Gerstein lab has considerable expertise in developing standardized pipelines and quality control metrics for RNA-Seq and evaluating them in many consortia including ENCODE, exRNA, KBase et al. The lab is very experienced in developing and setting up pipelines for the processing of RNA-seq data; specially for long RNA-seq data for ENCODE, long and short RNA-seq data for the Brainspan project as well as a custom pipeline developed for the analysis of small exRNA-seq data for the Extracellular RNA Communication Consortium (ERCC). We have already developed an efficient in-house data processing workflow for RNA-seq data that includes data organization, format conversion, and quality assessment. Specifically, we will use RSEQtools[1] to quantify expression profiles of each type of annotation entry retrieved from the latest release of the GENCODE project. RSeqTools is a modular tool the Gerstein lab developed for the processing of RNA-seq data and generating either transcript, gene or exon level quantifications; RSeqTools introduced a data format, MRF, that allows for the exclusion of the actually sequences from mapped RNA-seq data for the distribution of restricted access data.

Our lab has also developed IQSeq [2]which calculates the relative and absolute abundance of contributing transcript isoforms to a gene from RNA-seq data using a fast algorithm based on the FIsher information matrix. Another tool we developed called FusionSeq[3] was to detect fusion transcript in RNA-seq data, which can be important biomarker for diseases such as various types of cancer and mental diseases.

We have past experience in non-coding annotation, as part of our 10-year history with the ENCODE and modENCODE projects. As part of these projects, we developed methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers, which we have partially validated. We will use the better enhancer definition provided by the Epigenome Roadmap and more recently from ENCODE projects.

We have extensive experience in performing large scale integrative analysis in various consortia like ENCODE, modENCODE, 1000 Genomes, KBase and Brainspan. First, using the machine learning approaches we developed method for identifying individual proximal and distal edges together with miRNA target prediction (and other) algorithms, we have completed the highly ambitious goal of constructing highly integrated regulatory networks for humans and model organisms based on the ENCODE and modENCODE datasets. These integrated networks consist of three major types of regulation: TF-gene, TF-miRNA and miRNA-gene, showing rich statistical patterns. For instance, the human regulatory network uniquely displays distinct preferences for binding at proximal and distal regions. The distal binding preference is possibly due to the intergenic space in the human genome, which is much larger relative to the genomes of other model organisms. More recently, we have constructed co-expression networks from the extensive amount of RNA-Seq data generated by ENCODE and modENCODE consortia. [4]We have developed a novel cross-species multilayer network framework, OrthoClust, for analyzing the co-expression networks in an integrated fashion by utilizing the orthology relationships of genes between species. [5]OrthoClust revealed conserved modules across human, worm and fly that are important for development. We also introduced a framework to quantify differences between networks and by comparing matching networks across organisms, found a consistent ordering of rewiring rates of different network types.[6] We have extensive experience in using network framework to integrate data of human variation. We have developed NetSNP [7], an approach to quantify indispensability of each gene in the genome by incorporating multiple network and evolutionary properties. Based on network properties, as well as many other genomics features, we have developed FunSeq, and more recently FunSeq2[8] for prioritizing variants. Using 1000 genomes variants, our pipeline has demonstrated great potential in prioritizing mutations in non-coding regions that are related to cancer.

1. Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, Snyder M, Gerstein M: **RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries**. *Bioinformatics* 2011, **27**(2):281-283.
2. Du J, Leng J, Habegger L, Sboner A, McDermott D, Gerstein M: **IQSeq: integrated isoform quantification analysis based on next-generation sequencing**. *PloS one* 2012, **7**(1):e29175.
3. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A *et al*: **Integrative annotation of variants from 1092 humans: application to cancer genomics**. *Science* 2013, **342**(6154):1235587.
4. Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ *et al*: **Comparative analysis of the transcriptome across distant species**. *Nature* 2014, **512**(7515):445-448.
5. Yan KK, Wang D, Rozowsky J, Zheng H, Cheng C, Gerstein M: **OrthoClust: an orthology-based network framework for clustering data across multiple species**. *Genome biology* 2014, **15**(8):R100.
6. Shou C, Bhardwaj N, Lam HY, Yan KK, Kim PM, Snyder M, Gerstein MB: **Measuring the evolutionary rewiring of biological networks**. *PLoS computational biology* 2011, **7**(1):e1001050.
7. Khurana E, Fu Y, Chen J, Gerstein M: **Interpretation of genomic variants using a unified biological network approach**. *PLoS computational biology* 2013, **9**(3):e1002886.
8. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E, Gerstein M: **FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer**. *Genome biology* 2014, **15**(10):480.