Principal Investigator

Dr. Jennifer Listgarten

Address

Microsoft, Microsoft Research New England, One Memorial Drive, Cambridge MA 02142, USA

Lead Collaborators

Ms Hilary Finucane

Collaborating Institutions and Addresses

Harvard School of Public Health, Epidemiology, 677 Huntington Avenue, Boston MA 02115, USA

Summary of research

GWAS, statistical methods, heritability, noisy phenotypes

Application Lay Summary:

1a: Some case-control phenotypes, labelled only with 0/1, may actually contain distinct underlying sub-phenotypes; for example, depression is thought to be such a phenotype. Analysis of such imprecisely defined phenotypes has negative consequences. First, it obscures genetic signal, decreasing apparent heritability as well as power for association analysis. Second, it obscures understanding of disease, because only a union of causal genes and pathways can be uncovered. In this work, we apply latent variable models to extract sub-phenotypes jointly from both genotypic and phenotypic data in order overcome these problems. In particular, we seek to identify sub-phenotypes that have higher heritability individually than when grouped together as a single phenotype. Without trying various phenotypes, we can't be sure where this sub-structure may be lying, and therefore plan to investigate these methods on a wide variety of data sets--those as large as possible.

1b: Understanding the genetic underpinnings of disease has an ultimate downstream effect of helping to find drug targets, improve diagnosis, and treatment. Our work, if successful, will yield a more nuanced understanding of the genetic underpinnings of many diseases.

1c: Our goal is to discover from genetic and non-genetic data, whether a given disease, such as depression, is actually a combination of distinct underlying diseases, which we refer to as “sub-phenotypes.” Detecting these sub-phenotypes is important in order to more precisely find the underlying genetic causes of the distinct sub-phenotypes, enabling better diagnosis, treatment and drug development than if these were unknown, as is currently the case. We are developing several statistical and machine learning approaches to tackle this problem, which we investigate on both synthetic and real data thought to contain these sub-phenotypes. We also compare our approach to related methods.

1d: Because we seek to have scalable solutions, and to achieve as much power as possible, we seek access to as data from as many samples as possible from this repository (i.e., not a subset).