Title

Integrative polygenic score modeling with tissue-specific annotation improves polygenic scores transferability

Authors and affiliations

Xiaohe Tian[1], Tabassum Fabiha[3], William F Li [1, 2], Kushal K Dey[3], Manolis Kellis [1, 2], Yosuke Tanigawa [1, 2]

1. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA
2. Broad Institute of MIT and Harvard, Cambridge, MA, USA
3. Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Abstract

Systematic characterization of functional annotations like histone modification, tissue-specific expression, and transcription factor binding profiles could enhance PGS transferability across genetic ancestry groups by prioritizing putatively causal alleles in predictive models. However, consensus on the best practices for combining such large-scale resources in PGS modeling has yet been reached.

We hypothesize that large-scale integration of tissue-specific functional annotations through statistical learning improves PGS transferability by incorporating biological priors, by which we introduce candidate variant-to-function (cV2F) informed inclusive PGS (iPGS). We analyze 406,659 ancestrally diverse individuals in UK Biobank and develop predictive models for four traits with clear causal tissues, focusing on genome-wide functional annotations across 1.3 million genetic variants.

In cV2F-iPGS, we use gradient boosted trees to learn the optimal combination of ENCODE4 functional annotations from fine-mapped variants across ~100 traits. We accordingly prioritize likely causal variants when fitting penalized regression on individual-level data from ancestry-diverse individuals. We assess model improvement of those with cV2F-informed priors compared to the vanilla iPGS, which only considers statistical correlations in fitting PGS models.

We show ancestry- and tissue-matched cV2F scores are most effective in improving prediction. Specifically, for predicting lymphocyte count in Africans, ancestry- and tissue-matched cV2F-iPGS show the best predictive performance (R^2=.0073), a 35.13% improvement over the vanilla model (R^2=.0054), a 25.86% improvement over the ancestry-mismatched model (R^2=.0058), and a 23.72% improvement over the tissue-agnostic model (R^2=.0059). The highest improvement is seen in the spirometry measure FEV_1/FVC ratio, where tissue-matched cV2F-iPGS (R^2=0.0067) showed a 36.21% improvement over the vanilla model (R^2=0.0049). Overall, we found ancestry- and tissue-matched cV2F improve transferability of iPGS scores by an average of 13.5% (95% CI: [3.6%, 23.4%], p-value: .004) across the four selected traits.

Lastly, we provide locus-level biological interpretations in cV2F-iPGS models. We find, for example, tissue-matched eQTL signals help improve PGS transferability.

In summary, our approach is the first to leverage multimodal biological priors in fitting PGS models on individuals across the continuum of genetic ancestry, improving PGS accuracy and specificity.