Gerstein lab has considerable experience in ENCODE, modENCODE, 1000 Genomes and KBase in doing large scale cross project integrative and comparative analysis. In particular, 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 (1).

We implemented the standardized RNA-seq processing pipelines including data organization, format conversion, and quality control metrics to process the RNA-Seq data first. Specifically, we employ STAR (2) to align the reads to their reference genome and RSEM (3) to quantify expression profiles of each type of annotation entry retrieved from the latest release of the GENCODE project. Additional quality control measures were introduced to assess potential issues including sequencing error rate, ribosomal contamination and DNA contamination.

We also developed ENCODE ChIP-seq data processing pipeline together with Zhiping Weng's lab. 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. The Gerstein lab developed PeakSeq(4), a versatile tool for identification of TF binding sites and a standard peak calling program used by the ENCODE and modENCODE consortia for ChIP-Seq. We will also use a new peak caller MUSIC (5)recently 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.

We implemented a standard genotype imputation pipeline. Genotype imputation enables us to evaluate the evidence for association at genetic markers that are not directly genotyped and increases the power of eQTL analysis. Moreover, genotype imputation is very important for combining data from studies using different genotyping platforms. Firstly, for the Sample level quality control, we exam the call rate, heterozygosity and relatedness between genotyped individuals correspondence between sex chromosome genotypes and reported gender of the raw genotype calling using PLINK(6). Then we will perform ancestry analysis on the QCed genotype data to identify the ancestry vectors. In order to improve the genotype imputation accuracy, SHAPEIT2 (7) will be used to estimate haplotypes from genotype data. The estimated haplotypes will be used as input for IMPUTE2 for imputation using the selected reference panel. We will also use both 1000 Genome phase 1 or the recently released HRC Reference Panel for imputation on Michigan Imputation Server. The imputed genotypes will be filtered according to imputation confidence score (INFO), minor allele frequency (MAF), SNP missing rate and Hardy-Weinberg Equilibrium (HWE).

References

1. Yan K-K, 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. doi: 10.1186/gb-2014-15-8-r100.

2. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. doi: 10.1093/bioinformatics/bts635. PubMed PMID: 23104886; PMCID: PMC3530905.
3. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323. doi: 10.1186/1471-2105-12-323. PubMed PMID: 21816040; PMCID: PMC3163565.
4. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 2009;27(1):66-75. doi: 10.1038/nbt.1518. PubMed PMID: 19122651; PMCID: PMC2924752.
5. Harmanci A, Rozowsky J, Gerstein M. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. Genome Biol. 2014;15(10):474. doi: 10.1186/s13059-014-0474-3. PubMed PMID: 25292436; PMCID: PMC4234855.
6. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75. doi: 10.1086/519795. PubMed PMID: 17701901; PMCID: PMC1950838.
7. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I, McQuillan R, Fraser RM, Campbell H, Polasek O, Asiki G, Ekoru K, Hayward C, Wright AF, Vitart V, Navarro P, Zagury JF, Wilson JF, Toniolo D, Gasparini P, Soranzo N, Sandhu MS, Marchini J. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 2014;10(4):e1004234. doi: 10.1371/journal.pgen.1004234. PubMed PMID: 24743097; PMCID: PMC3990520.