Unified Statistical Methods for Sequence-based Association Studies

Abstract

Fast and economic next generation sequencing (NGS) technologies will generate unprecedentedly massive (thousands of individuals) and high-dimensional (ten millions) genomic and epigenomic variation data that allow nearly complete evaluation of genomic and epigenomic variation including common and rare variants, RNA-seq, mRNA-seq and methylation-seq data. As a consequence, these genomic variation data are so densely distributed across the genome that the genetic variants can be considered as genomic variation observations varying over a continuum. The emergence of NGS technologies is not only changing our view of genomics from independently segregating discrete model to hybrid (both discrete and continuous) models, but also causing great changing in analytic methods for genomic and epigenomic analysis from standard multivariate data analysis to functional data analysis, from independent sampling to dependent sampling, from low dimensional data analysis to high dimensional data analysis, from single genomic or epigenomic variant analysis to integrated genomic and epigenomic analysis. To address the great challenges we are facing in NGS data analysis, the goals of this proposal are to develop novel and powerful statistical methods for sequence-based association studies and QTL (eQTL) analysis which leverage high dimensional data reduction, and functional data analysis techniques to identify both common and rare risk variants across the genome, investigate their function via intermediate phenotypes and expressions, estimate the total effects (intervention effects) and direct effects of variants on the phenotypes, and unify family and population-based designs using de-identified data.