DOI 10.18699/SBB-2022-262

A platform for case-control matching enables association studies without genotype sharing

Loboda A.[1]*, Artomov M.[1, 2], Artyomov M.[3], Daly M.[2, 4]

[1] *ITMO University, St. Petersburg, Russia*

[2] *Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA*

[3] *Department of Immunology and Pathology, Washington University in St. Louis, St. Louis, MO, USA*

[4] *Institute for Molecular Medicine Finland, Helsinki, Finland*

** aleks.loboda@gmail.com*

Key words: GWAS, Privacy Preserving, Database, Exome Sequencing

Motivation and Aim: Acquiring a sufficiently powered cohort of control samples can be time consuming or, sometimes, impossible. Accordingly, an ability to leverage control samples that were already collected and sequenced elsewhere could dramatically improve power in genetic association studies. Majority of the genotyped and sequenced human DNA samples to date are subject to strict data sharing regulations, large-scale sharing of, in particular, control samples is extremely challenging. We developed a method allowing selection of the best-matching controls in an external pool of samples that is compliant with personal genotype data protection restrictions.

Methods and Algorithms: We provide a web platform that stores 39,844 exome sequencing samples available for control selection and a complimentary R-package to be used on a user side for generation of anonymous data from case genotypes that will be uploaded to the web-platform.

Results: Our approach uses singular value decomposition of the matrix of case genotypes to rank external controls by similarity to cases. We demonstrate that this recovers an accurate case-control association results for both ultra-rare and common variants. Finally, we provide a free access to a database of 39,844 exomes to be used as external controls that enables association studies for case cohorts lacking control subjects.

Conclusion: We present a freely accessible resource with a large-scale control database enabling association studies for “case-only” sequencing data with carefully selected controls and controllable error rates.

References

1. Bodea C.A., Neale B.M., Ripke S., Barclay M., Peyrin-Biroulet L., Chamaillard M., ..., Ferrier N. A method to exploit the structure of genetic ancestry space to enhance case-control studies. *Am J Hum Genet.* 2016;98(5):857-868.