The Gerstein lab has developed many methods to unravel gene regulatory networks in a variety of contexts. In the following is a sequential presentation of analyses developed by the lab:

Finding enhancers. Over the course of work in the ENCODE and modENCODE projects since 2003 [2, 3], the Gerstein lab gained extensive experience in annotating non-coding DNA. We have developed machine learning methods to integrate signals for histone modifications, DNA methylation, chromatin accessibility, sequence conservation, sequence motifs, and gene annotations to identify enhancers, including those that are distal to their target genes. We have also built robust computational pipelines for processing massive amounts of data and identifying enhancers, transcription factor binding sites, and regulatory modules [4].

Building networks. We have previously contributed a large body of work on regulatory networks. Via data integration, we have constructed gene networks of various regulators, including transcription factors (TF) and micro-RNAs (mRNA) and their target genes [6-12]. Upon analyzing the structures of these networks, we found that, relative to centrality, hierarchy levels are better predictors of regulator importance [6, 13-16]. Thus, we developed a general-purpose algorithm to measure the hierarchical structure of any type of regulatory network [17]. Our network analysis software tools include TopNet [18], tYNA[19], and PubNet [20]. In addition to the global attributes of regulatory networks (such as their hierarchy), we also analyzed local topological features, such as network motifs (e.g., feed-forward loops) [6, 9, 12]. We further integrated regulatory networks with gene expression to uncover functional modules [21-24]. We integrated ENCODE data on TF binding, histone modifications, and target gene expression to establish regulatory relationships using a probabilistic model named TIP [25]. Identifying potential enhancers from gene-distal regions, we used these modules to characterize the associations between TF binding and gene expression [26-29]. We further integrated these data with protein-protein interaction and transcriptional regulation networks [8, 9, 30, 31]. This enabled us to separate TFs into histone-sensitive and -insensitive classes, which refined the prediction of target gene expression levels. To analyze multiple interconnected networks simultaneously, we constructed co-expression networks from the extensive RNA-seq data in various consortia [3]. We further developed a novel framework consisting of a cross-species multi-layer network (OrthoClust) to analyze co-expression networks in an integrated fashion using orthologous genes across species [29].

Developing approaches for relating regulatory networks to human genomic variation. We have extensive experience in identifying expression quantitative loci (eQTL) and allelic sites. In particular, we have developed the AlleleSeq method, which uses RNA-seq and ChIP-seq data to detect allelic sites, including those associated with gene expression and TF binding [41]. Furthermore, AlleleSeq constructs personal diploid genomes. Using AlleleSeq, we have spearheaded allele-specific analyses as part of our efforts in several major consortia, including ENCODE and the 1000 Genomes Project[3, 12, 42]. We have further developed AlleleSeq and applied the new version to 1,139 RNA-seq and ChIP-seq datasets for 382 samples in the 1000 Genomes Project, which enabled us to annotate the 1000 Genomes Project SNP catalog with allelic information. We constructed a database (AlleleDB) to house all the results as a resource. Both AlleleSeq and AlleleDB have are widely used by the scientific community. Recently, we also developed PrivaSeq, a tool to quantify how much individual-characterizing information is leaked by eQTLs[43].

Integrative modeling. Based on machine learning and network approaches, we have developed various integrated methods to model gene regulatory mechanisms. For example, we applied statistical models to characterize the relationships between the extent of TF binding and gene expression by integrating ChIP-seq and RNA-seq data [45]. Recently, we developed DREISS, a method to integrate a state-space model with dimensionality reduction using matrix factorization to identify the temporal expression patterns for various biological processes, such as the oscillation and degradation expression patterns during the cell cycle, embryonic development, and cancer progression [46]. We have also developed Loregic, a method to characterize the gene regulatory logics in complex systems [17]. We used Loregic to identify the cooperative logic among TFs binding to promoters and enhancers in leukemia by integrating ENCODE and TCGA data. We also have extensive experience in using the network framework to integrate human variation data. Our NetSNP method quantifies the indispensability of each gene by incorporating multiple network and evolutionary properties. Based on network properties and other genomic features, we have developed FunSeq [47] and

FunSeq2 for prioritizing somatic variants. Using 1000 genomes data, we have prioritized mutations in non-coding regions that may cause diseases [47].

Peak-calling methods. The Gerstein Lab has developed two peak calling algorithms, PeakSeq [49] and MUSIC[5] . PeakSeq calls the peaks for transcription factor ChIP-seq data and is used by the ENCODE and modENCODE consortia. MUSIC performs multi-scale decomposition of ChIP signals to enable simultaneous and accurate detection of enrichment at a wide range of peak breadths. MUSIC is particularly applicable to histone modifications and some transcription factors that display both punctate and broad regions of enrichment.

Integrating data from other consortia. We have extensive experience in performing large-scale integrative analyses**.** We have played key or lead roles in the DOE KBase, Brainspan, ENCODE, modENCODE, 1000 Genomes, PCAWG, and exRNA consortia. We work in multi-disciplinary teams and interact with scientists and physicians of highly diverse backgrounds within these consortia. We have applied simulation, machine learning, and knowledgebase design for working with multi-layered datasets.

References

2. Yip, K.Y., et al., *Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data.* PLoS One, 2010. **5**(1): p. e8121.
3. Gerstein, M.B., et al., *Comparative analysis of the transcriptome across distant species.* Nature, 2014. **512**(7515): p. 445-8.
4. Yip, K.Y., et al., *Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors.* Genome Biol, 2012. **13**(9): p. R48.
5. Harmanci, A., J. Rozowsky, and M. Gerstein, *MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework.* Genome Biol, 2014. **15**(10): p. 474.
6. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.
7. Negre, N., et al., *A cis-regulatory map of the Drosophila genome.* Nature, 2011. **471**(7339): p. 527-31.
8. Cheng, C., et al., *Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors.* Genome Biology, 2011. **12**(11): p. R111.
9. Gerstein, M.B., et al., *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* Science, 2010. **330**(6012): p. 1775-87.
10. Yan, K.K., et al., *Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks.* Proc Natl Acad Sci U S A, 2010. **107**(20): p. 9186-91.
11. Cheng, C., et al., *Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data.* PLoS Comput Biol, 2011. **7**(11): p. e1002190.
12. Boyle, A.P., et al., *Comparative analysis of regulatory information and circuits across distant species.* Nature, 2014. **512**(7515): p. 453-6.
13. Yu, H. and M. Gerstein, *Genomic analysis of the hierarchical structure of regulatory networks.* Proc Natl Acad Sci U S A, 2006. **103**(40): p. 14724-31.
14. Bhardwaj, N., P.M. Kim, and M.B. Gerstein, *Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators.* Sci Signal, 2010. **3**(146): p. ra79.
15. Bhardwaj, N., et al., *Analysis of combinatorial regulation: scaling of partnerships between regulators with the number of governed targets.* PLoS Comput Biol, 2010. **6**(5): p. e1000755.
16. Bhardwaj, N., K.K. Yan, and M.B. Gerstein, *Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels.* Proc Natl Acad Sci U S A, 2010. **107**(15): p. 6841-6.
17. Wang, D., et al., *Loregic: a method to characterize the cooperative logic of regulatory factors.* PLoS Comput Biol, 2015. **11**(4): p. e1004132.

18. Yu, H., et al., *TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics.* Nucleic Acids Res, 2004. **32**(1): p. 328-37.
19. Yip, K.Y., et al., *The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks.* Bioinformatics, 2006. **22**(23): p. 2968-70.
20. Douglas, S.M., G.T. Montelione, and M. Gerstein, *PubNet: a flexible system for visualizing literature derived networks.* Genome Biol, 2005. **6**(9): p. R80.
21. Luscombe, N.M., et al., *Genomic analysis of regulatory network dynamics reveals large topological changes.* Nature, 2004. **431**(7006): p. 308-12.
22. Qian, J., et al., *Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data.* Bioinformatics, 2003. **19**(15): p. 1917-26.
23. Yu, H., et al., *Genomic analysis of gene expression relationships in transcriptional regulatory networks.* Trends Genet, 2003. **19**(8): p. 422-7.
24. Cheng, C., et al., *mRNA expression profiles show differential regulatory effects of microRNAs between estrogen receptor-positive and estrogen receptor-negative breast cancer.* Genome Biol, 2009. **10**(9): p. R90.
25. Cheng, C., R. Min, and M. Gerstein, *TIP: A probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles.* Bioinformatics, 2011. **27**(23): p. 3221-3227.
26. Cheng, C., et al., *Understanding transcriptional regulation by integrative analysis of transcription factor binding data.* Genome Research, 2012. **22**(9): p. 1658-1667.
27. Cheng, C. and M. Gerstein, *Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells.* Nucleic Acids Research, 2011. **40**(2): p. 553-568.
28. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
29. Yan, K.K., et al., *OrthoClust: an orthology-based network framework for clustering data across multiple species.* Genome Biol, 2014. **15**(8): p. R100.
30. Cheng, C., et al., *A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets.* Genome Biology, 2011. **12**(2): p. R15.
31. Dong, X., et al., *Modeling gene expression using chromatin features in various cellular contexts.* Genome Biology, 2012. **13**(9): p. R53.
41. Rozowsky, J., et al., *AlleleSeq: analysis of allele-specific expression and binding in a network framework.* Mol Syst Biol, 2011. **7**: p. 522.
42. Chen, J., et al., *A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals.* Nat Commun, 2016. **7**: p. 11101.
43. Harmanci, A. and M. Gerstein, *Quantification of private information leakage from phenotype-genotype data: linking attacks.* Nat Methods, 2016. **13**(3): p. 251-6.
44. van de Geijn, B., et al., *WASP: allele-specific software for robust molecular quantitative trait locus discovery.* Nat Methods, 2015. **12**(11): p. 1061-3.
45. Cheng, C., et al., *Understanding transcriptional regulation by integrative analysis of transcription factor binding data.* Genome Res, 2012. **22**(9): p. 1658-67.
46. Wang, D., et al., *DREISS: Using State-Space Models to Infer the Dynamics of Gene Expression Driven by External and Internal Regulatory Networks.* PLoS Comput Biol, 2016. **12**(10): p. e1005146.
47. Khurana, E., et al., *Integrative annotation of variants from 1092 humans: application to cancer genomics.* Science, 2013. **342**(6154): p. 1235587.
49. Rozowsky, J., et al., *PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.* Nat Biotechnol, 2009. **27**(1): p. 66-75.