Gerstein lab has considerable expertise in developing these standardized pipelines and evaluating them in many consortia including ENCODE, exRNA, KBase. We have developed tools like RSeqTools [1], IQSeq [2], FusionSeq [3] for the processing of RNA-seq data, and PeakSeq [4], which was the first peak caller to process ChIP-seq data relative to a correctly normalized input DNA control as well as accounting for variability in genome-wide sequence mappability, and MUSIC [5] for processing ChIP-seq data. The lab also played an important role in developing the so-called IDR method for determining reproducibility of target lists identified from replicate ChIP-seq experiments in order to correctly set thresholds uniformly across different ChIP-seq experiments for different TFs across different labs. These and other standards for performing and analyzing ChIP-seq experiments were published in Landt et al. 2012 [6].

Gerstein lab also have experience on setting the standards within ENCODE and other consortia. For example, capitalizing on the uniformly processed and matched experimental data obtained by mod/ENCODE consortia, we have performed a series of comparative studies across distant metazoan phyla. A comparative analysis of human, worm, and fly revealed remarkable conservation of general properties of regulatory networks [7]. Also, as part of the GENCODE project we carried out a comprehensive annotation of pseudogenes, which was further integrated with ENCODE and 1000 Genomes Project data. All the information was stored in an online resource called psiDR [8].

The Gerstein lab has a tremendous amount of experience in developing large databases of QTLs and allelic sites. AlleleSeq [9] is a tool developed specifically for the detection of allelic sites, including those associated with gene expression and transcription factor binding using RNA-seq and ChIP-seq datasets. AlleleSeq has been applied in several publications [10-12]. Notably, we have previously used AlleleSeq in allele-specific analyses associated with gene expression using ENCODE RNA-seq datasets from a single cell line [11]. Recently, we have further developed AlleleSeq and applied the new version to 1,139 RNA-seq and ChIP-seq datasets for 382 cell lines found in the 1000 Genomes Project. We harmonized and aggregated multiple RNA-seq and ChIP-seq datasets separately for each cell line and uniformly reprocessed them using the updated AlleleSeq. This allowed us to annotate the 1000 Genomes Project SNP catalog with allelic information. We constructed a database, AlleleDB, to house all the results. The database can be queried for specific genomic regions and visualized as a track in the UCSC browser [13] and visualizer such as the Integrated Genomics Viewer [14] or downloaded as flat files for downstream analyses for users that are more advanced in bioinformatics training. We continue to maintain and update AlleleSeq as a publicly available resource. It has been utilized considerably by the scientific community, as indicated by the number of citations and publications using our data and tool.

The Gerstein and Sestan labs have rich experience in the mapping of RNA-seq data and the construction of transcripts form human brain and non-brain tissues. For example, RSEQtools is a computational package that enables expression quantification of annotated RNAs and identification of splice sites and gene models. IQseq provides a computationally efficient method to quantify isoforms for alternatively spliced transcripts. Both of these tools employ a special sequence read format we developed that can dissociate genome sequence information from

RNA-Seq signal, maintaining the privacy of test subjects. FusionSeq is designed for detecting fusion transcripts generated from either trans-splicing or genomic translocations in paired-end RNA-sequencing.

We have developed such as RSEQtools, IQSeq and FusionSeq. The combination of these three tools allows us to efficiently discover brain specific spliced transcripts and their potential functions.

We have much experience with developing enhancer calling pipelines in the framework of ENCODE, which we will utilize here. For example, we have applied machine-learning methods that integrate multiple genomics features to classify human regulatory regions from ENCODE data of more than 100 transcription factor binding sites. A computational pipeline was developed to identify potential enhancers from regions classified as gene-distal regulatory modules [18]. Making use of the potential enhancers, we developed the Function-based Prioritization of Sequence Variants (FunSeq) tool [12] for identification of candidate drivers in tumor genomes, and more recently, a more elaborate and flexible framework, FunSeq2, integrating various genomic and cancer resources to prioritize cancer somatic variants, especially regulatory noncoding mutations [19].

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 [10] and modENCODE datasets [21,22]. 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 [23]. 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 [24]. 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 [25].

We have extensive experience in using network framework to integrate data of human variation. We have developed NetSNP [26], 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 [12], and more recently FunSeq2 [19] 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 [12].

References

1. Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, et al. (2011) RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. Bioinformatics 27: 281-283.
2. Du J, Leng J, Habegger L, Sboner A, McDermott D, et al. (2012) IQSeq: integrated isoform quantification analysis based on next-generation sequencing. PLoS One 7: e29175.
3. Sboner A, Habegger L, Pflueger D, Terry S, Chen DZ, et al. (2010) FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol 11: R104.
4. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, et al. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27: 66-75.
5. Harmanci A, Rozowsky J, Gerstein M (2014) MUSIC: Identification of Enriched Regions in ChIP-Seq Experiments using a Mappability-Corrected Multiscale Signal Processing Framework. Genome Biol 15: 474.
6. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22: 1813-1831.
7. Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, et al. (2014) Comparative analysis of regulatory information and circuits across distant species. Nature 512: 453-456.
8. Pei B, Sisu C, Frankish A, Howald C, Habegger L, et al. (2012) The GENCODE pseudogene resource. Genome Biol 13: R51.
9. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, et al. (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol 7: 522.
10. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, et al. (2012) Architecture of the human regulatory network derived from ENCODE data. Nature 489: 91-100.
11. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, et al. (2012) Landscape of transcription in human cells. Nature 489: 101-108.
12. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, et al. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342: 1235587.
13. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.
14. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, et al. (2011) Integrative genomics viewer. Nat Biotechnol 29: 24-26.
15. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105-1111.
16. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, et al. (2006) GENCODE: producing a reference annotation for ENCODE. Genome Biol 7 Suppl 1: S4 1-9.
17. Li G, Bahn JH, Lee JH, Peng G, Chen Z, et al. (2012) Identification of allele-specific alternative mRNA processing via transcriptome sequencing. Nucleic Acids Res 40: e104.
18. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, et al. (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol 13: R48.

19. Fu Y, Liu Z, Lou S, Bedford J, Mu X, et al. (2014) FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. Genome Biol 15: 480.
20. He B, Chen C, Teng L, Tan K (2014) Global view of enhancer-promoter interactome in human cells. Proc Natl Acad Sci U S A 111: E2191-2199.
21. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, et al. (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science 330: 1775-1787.
22. Negre N, Brown CD, Ma L, Bristow CA, Miller SW, et al. (2011) A cis-regulatory map of the Drosophila genome. Nature 471: 527-531.
23. Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, et al. (2014) Comparative analysis of the transcriptome across distant species. Nature 512: 445-448.
24. Yan KK, Wang D, Rozowsky J, Zheng H, Cheng C, et al. (2014) OrthoClust: an orthology-based network framework for clustering data across multiple species. Genome Biol 15: R100.
25. Shou C, Bhardwaj N, Lam HY, Yan KK, Kim PM, et al. (2011) Measuring the evolutionary rewiring of biological networks. PLoS Comput Biol 7: e1001050.
26. Khurana E, Fu Y, Chen J, Gerstein M (2013) Interpretation of genomic variants using a unified biological network approach. PLoS Comput Biol 9: e1002886.

27. Willsey AJ, Sanders SJ, Li M, Dong S, Tebbenkamp AT, et al. (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell 155:997-1007.

28: Cotney J, Muhle RA, Sanders SJ, Liu L, Willsey AJ, et al., (2015). The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. Nat Commun 6:6404.

29. Liu L, Lei J, Sanders SJ, Willsey AJ, Kou Y, et al. (2014). DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. Mol Autism 5:22.

30. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, et al., (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485: 237-241.

31. Zhu Y, Li M, Sousa AM, Sestan N (2014) XSAnno: a framework for building ortholog models in cross-species transcriptome comparisons. BMC Genomics15:343.