doi 10.18699/bgrs2024-4.3-06

Identification of the optimal reference panel for the imputation of genotype data from the Russian sample on the example of the RuDDS cohort

Berdnikova A.A.[1, 2]*, Tsepilov Y.A.[2], Nostaeva A.V.[1], Leonova O.N.[2, 3], Elgaeva E.E.[1, 2]

[1] *Novosibirsk State University, Novosibirsk, Russia*

[2] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[3] *Priorov Central Institute of Traumatology and Orthopedic, Moscow, Russia*

* *a.berdnikova1@g.nsu.ru*

Key words: imputation; GWAS; SNP; genotyping; DNA microarray

Motivation and Aim: Imputation is a method that allows you to restore missing information about genetic variants that were not genotyped directly from DNA microarrays. Its main idea is to find haplotypes that are similar between the sample of interest (which is usually genotyped with DNA microarrays) and a reference sample (which has whole-genome sequencing data). Imputation of genotype data is an important step in genome wide association studies (GWAS) because it leads to a significant increase in the number of analyzed variants, which therefore improves the resolution of GWAS and enhances the comparability of the data obtained from diverse cohorts and/or different microarrays. In foreign studies, TOPMed reference panel has been considered as a gold standard for imputation, but the transferability of these recommendations to Russian samples is poorly studied. In this regard, the search for an optimal reference panel for imputation of Russian population data is an essential problem. In this study we aimed to test three most popular reference panels adapted for individuals of European descent (HRC, 1000 Genomes, TOPMed) to find an optimal reference panel for imputation of genotype data from Russian population.

Methods and Algorithms: We used genotype data, obtained utilizing DNA microarray with whole-genome coverage (Illumina Infinium Global Screening Array-24 v3.0 Kit) in a sample of self-defined Russian individuals from the Russian Disc Degeneration Study [1]. These data passed preliminary quality control resulting in 132 samples and 449604 SNPs. We then performed imputation of processed data using three reference panels: HRC (Version r1.1 2016) [2], 1000 Genomes Phase 3 (Version 5) [3], and TOPMed (Version r2) [4]. To conduct imputation with the first two panels we utilized Michigan Imputation Server and to run imputation with the latter we used TOPMed Imputation Server [5]. Prior to imputation we conducted phasing with the software Eagle (version 2.4) [6]. After imputation, we compared the quality of imputed data based on the total number of SNPs with imputation quality statistic $R^2 > 0.7$. For the panel shown the best results we calculated a polygenic estimate of the height and body mass index (BMI) and collated them with the real height and BMI of the study participants as a proof of a high imputation quality. The polygenic scores were computed using PLINK v2 [7]. All the calculations were performed in R programming language (version 4.2.3).

Results: Imputation procedure allowed us to obtain 39131578 SNPs while using HRC reference panel, 47109535 SNPs when applying 1000 Genomes panel and

292136463 SNPs by utilizing TOPMed reference panel. After filtration of imputed data by $R^2 > 0.7$ the number of genetic variants decreased to 7358967, 7343241 and 7558595 for HRC, 1000 Genomes and TOPMed, respectively. The latter panel provided maximum number of high-quality imputed SNPs. Using the TOPMed-imputed genotypes we calculated the polygenic scores for height and BMI and compered them with the phenotypes from the study sample. The Pearson correlation coefficient for height and its estimate was equal to 0.35 (p-value = 3.8e–05) and the same coefficient for BMI and its polygenic score was 0.27 (p-value = 0.002).

Conclusion: TOPMed reference panel was considered as the most optimal for imputation of genotype data from the Russian sample.

Funding: This work was supported by the Russian Science Foundation (grant No. 22-15-20037) and the Government of the Novosibirsk region.

References

1. Leonova O.N. et al. A protocol for recruiting and analyzing the disease-oriented Russian disc degeneration study (RuDDS) biobank for functional omics studies of lumbar disc degeneration. *PloS One*. 2022;17(5):e0267384
2. Sudmant P.H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526(7571):75-81
3. McCarthy S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48(10):1279-1283
4. Taliun D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. 2021;590(7845):290-299
5. Das S. et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284-1287
6. Loh P.R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet*. 2016;48(11):1443-1448
7. Chang C.C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4(1):s13742-015-0047-8