Gerstein lab have considerable experience in ENCODE, modENCODE, 1000 Genomes and KBase in doing large scale cross project integrative and comparative analysis. We developed the standardized RNA-seq processing pipelines including data organization, format conversion, and quality assessment which will then be run in large-scale on the PDC (Protected Data Cloud) to process the RNA-Seq data in different projects first. 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 employ Tophat [1] to uniquely align the filtered reads to their reference genome and RSEQtools [2] to quantify expression profiles of each type of annotation entry retrieved from the latest release of the GENCODE project. RSeqTools [2] 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. Additional quality control measures was introduced to assess potential issues including sequencing error rate, ribosomal contamination, DNA contamination and gene coverage uniformity and the correlation between technical and/or biological replicates.

We have developed a novel cross-species multi-layer network framework, OrthoClust, for analyzing the co-expression networks in an integrated fashion by utilizing the orthology relationships of genes between species [3]

We have developed chipseq data processing pipeline in Gerstein lab to uniformly process data in ENCODE, Brainspan and Epigenomics Roadmap. This pipeline includes steps of quality assessment, trimming the contamination, alignment of the fastq files, peak calling and downstream analysis such as peak comparison, peak annotation, motif analysis and super-enhancers identification. In this pipeline, we added a new peak caller MUSIC developed in Gerstein lab. MUSIC performs multiscale decomposition of ChIP signals to enable simultaneous and accurate detection of enrichment at a range of narrow and broad peak breadths. This tool is particularly applicable to studies of histone modifications and previously uncharacterized transcription factors, both of which may display both broad and punctate regions of enrichment.

Moreover, we developed methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers based on our past experience in non-coding annotation, as part of our 10-year history with the ENCODE and modENCODE projects [6]. We use the better enhancer definition provided by the Epigenome Roadmap [7-9], and more recently from ENCODE projects.

1 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111, doi:btp120 [pii]
10.1093/bioinformatics/btp120 (2009).

2 Habegger, L. *et al.* RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* **27**, 281-283, doi:btq643 [pii]

10.1093/bioinformatics/btq643 (2011).
3 Yan, K. K. *et al.* OrthoClust: an orthology-based network framework for clustering data across multiple species. *Genome Biol* **15**, R100, doi:10.1186/gb-2014-15-8-r100 (2014).
6 Yip, K. Y., Alexander, R. P., Yan, K. K. & Gerstein, M. Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PLoS One* **5**, e8121, doi:10.1371/journal.pone.0008121 (2010).
7 Ziller, M. J. *et al.* Dissecting neural differentiation regulatory networks through epigenetic footprinting. *Nature* **518**, 355-359, doi:10.1038/nature13990 (2015).
8 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).
9 Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**, 350-354, doi:10.1038/nature14217 (2015).