Deep learning prediction of chemotherapy response and DNA methylation using histological image data

Fast and economic next generation sequencing (NGS) technologies have generated massive, high-dimensional genomic and epigenetic variation data that allow nearly complete evaluation of genomic and epigenetic variation including common and rare variants, miRNA-seq and mRNA-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 causing great changes in analytic methods for genomic 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 association analysis to causal inference, and from single genomic variant analysis to integrated systems biology approaches. 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, causal inference and functional data analysis techniques to identify both common and rare risk variants across the genome associated with disease, investigate their functional and clinical consequences via intermediate phenotypes and gene expression measured by RNA-seq, estimate the total (intervention effects) and direct effects of variants on the phenotypes, and unify family and population-based designs. In summary, this application will provide novel strategies and tools for unraveling connection between genetic variants, gene expressions, intermediate phenotypes and diseases. The specific aims of the proposal are as follows:

**Aim 1: Develop a general framework and innovative statistical methods for testing the association of both common and rare variants which unify family and population-based designs and integrate multiple variants across a given genomic region defined by a gene or other regulation unit.** Specifically, by theoretically deriving new formulas for calculation of the covariance matrix for dependent samples and quasi-likelihood score functional approaches, we develop a general framework and novel statistical methods for association studies with data sampled from pedigrees with any complex structure and unrelated individuals from any structured populations, and a universal procedure that can transform any population-based association tests to family-based association tests.

**Aim 2: Develop novel statistical methods that can be applied to a broad range of population-based and family-based QTL and eQTL analysis including longitudinal QTL analysis and eQTL analysis with RNA-seq data.** Two approaches will be used for QTL (eQTL) analysis. First approach is to develop a novel functional mixed-effect model with scalar and functional response, and functional predictors for family and population-based (adjusting for population stratification) QTL (eQTL) analysis. An alternative approach is to use causal inference for identifying risk variants, and intervention calculus coupled with graphical model and inverse regression coupled with kernel smoothing, respectively, for estimating their total (intervention) and direct effects on the phenotypes while avoiding confounding.

**Aim 3: Develop novel statistical methods for sequence-based pathway analysis in both population and family study designs.** Specifically, we will develop novel genome-information content-based and multivariate functional principal component analysis-based statistics for testing the association of pathways with the disease, including the entire allelic spectrum of genetic variants in the pathways and allowing the data sampled from both families and population in the presence or absence of population substructure. In addition, we will distribute publicly accessible software for all methods developed in the course of this proposal.

Bioinformatics Pipeline and Database

Coding regions were obtained based on GENCODE19 gene annotation model obtained from ftp://ftp.sanger.ac.uk/pub/gencode/Gencode\_human/release\_19/gencode.v19.annotation.gtf.gz

Protein\_coding transcripts that had an annotated START and STOP were used.