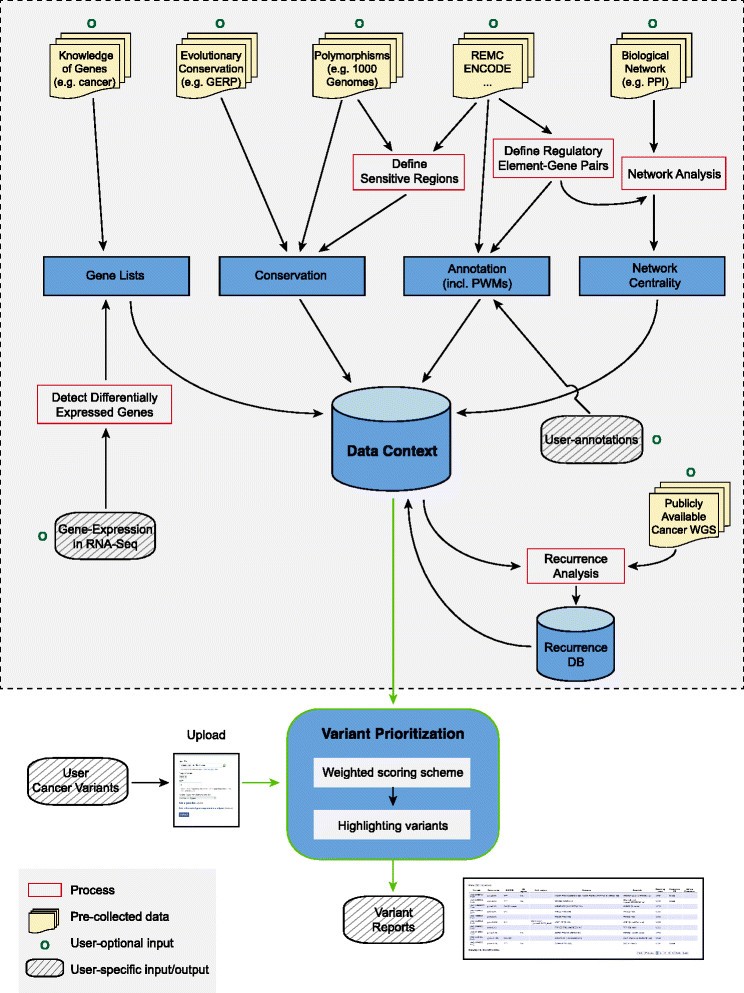
#### Integrative models for relating gene expression with TF binding and histone modifications.

Both transcription factors (TFs) and histone modifications (HMs) play critical roles in gene transcriptional regulation [12, 13]. 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 [4, 14, 15]. 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 [4, 15]. TF binding and HMs are more predictive of gene expression in the same cell type, indicating that the models are cell type specific. Moreover, differential gene expression between cell types 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 HMs and DNase hypersensitivity for determining expression [4]. Beyond promoters, we have demonstrated that enhancer activities and gene body epigenetic signals are also indicative of gene expression levels independent of promoter activities [6, 16]. Our recent GEEK framework [8] further models many types of gene regulatory mechanisms jointly using a network approach.



**Figure 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 [10, 11]. 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 (**Figure 2**). We also developed a generalized model named GRAM to predict cell-type-specific molecular effects of non-coding variants on their associated genes [17]. This tool has been used to predict the effects of fine-mapping causal variants from genome-wide association studies. In addition, we developed a variant-scoring framework named RADAR to pinpoint variants associated with RNA binding protein function dysregulation [18]. We also developed AlleleSeq, a tool for detecting candidate variants associated with allele-specific binding and allele-specific expression [19-21]. Recently, we have developed the MARVEL framework [7] that can detect convergent effects of different genomic variants on the same gene based on cell type specific regulatory element activities, thereby capturing higher-order relationships between genetic variants.

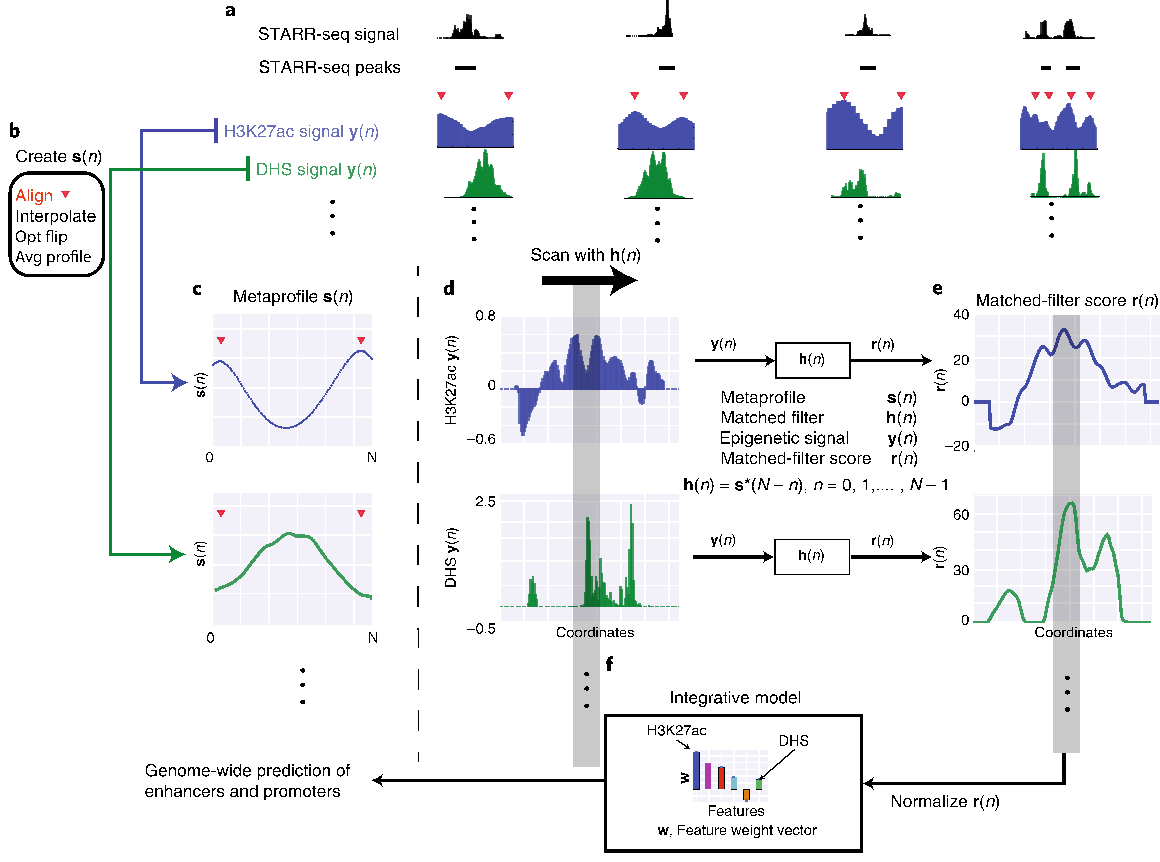
### *DSPN model as a foundation to build interpretable deep-learning models*

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 [9]. 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 [22]. 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 [9].

We have developed further specialized deep-learning and other 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 [23], learning higher-order features for TF-to-target gene linkages, and protein-protein interactions [24-26], and predicting splicing patterns from epigenomic signals [27]. 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 [8] and disease gene prediction [28].

#### 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 [39]. We developed a framework named matched-filter (**Figure 4**) using Drosophila STARR-seq to create shape-matching filters based on meta-profiles of epigenetic features [40].



**Figure 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.

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 [41].

#### 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 [42]. 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 [16]. We developed the iTAR tool for identifying target genes of TFs for reconstructing TF binding networks [43]. We have used high-order neural networks and kernel methods to reconstruct peptide-major histocompatibility complex (MHC) interactions [24]. We also developed the ProbRNA method for reconstructing protein-RNA interaction networks using structure-probing sequencing data [44]. 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, 45]. More recently, we reconstructed gene regulatory networks in human neurons for the study of psychiatric conditions [9].

#### 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 used network properties such as centrality to evaluate the functional significance of genomic variants [10]. 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 [46]. 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 [47]. 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 [48]. 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 [49]. In addition to studying individual networks, ultimately it is necessary to study multiple networks jointly to understand how they affect each other. We recently proposed the GEEK framework that enables this kind of network integration by learning an embedding of the nodes in a latent space according to random walks in an integrated heterogeneous network guided by metapaths, which are node-type sequences that encode prior biological knowledge about gene regulation [8]. We have also demonstrated the use of graph neural networks in functional gene annotation [28].

#### 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 [68-70] and modENCODE [45, 68].

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 [9], and the Pan-Cancer Analysis of Whole Genomes [71] and also participated in the GENCODE [72], DOE Kbase [73], exRNA [74], TCGA [75], and IHEC [76] consortia.

#### 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. 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 [64, 77, 78].

#### 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 [79, 80]. 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 CITED**

1. Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., 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. J., Lian, J.,

Monahan, H., O’Geen, H., Ouyang, Z., Partridge, E. C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L.,

Reddy, T. E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K. Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P. J., Myers, R. M., Weissman, S. M. & Snyder, M. Architecture of the human regulatory network derived from ENCODE data. ***Nature*** 489, 91–100 (2012).

1. 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, R.-I., 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, R. A., Babina, M., Baillie, J. K., Barnett, T. C., Beckhouse, A. G., Blumenthal, A., Bodega, B., Bonetti, A., Briggs, J., Brombacher, F., Carlisle, A. J., Clevers, H. C., Davis, C. A., Detmar, M., Dohi, T., Edge, A. S. B., Edinger, M., Ehrlund, A., Ekwall, K., Endoh, M., Enomoto, H., Eslami, A., Fagiolini, M., Fairbairn, L., Farach-Carson, M. C., Faulkner, G. J., Ferrai, C., Fisher, M. E., Forrester, L. M., Fujita, R., Furusawa, J.-I., Geijtenbeek, T. B., Gingeras, T., Goldowitz, D., Guhl, S., Guler, R., Gustincich, S., Ha, T. J., Hamaguchi, M., Hara, M., Hasegawa, Y., Herlyn, M., Heutink, P., Hitchens, K. J., Hume, D. A., Ikawa,

T., Ishizu, Y., Kai, C., Kawamoto, H., Kawamura, Y. I., Kempfle, J. S., Kenna, T. J., Kere, J., Khachigian,

L. M., Kitamura, T., Klein, S., Klinken, S. P., Knox, A. J., Kojima, S., Koseki, H., Koyasu, S., Lee, W., Lennartsson, A., Mackay-Sim, A., Mejhert, N., Mizuno, Y., Morikawa, H., Morimoto, M., Moro, K., Morris,

K. J., Motohashi, H., Mummery, C. L., 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, D. A., Passier, R., Patrikakis, M., Pombo, A., Pradhan-Bhatt, S., Qin, X.-Y., Rehli, M., Rizzu, P., Roy, S., Sajantila, A., Sakaguchi, S., Sato, H., Satoh, H., Savvi, S., Saxena, A., Schmidl, C., Schneider, C., Schulze-Tanzil, G. G., Schwegmann, A., Sheng, G., Shin, J. W., Sugiyama, D., Sugiyama, T., Summers, K. M., Takahashi, N., Takai, J., Tanaka, H., Tatsukawa, H., Tomoiu, A., Toyoda, H., van de Wetering, M., van den Berg, L. M., Verardo, R., Vijayan, D., Wells, C. A., Winteringham, L. N., Wolvetang, E., Yamaguchi, Y., Yamamoto, M., Yanagi-Mizuochi, C., Yoneda, M., Yonekura, Y., Zhang, P. G., 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, C. O., Kasukawa, T., Kawaji, H., Carninci, P., Forrest, A. R. R. & Hayashizaki, Y. FANTOM5 CAGE profiles of human and mouse samples. ***Sci. data*** 4, 170112 (2017).

1. Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen,

T. S. & Thomson, J. A. The NIH Roadmap Epigenomics Mapping Consortium. ***Nat. Biotechnol.*** 28, 1045–8 (2010).

1. Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K. Y., Rozowsky, J., Yan, K.-K., Dong, X., Djebali, S.,

Ruan, Y., Davis, C. A., Carninci, P., Lassman, T., Gingeras, T. R., Guigó, R., Birney, E., Weng, Z., Snyder, M. & Gerstein, M. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. ***Genome Res.*** 22, 1658–67 (2012).

1. Lou, S., Lee, H.-M., Qin, H., Li, J.-W., Gao, Z., Liu, X., Chan, L. L., Kl Lam, V., So, W.-Y., Wang, Y., Lok,

S., Wang, J., Ma, R. C., Tsui, S. K.-W., Chan, J. C., Chan, T.-F. & Yip, K. Y. Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. ***Genome Biol.*** 15, 408 (2014).

1. Li, L., Gao, Y., Wu, Q., Cheng, A. S. L. & Yip, K. Y. New guidelines for DNA methylome studies regarding 5-hydroxymethylcytosine for understanding transcriptional regulation. ***Genome Res.*** 29, 543– 553 (2019).
2. Fu, A. X., Lui, K. N.-C., Tang, C. S.-M., Ng, R. K., Lai, F. P.-L., Lau, S.-T., Li, Z., Garcia-Barcelo, M.-M.,

Sham, P.-C., Tam, P. K.-H., Ngan, E. S.-W. & Yip, K. Y. Whole-genome analysis of noncoding genetic variations identifies multiscale regulatory element perturbations associated with Hirschsprung disease. ***Genome Res.*** 30, 1618–1632 (2020).

1. Cao, Q., Zhang, Z., Fu, A. X., Wu, Q., Lee, T.-L., Lo, E., Cheng, A. S. L., Cheng, C., Leung, D. & Yip, K.

Y. A unified framework for integrative study of heterogeneous gene regulatory mechanisms. ***Nat. Mach. Intell.*** 2, 447–456 (2020).

9. Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C. P., Clarke, D., Gu, M., Emani, P., Yang, Y.

T., Xu, M., Gandal, M. J., Lou, S., Zhang, J., Park, J. J., Yan, C., Rhie, S. K., Manakongtreecheep, K., Zhou, H., Nathan, A., Peters, M., Mattei, E., Fitzgerald, D., Brunetti, T., Moore, J., Jiang, Y., Girdhar, K., Hoffman, G. E., Kalayci, S., Gümüş, Z. H., Crawford, G. E., PsychENCODE Consortium, Roussos, P., Akbarian, S., Jaffe, A. E., White, K. P., Weng, Z., Sestan, N., Geschwind, D. H., Knowles, J. A. & Gerstein, M. B. Comprehensive functional genomic resource and integrative model for the human brain. ***Science*** 362, (2018).

1. Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., 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, U. S., Flicek, P., Fragoza, R., Garrison, E., Gibbs, R., Gümüş, Z. H., Herrero, J., Kitabayashi, N., Kong, Y., Lage, K., Liluashvili, V., Lipkin, S. M., MacArthur, D. G., Marth, G., Muzny, D., Pers, T. H., Ritchie, G. R. S., Rosenfeld, J. A., Sisu, C., Wei, X., Wilson, M., Xue, Y., Yu, F., 1000 Genomes Project Consortium, Dermitzakis, E. T., Yu, H., Rubin, M. A., Tyler-Smith, C. & Gerstein, M. Integrative annotation of variants from 1092 humans: application to cancer genomics. ***Science*** 342, 1235587 (2013).
2. Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X. J., Yip, K. Y., Khurana, E. & Gerstein, M. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. ***Genome Biol.*** 15, 480 (2014).
3. Farnham, P. J. Insights from genomic profiling of transcription factors. ***Nat. Rev. Genet.*** 10, 605–16 (2009).
4. Berger, S. L. The complex language of chromatin regulation during transcription. ***Nature*** 447, 407–12 (2007).
5. Dong, X., Greven, M. C., Kundaje, A., Djebali, S., Brown, J. B., Cheng, C., Gingeras, T. R., Gerstein, M., Guigó, R., Birney, E. & Weng, Z. Modeling gene expression using chromatin features in various cellular contexts. ***Genome Biol.*** 13, R53 (2012).
6. 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.*** 40, 553–68 (2012).
7. Cao, Q., Anyansi, C., Hu, X., Xu, L., Xiong, L., Tang, W., Mok, M. T. S., Cheng, C., Fan, X., Gerstein, M., Cheng, A. S. L. & Yip, K. Y. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. ***Nat. Genet.*** 49, 1428–1436 (2017).
8. Lou, S., Cotter, K. A., Li, T., Liang, J., Mohsen, H., Liu, J., Zhang, J., Cohen, S., Xu, J., Yu, H., Rubin,

M. A. & Gerstein, M. GRAM: A GeneRAlized Model to predict the molecular effect of a non-coding variant in a cell-type specific manner. ***PLoS Genet.*** 15, e1007860 (2019).

1. Zhang, J., Liu, J., Lee, D., Feng, J.-J., 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.*** 21, 151 (2020).
2. Onuchic, V., Lurie, E., Carrero, I., Pawliczek, P., Patel, R. Y., Rozowsky, J., Galeev, T., Huang, Z., Altshuler, R. C., Zhang, Z., Harris, R. A., Coarfa, C., Ashmore, L., Bertol, J. W., Fakhouri, W. D., Yu, F., Kellis, M., Gerstein, M. & Milosavljevic, A. Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. ***Science*** 361, (2018).
3. Chen, J., Rozowsky, J., Galeev, T. R., 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.*** 7, 11101 (2016).
4. 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.*** 7, 522 (2011).
5. Warrell, J., Mohsen, H. & Gerstein, M. Rank Projection Trees for Multilevel Neural Network Interpretation. in ***Proceedings of the Machine Learning for Health (ML4H) Workshop at NeurIPS 2018*** (2018).
6. Lou, S., Li, T., Spakowicz, D., Yan, X., Chupp, G. L. & Gerstein, M. Latent-space embedding of expression data identifies gene signatures from sputum samples of asthmatic patients. ***BMC Bioinformatics*** 21, 457 (2020).
7. Kuksa, P. P., Min, M. R., Dugar, R. & Gerstein, M. High-order neural networks and kernel methods for peptide-MHC binding prediction. ***Bioinformatics*** 31, 3600–7 (2015).
8. Min, M. R., Xia, N., Cheng, C. & Gerstein, M. Interpretable sparse high-order Boltzmann machines. in ***Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*** (2014).
9. Min, M. R., Ning, X., Qi, Y., Cheng, C., Bonner, A. & Gerstein, M. Ensemble learning based sparse high- order Boltzmann machine for unsupervised feature interaction identification. in ***NIPS Workshop on Machine Learning in Computational Biology*** (2014).
10. Lee, D., Zhang, J., Liu, J. & Gerstein, M. Epigenome-based splicing prediction using a recurrent neural network. ***PLoS Comput. Biol.*** 16, e1008006 (2020).
11. Hong, C., Cao, Q., Zhang, Z., Tsui, S. K.-W. & Yip, K. Y. Reusability report: Capturing properties of biological objects and their relationships using graph neural networks. ***Nat. Mach. Intell.*** 4, 222–226 (2022).
12. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. in ***Proceedings of the Workshop of the 5th International Conference on Learning Representations (ICLR)*** (2017).
13. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. An introduction to variational methods for graphical models. ***Mach. Learn.*** 37, 183–233 (1999).
14. Warrell, J., Mohsen, H. & Gerstein, M. Compression-based network interpretability schemes. ***bioRxiv***

(2020). doi:10.1101/2020.10.27.358226

1. Zhou, W., Veitch, V., Austern, M., Adams, R. P. & Orbanz, P. Non-vacuous generalization bounds at the ImageNet scale: A PAC-Bayesian compression approach. in ***The 7th International Conference on Learning Representations*** (2019). doi:https://doi.org/10.48550/arXiv.1804.05862
2. Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V, Lo, C. C., McDonald, E. R., Barretina, J., Gelfand, E. T., Bielski, C. M., Li, H., Hu, K., Andreev-Drakhlin, A. Y., Kim, J., Hess, J. M., Haas, B. J.,

Aguet, F., Weir, B. A., Rothberg, M. V, Paolella, B. R., Lawrence, M. S., Akbani, R., Lu, Y., Tiv, H. L.,

Gokhale, P. C., de Weck, A., Mansour, A. A., Oh, C., Shih, J., Hadi, K., Rosen, Y., Bistline, J.,

Venkatesan, K., Reddy, A., Sonkin, D., Liu, M., Lehar, J., Korn, J. M., Porter, D. A., Jones, M. D., Golji,

J., Caponigro, G., Taylor, J. E., Dunning, C. M., Creech, A. L., Warren, A. C., McFarland, J. M., Zamanighomi, M., Kauffmann, A., Stransky, N., Imielinski, M., Maruvka, Y. E., Cherniack, A. D., Tsherniak, A., Vazquez, F., Jaffe, J. D., Lane, A. A., Weinstock, D. M., Johannessen, C. M., Morrissey,

M. P., Stegmeier, F., Schlegel, R., Hahn, W. C., Getz, G., Mills, G. B., Boehm, J. S., Golub, T. R., Garraway, L. A. & Sellers, W. R. Next-generation characterization of the Cancer Cell Line Encyclopedia. ***Nature*** 569, 503–508 (2019).

1. Denzler, J. & Brown, C. M. Information theoretic sensor data selection for active object recognition and state estimation. ***IEEE Trans. Pattern Anal. Mach. Intell.*** 24, 145–157 (2002).
2. Sachs, K., Perez, O., Pe’er, D., Lauffenburger, D. A. & Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. ***Science*** 308, 523–9 (2005).
3. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. ***PLoS Comput. Biol.*** 11, e1004219 (2015).
4. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. in ***Proceedings of the 31st International Conference on Neural Information Processing Systems*** 4768–4777 (2017).
5. Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M. & Stark, A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. ***Science*** 339, 1074–7 (2013).
6. Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., 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.*** 13, R48 (2012).
7. Sethi, A., Gu, M., Gumusgoz, E., Chan, L., Yan, K.-K., Rozowsky, J., Barozzi, I., Afzal, V., Akiyama, J.

A., Plajzer-Frick, I., Yan, C., Novak, C. S., Kato, M., Garvin, T. H., Pham, Q., Harrington, A., Mannion, B.

J., Lee, E. A., Fukuda-Yuzawa, Y., Visel, A., Dickel, D. E., Yip, K. Y., Sutton, R., Pennacchio, L. A. & Gerstein, M. Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. ***Nat. Methods*** 17, 807–814 (2020).

1. Chen, Z., Zhang, J., Liu, J., Zhang, Z., Zhu, J., Lee, D., Xu, M. & Gerstein, M. SCAN-ATAC-Sim: a scalable and efficient method for simulating single-cell ATAC-seq data from bulk-tissue experiments. ***Bioinformatics*** (2021). doi:10.1093/bioinformatics/btaa1039
2. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F. & Gerstein, M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. ***Science*** 302, 449–53 (2003).
3. Yang, C.-C., Andrews, E. H., Chen, M.-H., Wang, W.-Y., Chen, J. J. W., Gerstein, M., Liu, C.-C. & Cheng, C. iTAR: a web server for identifying target genes of transcription factors using ChIP-seq or

ChIP-chip data. ***BMC Genomics*** 17, 632 (2016).

1. Hu, X., Wong, T. K. F., Lu, Z. J., Chan, T. F., Lau, T. C. K., Yiu, S. M. & Yip, K. Y. Computational identification of protein binding sites on RNAs using high-throughput RNA structure-probing data. ***Bioinformatics*** 30, 1049–1055 (2014).
2. Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R. K., Feng, X., Leng, J., Vielle, A., Niu, W., Rhrissorrakrai, K., Agarwal, A., Alexander, R. P., Barber, G., Brdlik, C. M., Brennan, J., Brouillet, J. J., Carr, A., Cheung, M.-S., Clawson, H., Contrino, S., Dannenberg, L. O.,

Dernburg, A. F., Desai, A., Dick, L., Dosé, A. C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing,

B., Feingold, E. A., Gassmann, R., Good, P. J., Green, P., Gullier, F., Gutwein, M., Guyer, M. S.,

Habegger, L., Han, T., Henikoff, J. G., Henz, S. R., Hinrichs, A., Holster, H., Hyman, T., Iniguez, A. L.,

Janette, J., Jensen, M., Kato, M., Kent, W. J., Kephart, E., Khivansara, V., Khurana, E., Kim, J. K., Kolasinska-Zwierz, P., Lai, E. C., Latorre, I., Leahey, A., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R. F., Lubling, Y., Lyne, R., MacCoss, M., Mackowiak, S. D., Mangone, M., McKay, S., Mecenas, D.,

Merrihew, G., Miller, D. M., Muroyama, A., Murray, J. I., Ooi, S.-L., Pham, H., Phippen, T., Preston, E. A., Rajewsky, N., Rätsch, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, A., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F. J., Slightam, C., Smith, R., Spencer, W. C., Stinson, E. O., Taing, S., Takasaki, T., Vafeados, D., Voronina, K., Wang, G.,

Washington, N. L., Whittle, C. M., Wu, B., Yan, K.-K., Zeller, G., Zha, Z., Zhong, M., Zhou, X., modENCODE Consortium, Ahringer, J., Strome, S., Gunsalus, K. C., Micklem, G., Liu, X. S., Reinke, V., Kim, S. K., Hillier, L. W., Henikoff, S., Piano, F., Snyder, M., Stein, L., Lieb, J. D. & Waterston, R. H. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. ***Science*** 330, 1775–87 (2010).

1. Cheng, C., Andrews, E., Yan, K.-K., Ung, M., Wang, D. & Gerstein, M. An approach for determining and measuring network hierarchy applied to comparing the phosphorylome and the regulome. ***Genome Biol.*** 16, 63 (2015).
2. Zhang, J., Liu, J., Lee, D., Lou, S., Chen, Z., Gürsoy, G. & Gerstein, M. DiNeR: a Differential graphical model for analysis of co-regulation Network Rewiring. ***BMC Bioinformatics*** 21, 281 (2020).
3. Lou, S., Li, T., Kong, X., Zhang, J., Liu, J., Lee, D. & Gerstein, M. TopicNet: a framework for measuring transcriptional regulatory network change. ***Bioinformatics*** 36, i474–i481 (2020).
4. 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, K.- K., Sethi, A., Gürsoy, G., Schoenberg, M. R., Rozowsky, J., Warrell, J., Emani, P., Yang, Y. T., Galeev, T., Kong, X., Liu, S., Li, X., Krishnan, J., Feng, Y., Rivera-Mulia, J. C., Adrian, J., Broach, J. R., 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, R. J., Snyder, M., Gilbert, D. M., Yip, K., Cheng, C., Yue, F., Liu, X. S., White,

K. P. & Gerstein, M. An integrative ENCODE resource for cancer genomics. ***Nat. Commun.*** 11, 3696 (2020).

1. Pang, B. & Snyder, M. P. Systematic identification of silencers in human cells. ***Nat. Genet.*** 52, 254–263 (2020).
2. Bao, L., Zhou, M. & Cui, Y. CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators. ***Nucleic Acids Res.*** 36, D83-7 (2008).
3. Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., Zhang, M. Q., Lobanenkov, V. V & Ren, B. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. ***Cell*** 128, 1231–45 (2007).
4. Zhuang, Z., Shen, X. & Pan, W. A simple convolutional neural network for prediction of enhancer- promoter interactions with DNA sequence data. ***Bioinformatics*** 35, 2899–2906 (2019).
5. Singh, S., Yang, Y., Póczos, B. & Ma, J. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. ***Quant. Biol. (Beijing, China)*** 7, 122–137 (2019).
6. Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., Nguyen, T. H., Kane, M., Perez, E. M., Durand, N. C., Lareau, C. A., Stamenova, E. K., Aiden, E. L., Lander, E. S. & Engreitz, J. M. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. ***Nat. Genet.*** 51, 1664–1669 (2019).
7. Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. Bin, Orlov, Y. L., Velkov, S., Ho, A.,

Mei, P. H., Chew, E. G. Y., Huang, P. Y. H., Welboren, W.-J., Han, Y., Ooi, H. S., Ariyaratne, P. N.,

Vega, V. B., Luo, Y., Tan, P. Y., Choy, P. Y., Wansa, K. D. S. A., Zhao, B., Lim, K. S., Leow, S. C., Yow,

J. S., Joseph, R., Li, H., Desai, K. V, Thomsen, J. S., Lee, Y. K., Karuturi, R. K. M., Herve, T., Bourque,

G., Stunnenberg, H. G., Ruan, X., Cacheux-Rataboul, V., Sung, W.-K., Liu, E. T., Wei, C.-L., Cheung, E. & Ruan, Y. An oestrogen-receptor-alpha-bound human chromatin interactome. ***Nature*** 462, 58–64 (2009).

1. Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J. & Chang, H. Y. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. ***Nat. Methods*** 13, 919– 922 (2016).
2. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. in ***International Conference on Learning Representations Workshops*** (2013).
3. Yun, S., Jeong, M., Kim, R., Kang, J. & Kim, H. J. Graph Transformer Networks. in ***Advances in Neural Information Processing Systems*** (2019).
4. Schreiber, J., Singh, R., Bilmes, J. & Noble, W. S. A pitfall for machine learning methods aiming to predict across cell types. ***Genome Biol.*** 21, 282 (2020).
5. Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. ***Nat. Rev. Genet.*** 23, 169–181 (2022).
6. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C. & Sun, M. Graph neural networks: A review of methods and applications. ***AI Open*** 1, 57–81 (2020).
7. Lee, D., Shi, M., Moran, J., Wall, M., Zhang, J., Liu, J., Fitzgerald, D., Kyono, Y., Ma, L., White, K. P. & Gerstein, M. STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. ***Genome Biol.*** 21, 298 (2020).
8. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments.

***Ann. Appl. Stat.*** 5, (2011).

1. Arbel, H., Basu, S., Fisher, W. W., Hammonds, A. S., Wan, K. H., Park, S., Weiszmann, R., Booth, B. W., Keranen, S. V, Henriquez, C., Shams Solari, O., Bickel, P. J., Biggin, M. D., Celniker, S. E. & Brown,

J. B. Exploiting regulatory heterogeneity to systematically identify enhancers with high accuracy. ***Proc. Natl. Acad. Sci. U. S. A.*** 116, 900–908 (2019).

1. Breiman, L. Random forests. ***Mach. Learn.*** 45, 5–32 (2001).
2. Byun, J., Han, Y., Li, Y., Xia, J., Long, E., Choi, J., Xiao, X., Zhu, M., Zhou, W., Sun, R., Bossé, Y.,

Song, Z., Schwartz, A., Lusk, C., Rafnar, T., Stefansson, K., Zhang, T., Zhao, W., Pettit, R. W., Liu, Y.,

Li, X., Zhou, H., Walsh, K. M., Gorlov, I., Gorlova, O., Zhu, D., Rosenberg, S. M., Pinney, S., Bailey- Wilson, J. E., Mandal, D., de Andrade, M., Gaba, C., Willey, J. C., You, M., Anderson, M., Wiencke, J. K., Albanes, D., Lam, S., Tardon, A., Chen, C., Goodman, G., Bojeson, S., Brenner, H., Landi, M. T., Chanock, S. J., Johansson, M., Muley, T., Risch, A., Wichmann, H.-E., Bickeböller, H., Christiani, D. C., Rennert, G., Arnold, S., Field, J. K., Shete, S., Le Marchand, L., Melander, O., Brunnstrom, H., Liu, G., Andrew, A. S., Kiemeney, L. A., Shen, H., Zienolddiny, S., Grankvist, K., Johansson, M., Caporaso, N., Cox, A., Hong, Y.-C., Yuan, J.-M., Lazarus, P., Schabath, M. B., Aldrich, M. C., Patel, A., Lan, Q.,

Rothman, N., Taylor, F., Kachuri, L., Witte, J. S., Sakoda, L. C., Spitz, M., Brennan, P., Lin, X., McKay, J., Hung, R. J. & Amos, C. I. Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer. ***Nat. Genet.*** 54, 1167–1177 (2022).

1. Boyle, A. P., Araya, C. L., Brdlik, C., Cayting, P., Cheng, C., Cheng, Y., Gardner, K., Hillier, L. W.,

Janette, J., Jiang, L., Kasper, D., Kawli, T., Kheradpour, P., Kundaje, A., Li, J. J., Ma, L., Niu, W., Rehm,

E. J., Rozowsky, J., Slattery, M., Spokony, R., Terrell, R., Vafeados, D., Wang, D., Weisdepp, P., Wu, Y.-C., Xie, D., Yan, K.-K., Feingold, E. A., Good, P. J., Pazin, M. J., Huang, H., Bickel, P. J., Brenner, S. E., Reinke, V., Waterston, R. H., Gerstein, M., White, K. P., Kellis, M. & Snyder, M. Comparative analysis of regulatory information and circuits across distant species. ***Nature*** 512, 453–6 (2014).

1. Gerstein, M. B., Rozowsky, J., Yan, K.-K., Wang, D., Cheng, C., Brown, J. B., Davis, C. A., Hillier, L.,

Sisu, C., Li, J. J., Pei, B., Harmanci, A. O., Duff, M. O., Djebali, S., Alexander, R. P., Alver, B. H.,

Auerbach, R., Bell, K., Bickel, P. J., Boeck, M. E., Boley, N. P., Booth, B. W., Cherbas, L., Cherbas, P.,

Di, C., Dobin, A., Drenkow, J., Ewing, B., Fang, G., Fastuca, M., Feingold, E. A., Frankish, A., Gao, G.,

Good, P. J., Guigó, R., Hammonds, A., Harrow, J., Hoskins, R. A., Howald, C., Hu, L., Huang, H.,

Hubbard, T. J. P., Huynh, C., Jha, S., Kasper, D., Kato, M., Kaufman, T. C., Kitchen, R. R., Ladewig, E.,

Lagarde, J., Lai, E., Leng, J., Lu, Z., MacCoss, M., May, G., McWhirter, R., Merrihew, G., Miller, D. M., Mortazavi, A., Murad, R., Oliver, B., Olson, S., Park, P. J., Pazin, M. J., Perrimon, N., Pervouchine, D., Reinke, V., Reymond, A., Robinson, G., Samsonova, A., Saunders, G. I., Schlesinger, F., Sethi, A., Slack, F. J., Spencer, W. C., Stoiber, M. H., Strasbourger, P., Tanzer, A., Thompson, O. A., Wan, K. H.,

Wang, G., Wang, H., Watkins, K. L., Wen, J., Wen, K., Xue, C., Yang, L., Yip, K., Zaleski, C., Zhang, Y., Zheng, H., Brenner, S. E., Graveley, B. R., Celniker, S. E., Gingeras, T. R. & Waterston, R. Comparative analysis of the transcriptome across distant species. ***Nature*** 512, 445–8 (2014).

1. Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., 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, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk,

B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., 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, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R. & Gingeras, T. R. Landscape of transcription in human cells. ***Nature*** 489, 101–8 (2012).

1. Kumar, S., Warrell, J., Li, S., McGillivray, P. D., Meyerson, W., Salichos, L., Harmanci, A., Martinez- Fundichely, A., Chan, C. W. Y., Nielsen, M. M., Lochovsky, L., Zhang, Y., Li, X., Lou, S., Pedersen, J. S., Herrmann, C., Getz, G., Khurana, E. & Gerstein, M. B. Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. ***Cell*** 180, 915-927.e16 (2020).
2. Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Mu, X. J., 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, J. M., Kong, Y.,

Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny,

D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll,

S. A., 1000 Genomes Project Consortium, Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E. & Korbel, J. O. An integrated map of structural variation in 2,504 human genomes. ***Nature*** 526, 75–81 (2015).

1. Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., Dehal, P., Ware,

D., Perez, F., Canon, S., Sneddon, M. W., Henderson, M. L., Riehl, W. J., Murphy-Olson, D., Chan, S.

Y., Kamimura, R. T., Kumari, S., Drake, M. M., Brettin, T. S., Glass, E. M., Chivian, D., Gunter, D.,

Weston, D. J., Allen, B. H., Baumohl, J., Best, A. A., Bowen, B., Brenner, S. E., Bun, C. C., Chandonia,

J.-M., Chia, J.-M., Colasanti, R., Conrad, N., Davis, J. J., Davison, B. H., DeJongh, M., Devoid, S., Dietrich, E., Dubchak, I., Edirisinghe, J. N., Fang, G., Faria, J. P., Frybarger, P. M., Gerlach, W., Gerstein, M., Greiner, A., Gurtowski, J., Haun, H. L., He, F., Jain, R., Joachimiak, M. P., Keegan, K. P.,

Kondo, S., Kumar, V., Land, M. L., Meyer, F., Mills, M., Novichkov, P. S., Oh, T., Olsen, G. J., Olson, R., Parrello, B., Pasternak, S., Pearson, E., Poon, S. S., Price, G. A., Ramakrishnan, S., Ranjan, P., Ronald, P. C., Schatz, M. C., Seaver, S. M. D., Shukla, M., Sutormin, R. A., Syed, M. H., Thomason, J., Tintle, N. L., Wang, D., Xia, F., Yoo, H., Yoo, S. & Yu, D. KBase: The United States Department of Energy Systems Biology Knowledgebase. ***Nat. Biotechnol.*** 36, 566–569 (2018).

1. Murillo, O. D., Thistlethwaite, W., Rozowsky, J., Subramanian, S. L., Lucero, R., Shah, N., Jackson, A. R., Srinivasan, S., Chung, A., Laurent, C. D., Kitchen, R. R., Galeev, T., Warrell, J., Diao, J. A., Welsh,

J. A., Hanspers, K., Riutta, A., Burgstaller-Muehlbacher, S., Shah, R. V, Yeri, A., Jenkins, L. M., Ahsen,

M. E., Cordon-Cardo, C., Dogra, N., Gifford, S. M., Smith, J. T., Stolovitzky, G., Tewari, A. K., Wunsch,

B. H., Yadav, K. K., Danielson, K. M., Filant, J., Moeller, C., Nejad, P., Paul, A., Simonson, B., Wong, D.

K., Zhang, X., Balaj, L., Gandhi, R., Sood, A. K., Alexander, R. P., Wang, L., Wu, C., Wong, D. T. W.,

Galas, D. J., Van Keuren-Jensen, K., Patel, T., Jones, J. C., Das, S., Cheung, K.-H., Pico, A. R., Su, A. I., Raffai, R. L., Laurent, L. C., Roth, M. E., Gerstein, M. B. & Milosavljevic, A. exRNA Atlas Analysis Reveals Distinct Extracellular RNA Cargo Types and Their Carriers Present across Human Biofluids.

***Cell*** 177, 463-477.e15 (2019).

1. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. ***Nature*** 578, 82–93 (2020).
2. Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. ***Cell*** 167, 1145–1149 (2016).
3. 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.*** 15, 474 (2014).
4. Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M. & Gerstein, M. B. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. ***Nat. Biotechnol.*** 27, 66–75 (2009).
5. Du, J., Bjornson, R. D., Zhang, Z. D., Kong, Y., Snyder, M. & Gerstein, M. B. Integrating sequencing technologies in personal genomics: optimal low cost reconstruction of structural variants. ***PLoS Comput. Biol.*** 5, e1000432 (2009).
6. Emanuelsson, O., Nagalakshmi, U., Zheng, D., Rozowsky, J. S., Urban, A. E., Du, J., Lian, Z., Stolc, V., Weissman, S., Snyder, M. & Gerstein, M. B. Assessing the performance of different high-density tiling microarray strategies for mapping transcribed regions of the human genome. ***Genome Res.*** 17, 886–97 (2007).
7. Tran, D., Hoffman, M. D., Saurous, R. A., Brevdo, E., Murphy, K. & Blei, D. M. Deep probabilistic programming. in ***Proceedings of the 5th International Conference on Learning Representations (ICLR)*** (2017). doi:https://doi.org/10.48550/arXiv.1701.0375