DOI 10.18699/SBB-2018-15

A database and analytical platform for mining billions of genetic associations

D.D. Gorev[1, 2, 3]*, T.I. Shashkova[1, 2, 3], E. Pakhomov[2], A. Torgasheva[4], L. Klaric[5, 6], A. Severinov [1, 3], S. Sharapov [2, 4], D.G. Alexeev[1, 2], Y.S. Aulchenko[2, 3, 4]

[1] *LLC 'Knomics', Moscow, Russia*

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

[3] *Moscow Institute of Physics and Technology, Moscow, Russia*

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

[5] *Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK*

[6] *Genos Glycoscience Laboratory, Zagreb, Croatia*

** e-mail: gorev.d@phystech.edu*

Key words: genome-wide association study, snp, database, web-service

Motivation and Aim: Hundreds of genome-wide association studies (GWAS) of human traits are performed each year, and are, together with results from tens of thousands of previously reported GWAS, freely available. These results are published in the form of summary statistics (where for each SNP the allelic frequencies, estimates of the coefficients of regression and their standard errors are typically reported). This information can be used for multiple purposes – from research in fundamental biology and genetics, to biomarker and therapeutic intervention of targets discovery. While the amount of information accumulated by the scientific community is very large, the use of this valuable information is restricted by lack of reporting guidelines and facilities that would allow for quality control (QC), storage, and analysis of such data. This situation forces researchers to spend a lot of time and efforts on data collection, data preprocessing to accommodate different analytical tools, and QC. In this work, we designed a platform for storage, QC, and analysis of GWAS summary statistics.

Results: The original data harmonization algorithm was developed to effectively store and quickly access GWAS data. For data storage and manipulations on our platform we use two database management systems, ClickHouse, to store harmonized GWAS results and PostgreSQL for meta-data storage. Clickhouse provides us with rapid-access storage accessible via powerful and flexible SQL interface. The platform implements several GWAS QC algorithms. It also embeds several methods often used for analysis of GWAS summary statistics, such as LDsr and MRbase libraries that facilitate genetic correlations and mendelian randomization analyses, respectively, and our own implementation of the summary data-based mendelian randomization and heterogeneity in dependent instruments (SMR-HEIDI) testing that allows for analysis of pleiotropy. As a proof of concept, 429 GWASs, totaling to ~3 billion of SNP-trait associations, were uploaded to the platform. On average, the selection of all SNPs in a 500 kbp range from all GWAS in the database to a Python Dataframe takes ~17 seconds, while selection of a specific GWAS ~34 seconds.

Conclusion: We have developed a platform for storage, quality control and analysis of summary GWAS data. The platform is capable of storage and high throughput retrieval and analysis of results of thousands of GWAS.

Availability: GWAS-MAP will soon be available as a web-service platform.