Specific Aims

Polygenic risk scores use genetic data to quantify disease risk and thus provide clinicians and public health workers with additional risk factors to inform preventive efforts with the goals of enhancing health, lengthening life, and reducing illness and disability. Recent technology advances have enabled genotyping and phenotyping of hundreds of thousands of individuals as biomedical scientists seek to identify the genetic underpinnings of complex diseases in studies called genome-wide association studies. Genome-wide association studies estimate associations between SNPs, one at a time, and a complex disease. Polygenic risk scores leverage genome-wide SNP effects from genome-wide association studies to quantify disease risk. While scientists have developed many statistical methods to calculate polygenic risk scores, current methods ignore information that may improve predictive performance. Currently, a major obstacle in the field is that polygenic risk scores neglect gene-gene interactions.

To solve this problem, we will develop a polygenic risk score method that accounts for gene-gene interactions in a Bayesian statistical model. In our Bayesian model, we will specify a spike-and-slab prior distribution for genome-wide SNP effects and SNP-SNP interactions. A spike-and-slab prior distribution assumes that most effects are zero, while a small proportion of effects arise from a normal distribution. We'll use mean field variational inference methods to approximate the posterior distribution in an accurate and computationally efficient manner. The feasibility of our strategy is supported by recent advances in variational inference and in large-scale computing. Our proposed statistical method is expected to outperform current polygenic risk score methods that ignore gene-gene interactions. Our new, more accurate method for calculating polygenic risk scores will enable better stratification of individuals by genetic risk, which, in turn, will allow for more accurate targeting of preventive public health interventions.

Aim 1 will develop a Bayesian statistical model for polygenic risk scores that uses only SNP effect estimates (and ignores SNP-SNP interactions). We'll use a spike-and-slab prior for SNP effects across the genome to induce sparsity. We'll use mean field variational methods to fit the posterior distribution for SNP effects. With the posterior estimates of SNP effects, we'll construct polygenic risk scores as weighted sums of SNP minor allele counts, where the weights are the posterior SNP effect estimates.

Aim 2 will develop a Bayesian statistical model for polygenic risk scores based on SNP effect estimates and estimates for SNP-SNP interactions. We'll use a spike-and-slab prior to induce sparsity of both SNP effects and SNP-SNP interactions. We'll fit our Bayesian model by using approximations from mean field variational methods. We'll write open-source computer code to implement our method, then assess its statistical performance with simulated traits.

Our proposed studies will establish mean field variational methods as a computationally efficient and accurate method for posterior inference in biobank-scale genetics studies, an approach that current methods lack. Furthermore, our new methods will enable the modeling of gene-gene interactions to improve prediction accuracy. Our open-source, well documented, and thoroughly tested software will provide the genetics research community with a valuable resource and enable future advances in statistical methods for polygenic risk score construction. Our studies will improve accuracy of ongoing public health efforts to identify people at high risk for deadly diseases and will enable early, preventive interventions to improve the lives of millions around the world.