Chinese population). The source BAM files with high coverage for each such sample were subjected to downsampling to five coverage depth levels: 0.1x, 0.3x, 0.5x, 0.7x, and 0.9x, which made it possible to simulate the low-coverage sequencing conditions.

Genotype imputation for simulated data was performed using the GLIMPSE2 v2.0.0 pipeline [26]. The procedure included sequential steps: genome splitting into overlapping windows using the GLIMPSE2_chunk utility, creation of the reference panel in the GLIMPSE2 format by means of GLIMPSE2_split_reference, the imputation itself and phasing via GLIMPSE2_phase, as well as combining the results from all windows to restore full-chromosome files using GLIMPSE2_ligate.

The imputation accuracy was assessed by matching imputed genotypes with reference ones, for which the original high-confidence genotypes of the same samples before downsampling were used. The GLIMPSE2_concordance tool calculating two indicators was used for calculation: concordance (percentage of completely matched genotypes) and R^2 (square of the correlation coefficient between imputed and reference allele dosages). Such an approach allowed for quantitative characterization of the imputation accuracy as a function of the target sample coverage depth and the degree of its similarity to the reference panel. The reference panel is available from the Zenodo open repository: https://doi.org/10.5281/zenodo.21296424.

RESULTS

Characteristics of the created reference panel

As a result of consolidation and unified processing of 618 *M. mulatta* macaques the phased reference haplotype panel has been produced that includes 33,457,491 highly reliable SNP markers. The panel population structure is determined by the composition of projects included. Most samples are of Indian origin (418 animals; 67.6%), which is due to the PRJNA382404 project predominance in the sample (OHSU, $n = 397$). Samples of the project belong to the Indian population. The Chinese origin is reported for 200 animals (32.4%; project CRA014717, KIZ). Such ratio guarantees the representation of both major geographic lineages in the panel, thereby expanding its application area to laboratory populations of mixed or undefined origin. The sex composition of the sample is characterized by the predominance of females: 410 females (66.3%) and 208 males (33.7%), which reflects the structure of the source publicly available WGS data.

Imputation accuracy evaluation

The imputation accuracy was assessed based on two metrics: concordance (percentage of completely matched genotypes) and R^2 (square of the correlation coefficient between imputed and reference allele dosages). In contrast to concordance, R^2 is more sensitive to errors in cases of rare variants, since it reflects the correlation between the allele dosages. As expected, both indicators were directly dependent on the target sample coverage: the imputation accuracy consistently increased with greater sequencing depth. In addition, the imputation accuracy was assessed depending on the minor allele frequency (MAF) within the range of 1–45%.

As for the M00016 sample of Indian origin, the R^2 value increased from 0.56 to 0.96 when coverage increased from 0.1× to 0.9× (Fig. 1A). Stratification by MAF showed that the R^2 values for common variants (MAF ≥ 10%) reached 0.95–0.96 even with the coverage of 0.7×, while the R^2 values for rare variants (MAF < 1%) was 0.56 at 0.1x and 0.84 at 0.9x. At 0.9x, concordance for this sample varied between 0.61 (MAF 1%) and 0.96–0.97 (MAF ≥ 10%).

As for the CRR1023808 sample from the Chinese population, the R^2 value increased from 0.58 at 0.1x to 0.95 at 0.9x (Fig. 1B). The R^2 values for common variants (MAF ≥ 10%) reached 0.93–0.95 with the coverage of 0.7×, and that for rare variants (MAF 1%) reached 0.58 at 0.1x and 0.82 at 0.9×. At 0.9×, concordance for this sample varied between 0.63 (MAF 1%) and 0.95–0.97 (MAF ≥ 10%).

The R^2 and concordance values obtained suggest high efficiency of the produced reference panel for genotype imputation in rhesus macaques. The R^2 values for common and mid-frequency variants (MAF ≥ 10%) reach 0.93–0.96 even with the coverage of 0.7×, and concordance at 0.9× exceeds 0.95 for both test samples, confirming the genotype reconstruction reliability for this category of markers. The imputation accuracy for rare variants (MAF 1%) was predictably lower: R^2 was 0.82–0.84 at 0.9×, concordance was 0.86–0.88, which suggests the need for deeper coverage (at least 0.7×–0.9×) when conducting the studies focused on the analysis of rare genetic variants. Thus, the panel created ensures high imputation accuracy for a broad range of variants and can be used in population and association studies of rhesus macaques.

Population structure analysis

The principal component analysis (PCA) was conducted based on 33,457,491 SNP markers in order to assess the reference panel genetic structure. The first principal component (PC1, 68.5% of total variance) clearly divides samples of Indian and Chinese origin into two distinct clusters (Fig. 2), which is consistent with the earlier published data in the considerable genetic divergence among two major geographic lineages of *M. mulatta* [27]. The second principal component (PC2, 5.2% of dispersion) reveals an extra structure within each group: samples of Indian origin demonstrate a relatively compact distribution, while Chinese samples are characterized by the

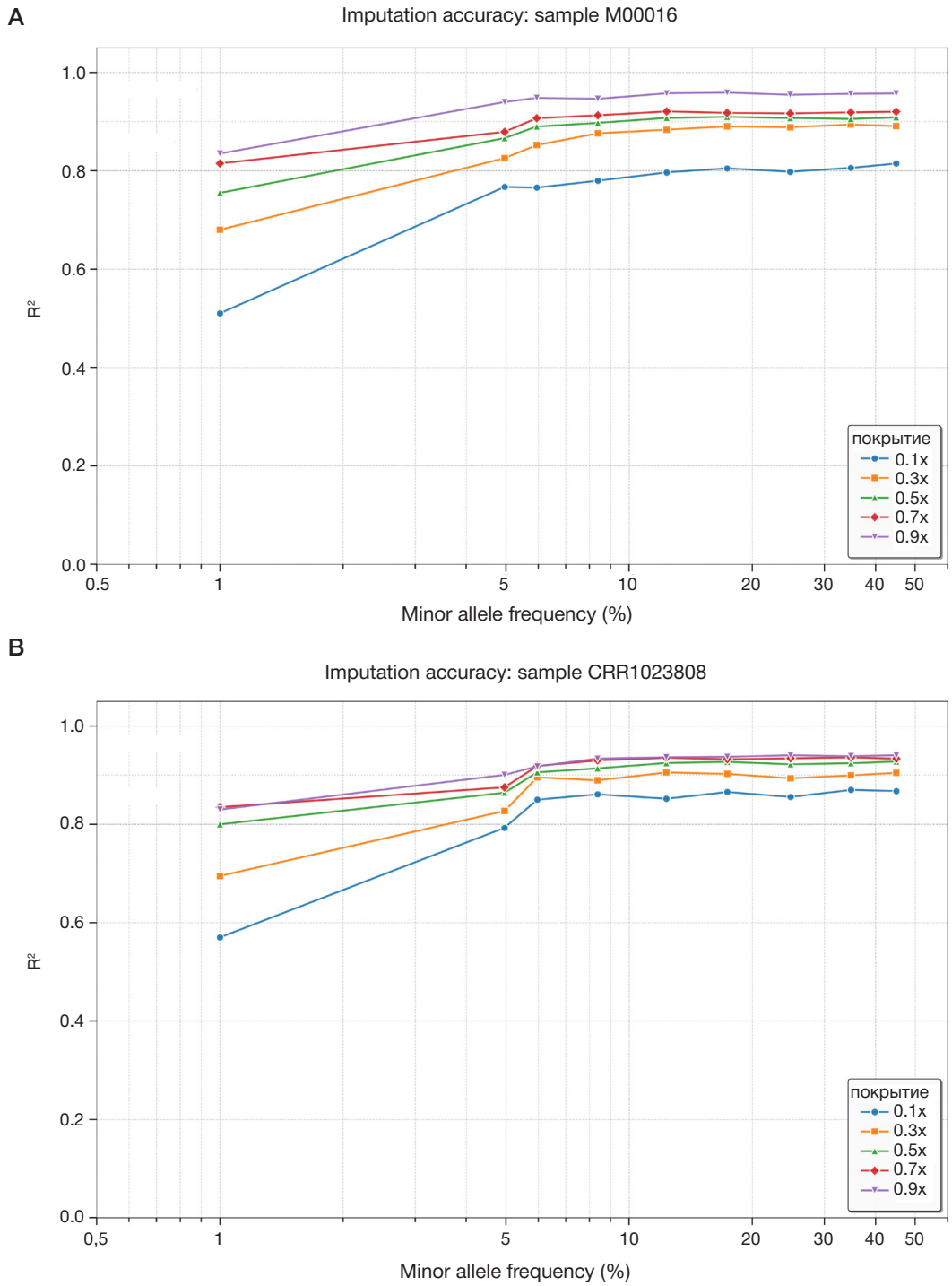

Fig. 1. Genotype imputation accuracy (R^2) depending on the coverage depth and minor allele frequency (MAF) for two test samples. The horizontal axis shows the minor allele frequency (%), the vertical axis shows the R^2 coefficient between imputed and reference allele dosages. Each curve corresponds to one of five coverage levels (0.1×; 0.3×; 0.5×; 0.7×; 0.9×). **A**. M00016 sample of Indian origin. **B**. CRR1023808 sample of Chinese origin

wider spread along the PC2, which corresponds to the higher genetic heterogeneity of the East Asian lineage resulting from its subspecies-level diversity [28].

Panel testing on independent samples

To test the panel using independent data, imputation of genotypes of 10 *M. mulatta* macaques from the collection of the Kurchatov Medical Primatology Complex, National Research Centre "Kurchatov Institute", not included in the reference panel, was performed. Sequencing was conducted with the average coverage of 0.6×, which was compliant with the experimental design ensuring the optimal balance between the cost and reliability when using low-coverage sequencing [7, 12]. Imputation was performed by applying the GLIMPSE2 algorithm to the reference panel created.

The imputation quality was assessed based on the INFO/RSQ indicator which reflects the confidence level for each imputed variant and ranges from 0 to 1. In the studied sample, the INFO/RSQ values varied between 0.862 and 0.934, and the average value was 0.898. All the variants identified were characterized by INFO values > 0.8, corresponding to the strict selection threshold for frequent and mid-frequency variants [29, 30].

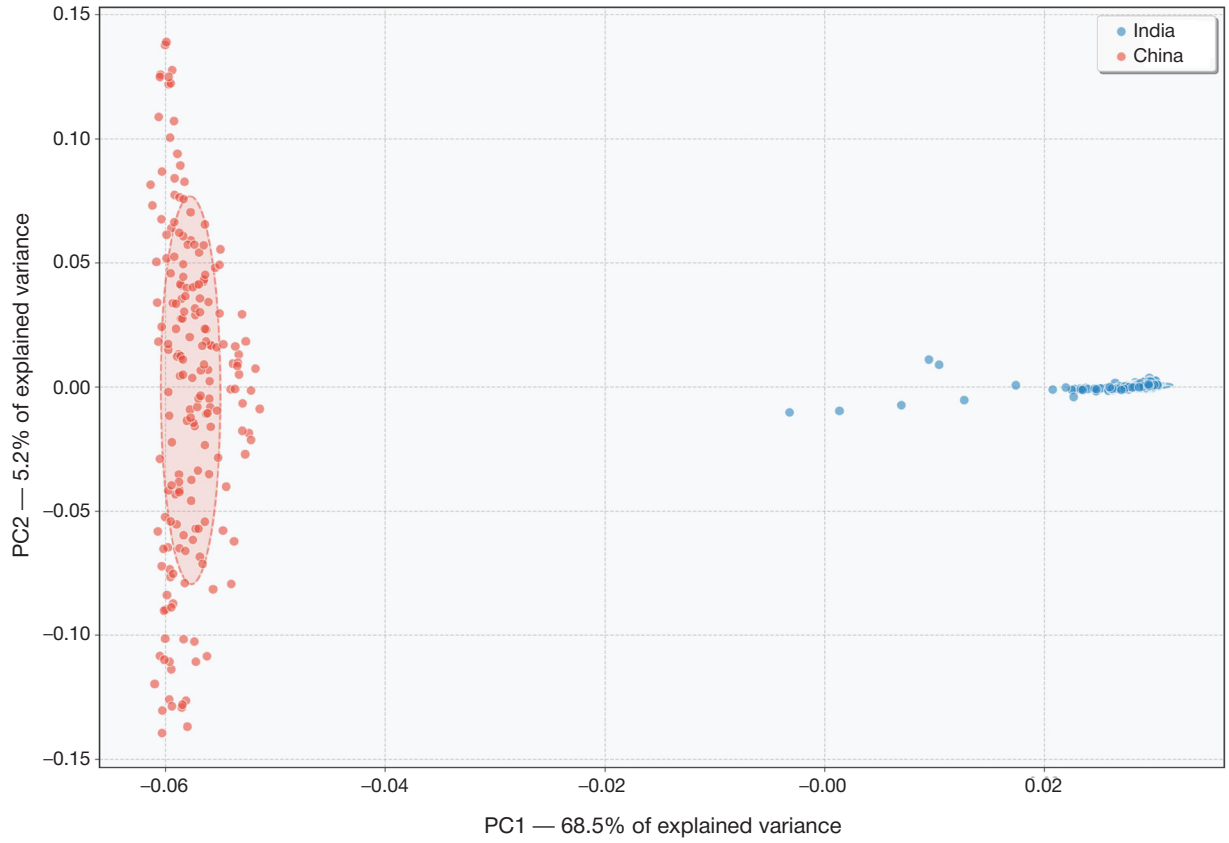

Fig. 2. *M. mulatta* reference haplotype panel principal component analysis (*n* = 618). *Blue* indicates samples of Indian origin (*n* = 418), *red* indicates samples of Chinese origin (*n* = 200)

DISCUSSION

In this study the phased reference haplotype panel for genotype imputation in *Macaca mulatta* that included 618 animals and 33.4 million SNP markers processed using the common standardized pipeline was produced and validated. Among available resources for imputation of genotypes of this species, the panel of 741 animals of the Indian lineage should be highlighted [13], as well as the large cohort of 919 Chinese macaques with phenotypic data representing the GWAS-oriented resource, but not a single phased imputation panel [31]. Neither resource covers both geographic lineages of the species within the framework of the common standardized processing pipeline. At the same time, many laboratory *M. mulatta* populations in the global research centers were formed over decades through the importation of animals from different regions and subsequent uncontrolled crossbreeding, which often led to the unknown or mixed origin of animals [14, 27, 32]. This determines the practical need in the integrated resource covering both lineages, developed in this study.

The imputation accuracy indicators obtained are fully in line with the published data. The highest accuracy was achieved for samples of Indian origin: R^2 0.95–0.96 for common variants even with the coverage of 0.7x, which is comparable with the results obtained based on the Indian panel [13]. Indicators for the Chinese lineage are slightly lower, which naturally reflects its lower representation. The panel developed is far beyond the earlier available resources in marker density, ensuring the coverage of over 33 million SNPs, which corresponds to the level of contemporary reference panels for livestock and opens the prospects for the high-resolution population and association analysis. The panel ensures the imputation quality comparable with that of commercially available SNP chips even with the coverage of 0.6x, which makes the approach cost-effective when used for genotyping of large cohorts.

The practical value of the resource created is determined by the broad range of biomedical tasks, for which *M. mulatta* is the main model: research focused on infectious diseases, research in the field of neurobiology, pharmacology, and toxicology. The imputation-based whole-genome sequencing enables the population screening of animals for carriage of variants orthologous to human pathogenic mutations [33], as well as for controlling the inter-individual pharmacokinetic variability due to polymorphic variants of drug metabolism enzymes [34]. This approach has proven its efficacy in other model and livestock species. A multi-breed panel was created for cattle, including 61.8 million SNP markers. When applying the GLIMPSE2 algorithm, which was also used in our study, the panel showed the concordance over 99% even with the coverage of 0.1× [35]. The SWIM panel (2259 animals of 44 breeds) [36] and the PHARP panel constructed based on the data of 1181 animals from 71 populations [37] were reported for swine. All these panels have been generated by combining the data of several publicly available projects focused on genotyping of various populations, which is consistent with the approach implemented in our study. The growing interest in this area has been also reported in domestic studies: the prospects of low-coverage sequencing for genomic selection of cattle are discussed in the review by Russian authors [38], and the imputed whole-genome data are already used for the association analysis of economically significant traits in Russian herds [39]. The successful experience of creating such resources for other species confirms that the bioinformatics pipeline developed in our study is reproducible and can be adapted for the development of similar panels for other primate species used in biomedical research.

The main directions of the resource development are the panel Chinese component expansion based on the publicly available data and the inclusion of samples from domestic primatology collection, which will make it possible to increase the imputation accuracy for populations of mixed origin. The reference panel created and the bioinformatics approach developed provide the basis for conducting large genome-wide association and population genetic studies involving rhesus macaques in domestic laboratories.

Limitations of the study

The main limitation of the study is uneven representation of geographic lineages in the panel: the number of samples of Indian origin (418) is more than double the number of Chinese ones (200). This has determined the higher imputation accuracy for the Indian one. Furthermore, the panel has been constructed based on the data from foreign repositories (BCM, OHSU, KIZ), so further inclusion of samples from domestic primatology collection is required to more accurately reflect local populations. Accuracy assessment by artificial downsampling was performed for two samples within the coverage range of 0.1–0.9x, while testing on real low-coverage data involved 10 independent samples with the average coverage of 0.6x. In this regard, further testing in the larger heterogeneous sample will make it possible to further confirm reproducibility of the results.

CONCLUSIONS

The phased reference haplotype panel for genotype imputation in *Macaca mulatta* including 618 animals and 33.4 million SNP markers of both major geographic lineages of the species processed using the common standardized pipeline has been created and validated. The panel ensuring high imputation accuracy with the ultralow coverage is ready for practical use in population genetic and association studies. The pipeline developed can serve as a template for the creation of similar resources for other primate species used in biomedical research.

References

1. Gardner MB, Luciw PA. Macaque models of human infectious disease. ILAR J. 2008; 49: 220–55. Available from: https://doi.org/10.1093/ilar.49.2.220.
2. Joho D, Sato K, Kumita W, Saido TC, Sasaki E, Sasaguri H. Advances in non-human primate models for Alzheimer's disease research. Neurosci Res. 2025; 221: 104984. Available from: https://doi.org/10.1016/j.neures.2025.104984.
3. Vallender EJ, Hotchkiss CE, Lewis AD, Rogers J, Stern JA, Peterson SM, et al. Nonhuman primate genetic models for the study of rare diseases. Orphanet J Rare Dis. 2023; 18: 20. Available from: https://doi.org/10.1186/s13023-023-02619-3.
4. Warren WC, Harris RA, Haukness M, Fiddes IT, Murali SC, Fernandes J, et al. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. Science, N.Y. 2020; 370: eabc6617. Available from: https://doi.org/10.1126/science.abc6617.
5. Bimber BN, Yan MY, Peterson SM, Ferguson B. mGAP: the macaque genotype and phenotype resource, a framework for accessing and interpreting macaque variant data, and identifying new models of human disease. BMC Genomics. 2019; 20: 176. Available from: https://doi.org/10.1186/s12864-019-5559-7.
6. Bimber BN, Raboin MJ, Letaw J, Nevonen KA, Spindel JE, McCouch SR, et al. Whole-genome characterization in pedigreed non-human primates using genotyping-by-sequencing (GBS) and imputation. BMC Genomics. 2016; 17: 676. Available from: https://doi.org/10.1186/s12864-016-2966-x.
7. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat Genet. 2021; 53: 120–6. Available from: https://doi.org/10.1038/s41588-020-00756-0.
8. Mitt M, Kals M, Pärn K, Gabriel SB, Lander ES, Palotie A, et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet. 2017; 25: 869–76. Available from: https://doi.org/10.1038/ejhg.2017.51.
9. Zhong X, Peng J, Shen QS, Chen J-Y, Gao H, Luan X, et al. RhesusBase PopGateway: Genome-Wide Population Genetics Atlas in Rhesus Macaque. Mol Biol Evol. 2016; 33: 1370–5. Available from: https://doi.org/10.1093/molbev/msw025.
10. Du L, Guo T, Liu Q, Li J, Zhang X, Xing J, et al. MACSNVdb: a high-quality SNV database for interspecies genetic divergence investigation among macaques. Database (Oxford). 2020; 2020: baaa027. Available from: https://doi.org/10.1093/database/baaa027.
11. Quick C, Anugu P, Musani S, Weiss ST, Burchard EG, White MJ, et al. Sequencing and imputation in GWAS: Cost-effective strategies to increase power and genomic coverage across diverse populations. Genet Epidemiol. 2020; 44: 537–49. Available from: https://doi.org/10.1002/gepi.22326.
12. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet. 2012; 44: 631–5. Available from: https://doi.org/10.1038/ng.2283.
13. Watowich MM, Chiou KL, Graves B, Montague MJ, Brent LJN, Higham JP, et al. Best practices for genotype imputation from low-coverage sequencing data in natural populations. Mol Ecol Resour. 2025; 25: e13854. Available from: https://doi.org/10.1111/1755-0998.13854.
14. Kanthaswamy S, Trask JS, Ross CT, Kou A, Houghton P, Smith DG, et al. A large-scale SNP-based genomic admixture analysis of the captive rhesus macaque colony at the California National Primate Research Center. Am J Primatol. 2012; 74: 747–57. Available from: https://doi.org/10.1002/ajp.22025.
15. Kanthaswamy S, Satkoski J, Kou A, Malladi V, Glenn Smith D. Detecting signatures of inter-regional and inter-specific hybridization among the Chinese rhesus macaque specific pathogen-free (SPF) population using single nucleotide polymorphic (SNP) markers. J Med Primatol. 2010; 39: 252–65. Available from: https://doi.org/10.1111/j.1600-0684.2010.00430.x.
16. Wang J, Wang M, Moshiri A, Harris RA, Raveendran M, Nguyen T, et al. Genetic diversity of 1,845 rhesus macaques improves genetic variation interpretation and identifies disease models. Nat Commun. 2024; 15: 5658. Available from: https://doi.org/10.1038/s41467-024-49922-6.
17. Rogers J. Genomic resources for rhesus macaques (Macaca mulatta). Mamm Genome. 2022; 33: 91–9. Available from: https://doi.org/10.1007/s00335-021-09922-z.
18. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Res. 2022; 50: D387–90. Available from: https://doi.org/10.1093/nar/gkab1053.
19. David Y, Alisha A, Awais A, Rajkumar D, Dipayan G, Muhammad H, et al. The European Nucleotide Archive in 2025. Nucleic Acids Res. 2026; 54: D120–7. Available from: https://doi.org/10.1093/nar/gkaf1295.
20. CNCB-NGDC Members and Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2025. Nucleic Acids Res. 2025; 53: D30–44. Available from: https://doi.org/10.1093/nar/gkae978.
21. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ

preprocessor. Bioinformatics. 2018; 34: 1884–90. Available from: https://doi.org/10.1093/bioinformatics/bty560.
22. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9: 357–9. Available from: https://doi.org/10.1038/nmeth.1923.
23. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021; 10: giab008. Available from: https://doi.org/10.1093/gigascience/giab008.
24. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156–8. Available from: https://doi.org/10.1093/bioinformatics/btr330.
25. Hofmeister RJ, Ribeiro DM, Rubinacci S, Delaneau O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. Nat Genet. Nature Publishing Group. 2023; 55: 1243–9. Available from: https://doi.org/10.1038/s41588-023-01415-w.
26. Rubinacci S, Hofmeister RJ, Sousa da Mota B, Delaneau O. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. Nat Genet. 2023; 55: 1088–90. Available from: https://doi.org/10.1038/s41588-023-01438-3.
27. Smith DG. Genetic characterization of Indian-origin and Chinese-origin rhesus macaques (Macaca mulatta). Comp Med. 2005; 55: 227–30.
28. Liu Z, Tan X, Orozco-terWengel P, Zhou X, Zhang L, Tian S, et al. Population genomics of wild Chinese rhesus macaques reveals a dynamic demographic history and local adaptation, with implications for biomedical research. Gigascience. 2018; 7: giy106. Available from: https://doi.org/10.1093/gigascience/giy106.
29. Stahl K, Papiol S, Budde M, Heilbronner M, Oraki Kohshour M, Falkai P, et al. Aggregating single nucleotide polymorphisms improves filtering for false-positive associations postimputation. G3 (Bethesda). Bethesda, Md. 2025; 15: jkaf043. Available from: https://doi.org/10.1093/g3journal/jkaf043.
30. Fu J, Zhang Q, Wang J, Wang M, Zhang B, Zhu W, и др. Cross-ancestry genome-wide association studies of brain imaging phenotypes. Nat Genet. 2024; 56: 1110–20. Available from: https://doi.org/10.1038/s41588-024-01766-y.
31. Zhang B-L, Chen Y, Zhang Y, Qiao Y, Wu Y, Zhang Y, et al. Forward and reverse genomic screens enhance the understanding of phenotypic variation in a large Chinese rhesus macaque cohort. Nat Commun. Nature Publishing Group. 2025; 16: 8703. Available from: https://doi.org/10.1038/s41467-025-63747-x.
32. Lyke MM, Bagwell A, Newman D, Galindo S, Church T, Christensen C, et al. Mergers and migrations: drivers of population genetic structure in a captive rhesus macaque colony. BMC Genomics. 2026; Available from: https://doi.org/10.1186/s12864-026-13090-2.
33. Wang J, Wang M, Moshiri A, Harris RA, Raveendran M, Nguyen T, et al. Genetic diversity of 1,845 rhesus macaques improves genetic variation interpretation and identifies disease models. Nat Commun. 2024; 15: 5658. Available from: https://doi.org/10.1038/s41467-024-49922-6.
34. Uno Y, Uehara S, Yamazaki H. Genetic polymorphisms of drug-metabolizing cytochrome P450 enzymes in cynomolgus and rhesus monkeys and common marmosets in preclinical studies for humans. Biochem Pharmacol. 2018; 153: 184–95. Available from: https://doi.org/10.1016/j.bcp.2017.12.015.
35. Zhang Z, Wang A, Hu H, Wang L, Gong M, Yang Q, et al. The efficient phasing and imputation pipeline of low-coverage whole genome sequencing data using a high-quality and publicly available reference panel in cattle. Animal Research and One Health. 2023; 1: 4–16. Available from: https://doi.org/10.1002/aro2.8.
36. Ding R, Savegnago R, Liu J, Long N, Tan C, Cai G, et al. The SWine IMputation (SWIM) haplotype reference panel enables nucleotide resolution genetic mapping in pigs. Commun Biol. 2023; 6: 577. Available from: https://doi.org/10.1038/s42003-023-04933-9.
37. Wang Z, Zhang Z, Chen Z, Sun J, Cao C, Wu F, et al. PHARP: a pig haplotype reference panel for genotype imputation. Sci Rep. 2022; 12: 12645. Available from: https://doi.org/10.1038/s41598-022-15851-x.
38. Kostyunina O, Koldichev N, Nemkovskiy G, Traspov A, Ermilov A, Bakoev F, et al. Low-Coverage Whole-Genome Sequencing (lcWGS) in Cattle: Analysis of Potential and Prospects for Application. Animals (Basel). 2025; 15: 3538. Available from: https://doi.org/10.3390/ani15243538.
39. Genome-wide association study of carcass traits in an Angus beef cattle Russian herd using imputed Whole-Genome Sequence data. The Thirteenth International Multiconference. Institute of Cytology and Genetics, the Siberian Branch of the Russian Academy of Sciences; 2022 [дата обращения: 4 август 2026 г.]. Available from: https://doi.org/10.18699/SBB-2022-397.

Литература

1. Gardner MB, Luciw PA. Macaque models of human infectious disease. ILAR J. 2008; 49: 220–55. Available from: https://doi.org/10.1093/ilar.49.2.220.
2. Joho D, Sato K, Kumita W, Saido TC, Sasaki E, Sasaguri H. Advances in non-human primate models for Alzheimer's disease research. Neurosci Res. 2025; 221: 104984. Available from: https://doi.org/10.1016/j.neures.2025.104984.
3. Vallender EJ, Hotchkiss CE, Lewis AD, Rogers J, Stern JA, Peterson SM, et al. Nonhuman primate genetic models for the study of rare diseases. Orphanet J Rare Dis. 2023; 18: 20. Available from: https://doi.org/10.1186/s13023-023-02619-3.
4. Warren WC, Harris RA, Haukness M, Fiddes IT, Murali SC, Fernandes J, et al. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. Science, N.Y. 2020; 370: eabc6617. Available from: https://doi.org/10.1126/science.abc6617.
5. Bimber BN, Yan MY, Peterson SM, Ferguson B. mGAP: the macaque genotype and phenotype resource, a framework for accessing and interpreting macaque variant data, and identifying new models of human disease. BMC Genomics. 2019; 20: 176. Available from: https://doi.org/10.1186/s12864-019-5559-7.
6. Bimber BN, Raboin MJ, Letaw J, Nevonen KA, Spindel JE, McCouch SR, et al. Whole-genome characterization in pedigreed non-human primates using genotyping-by-sequencing (GBS) and imputation. BMC Genomics. 2016; 17: 676. Available from: https://doi.org/10.1186/s12864-016-2966-x.
7. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat Genet. 2021; 53: 120–6. Available from: https://doi.org/10.1038/s41588-020-00756-0.
8. Mitt M, Kals M, Pärn K, Gabriel SB, Lander ES, Palotie A, et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet. 2017; 25: 869–76. Available from: https://doi.org/10.1038/ejhg.2017.51.
9. Zhong X, Peng J, Shen QS, Chen J-Y, Gao H, Luan X, et al. RhesusBase PopGateway: Genome-Wide Population Genetics Atlas in Rhesus Macaque. Mol Biol Evol. 2016; 33: 1370–5. Available from: https://doi.org/10.1093/molbev/msw025.
10. Du L, Guo T, Liu Q, Li J, Zhang X, Xing J, et al. MACSNVdb: a high-quality SNV database for interspecies genetic divergence investigation among macaques. Database (Oxford). 2020; 2020: baaa027. Available from: https://doi.org/10.1093/database/baaa027.
11. Quick C, Anugu P, Musani S, Weiss ST, Burchard EG, White MJ, et al. Sequencing and imputation in GWAS: Cost-effective strategies to increase power and genomic coverage across diverse populations. Genet Epidemiol. 2020; 44: 537–49. Available from: https://doi.org/10.1002/gepi.22326.
12. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet. 2012; 44: 631–5. Available from: https://doi.org/10.1038/ng.2283.
13. Watowich MM, Chiou KL, Graves B, Montague MJ, Brent LJN, Higham JP, et al. Best practices for genotype imputation from low-coverage sequencing data in natural populations. Mol Ecol Resour. 2025; 25: e13854. Available from: https://doi.org/10.1111/1755-0998.13854.
14. Kanthaswamy S, Trask JS, Ross CT, Kou A, Houghton P, Smith DG,

et al. A large-scale SNP-based genomic admixture analysis of the captive rhesus macaque colony at the California National Primate Research Center. Am J Primatol. 2012; 74: 747–57. Available from: https://doi.org/10.1002/ajp.22025.
15. Kanthaswamy S, Satkoski J, Kou A, Malladi V, Glenn Smith D. Detecting signatures of inter-regional and inter-specific hybridization among the Chinese rhesus macaque specific pathogen-free (SPF) population using single nucleotide polymorphic (SNP) markers. J Med Primatol. 2010; 39: 252–65. Available from: https://doi.org/10.1111/j.1600-0684.2010.00430.x.
16. Wang J, Wang M, Moshiri A, Harris RA, Raveendran M, Nguyen T, et al. Genetic diversity of 1,845 rhesus macaques improves genetic variation interpretation and identifies disease models. Nat Commun. 2024; 15: 5658. Available from: https://doi.org/10.1038/s41467-024-49922-6.
17. Rogers J. Genomic resources for rhesus macaques (Macaca mulatta). Mamm Genome. 2022; 33: 91–9. Available from: https://doi.org/10.1007/s00335-021-09922-z.
18. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Res. 2022; 50: D387–90. Available from: https://doi.org/10.1093/nar/gkab1053.
19. David Y, Alisha A, Awais A, Rajkumar D, Dipayan G, Muhammad H, et al. The European Nucleotide Archive in 2025. Nucleic Acids Res. 2026; 54: D120–7. Available from: https://doi.org/10.1093/nar/gkaf1295.
20. CNCB-NGDC Members and Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2025. Nucleic Acids Res. 2025; 53: D30–44. Available from: https://doi.org/10.1093/nar/gkae978.
21. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018; 34: 1884–90. Available from: https://doi.org/10.1093/bioinformatics/bty560.
22. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9: 357–9. Available from: https://doi.org/10.1038/nmeth.1923.
23. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021; 10: giab008. Available from: https://doi.org/10.1093/gigascience/giab008.
24. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156–8. Available from: https://doi.org/10.1093/bioinformatics/btr330.
25. Hofmeister RJ, Ribeiro DM, Rubinacci S, Delaneau O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. Nat Genet. Nature Publishing Group. 2023; 55: 1243–9. Available from: https://doi.org/10.1038/s41588-023-01415-w.
26. Rubinacci S, Hofmeister RJ, Sousa da Mota B, Delaneau O. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. Nat Genet. 2023; 55: 1088–90. Available from: https://doi.org/10.1038/s41588-023-01438-3.
27. Smith DG. Genetic characterization of Indian-origin and Chinese-origin rhesus macaques (Macaca mulatta). Comp Med. 2005; 55: 227–30.
28. Liu Z, Tan X, Orozco-terWengel P, Zhou X, Zhang L, Tian S, et al. Population genomics of wild Chinese rhesus macaques reveals a dynamic demographic history and local adaptation, with implications for biomedical research. Gigascience. 2018; 7: giy106. Available from: https://doi.org/10.1093/gigascience/giy106.
29. Stahl K, Papiol S, Budde M, Heilbronner M, Oraki Kohshour M, Falkai P, et al. Aggregating single nucleotide polymorphisms improves filtering for false-positive associations postimputation. G3 (Bethesda). Bethesda, Md. 2025; 15: jkaf043. Available from: https://doi.org/10.1093/g3journal/jkaf043.
30. Fu J, Zhang Q, Wang J, Wang M, Zhang B, Zhu W, и др. Cross-ancestry genome-wide association studies of brain imaging phenotypes. Nat Genet. 2024; 56: 1110–20. Available from: https://doi.org/10.1038/s41588-024-01766-y.
31. Zhang B-L, Chen Y, Zhang Y, Qiao Y, Wu Y, Zhang Y, et al. Forward and reverse genomic screens enhance the understanding of phenotypic variation in a large Chinese rhesus macaque cohort. Nat Commun. Nature Publishing Group. 2025; 16: 8703. Available from: https://doi.org/10.1038/s41467-025-63747-x.
32. Lyke MM, Bagwell A, Newman D, Galindo S, Church T, Christensen C, et al. Mergers and migrations: drivers of population genetic structure in a captive rhesus macaque colony. BMC Genomics. 2026; Available from: https://doi.org/10.1186/s12864-026-13090-2.
33. Wang J, Wang M, Moshiri A, Harris RA, Raveendran M, Nguyen T, et al. Genetic diversity of 1,845 rhesus macaques improves genetic variation interpretation and identifies disease models. Nat Commun. 2024; 15: 5658. Available from: https://doi.org/10.1038/s41467-024-49922-6.
34. Uno Y, Uehara S, Yamazaki H. Genetic polymorphisms of drug-metabolizing cytochrome P450 enzymes in cynomolgus and rhesus monkeys and common marmosets in preclinical studies for humans. Biochem Pharmacol. 2018; 153: 184–95. Available from: https://doi.org/10.1016/j.bcp.2017.12.015.
35. Zhang Z, Wang A, Hu H, Wang L, Gong M, Yang Q, et al. The efficient phasing and imputation pipeline of low-coverage whole genome sequencing data using a high-quality and publicly available reference panel in cattle. Animal Research and One Health. 2023; 1: 4–16. Available from: https://doi.org/10.1002/aro2.8.
36. Ding R, Savegnago R, Liu J, Long N, Tan C, Cai G, et al. The SWine IMputation (SWIM) haplotype reference panel enables nucleotide resolution genetic mapping in pigs. Commun Biol. 2023; 6: 577. Available from: https://doi.org/10.1038/s42003-023-04933-9.
37. Wang Z, Zhang Z, Chen Z, Sun J, Cao C, Wu F, et al. PHARP: a pig haplotype reference panel for genotype imputation. Sci Rep. 2022; 12: 12645. Available from: https://doi.org/10.1038/s41598-022-15851-x.
38. Kostyunina O, Koldichev N, Nemkovskiy G, Traspov A, Ermilov A, Bakoev F, et al. Low-Coverage Whole-Genome Sequencing (lcWGS) in Cattle: Analysis of Potential and Prospects for Application. Animals (Basel). 2025; 15: 3538. Available from: https://doi.org/10.3390/ani15243538.
39. Genome-wide association study of carcass traits in an Angus beef cattle Russian herd using imputed Whole-Genome Sequence data. The Thirteenth International Multiconference. Institute of Cytology and Genetics, the Siberian Branch of the Russian Academy of Sciences; 2022 [дата обращения: 4 август 2026 г.]. Available from: https://doi.org/10.18699/SBB-2022-397.