Gerstein lab experience in leading consortium data analysis centers to build data processing pipelines. We have extensive experience developing QC metrics and uniform processing pipelines for consortia (ENCODE[1, 2], PyschENCODE[3, 4], and the Extracellular RNA Communication Project (exRNA))[5, 6]. We have lead bake-off style analysis comparisons for applications such as enhancer prediction for ENCODE and RNA-seq quantification for the RNA-seq Genome Annotation Assessment Project[7]. As part of the exRNA consortium, we developed the extracellular RNA processing toolkit (exceRpt) pipeline (**Fig. 1**) for uniform processing throughout the consortium[5, 6]. exceRpt has been used over 80,000 times. exceRpt performs sequential alignment of RNA to contaminants, to human transcriptome and genome sequences, to human repetitive elements, and finally to exogenous sequences.

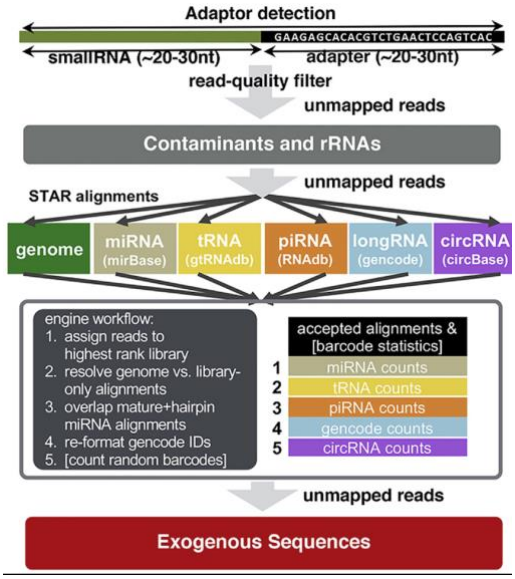

Figure 1. Flowchart of exceRpt.

Prior lab experience processing bulk and single-cell chromatin accessibility data analyses. We have extensive experience processing epigenetic profiling data in big consortia. For example, we developed (1) PeakSeq[8] for the genome-wide identification of transcription factor (TF) binding sites from ChIP-Seq data, which is used by ENCODE; and (2) MUSIC[9], a peak caller that performs multiscale decomposition of ChIP-Seq signal. For scATAC-seq data, we have recently developed a prototype of a stand-alone scATAC-seq data processing pipeline.

Experience with constructing the PsychENCODE brain cell atlas. We led the PsychENCODE data analysis center's effort to construct a cell atlas of prefrontal cortex in human brain[3]. Specifically, we carefully developed a way of clustering cells by incorporating pre-existing clusters to ensure that our new clusters were consistent with the old clusters.

Experience with deconvolution methods to detect cell proportion changes. We previously investigated how changes in cell proportions contribute to variations in tissue-level gene expression across individuals[3]. Adding publicly available data, we first used NMF to decompose bulk tissue data. We found that the top principal components correlated with cell expression signatures, suggesting single-cell contributions to the bulk tissue gene expression. We then deconvolved the bulk tissue expression across 1,866 individuals in PsychENCODE and GTEx using single-cell data via non-negative least squares and found that the cell fraction changes were associated with aging and disorders.

Tools to detect rewiring of regulatory networks. We have extensive experience conducting network comparisons. We developed many tools for comparative gene regulatory network analyses, including Loregic[10], which analyzes regulatory cooperativity; OrthoClust[11], which discovers novel human gene functions via clustering cross-species gene co-expression networks; DREISS[12], which analyzes the dynamics of gene regulatory networks using dynamic models. We also identified the rewiring pattern for human gene regulatory networks to prioritize genes for human diseases[13].

Our deep learning models for GWAS analysis. We developed an interpretable Deep Structured Phenotype Network (DSPN)[3], a deep learning model based conditional Deep Boltzmann Machine architecture with multiple layers. DSPN improved GWAS disease prediction *by sixfold compared to additive polygenic risk scores*. DSPN highlights key genes and allows imputation of missing transcriptome from genotype.

We have extensive experience developing pipelines and characterizing data quality for transcriptomic[14-17], extracellular RNA[5], proteomic[18-20], genomic[21-23], and protein interaction data[24], which have been adopted by major consortia[25-27]. We also have experience with developing software for processing RNA-seq data in

general. For instance, our software RSEQtools[14] provides an efficient package for basic RNA-Seq data that uses a compact data summary format, the Mapped Read Format (MRF) that enables anonymization; and, more recently, a privacy BAM format that may minimize the leakage of genetic information from single cell RNA-Seq data[28].

Finally, the Gerstein lab has extensive experience developing and actively maintaining a well-catalogued and well-organized lab frequently asked questions (FAQ) page using WordPress (faq.gersteinlab.org), wherein queries are emailed to the PI by other research groups, and the anonymized questions are then publicly posted (along with detailed answers) on this FAQ page.

References

1. Consortium, E.P., An integrated encyclopedia of DNA elements in the human genome. Nature, 2012. 489(7414): p. 57-74.

2. Gerstein, M.B., A. Kundaje, M. Hariharan, S.G. Landt, et al., Architecture of the human regulatory network derived from ENCODE data. Nature, 2012. 489(7414): p. 91-100.

3. Wang, D., S. Liu, J. Warrell, H. Won, et al., Comprehensive functional genomic resource and integrative model for the human brain. Science, 2018. 362(6420).

4. Akbarian, S., C. Liu, J.A. Knowles, F.M. Vaccarino, et al., The PsychENCODE project. Nat Neurosci, 2015. 18(12): p. 1707-12.

5. Rozowsky, J., R.R. Kitchen, J.J. Park, T.R. Galeev, et al., exceRpt: A Comprehensive Analytic Platform for Extracellular RNA Profiling. Cell Syst, 2019. 8(4): p. 352-357 e3.

6. Das, S., R.N.A.C.C. Extracellular, K.M. Ansel, M. Bitzer, et al., The Extracellular RNA Communication Consortium: Establishing Foundational Knowledge and Technologies for Extracellular RNA Research. Cell, 2019. 177(2): p. 231-242.

7. Steijger, T., J.F. Abril, P.G. Engstrom, F. Kokocinski, et al., Assessment of transcript reconstruction methods for RNA-seq. Nat Methods, 2013. 10(12): p. 1177-84.

8. Rozowsky, J., G. Euskirchen, R.K. Auerbach, Z.D. Zhang, et al., PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol, 2009. 27(1): p. 66-75.

9. Harmanci, A., J. Rozowsky, and M. Gerstein, MUSIC: Identification of Enriched Regions in ChIP-Seq Experiments using a Mappability-Corrected Multiscale Signal Processing Framework. Genome Biol, 2014. 15(10): p. 474.

10. Wang, D., K.K. Yan, C. Sisu, C. Cheng, et al., Loregic: a method to characterize the cooperative logic of regulatory factors. PLoS Comput Biol, 2015. 11(4): p. e1004132.

11. Yan, K.K., D. Wang, J. Rozowsky, H. Zheng, et al., OrthoClust: an orthology-based network framework for clustering data across multiple species. Genome Biol, 2014. 15(8): p. R100.

12. Wang, D., F. He, S. Maslov, and M. Gerstein, DREISS: Using State-Space Models to Infer the Dynamics of Gene Expression Driven by External and Internal Regulatory Networks. PLoS Comput Biol, 2016. 12(10): p. e1005146.

13. Zhang, J., D. Lee, V. Dhiman, P. Jiang, et al., 2019.

14. Habegger, L., A. Sboner, T.A. Gianoulis, J. Rozowsky, et al., RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. Bioinformatics, 2011. 27(2): p. 281-3.

15. Jee, J., J. Rozowsky, K.Y. Yip, L. Lochovsky, et al., ACT: aggregation and correlation toolbox for analyses of genome tracks. Bioinformatics, 2011. 27(8): p. 1152-4.

16. Rozowsky, J.S., D. Newburger, F. Sayward, J. Wu, et al., The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci. Genome Res, 2007. 17(6): p. 732-45.

17. Lu, Z.J., K.Y. Yip, G. Wang, C. Shou, et al., 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): p. 276-85.

18. Sboner, A., A. Karpikov, G. Chen, M. Smith, et al., Robust-linear-model normalization to reduce technical variability in functional protein microarrays. J Proteome Res, 2009. 8(12): p. 5451-64.

19. Vidal, M., D.W. Chan, M. Gerstein, M. Mann, et al., The human proteome - a scientific opportunity for transforming diagnostics, therapeutics, and healthcare. Clin Proteomics, 2012. 9(1): p. 6.

20. Carlyle, B.C., R.R. Kitchen, J.E. Kanyo, E.Z. Voss, et al., A multiregional proteomic survey of the postnatal human brain. Nat Neurosci, 2017. 20(12): p. 1787-1795.

21. Kitchen, R.R., J.S. Rozowsky, M.B. Gerstein, and A.C. Nairn, Decoding neuroproteomics: integrating the genome, translatome and functional anatomy. Nat Neurosci, 2014. 17(11): p. 1491-9.

22. Carlyle, B.C., R.R. Kitchen, J. Zhang, R.S. Wilson, et al., Isoform-Level Interpretation of High-Throughput Proteomics Data Enabled by Deep Integration with RNA-seq. J Proteome Res, 2018. 17(10): p. 3431-3444.

23. Yu, E.Z., A.E. Burba, and M. Gerstein, PARE: a tool for comparing protein abundance and mRNA expression data. BMC Bioinformatics, 2007. 8: p. 309.

24. Lin, N., B. Wu, R. Jansen, M. Gerstein, et al., Information assessment on predicting protein-protein interactions. BMC Bioinformatics, 2004. 5: p. 154.

25. Gerstein, M.B., Z.J. Lu, E.L. Van Nostrand, C. Cheng, et al., Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science, 2010. 330(6012): p. 1775-87.

26. Gerstein, M.B., J. Rozowsky, K.K. Yan, D. Wang, et al., Comparative analysis of the transcriptome across distant species. Nature, 2014. 512(7515): p. 445-8.

27. Greenbaum, D., C. Colangelo, K. Williams, and M. Gerstein, Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol, 2003. 4(9): p. 117.

28. Gürsoy, G., P. Emani, O.A. Jolanki, C.M. Brannon, et al., 2019.