dbGaP Data Preprocessing

Below is the citation of the description of dbGaP preprocessing described previously in the research [1].

The individual genotypic and phenotypic data were downloaded from the international database of genotypes and phenotypes, dbGaP, under accession number phs000615.v1.p1. According to the project description, the data were obtained from a large study of about 6000 patients with ischemic stroke and about the same number of people without the disease conducted by 13 genetic centers in the US and 11 in Europe [2]. From them, individuals aged over 55 years who identified themselves as white were selected by us.

Only the polymorphisms located on autosomes were considered in the study. To check the population structure, the genotypic data were processed with the ADMIXTURE program [3] and the PCA method implemented in the Plink 1.9 program [4]. ADMIXTURE showed the presence of individuals with different ancestral components in cases and controls. A two-component model was applied and 1035 people with a substantial contribution from the second ancestral component (i.e., greater than 0.046) were removed from the dataset. Twenty individuals with eigenvector values greater than six standard deviations from the mean were additionally removed according to the results of the PCA analysis of the first twenty principal components. The final numbers of individuals were 4929 and 652 in the case and control groups, respectively.

The quality of genotyping was evaluated using the Plink 1.9 program. Polymorphisms and individuals with a proportion of missing genotypes of more than 20% were filtered out. The genotypes of people from the control group were tested for concordance with the Hardy–Weinberg law using the Plink 1.9 program. The polymorphisms with p-value < 1 × 10−5 were excluded from further consideration. The total number of polymorphisms left equaled 883,908, of which 883,749 were SNPs and 159 short insertion/deletions. The missing genotypes were restored statistically using the LinkImpute program [5]. The original genotypic data were under human genome assembly GRCh37, which was kept while processing the data.

1. References

1. Khvorykh, G.V.; Sapozhnikov, N.A.; Limborska, S.A.; Khrunin, A.V. Evaluation of Density-Based Spatial Clustering for Identifying Genomic Loci Associated with Ischemic Stroke in Genome-Wide Data. Int. J. Mol. Sci. 2023, 24, 15355. https://doi.org/10.3390/ijms242015355

1. Meschia, J.F.; Arnett, D.K.; Ay, H.; Brown, R.D.; Benavente, O.R.; Cole, J.W.; Bakker, P.I.W.; Dichgans, M.; Doheny, K.F.; Fornage, M. Stroke Genetics Network (SiGN) Study. Stroke 2013, 44, 2694–2702. https://doi.org/10.1161/STROKEAHA.113.001857.

2. Alexander, D.H.; Novembre, J.; Lange, K. Fast Model-Based Estimation of Ancestry in Unrelated Individuals. Genome Res. 2009, 19, 1655–1664. https://doi.org/10.1101/gr.094052.109.

3. Chang, C.C.; Chow, C.C.; Tellier, L.C.A.M.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets. Gigascience 2015, 4, 7. https://doi.org/10.1186/s13742-015-0047-8.

4. Money, D.; Gardner, K.; Migicovsky, Z.; Schwaninger, H.; Zhong, G.-Y.; Myles, S.L. LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. G3 Genes|Genomes|Genetics 2015, 5, 2383–2390. https://doi.org/10.1534/g3.115.021667.