DOI 10.18699/BGRSSB-2018-050

Mining large database of genome-wide associations to identify biomarkers and intervention targets

T.I. Shashkova[1, 2, 3]*, A. Torgasheva[4], D.D. Gorev[1, 2, 3], E. Pakhomov[2], L. Klaric[5, 6], A.V. Severinov[1, 3], S. Sharapov[2, 4], Y.A. Tsepilov[2, 4], D.G. Alexeev[1, 2], J.F. Wilson[5], P. Joshi[5], Y.S. Aulchenko[2, 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] *University of Edinburgh, Edinburgh, UK*
[6] *Genos Glycoscience Laboratory, Zagreb, Croatia*
** e-mail: shashkova@phystech.edu*

Key words: genome-wide association studies, coronary artery disease, biomarker, intervention target

Motivation and Aim: Results from tens of thousands of GWAS that have been performed over the last decade are publicly available. The results are typically presented in the form of genome-wide summary statistics (for each SNP the allelic frequencies, estimates of the coefficients of regression and their standard errors are usually reported). This information can be used for multiple purposes – from research in fundamental biology and genetics, to biomarker and target discovery for therapeutic intervention. The aim of this work is to demonstrate that mining large databases of GWAS results allows identification of biomarkers and intervention targets. We focussed on coronary artery disease (CAD), one of the most economically and socially significant and one of the most studied complex common human diseases.

Methods and Algorithms: We developed a system, 'GWAS-MAP', that allows for a platform for storage, quality control and analysis of GWAS summary statistics. Our platform embeds LDsr and MRbase libraries, facilitating genetic correlations and mendelian randomization (MR) analyses, respectively. The analysis of pleiotropy is possible via our own implementation of summary-level mendelian randomization (SMR)/heterogeneity in dependent instruments (HEIDI) testing. We populated our database with about 220 GWASes. These included GWASes of lipid levels, GWASes for 128 metabolites, and 82 proteins from OLINK panel. We have also used eQTL data from a range of sources.

Results: Observed genetic correlations were consistent with previous studies. We selected 51 loci that were associated with CAD at genome-wide significant level in published GWASes. To define loci profiles in "omics" space we considered SMR results for CAD in metabolomic and proteomic space, which allowed us to cluster loci in several biologically meaningful groups. To understand the biological bases of the locus action, and to potentially provide a drug target, within each locus, we have prioritised the genes using the SMR/HEIDI test between CAD and gene expression. Our results confirmed existing knowledge of CAD mechanisms and suggested several CAD biomarkers and intervention targets.

Conclusion: The analysis of CAD using GWAS-MAP allowed for identification of biomarkers and potential intervention targets. Some of these biomarkers and targets are already well known and are used in clinical practice, validating the approach, whilst some are new – showing our approach to biomarker and target discovery is effective.