DOCTORAL THESIS

Statistical methods for integrative analysis of genomic data

Ming, Jingsi

Date of Award:
2018

Abstract

Thousands of risk variants underlying complex phenotypes (quantitative traits and diseases) have been identified in genome-wide association studies (GWAS). However, there are still several challenges towards deepening our understanding of the genetic architectures of complex phenotypes. First, the majority of GWAS hits are in non-coding region and their biological interpretation is still unclear. Second, most complex traits are suggested to be highly polygenic, i.e., they are affected by a vast number of risk variants with individually small or moderate effects, whereas a large proportion of risk variants with small effects remain unknown. Third, accumulating evidence from GWAS suggests the pervasiveness of pleiotropy, a phenomenon that some genetic variants can be associated with multiple traits, but there is a lack of unified framework which is scalable to reveal relationship among a large number of traits and prioritize genetic variants simultaneously with functional annotations integrated. In this thesis, we propose two statistical methods to address these challenges using integrative analysis of summary statistics from GWASs and functional annotations.

In the first part, we propose a latent sparse mixed model (LSMM) to integrate functional annotations with GWAS data. Not only does it increase the statistical power of identifying risk variants, but also offers more biological insights by detecting relevant functional annotations. To allow LSMM scalable to millions of variants and hundreds of functional annotations, we developed an efficient variational expectation-maximization (EM) algorithm for model parameter estimation and statistical inference. We first conducted comprehensive simulation studies to evaluate the performance of LSMM. Then we applied it to analyze 30 GWASs of complex phenotypes integrated with nine genic category annotations and 127 cell-type specific functional annotations from the Roadmap project. The results demonstrate that our method possesses more statistical power than conventional methods, and can help researchers achieve deeper understanding of genetic architecture of these complex phenotypes.

In the second part, we propose a latent probit model (LPM) which combines summary statistics from multiple GWASs and functional annotations, to characterize relationship and increase statistical power to identify risk variants. LPM can also perform hypothesis testing for pleiotropy and annotations enrichment. To enable the scalability of LPM as the number of GWASs increases, we developed an efficient parameter-expanded EM (PX-EM) algorithm which can execute parallelly. We first validated the performance of LPM through comprehensive simulations, then applied it to analyze 44 GWASs with nine genic category annotations. The results demonstrate the benefits of LPM and can offer new insights of disease etiology.

Table of Contents