Specific Aims

Many common diseases are genetically complex, meaning that the inheritence of the disease does not follow simple Mendelian patterns. In the past decade, an enormous amount of resources have been directed toward identifying the mutations responsible for variation in disease risk within the human population, and toward understanding the biology of complex genetic disease more broadly. Genome wide association studies (GWAS) have shown that for many such diseases, the number of mutations which contribute to disease risk is likley to be truly vast. Typically, the effect of any individual mutation on disease risk is quite small (e.g. odds ratios < 1.5), but a growing body of evidence indicates that the collective contributions of these many small effect loci can account for the bulk of the heritability of common complex disease.

A recurring motif in the last quarter century of research in disease genetics has been the propensity for emerging data to upend existing intuition in the field, and there is still little concensus in explaining why the prevalence of disease or the number, effect sizes, and frequencies of disease predisposing mutations (i.e. the "genetic architecture") take the values they do. As the pace of data production accelerates and we gain unprecedented insights into the identities and properties of the mutations which drive disease risk, we will need principled and quantitative models in order to address these questions. Ultimately, *the genetic architecture and the population prevalence of complex disease are the result of an interplay between internal biological forces, such as the mutation rate, the distribution of mutational effects, and the nature of the pleiotropy, and external population level forces, such as natural selection, population size changes, or variation in diet and lifestyle.* An understanding of how the interplay between these forces shapes the genetic architecture and prevalence of disease requires a population genetic treatment, and while the evolution of deleterious (i.e. disease pre-disposing) alleles has long been of interest in population genetics, most work to date has focused either on Mendelian diseases or on the behavior of individual disease alleles in a vaccum, without considering their connection to the broader disease phenotype.

I therefore propose to study how population genetic processes are responsible for shaping the genetic architecture and prevalence of complex disease. To this end, I will develop new theory which will include population genetic models of disease which include an explicit model for how selection on the disease and other pleiotropic phenotypes influence architecture and prevalence. I will also develop a principled likelihood based inference framework to infer the parameters that govern these features. A user friendly software package implementing this inference approach will be made publicly available.

Specific Aim 1: Models Relating Population Genetic Processes with the Architecture of Complex Diseases
I will develop a set of flexible population genetic models which describe how features such as the mutational target size and distribution of effects, the nature of pleiotropy, and the fitness cost of the disease (among others) contribute to shape disease architecture and prevalence. Specifically, this modeling work will demonstrate how observable quantities such as the disease prevalence, number of mutations discovered in GWAS and the joint distribution of their effect sizes and allele frequencies depend on the population genetic parameters. These models will represent the first generative models for complex disease which include a biologically grounded parameterization of pleiotropy, and will also consider the impact of factors such as recent environmental change and changes in population size over the course of human evolution.

Specific Aim 2: Inference of Model Parameters from GWAS of Complex Diseases   Using the models developed in Aim 1, I will develop a likelihood based approach to estimate the evolutionary parameters governing disease susceptibility, such as the mutational target size, the fitness cost of the disease, and the extent to which selection on variants affecting disease risk arises from their pleiotropic effects on other traits. I will begin with methods that rely on the frequencies and effect sizes of genome wide significant associations from GWAS, but will also explore the use of information about how genetic variance is distributed among alleles with effect too small to be detected in GWAS (which can be inferred by other means). I will apply the inference machinery to datasets from GWAS for at least 10 complex disease. In addition to estimating parameters based for a given model, I will compare the fit among models, in order to learn, for example, whether recent changes to the human environment have impacted the architecture and/or prevalence of a given disease, or whether stabilizing selection on pleiotropically related quantitative traits plays a role in shaping the architecture. The inference tools I develop will be made available in a user-friendly software package.