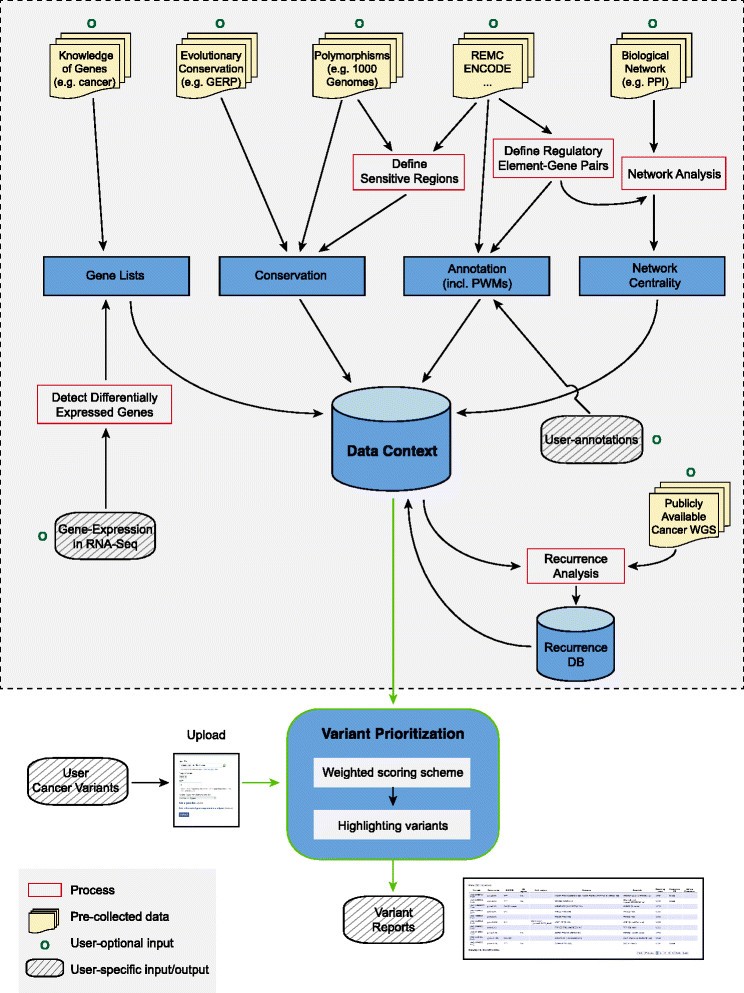
#### Integrative models for relating gene expression with TF binding and histone modifications.

Both transcription factors (TFs) and histone modifications (HMs) play critical roles in gene transcriptional regulation [7, 8]. Taking advantage of the ENCODE, modENCODE and other data sources, we constructed statistical models to quantify the relationship between TF/HM signals and gene expression level in human and model organisms [9-11]. Our models indicated that over 60% of variation in gene expression can be explained by either TF binding or HM signals in the promoter regions of genes [9, 11]. TF binding and HMs are more predictive of gene expression in the same cell line, indicating that the models are cell line specific. Moreover, differential gene expression between cell lines is predictable by differential TF binding and HM signals. We also developed models to quantify the relationships between TF-binding signals and other chromatin features such as histone modifications and DNase hypersensitivity for determining expression [9].



**Fig. 2: Schematic workflow of FunSeq2.** FunSeq2 consists of two components: creation of data context and variant prioritization.

#### Integrative models quantifying the impact of genomic variants.

We have extensive experience with quantifying variant impacts from various aspects. For instance, we developed a variant-prioritization pipeline named FunSeq that included an adjustable data context [6, 12]. This tool has been widely used to identify disease-causing mutations for further in-depth analyses to understand the mechanisms underlying disease pathogenesis. FunSeq links each noncoding mutation to target genes, and prioritizes such variants based on functional annotation, sequence features, conservation, network connectivity, and mutation frequency in diseases (**Fig. 2**). We also developed a generalized model named GRAM to predict cell-type-specific molecular effects of non-coding variants on their associated genes [13]. This tool has been used to predict the effects of fine-mapping causal variants from genome-wide association studies. Finally, we developed a variant-scoring framework named RADAR to pinpoint variants associated with RNA binding protein function dysregulation [14]. In addition, we developed AlleleSeq, a tool for detecting candidate variants associated with allele-specific binding and allele-specific expression [15-17].

### *DSPN model as a foundation to build interpretable deep-learning models*

In order to probe the functional effects of genetic variation and the mechanistic underpinnings of disease, it is necessary to build integrated models that relate molecular- and cellular-level phenotypes to high-level traits. In our previous work, we developed an interpretable integrated modeling framework for this purpose, within the context of psychiatric genomics [5]. Our Deep Structured Phenotype Network (DSPN) framework allowed us to model the joint distribution of all phenotypes of interest conditioned on genetic variation; a joint energy function enabled us to embed prior knowledge in the connectivity of the network and interpret new relationships during and after training. We used a conditional deep Boltzmann machine architecture with multiple layers, including genotype, gene expression, epigenetics, and cell fraction layers, and introduced lateral connectivity at the visible layer to embed the gene regulatory network (GRN) and quantitative trait locus (QTL) linkages. Further, we developed a rank-statistic-based interpretation scheme that allows us to functionally annotate hidden nodes and prioritize them relative to disorders [18]. Our model improved disease prediction by 6-fold compared to additive polygenic risk scores for schizophrenia, highlighted key genes for schizophrenia and other disorders, and allowed imputation of missing transcriptome information from genotype alone [5].

We have developed further specialized deep-learning and machine-learning approaches for extracting complex higher-order structure from genetics and genomics data. These include extracting latent signatures from gene expression data as biomarkers in asthma [19], learning higher-order features for TF to target gene linkages, and protein-protein interactions [20-22], and predicting splicing patterns from epigenomic signals [23]. Further, we have adapted deep-learning-based methods for natural language processing to learn embeddings of multiple biological networks in a shared latent space for integrated cross-network interpretation [24]



**Fig. 4: Flowchart of the matched-filter model.** Meta-profiles of epigenetic features are produced and integrated by a machine-learning model to predict active promoters and enhancers in a genome-wide fashion.

#### Identifying enhancers from ChIP-seq, STARR-seq, and single-cell data

We have developed integrative models to predict human enhancers based on TF binding data from the ENCODE project [36]. We developed a framework named MatchedFilter (**Fig. 4**) using Drosophila STARR-seq to create shape-matching filters based on meta-profiles of epigenetic features [37].

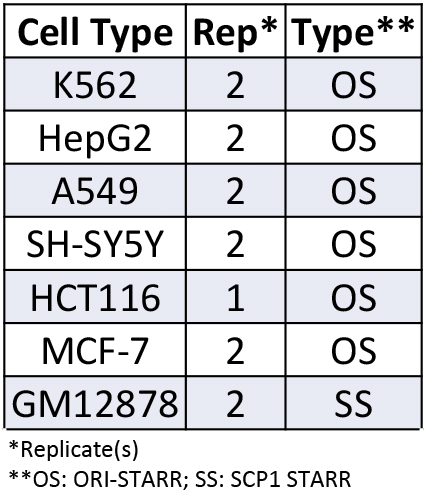
We integrated the resultant features with supervised machine-learning algorithms to predict enhancers in *Drosophila*. We demonstrated that we could extend our model to predict enhancers in mammals. We have also tried to define enhancers from single-cell multi-omics data. For instance, we have developed an initial single-cell ATAC-seq pipeline and simulator to identify cell-type-specific cis-regulatory elements [38].

#### Reconstructing biological networks by combining high- and low-throughput data

Network reconstruction requires integrating both large-scale yet noisy data from high-throughput experiments and reliable yet incomplete data from small-scale experiments. Our team has a very strong track record in network reconstruction. In fact, we were among the first to use statistical models to perform systematic network reconstruction of protein-protein interaction networks [39]. Since then, we have developed many computational methods for reconstructing a variety of biological networks. For example, we developed the JEME method for reconstructing the enhancer-target gene interaction network by integrating epigenetic data from hundreds of human samples [35]. We developed the iTAR tool for identifying target genes of TFs for reconstructing TF binding networks [40]. We have used high-order neural networks and kernel methods to reconstruct peptide-major histocompatibility complex interactions [20]. We also developed the ProbRNA method for reconstructing protein-RNA interaction networks using structure-probing sequencing data [41]. These are just some examples of the many methods we have developed.

Using our methods, we have successfully reconstructed different types of networks in humans and model organisms, including large-scale efforts aimed at producing reference networks. As some notable examples, we have reconstructed TF binding and miRNA binding networks in human cell lines and in *C. elegans*, respectively, based on data from ENCODE and modENCODE [1, 42]. More recently, we reconstructed gene regulatory networks in human neurons for the studying psychiatric conditions [5].

Table 2: STARR-seq data



#### Analyzing biological networks to elucidate the effects of genomic variants

Reconstructed networks can help infer the direct and indirect effects of genomic variants. For example, we have made use of network properties such as centrality to evaluate the functional significance of genomic variants [6]. Network hierarchy represents another useful concept for predicting the impacts of genomic variants, where the perturbation of elements at upper layers generally has broader effects. We have developed different methods for determining network hierarchies, such as HirNet [43]. Genomic variants can also lead to disruptions of network connections. We developed DiNeR for identifying disruptions of TF co-regulation by variants and analyzing their consequences [44]. On a larger scale, some network perturbations may propagate to cause major network rewiring. We developed the TopicNet method to measure such rewiring in transcriptional regulatory networks [45]. We have also applied this idea to study network rewiring in cancer cells, as part of our efforts toward producing a general resource for cancer research based on ENCODE data [46].

#### As listed in **Table 2**, we have collected whole-genome STARR-seq data across seven cell types (37). We carefully investigated the epigenetics signals around the STARR-seq peaks, which are either active enhancers or promoters. In addition to the combinatory existence of epigenetic marks, we found that strong STARR-seq peaks further require an enriched peak-trough-peak (“double peak”) signal shape in several histone modification marks such as H3k27ac and H3k4me3. This confirms the conclusion that enhancer regions tend to be depleted of histone proteins and contain accessible DNA where various TFs and cofactors can bind (37). The troughs in the double-peak ChIP-seq signal represent the accessible chromatin regions, which usually lead to a peak from the ATAC-seq data at the center.

#### Experience in consortium research, especially in all 4 phases of ENCODE

Our team has extensive experience participating in consortium research, collaborating with computational and experimental groups, and leading different working groups. Specifically, we participated in every phase of ENCODE to annotate the noncoding genome. These efforts led to several flagship consortium papers. For instance, we constructed regulatory networks from extensive ChIP-seq profiles, which was important for interpreting personal genome sequences and understanding basic principles of human biology and disease [1]. We played a leading role in the integrative analysis for ENCODE [52-54] and modENCODE [42, 52].

In addition to ENCODE, we have played a significant role in a number of consortia. Specifically, we were part of the 1,000 Genomes Project and led the analysis of germline variant patterns within the regulatory regions to quantify variant impacts. Recently, we led analysis projects of PsychENCODE, Brainspan [5], and the Pan-Cancer Analysis of Whole Genomes [55] and also participated in the GENCODE [56, 57], DOE Kbase [58], exRNA [59] and TCGA consortia [60].

#### Extensive experience in developing computational pipelines and QC metrics for genomics.

We have extensive experience developing data quality standards and data processing strategies for numerous state-of-the-art sequencing technologies, some of which are widely used by the above consortia. As part of the ENCODE project, our group was responsible for developing the irreproducible discovery rate (IDR) [61], a method to determine consistency across biological replicates of ChIP-seq peaks, which is now an essential element of the ENCODE quality control pipeline. We have developed tools to analyze a variety of high-throughput sequencing data in ENCODE, such as RNA-seq, ChIP-seq, and STARR-seq data [61-63].

#### Experience in prioritizing functional elements for validations in a collaborative group fashion.

We are one of the leading analytical groups that are prioritizing functional elements and variants. Our team has extensive experience prioritizing elements for experimental follow-up [64, 65]. We have coordinated multiple computational groups to participate in the ENCODE3 enhancer challenge. We also co-led the analysis of the ENCODE cancer working group, including prioritizing regions for detailed follow-up with novel functional genomic assays. In addition, we provided the PsychENCODE FCC target region list for validations.

**REFERENCES**

1. 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 JJ, 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. Epub 2012/09/08. doi: 10.1038/nature11245. PubMed PMID: 22955619; PMCID: PMC4154057.

2. Noguchi S, Arakawa T, Fukuda S, Furuno M, Hasegawa A, Hori F, Ishikawa-Kato S, Kaida K, Kaiho A, Kanamori-Katayama M, Kawashima T, Kojima M, Kubosaki A, Manabe RI, Murata M, Nagao-Sato S, Nakazato K, Ninomiya N, Nishiyori-Sueki H, Noma S, Saijyo E, Saka A, Sakai M, Simon C, Suzuki N, Tagami M, Watanabe S, Yoshida S, Arner P, Axton RA, Babina M, Baillie JK, Barnett TC, Beckhouse AG, Blumenthal A, Bodega B, Bonetti A, Briggs J, Brombacher F, Carlisle AJ, Clevers HC, Davis CA, Detmar M, Dohi T, Edge ASB, Edinger M, Ehrlund A, Ekwall K, Endoh M, Enomoto H, Eslami A, Fagiolini M, Fairbairn L, Farach-Carson MC, Faulkner GJ, Ferrai C, Fisher ME, Forrester LM, Fujita R, Furusawa JI, Geijtenbeek TB, Gingeras T, Goldowitz D, Guhl S, Guler R, Gustincich S, Ha TJ, Hamaguchi M, Hara M, Hasegawa Y, Herlyn M, Heutink P, Hitchens KJ, Hume DA, Ikawa T, Ishizu Y, Kai C, Kawamoto H, Kawamura YI, Kempfle JS, Kenna TJ, Kere J, Khachigian LM, Kitamura T, Klein S, Klinken SP, Knox AJ, Kojima S, Koseki H, Koyasu S, Lee W, Lennartsson A, Mackay-Sim A, Mejhert N, Mizuno Y, Morikawa H, Morimoto M, Moro K, Morris KJ, Motohashi H, Mummery CL, Nakachi Y, Nakahara F, Nakamura T, Nakamura Y, Nozaki T, Ogishima S, Ohkura N, Ohno H, Ohshima M, Okada-Hatakeyama M, Okazaki Y, Orlando V, Ovchinnikov DA, Passier R, Patrikakis M, Pombo A, Pradhan-Bhatt S, Qin XY, Rehli M, Rizzu P, Roy S, Sajantila A, Sakaguchi S, Sato H, Satoh H, Savvi S, Saxena A, Schmidl C, Schneider C, Schulze-Tanzil GG, Schwegmann A, Sheng G, Shin JW, Sugiyama D, Sugiyama T, Summers KM, Takahashi N, Takai J, Tanaka H, Tatsukawa H, Tomoiu A, Toyoda H, van de Wetering M, van den Berg LM, Verardo R, Vijayan D, Wells CA, Winteringham LN, Wolvetang E, Yamaguchi Y, Yamamoto M, Yanagi-Mizuochi C, Yoneda M, Yonekura Y, Zhang PG, Zucchelli S, Abugessaisa I, Arner E, Harshbarger J, Kondo A, Lassmann T, Lizio M, Sahin S, Sengstag T, Severin J, Shimoji H, Suzuki M, Suzuki H, Kawai J, Kondo N, Itoh M, Daub CO, Kasukawa T, Kawaji H, Carninci P, Forrest ARR, Hayashizaki Y. FANTOM5 CAGE profiles of human and mouse samples. Sci Data. 2017;4:170112. Epub 2017/08/30. doi: 10.1038/sdata.2017.112. PubMed PMID: 28850106; PMCID: PMC5574368.

3. Tran D, Hoffman MD, Saurous RA, Brevdo E, Murphy K, Blei DM. Deep probabilistic programming. arXiv preprint. 2017.

4. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol. 2010;28(10):1045-8. Epub 2010/10/15. doi: 10.1038/nbt1010-1045. PubMed PMID: 20944595; PMCID: PMC3607281.

5. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, Clarke D, Gu M, Emani P, Yang YT, Xu M, Gandal MJ, Lou S, Zhang J, Park JJ, Yan C, Rhie SK, Manakongtreecheep K, Zhou H, Nathan A, Peters M, Mattei E, Fitzgerald D, Brunetti T, Moore J, Jiang Y, Girdhar K, Hoffman GE, Kalayci S, Gumus ZH, Crawford GE, Psych EC, Roussos P, Akbarian S, Jaffe AE, White KP, Weng Z, Sestan N, Geschwind DH, Knowles JA, Gerstein MB. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018;362(6420). Epub 2018/12/14. doi: 10.1126/science.aat8464. PubMed PMID: 30545857; PMCID: PMC6413328.

6. 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 GRS, 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. Epub 2013/10/05. doi: 10.1126/science.1235587. PubMed PMID: 24092746; PMCID: PMC3947637.

7. Farnham PJ. Insights from genomic profiling of transcription factors. Nat Rev Genet. 2009;10(9):605-16. Epub 2009/08/12. doi: 10.1038/nrg2636. PubMed PMID: 19668247; PMCID: PMC2846386.

8. Berger SL. The complex language of chromatin regulation during transcription. Nature. 2007;447(7143):407-12. Epub 2007/05/25. doi: 10.1038/nature05915. PubMed PMID: 17522673.

9. 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 Res. 2012;22(9):1658-67. Epub 2012/09/08. doi: 10.1101/gr.136838.111. PubMed PMID: 22955978; PMCID: PMC3431483.

10. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M, Guigo R, Birney E, Weng Z. Modeling gene expression using chromatin features in various cellular contexts. Genome Biol. 2012;13(9):R53. Epub 2012/09/07. doi: 10.1186/gb-2012-13-9-r53. PubMed PMID: 22950368; PMCID: PMC3491397.

11. 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 Res. 2012;40(2):553-68. Epub 2011/09/20. doi: 10.1093/nar/gkr752. PubMed PMID: 21926158; PMCID: PMC3258143.

12. 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. Epub 2014/10/03. doi: 10.1186/s13059-014-0480-5. PubMed PMID: 25273974; PMCID: PMC4203974.

13. Lou S, Cotter KA, Li T, Liang J, Mohsen H, Liu J, Zhang J, Cohen S, Xu J, Yu H, Rubin MA, Gerstein M. GRAM: A GeneRAlized Model to predict the molecular effect of a non-coding variant in a cell-type specific manner. PLoS Genet. 2019;15(8):e1007860. Epub 2019/08/31. doi: 10.1371/journal.pgen.1007860. PubMed PMID: 31469829; PMCID: PMC6742416.

14. Zhang J, Liu J, Lee D, Feng JJ, Lochovsky L, Lou S, Rutenberg-Schoenberg M, Gerstein M. RADAR: annotation and prioritization of variants in the post-transcriptional regulome of RNA-binding proteins. Genome Biol. 2020;21(1):151. Epub 2020/07/31. doi: 10.1186/s13059-020-01979-4. PubMed PMID: 32727537; PMCID: PMC7391703.

15. Onuchic V, Lurie E, Carrero I, Pawliczek P, Patel RY, Rozowsky J, Galeev T, Huang Z, Altshuler RC, Zhang Z, Harris RA, Coarfa C, Ashmore L, Bertol JW, Fakhouri WD, Yu F, Kellis M, Gerstein M, Milosavljevic A. Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. Science. 2018;361(6409). Epub 2018/08/25. doi: 10.1126/science.aar3146. PubMed PMID: 30139913; PMCID: PMC6198826.

16. Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan L, Gerstein M. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat Commun. 2016;7:11101. Epub 2016/04/19. doi: 10.1038/ncomms11101. PubMed PMID: 27089393; PMCID: PMC4837449.

17. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, Bhardwaj N, Rubin M, Snyder M, Gerstein M. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011;7:522. Epub 2011/08/04. doi: 10.1038/msb.2011.54. PubMed PMID: 21811232; PMCID: PMC3208341.

18. Warrell J, Mohsen H, Gerstein M. Rank Projection Trees for Multilevel Neural Network Interpretation. arXiv preprint. 2018.

19. Lou S, Li T, Spakowicz D, Chupp GL, Gerstein M. Latent-space embedding of expression data identifies gene signatures from sputum samples of asthmatic patients. bioRxiv. 2019:646976. doi: 10.1101/646976.

20. Kuksa PP, Min MR, Dugar R, Gerstein M. High-order neural networks and kernel methods for peptide-MHC binding prediction. Bioinformatics. 2015;31(22):3600-7. Epub 2015/07/25. doi: 10.1093/bioinformatics/btv371. PubMed PMID: 26206306.

21. Min MR, Ning X, Cheng C, Gerstein M, editors. Interpretable sparse high-order boltzmann machines. Artificial Intelligence and Statistics; 2014.

22. Min MR, Ning X, Qi Y, Cheng C, Bonner A, Gerstein M, editors. Ensemble Learning Based Sparse High-Order Boltzmann Machine for Unsupervised Feature Interaction Identification. NIPS Workshop on Machine Learning in Computational Biology; 2014.

23. Lee D, Zhang J, Liu J, Gerstein M. Epigenome-based splicing prediction using a recurrent neural network. PLoS Comput Biol. 2020;16(6):e1008006. Epub 2020/06/26. doi: 10.1371/journal.pcbi.1008006. PubMed PMID: 32584815; PMCID: PMC7343189.

24. Cao Q, Zhang Z, Fu AX, Wu Q, Lee T-L, Lo E, Cheng ASL, Cheng C, Leung D, Yip KY. A unified framework for integrative study of heterogeneous gene regulatory mechanisms. Nature Machine Intelligence. 2020;2(8):447-56. doi: 10.1038/s42256-020-0205-2.

25. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint. 2016.

26. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An Introduction to Variational Methods for Graphical Models. Machine Learning. 1999;37(2):183-233. doi: 10.1023/A:1007665907178.

27. Warrell J, Mohsen H, Gerstein M. Compression-based Network Interpretability Schemes. bioRxiv. 2020.

28. Zhou W, Veitch V, Austern M, Adams RP, Orbanz P. Non-vacuous generalization bounds at the imagenet scale: a PAC-Bayesian compression approach. arXiv preprint. 2018.

29. Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER, 3rd, Barretina J, Gelfand ET, Bielski CM, Li H, Hu K, Andreev-Drakhlin AY, Kim J, Hess JM, Haas BJ, Aguet F, Weir BA, Rothberg MV, Paolella BR, Lawrence MS, Akbani R, Lu Y, Tiv HL, Gokhale PC, de Weck A, Mansour AA, Oh C, Shih J, Hadi K, Rosen Y, Bistline J, Venkatesan K, Reddy A, Sonkin D, Liu M, Lehar J, Korn JM, Porter DA, Jones MD, Golji J, Caponigro G, Taylor JE, Dunning CM, Creech AL, Warren AC, McFarland JM, Zamanighomi M, Kauffmann A, Stransky N, Imielinski M, Maruvka YE, Cherniack AD, Tsherniak A, Vazquez F, Jaffe JD, Lane AA, Weinstock DM, Johannessen CM, Morrissey MP, Stegmeier F, Schlegel R, Hahn WC, Getz G, Mills GB, Boehm JS, Golub TR, Garraway LA, Sellers WR. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569(7757):503-8. Epub 2019/05/10. doi: 10.1038/s41586-019-1186-3. PubMed PMID: 31068700; PMCID: PMC6697103.

30. Denzler J, Brown CM. Information theoretic sensor data selection for active object recognition and state estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002;24(2):145-57. doi: 10.1109/34.982896.

31. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005;308(5721):523-9. Epub 2005/04/23. doi: 10.1126/science.1105809. PubMed PMID: 15845847.

32. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015;11(4):e1004219. Epub 2015/04/18. doi: 10.1371/journal.pcbi.1004219. PubMed PMID: 25885710; PMCID: PMC4401657.

33. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. arXiv preprint. 2017.

34. Lundberg SM, Lee S-I, editors. A unified approach to interpreting model predictions. Advances in neural information processing systems; 2017.

35. Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, Mok MTS, Cheng C, Fan X, Gerstein M, Cheng ASL, Yip KY. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. Nat Genet. 2017;49(10):1428-36. Epub 2017/09/05. doi: 10.1038/ng.3950. PubMed PMID: 28869592.

36. 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. Epub 2012/09/07. doi: 10.1186/gb-2012-13-9-r48. PubMed PMID: 22950945; PMCID: PMC3491392.

37. Sethi A, Gu M, Gumusgoz E, Chan L, Yan KK, Rozowsky J, Barozzi I, Afzal V, Akiyama JA, Plajzer-Frick I, Yan C, Novak CS, Kato M, Garvin TH, Pham Q, Harrington A, Mannion BJ, Lee EA, Fukuda-Yuzawa Y, Visel A, Dickel DE, Yip KY, Sutton R, Pennacchio LA, Gerstein M. Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. Nat Methods. 2020;17(8):807-14. Epub 2020/08/02. doi: 10.1038/s41592-020-0907-8. PubMed PMID: 32737473.

38. Chen Z, Zhang J, Liu J, Zhang Z, Zhu J, Xu M, Gerstein M. SCAN-ATAC Sim: a scalable and efficient method to simulate single-cell ATAC-seq from bulk-tissue experiments. bioRxiv. 2020:2020.05.29.123638. doi: 10.1101/2020.05.29.123638.

39. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 2003;302(5644):449-53. Epub 2003/10/18. doi: 10.1126/science.1087361. PubMed PMID: 14564010.

40. Yang CC, Andrews EH, Chen MH, Wang WY, Chen JJ, Gerstein M, Liu CC, Cheng C. iTAR: a web server for identifying target genes of transcription factors using ChIP-seq or ChIP-chip data. BMC Genomics. 2016;17(1):632. Epub 2016/08/16. doi: 10.1186/s12864-016-2963-0. PubMed PMID: 27519564; PMCID: PMC4983039.

41. Hu X, Wong TK, Lu ZJ, Chan TF, Lau TC, Yiu SM, Yip KY. Computational identification of protein binding sites on RNAs using high-throughput RNA structure-probing data. Bioinformatics. 2014;30(8):1049-55. Epub 2014/04/15. doi: 10.1093/bioinformatics/btt757. PubMed PMID: 24376038.

42. 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. Epub 2010/12/24. doi: 10.1126/science.1196914. PubMed PMID: 21177976; PMCID: PMC3142569.

43. Cheng C, Andrews E, Yan KK, Ung M, Wang D, Gerstein M. An approach for determining and measuring network hierarchy applied to comparing the phosphorylome and the regulome. Genome Biol. 2015;16:63. Epub 2015/04/17. doi: 10.1186/s13059-015-0624-2. PubMed PMID: 25880651; PMCID: PMC4404648.

44. Zhang J, Liu J, Lee D, Lou S, Chen Z, Gursoy G, Gerstein M. DiNeR: a Differential graphical model for analysis of co-regulation Network Rewiring. BMC Bioinformatics. 2020;21(1):281. Epub 2020/07/04. doi: 10.1186/s12859-020-03605-3. PubMed PMID: 32615918; PMCID: PMC7333332.

45. Lou S, Li T, Kong X, Zhang J, Liu J, Lee D, Gerstein M. TopicNet: a framework for measuring transcriptional regulatory network change. Bioinformatics. 2020;36(Supplement\_1):i474-i81. Epub 2020/07/14. doi: 10.1093/bioinformatics/btaa403. PubMed PMID: 32657410; PMCID: PMC7355251.

46. Zhang J, Lee D, Dhiman V, Jiang P, Xu J, McGillivray P, Yang H, Liu J, Meyerson W, Clarke D, Gu M, Li S, Lou S, Xu J, Lochovsky L, Ung M, Ma L, Yu S, Cao Q, Harmanci A, Yan KK, Sethi A, Gursoy G, Schoenberg MR, Rozowsky J, Warrell J, Emani P, Yang YT, Galeev T, Kong X, Liu S, Li X, Krishnan J, Feng Y, Rivera-Mulia JC, Adrian J, Broach JR, Bolt M, Moran J, Fitzgerald D, Dileep V, Liu T, Mei S, Sasaki T, Trevilla-Garcia C, Wang S, Wang Y, Zang C, Wang D, Klein RJ, Snyder M, Gilbert DM, Yip K, Cheng C, Yue F, Liu XS, White KP, Gerstein M. An integrative ENCODE resource for cancer genomics. Nat Commun. 2020;11(1):3696. Epub 2020/07/31. doi: 10.1038/s41467-020-14743-w. PubMed PMID: 32728046; PMCID: PMC7391744.

47. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013;339(6123):1074-7. Epub 2013/01/19. doi: 10.1126/science.1232542. PubMed PMID: 23328393.

48. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D, editors. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision; 2017.

49. Pang B, Snyder MP. Systematic identification of silencers in human cells. Nat Genet. 2020;52(3):254-63. Epub 2020/02/26. doi: 10.1038/s41588-020-0578-5. PubMed PMID: 32094911; PMCID: PMC7148122.

50. Bao L, Zhou M, Cui Y. CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators. Nucleic Acids Res. 2008;36(Database issue):D83-7. Epub 2007/11/06. doi: 10.1093/nar/gkm875. PubMed PMID: 17981843; PMCID: PMC2238977.

51. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell. 2007;128(6):1231-45. Epub 2007/03/27. doi: 10.1016/j.cell.2006.12.048. PubMed PMID: 17382889; PMCID: PMC2572726.

52. Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, Jiang L, Kasper D, Kawli T, Kheradpour P, Kundaje A, Li JJ, Ma L, Niu W, Rehm EJ, Rozowsky J, Slattery M, Spokony R, Terrell R, Vafeados D, Wang D, Weisdepp P, Wu YC, Xie D, Yan KK, Feingold EA, Good PJ, Pazin MJ, Huang H, Bickel PJ, Brenner SE, Reinke V, Waterston RH, Gerstein M, White KP, Kellis M, Snyder M. Comparative analysis of regulatory information and circuits across distant species. Nature. 2014;512(7515):453-6. Epub 2014/08/29. doi: 10.1038/nature13668. PubMed PMID: 25164757; PMCID: PMC4336544.

53. Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, Pei B, Harmanci AO, Duff MO, Djebali S, Alexander RP, Alver BH, Auerbach R, Bell K, Bickel PJ, Boeck ME, Boley NP, Booth BW, Cherbas L, Cherbas P, Di C, Dobin A, Drenkow J, Ewing B, Fang G, Fastuca M, Feingold EA, Frankish A, Gao G, Good PJ, Guigo R, Hammonds A, Harrow J, Hoskins RA, Howald C, Hu L, Huang H, Hubbard TJ, Huynh C, Jha S, Kasper D, Kato M, Kaufman TC, Kitchen RR, Ladewig E, Lagarde J, Lai E, Leng J, Lu Z, MacCoss M, May G, McWhirter R, Merrihew G, Miller DM, Mortazavi A, Murad R, Oliver B, Olson S, Park PJ, Pazin MJ, Perrimon N, Pervouchine D, Reinke V, Reymond A, Robinson G, Samsonova A, Saunders GI, Schlesinger F, Sethi A, Slack FJ, Spencer WC, Stoiber MH, Strasbourger P, Tanzer A, Thompson OA, Wan KH, Wang G, Wang H, Watkins KL, Wen J, Wen K, Xue C, Yang L, Yip K, Zaleski C, Zhang Y, Zheng H, Brenner SE, Graveley BR, Celniker SE, Gingeras TR, Waterston R. Comparative analysis of the transcriptome across distant species. Nature. 2014;512(7515):445-8. Epub 2014/08/29. doi: 10.1038/nature13424. PubMed PMID: 25164755; PMCID: PMC4155737.

54. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Roder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigo R, Gingeras TR. Landscape of transcription in human cells. Nature. 2012;489(7414):101-8. Epub 2012/09/08. doi: 10.1038/nature11233. PubMed PMID: 22955620; PMCID: PMC3684276.

55. Kumar S, Warrell J, Li S, McGillivray PD, Meyerson W, Salichos L, Harmanci A, Martinez-Fundichely A, Chan CWY, Nielsen MM, Lochovsky L, Zhang Y, Li X, Lou S, Pedersen JS, Herrmann C, Getz G, Khurana E, Gerstein MB. Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. Cell. 2020;180(5):915-27 e16. Epub 2020/02/23. doi: 10.1016/j.cell.2020.01.032. PubMed PMID: 32084333; PMCID: PMC7210002.

56. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, Fan X, Wen J, Handsaker RE, Fairley S, Kronenberg ZN, Kong X, Hormozdiari F, Lee D, Wenger AM, Hastie AR, Antaki D, Anantharaman T, Audano PA, Brand H, Cantsilieris S, Cao H, Cerveira E, Chen C, Chen X, Chin CS, Chong Z, Chuang NT, Lambert CC, Church DM, Clarke L, Farrell A, Flores J, Galeev T, Gorkin DU, Gujral M, Guryev V, Heaton WH, Korlach J, Kumar S, Kwon JY, Lam ET, Lee JE, Lee J, Lee WP, Lee SP, Li S, Marks P, Viaud-Martinez K, Meiers S, Munson KM, Navarro FCP, Nelson BJ, Nodzak C, Noor A, Kyriazopoulou-Panagiotopoulou S, Pang AWC, Qiu Y, Rosanio G, Ryan M, Stutz A, Spierings DCJ, Ward A, Welch AE, Xiao M, Xu W, Zhang C, Zhu Q, Zheng-Bradley X, Lowy E, Yakneen S, McCarroll S, Jun G, Ding L, Koh CL, Ren B, Flicek P, Chen K, Gerstein MB, Kwok PY, Lansdorp PM, Marth GT, Sebat J, Shi X, Bashir A, Ye K, Devine SE, Talkowski ME, Mills RE, Marschall T, Korbel JO, Eichler EE, Lee C. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10(1):1784. Epub 2019/04/18. doi: 10.1038/s41467-018-08148-z. PubMed PMID: 30992455; PMCID: PMC6467913.

57. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, Konkel MK, Malhotra A, Stutz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Mu XJ, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer EW, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Genomes Project C, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75-81. Epub 2015/10/04. doi: 10.1038/nature15394. PubMed PMID: 26432246; PMCID: PMC4617611.

58. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D, Perez F, Canon S, Sneddon MW, Henderson ML, Riehl WJ, Murphy-Olson D, Chan SY, Kamimura RT, Kumari S, Drake MM, Brettin TS, Glass EM, Chivian D, Gunter D, Weston DJ, Allen BH, Baumohl J, Best AA, Bowen B, Brenner SE, Bun CC, Chandonia JM, Chia JM, Colasanti R, Conrad N, Davis JJ, Davison BH, DeJongh M, Devoid S, Dietrich E, Dubchak I, Edirisinghe JN, Fang G, Faria JP, Frybarger PM, Gerlach W, Gerstein M, Greiner A, Gurtowski J, Haun HL, He F, Jain R, Joachimiak MP, Keegan KP, Kondo S, Kumar V, Land ML, Meyer F, Mills M, Novichkov PS, Oh T, Olsen GJ, Olson R, Parrello B, Pasternak S, Pearson E, Poon SS, Price GA, Ramakrishnan S, Ranjan P, Ronald PC, Schatz MC, Seaver SMD, Shukla M, Sutormin RA, Syed MH, Thomason J, Tintle NL, Wang D, Xia F, Yoo H, Yoo S, Yu D. KBase: The United States Department of Energy Systems Biology Knowledgebase. Nat Biotechnol. 2018;36(7):566-9. Epub 2018/07/07. doi: 10.1038/nbt.4163. PubMed PMID: 29979655; PMCID: PMC6870991.

59. Murillo OD, Thistlethwaite W, Rozowsky J, Subramanian SL, Lucero R, Shah N, Jackson AR, Srinivasan S, Chung A, Laurent CD, Kitchen RR, Galeev T, Warrell J, Diao JA, Welsh JA, Hanspers K, Riutta A, Burgstaller-Muehlbacher S, Shah RV, Yeri A, Jenkins LM, Ahsen ME, Cordon-Cardo C, Dogra N, Gifford SM, Smith JT, Stolovitzky G, Tewari AK, Wunsch BH, Yadav KK, Danielson KM, Filant J, Moeller C, Nejad P, Paul A, Simonson B, Wong DK, Zhang X, Balaj L, Gandhi R, Sood AK, Alexander RP, Wang L, Wu C, Wong DTW, Galas DJ, Van Keuren-Jensen K, Patel T, Jones JC, Das S, Cheung KH, Pico AR, Su AI, Raffai RL, Laurent LC, Roth ME, Gerstein MB, Milosavljevic A. exRNA Atlas Analysis Reveals Distinct Extracellular RNA Cargo Types and Their Carriers Present across Human Biofluids. Cell. 2019;177(2):463-77 e15. Epub 2019/04/06. doi: 10.1016/j.cell.2019.02.018. PubMed PMID: 30951672; PMCID: PMC6616370.

60. Consortium ITP-CAoWG. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82-93. Epub 2020/02/07. doi: 10.1038/s41586-020-1969-6. PubMed PMID: 32025007; PMCID: PMC7025898.

61. Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. The annals of applied statistics. 2011;5(3):1752-79.

62. 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. Epub 2014/10/09. doi: 10.1186/s13059-014-0474-3. PubMed PMID: 25292436; PMCID: PMC4234855.

63. 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. Epub 2009/01/06. doi: 10.1038/nbt.1518. PubMed PMID: 19122651; PMCID: PMC2924752.

64. Du J, Bjornson RD, Zhang ZD, Kong Y, Snyder M, Gerstein MB. Integrating sequencing technologies in personal genomics: optimal low cost reconstruction of structural variants. PLoS Comput Biol. 2009;5(7):e1000432. Epub 2009/07/14. doi: 10.1371/journal.pcbi.1000432. PubMed PMID: 19593373; PMCID: PMC2700963.

65. Emanuelsson O, Nagalakshmi U, Zheng D, Rozowsky JS, Urban AE, Du J, Lian Z, Stolc V, Weissman S, Snyder M, Gerstein MB. Assessing the performance of different high-density tiling microarray strategies for mapping transcribed regions of the human genome. Genome Res. 2007;17(6):886-97. Epub 2006/11/23. doi: 10.1101/gr.5014606. PubMed PMID: 17119069; PMCID: PMC1891347.

66. Lee D, Shi M, Moran J, Wall M, Zhang J, Liu J, Fitzgerald D, Kyono Y, Ma L, White KP, Gerstein M. STARRPeaker: Uniform processing and accurate identification of STARR-seq active regions. bioRxiv. 2020:694869. doi: 10.1101/694869.

67. Arbel H, Basu S, Fisher WW, Hammonds AS, Wan KH, Park S, Weiszmann R, Booth BW, Keranen SV, Henriquez C, Shams Solari O, Bickel PJ, Biggin MD, Celniker SE, Brown JB. Exploiting regulatory heterogeneity to systematically identify enhancers with high accuracy. Proc Natl Acad Sci U S A. 2019;116(3):900-8. Epub 2019/01/02. doi: 10.1073/pnas.1808833115. PubMed PMID: 30598455; PMCID: PMC6338827.

68. Breiman L. Random Forests. Machine Learning 2001;45:5-32.