We have ENCODE ChIP-seq data processing pipeline developed by both Gerstein lab and 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 (1), 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 datasets (1). We also developed a new peak caller MUSIC (2) 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 have already implemented this pipeline to process ChIP-Seq data from both PsychENCODE and BrainSpan. Moreover, we have developed methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers based on our experience in non-coding annotation, as part of our 10-year history with the ENCODE and modENCODE projects (3).

We have also implemented a standard eQTL analysis pipeline in Gerstein lab for PsychENCODE and genomic privacy paper (4). We use Matrix eQTL and/or fastQTL package for eQTL analysis.

We have substantial experience in developing computational approaches to identify specific dynamic patterns of gene expression. We have developed a novel clustering algorithm, OrthoClust to simultaneously cluster multi-layer networks (5). We applied OrthoClust to developmental gene expression datasets of worm (C. elegans) and fruitfly (D. melanogaster), and discovered the cross-species and species-specific gene co-expression modules (Figure 1). We also found the modular eigengenes, revealing the systematically gene expression and regulation dynamics during embryonic development. In 2016, we also developed another novel computational method, DREISS to identify the gene expression dynamics driven by internal and external regulatory networks (6). We applied DREISS to the time-series gene expression datasets of C. elegans and D. melanogaster during their embryonic development (Figure 2). We analyzed the expression dynamics of the conserved, orthologous genes (orthologs), seeing the degree to which these can be accounted for by orthologous (internal) versus species-specific (external) TFs. We found that between two species, the orthologs have matched, internally driven expression patterns, but very different species-specific, externally driven ones. This is particularly true for genes with evolutionarily ancient functions (e.g. the ribosomal proteins), in contrast to those with more recently evolved functions (e.g., cell-cell communication).

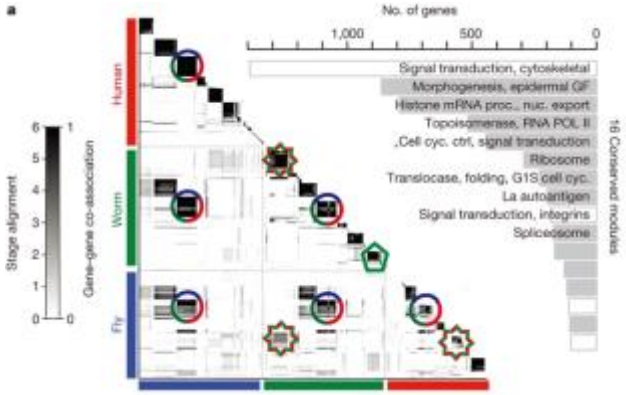

Figure 1 Cross-species gene co-expression network clustering. Left, human, worm and fly gene–gene co-association matrix; darker colouring reflects the increased likelihood that a pair of genes are assigned to the same module. A dark block along the diagonal represents a group of genes within a species. If this is associated with an off-diagonal block then it is a cross-species module (for example, a three-species conserved module is shown with a circle and a worm–fly module, with a star). However, if a diagonal block has no off-diagonal associations, then it forms a species-specific module (for example, green pentagon). Right, the Gene Ontology functional enrichment of genes within the 16 conserved modules is shown. GF, growth factor; nuc., nuclear; proc., processing.

We have developed a number of advanced methods for normalization, analysis, and comparison of RNA-seq profiles. In particular: 1) incRNA, a method that predicts novel ncRNAs

using known ncRNAs of various biotypes as a training set (7); 2) FusionSeq, a pipeline to detect transcripts that arise due to trans-splicing or chromosomal translocations (8, 9); 3) IQSeq, a transcript isoform quantification tool that uses an EM algorithm to resolve the maximum likelihood expression level of individual transcript isoforms (10); 4) Pseudo-seq which addresses the issue of quantification of pseudogene and repetitive region expression (11); and 5) the Aggregation and Correlation Toolbox (ACT), which is a general purpose tool for comparing genomic signal tracks (12). In addition, we contributed to the development of a classification and analysis scheme for “spike” event patterns in omics data with longitudinal profiles(13).

We have comprehensive experience integrating transcriptomic, metabolomics, and proteomic data. We integrated unknown metabolites, which can constitute as much as 50% of spectral features (13), with transcriptomics profiles from different experimental conditions (14). By defining statistics to correlate the co-occurrence patterns of metabolites and genes we generated hypotheses about the identities of unannotated biosynthetic pathways. In addition, we have experience with the analysis of proteomic data and its integration with transcriptomics (15-18). This allowed us to identify previously uncharacterized proteins in a temporally and spatially resolved manner(18).

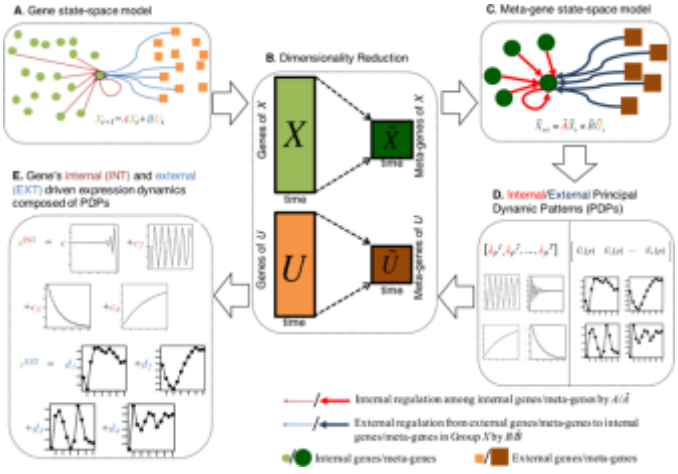

Figure 2 DREISS: Using State-Space Models to Infer the Dynamics of Gene Expression Driven by External and Internal Regulatory Networks. (A) DREISS models temporal gene expression dynamics using state-space models in control theory. The “state” refers to the expressions for a large group of genes of interest, such as the worm-fly orthologous genes investigated here. The “control” refers to any other group of genes that contribute to gene expressions of the “state”, such as the species-specific TF studied here. (B) it then projects high-dimensional gene expression space to lower-dimensional meta-gene expression spaces using dimensionality reduction techniques. (C) it derives the effective state-space models for meta-genes so that model parameters can be estimated. (D) it then identifies the meta-gene expression dynamic patterns; i.e., canonical temporal expression trajectories driven by “state” (internal) and by “control” (external) based on the analytic solutions to estimated models. (E) it finally calculates the coefficients of genes for the dynamic patterns of linear transformations between genes and meta-genes.

We also have made extensive use of machine-learning to generate models from integrated datasets. For example, we integrated ENCODE data on transcription factor (TF) binding, histone modifications, and target gene expression to establish regulatory relationships using a probabilistic model we named TIP (Target Identification from Profiles) (19). We identified potential enhancers from distal gene regions and we used these modules to quantify the relationship between TF binding and gene expression(5, 20-22). We integrated these data types with protein-protein interaction and transcriptional regulation networks (23-26). This allowed us to group TFs into histone-sensitive and -insensitive classes that refined the prediction of gene-regulation targets and effects. Finally, we were able to build cross-organism integrative chromatin models (5).

We have extensively analyzed patterns of variation in non-coding regions, along with their coding targets (27-29). We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations (28). In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region (22). Further studies showed relationships between selection and protein network topology (for instance, quantifying selection in hubs relative to proteins on the network periphery (30, 31). In recent studies (32, 33), we have integrated and extended these methods to develop a prioritization pipeline called FunSeq. It identifies sensitive and ultra-sensitive regions (i.e., those annotations under strong selective pressure, as determined using genomes from many individuals from diverse populations). It then identifies potentially deleterious variants in many

non-coding functional elements, including TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitive sites. It also detects their disruptiveness to TF binding sites (both loss-of and gain-of function events). Integrating large-scale data from various resources (including ENCODE and The 1000 Genomes Project) with cancer genomics data, our method is able to prioritize the known TERT promoter driver mutations, and it scores somatic recurrent mutations higher than those that are non-recurrent. Using FunSeq, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast, and prostate cancer samples (33). We developed Loregic, a general-purpose method to characterize the cooperatively of such regulatory factors (34).

References

1. 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.
2. 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.
3. Yip KY, Alexander RP, Yan KK, Gerstein M. Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. PLoS One. 2010;5(1):e8121. doi: 10.1371/journal.pone.0008121. PubMed PMID: 20126643; PMCID: PMC2811182.
4. Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. Nat Methods. 2016;13(3):251-6. doi: 10.1038/nmeth.3746. PubMed PMID: 26828419; PMCID: PMC4834871.
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 Biol. 2014;15(8):R100. doi: 10.1186/gb-2014-15-8-r100. PubMed PMID: 25249401; PMCID: PMC4289247.
6. Wang D, He F, Maslov S, Gerstein M. 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):e1005146. doi: 10.1371/journal.pcbi.1005146. PubMed PMID: 27760135; PMCID: PMC5070849.
7. Lu ZJ, Yip KY, Wang G, Shou C, Hillier LW, Khurana E, Agarwal A, Auerbach R, Rozowsky J, Cheng C, Kato M, Miller DM, Slack F, Snyder M, Waterston RH, Reinke V, Gerstein MB. Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. Genome Res. 2011;21(2):276-85. doi: 10.1101/gr.110189.110. PubMed PMID: 21177971; PMCID: PMC3032931.
8. Pflueger D, Terry S, Sboner A, Habegger L, Esgueva R, Lin PC, Svensson MA, Kitabayashi N, Moss BJ, MacDonald TY, Cao X, Barrette T, Tewari AK, Chee MS, Chinnaiyan AM, Rickman DS, Demichelis F, Gerstein MB, Rubin MA. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res. 2011;21(1):56-67. doi: 10.1101/gr.110684.110. PubMed PMID: 21036922; PMCID: PMC3012926.

9. Sboner A, Habegger L, Pflueger D, Terry S, Chen DZ, Rozowsky JS, Tewari AK, Kitabayashi N, Moss BJ, Chee MS, Demichelis F, Rubin MA, Gerstein MB. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol. 2010;11(10):R104. doi: 10.1186/gb-2010-11-10-r104. PubMed PMID: 20964841; PMCID: PMC3218660.
10. 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. doi: 10.1371/journal.pone.0029175. PubMed PMID: 22238592; PMCID: PMC3253133.
11. Sisu C, Pei B, Leng J, Frankish A, Zhang Y, Balasubramanian S, Harte R, Wang D, Rutenberg-Schoenberg M, Clark W, Diekhans M, Rozowsky J, Hubbard T, Harrow J, Gerstein MB. Comparative analysis of pseudogenes across three phyla. Proc Natl Acad Sci U S A. 2014;111(37):13361-6. doi: 10.1073/pnas.1407293111. PubMed PMID: 25157146; PMCID: PMC4169933.
12. Jee J, Rozowsky J, Yip KY, Lochovsky L, Bjornson R, Zhong G, Zhang Z, Fu Y, Wang J, Weng Z, Gerstein M. ACT: aggregation and correlation toolbox for analyses of genome tracks. Bioinformatics. 2011;27(8):1152-4. doi: 10.1093/bioinformatics/btr092. PubMed PMID: 21349863; PMCID: PMC3072554.
13. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, Blasco MA, Greenberg PL, Snyder P, Klein TE, Altman RB, Butte AJ, Ashley EA, Gerstein M, Nadeau KC, Tang H, Snyder M. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012;148(6):1293-307. doi: 10.1016/j.cell.2012.02.009. PubMed PMID: 22424236; PMCID: PMC3341616.
14. Gianoulis TA, Griffin MA, Spakowicz DJ, Dunican BF, Alpha CJ, Sboner A, Sismour AM, Kodira C, Egholm M, Church GM, Gerstein MB, Strobel SA. Genomic Analysis of the Hydrocarbon-Producing, Cellulolytic, Endophytic Fungus Ascocoryne sarcoides. PLoS Genetics. 2012;8(3):e1002558. doi: 10.1371/journal.pgen.1002558.
15. Kitchen RR, Rozowsky JS, Gerstein MB, Nairn AC. Decoding neuroproteomics: integrating the genome, translatome and functional anatomy. Nature Neuroscience. 2014;17(11):1491-9. doi: 10.1038/nn.3829.
16. Sboner A, Karpikov A, Chen G, Smith M, Dawn M, Freeman-Cook L, Schweitzer B, Gerstein MB. Robust-Linear-Model Normalization To Reduce Technical Variability in Functional Protein Microarrays. Journal of Proteome Research. 2009;8(12):5451-64. doi: 10.1021/pr900412k.
17. Smith A, Cheung K, Krauthammer M, Schultz M, Gerstein M. Leveraging the structure of the Semantic Web to enhance information retrieval for proteomics. Bioinformatics. 2007;23(22):3073-9. doi: 10.1093/bioinformatics/btm452.
18. Wu L, Hwang SI, Rezaul K, Lu LJ, Mayya V, Gerstein M, Eng JK, Lundgren DH, Han DK. Global Survey of Human T Leukemic Cells by Integrating Proteomics and Transcriptomics Profiling. Molecular & Cellular Proteomics. 2007;6(8):1343-53. doi: 10.1074/mcp.m700017-mcp200.

19. Cheng C, Min R, Gerstein M. TIP: A probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. Bioinformatics. 2011;27(23):3221-7. doi: 10.1093/bioinformatics/btr552.
20. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, Yan KK, Dong X, Djebali S, Ruan Y, Davis CA, Carninci P, Lassman T, Gingeras TR, Guigo R, Birney E, Weng Z, Snyder M, Gerstein M. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Research. 2012;22(9):1658-67. doi: 10.1101/gr.136838.111.
21. Cheng C, Gerstein M. 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):553-68. doi: 10.1093/nar/gkr752.
22. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74. doi: 10.1038/nature11247. PubMed PMID: 22955616; PMCID: PMC3439153.
23. Cheng C, Shou C, Yip KY, Gerstein MB. Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors. Genome Biology. 2011;12(11):R111. doi: 10.1186/gb-2011-12-11-r111.
24. Cheng C, Yan K-K, Yip KY, Rozowsky J, Alexander R, Shou C, Gerstein M. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biology. 2011;12(2):R15. doi: 10.1186/gb-2011-12-2-r15.
25. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M, Guigó R, Birney E, Weng Z. Modeling gene expression using chromatin features in various cellular contexts. Genome Biology. 2012;13(9):R53. doi: 10.1186/gb-2012-13-9-r53.
26. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K, Agarwal A, Alexander RP, Barber G, Brdlik CM, Brennan J, Brouillet JJ, Carr A, Cheung MS, Clawson H, Contrino S, Dannenberg LO, Dernburg AF, Desai A, Dick L, Dose AC, Du J, Egelhofer T, Ercan S, Euskirchen G, Ewing B, Feingold EA, Gassmann R, Good PJ, Green P, Gullier F, Gutwein M, Guyer MS, Habegger L, Han T, Henikoff JG, Henz SR, Hinrichs A, Holster H, Hyman T, Iniguez AL, Janette J, Jensen M, Kato M, Kent WJ, Kephart E, Khivansara V, Khurana E, Kim JK, Kolasinska-Zwierz P, Lai EC, Latorre I, Leahey A, Lewis S, Lloyd P, Lochovsky L, Lowdon RF, Lubling Y, Lyne R, MacCoss M, Mackowiak SD, Mangone M, McKay S, Mecenas D, Merrihew G, Miller DM, 3rd, Muroyama A, Murray JI, Ooi SL, Pham H, Phippen T, Preston EA, Rajewsky N, Ratsch G, Rosenbaum H, Rozowsky J, Rutherford K, Ruzanov P, Sarov M, Sasidharan R, Sboner A, Scheid P, Segal E, Shin H, Shou C, Slack FJ, Slightam C, Smith R, Spencer WC, Stinson EO, Taing S, Takasaki T, Vafeados D, Voronina K, Wang G, Washington NL, Whittle CM, Wu B, Yan KK, Zeller G, Zha Z, Zhong M, Zhou X, mod EC, Ahringer J, Strome S, Gunsalus KC, Micklem G, Liu XS, Reinke V, Kim SK, Hillier LW, Henikoff S, Piano F, Snyder M, Stein L, Lieb JD, Waterston RH. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010;330(6012):1775-87. doi: 10.1126/science.1196914. PubMed PMID: 21177976; PMCID: PMC3142569.
27. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B,

Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489(7414):91-100. doi: 10.1038/nature11245. PubMed PMID: 22955619; PMCID: PMC4154057.
28. Mu XJ, Lu ZJ, Kong Y, Lam HY, Gerstein MB. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. Nucleic Acids Res. 2011;39(16):7058-76. doi: 10.1093/nar/gkr342. PubMed PMID: 21596777; PMCID: PMC3167619.
29. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M, Gerstein M. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol. 2012;13(9):R48. doi: 10.1186/gb-2012-13-9-r48. PubMed PMID: 22950945; PMCID: PMC3491392.
30. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27. doi: 10.1093/biostatistics/kxj037. PubMed PMID: 16632515.
31. Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of genomic variants using a unified biological network approach. PLoS Comput Biol. 2013;9(3):e1002886. doi: 10.1371/journal.pcbi.1002886. PubMed PMID: 23505346; PMCID: PMC3591262.
32. 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 Biol. 2014;15(10):480. doi: 10.1186/s13059-014-0480-5. PubMed PMID: 25273974; PMCID: PMC4203974.
33. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J, Abyzov A, Balasubramanian S, Beal K, Chakravarty D, Challis D, Chen Y, Clarke D, Clarke L, Cunningham F, Evani US, Flicek P, Fragoza R, Garrison E, Gibbs R, Gumus ZH, Herrero J, Kitabayashi N, Kong Y, Lage K, Liluashvili V, Lipkin SM, MacArthur DG, Marth G, Muzny D, Pers TH, Ritchie GR, Rosenfeld JA, Sisu C, Wei X, Wilson M, Xue Y, Yu F, Genomes Project C, Dermitzakis ET, Yu H, Rubin MA, Tyler-Smith C, Gerstein M. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 2013;342(6154):1235587. doi: 10.1126/science.1235587. PubMed PMID: 24092746; PMCID: PMC3947637.
34. Wang D, Yan KK, Sisu C, Cheng C, Rozowsky J, Meyerson W, Gerstein MB. Loregic: a method to characterize the cooperative logic of regulatory factors. PLoS Comput Biol. 2015;11(4):e1004132. doi: 10.1371/journal.pcbi.1004132. PubMed PMID: 25884877; PMCID: PMC4401777.