Abstract   One of the central goals of modern human genetics is to understand why complex genetic diseases are as prevalent as they are, and why genetic risk is distributed among individuals and across the genome in the way that it is. Over the past decade, genome wide association studies (GWAS) have generated a deluge of information about the mutations that underlie the variation in susceptibility for complex disease. These findings show that for many diseases, variation in susceptibility arises from many hundreds or even thousands of variants, many of which segregate at appreciable frequencies in the population but have vanishingly small penetrance. Yet we still lack a good understanding of why there is so much genetic variation affecting the susceptibility to diseases that often involve a severe fitness cost, and what shapes this genetic variation (e.g., the distribution of variant frequencies and effect sizes)

Despite the basic and practical importance of these question, there has been surprisingly little work aimed at answering them, and specifically at understanding how population genetics processes give rise to the genetic basis of disease susceptibility being uncovered by GWAS.

The goal of the proposed research is to fill this gap. The first aim is to develop models describing how the genetic architecture and the population prevalence of complex disease results from an interplay between internal biological forces, such as the mutation rate, the distribution of mutational effects on the disease, and on other traits, and external population level forces, such as natural selection, population size changes, or variation in diet and lifestyle. The second aim is to develop a likelihood based statistical framework for inferring the parameters corresponding to these factors from the results of GWAS, and applying the inference to data for at least 10 complex disease in order to learn about the processes and parameters that shape their genetic architecture and determine their prevalence. An open access and well documented software package implementing the statistical inference will be made freely available to the research community. The proposed models and statistical inferences will the first to address these questions based on a principled biological model of disease, and are expected to substantially advance our understanding of the processes that shape complex disease susceptibility in humans.