Small-cohort GWAS discovery with AI over massive functional genomics knowledge graph

Anonymous Author(s)
Affiliation
Address
email

Abstract

Genome-Wide Association Studies (GWAS) links genetic markers with diseases and is the cornerstone for the development of effective therapeutics. However, for a long tail of many uncommon diseases, the small GWAS sample sizes limit detection power and hamper development of effective treatments. The recent substantial growth in the size of functional genomics data presents a fresh opportunity to tackle these challenges. Here, we introduce KGWAS, a novel geometric deep learning method that leverages a knowledge graph to integrate massive functional information about variants, genes, gene programs, and their interactions, assessing variant-disease associations. Unlike conventional GWAS, which treats variants independently, our approach recognizes that variants influence disease through complex cellular networks. Our realistic simulations show that KGWAS is well-calibrated and powerful in identifying disease variants. We applied KGWAS to 21 independent UK Biobank diseases/traits from small subsampled cohorts (N=1-10K), and KGWAS produced significantly more independent associations that were replicable in the full cohort (average N=374K), 22.0%-89.9% higher than state-of-the-art baselines. Next, we applied KGWAS to 554 less common UK Biobank diseases (N_{case} <5K) and identified 183 novel loci, 46.9% higher than the original GWAS, including rs2155219 associated with ulcerative colitis potentially via regulating *LRRC32* expression in CD4+ regulatory T cells, and rs73127651 associated with myasthenia gravis potentially via regulating *PPHLN1* expression in brain cell types. Overall, KGWAS is a flexible and powerful AI model to integrate the growing functional genomics data to discover novel variants for small cohort diseases.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.