Experimental signatures-guided framework for subcohort discovery in Biobank data

Beilei Bian[1,2,3], Thomas Lloyd[1,2], Jean Yang[1,2,3*]

1. School of Mathematics and Statistics, The University of Sydney, Sydney, New South Wales, Australia
2. Sydney Precision Data Science Centre, University of Sydney, Sydney, New South Wales, Australia
3. Charles Perkins Centre, The University of Sydney, Sydney, New South Wales, Australia

Experimental signatures derived from transcriptomic or proteomic measurements offer critical insights into disease biology. However, it is often hard to transfer such signatures to observational studies (profiling) in genetic and population-level studies. Here, we present a mechanism-informed approach to address this problem. The approach contains 4 components: (a) uncover gene modules and gene-phenotype relationships via network analysis. (b) select proxy phenotypes. (c) using gene modules to construct polygenic risk score (PRS) (d) subcohort selection and validation. Using COVID severity signatures and the Open Target Platform, we identify proxy phenotypes that share the same gene modules, serving as proxy diseases that capture overlapping genetic mechanisms. For each proxy phenotype, we construct pathway-restricted PRS by aggregating genetic variants within the gene modules. We further ensemble these PRSs to represent each individual's genetic risk. To evaluate, we develop an enrichment score that measures the immune characterisation of the top highest PRS individuals. Applying this framework and severe COVID gene signatures to the UK Biobank enables the identification of a subcohort showing enrichment for immune-related conditions vs low-risk individuals. This approach shows the promise of bringing signature perturbation experiments to a large cohort study.