**Gerstein lab experience in studying QTLs**

The Gerstein lab has extensive experience in eQTL analysis. Dr. Gerstein’s team is part of the PsychENCODE Consortium which has generated a comprehensive online resource for the adult brain across 1,866 individuals (Psych *et al.* 2015). This resource contains ~79,000 brain-active enhancers, sets of Hi-C linkages, and topologically associating domains; single-cell expression profiles for many cell types; eQTLs; and further QTLs associated with chromatin, splicing, and cell-type proportions (so-called cell fraction QTLs, or fQTLs).

We have extensive experience in analyzing single-cell data for inferring cell-type-specific eQTLs. In a paper under revision, we integrated single-nucleus, multi-omics datasets from the PsychENCODE consortium to create a uniformly processed resource comprising >2.8M nuclei from the prefrontal cortex across 388 individuals. Across 17 distinct cell types, we identified >1.4M single-cell eQTLs (termed “scQTLs”), with an average of >85K cis-eQTLs per cell type and ~690 eGenes per cell type. Our scQTLs are strongly enriched in narrow regions around the transcription start sites, and many of them are cell-type-specific (though ~47% of them appear in more than one cell type). To perform our scQTLs search, we followed the same general procedure used by GTEx, including conservative filtering on the cell-type level when generating pseudobulk data. We used this set of scQTLs as our "core callset" for downstream analyses. Furthermore, ~30% of the scQTLs overlap with bulk cis-eQTLs

Note the sparsity intrinsic to snRNA-seq data reduces power, particularly for the rarer cell types. Thus, in our single-cell eQTL analysis, we implemented a Bayesian linear mixed-effects model to identify more scQTLs for these rare cell types, which we term “Bayesian scQTLs”. Specifically, we quantified the relationship between genotype dosage and gene expression using a Bayesian linear mixed effects model, as shown in Figure 1.



**Figure 1. Hierarchical Bayesian model for single-cell eQTL analysis**

In this model, the QTL effect size from each cell type (for example, the effect size *βAstro* in the cell type Astrocytes) is estimated by considering a prior distribution p(*βAstro*) ~ Normal (*Θ*, *Σ*). In turn, *Θ* and Σ also have their respective prior distributions, and our objective is to approximate the joint posterior distribution p(*βAstro, βOPC, βEndothelial, … βOligo, Θ, Σ, σ-2 | X, Y*), with *X* denoting the genotypes and Y the gene expression within a given cell type. Because the joint posterior distribution has no closed-form expression, we used Gibbs sampling to approximate this distribution. Under this scheme, the hierarchical nature of the system’s setup allows for effects to be shared between cell types via global parameters *Θ* and *Σ*.

**Gerstein lab experience in isoform analysis.**

We have broad experience in identifying genetic variants that regulate isoform-specific expression in various data modalities and disease contexts. For example, we identified >2.6 million isoQTLs from short-read bulk RNA-Seq of the adult brain using a standardized GTEx pipeline (Wang *et al.* 2018). We calculated QTL callsets with multiple strict quality control levels, including >600k isoQTLs in high-confidence expressed genes (>5 FPKM). We recently sought to identify isoQTLs in 388 short read single-cell RNA-Seq datasets from the adult dorsolateral prefrontal cortex. We implemented a novel pipeline that integrated single-cell isoform quantification using the SCASA package, strict quality control to create normalized pseudobulk expression matrices, and isoQTL calculation with permuted multiple testing correction using the bulk RNA-based sQTLseekeR2 package (Garrido-Martin *et al.* 2021; Pan *et al.* 2022). Due to limitations with 10X single-cell sequencing methods, many isoforms are not captured in single-cell datasets (Pan *et al.* 2022). We could still identify ~750 genes with putative isoQTLs across >20 cell types, such as for the GABA receptor *GABRAPL1* in L5/6 excitatory neurons (Ye *et al.* 2021). Finally, we have developed several software packages to both identify alternative splicing events (IQSeq) and prioritize variants that affect splicing (VAT, RADAR, ESPRNN) (Du *et al.* 2012; Habegger *et al.* 2012; Lee *et al.* 2020; Zhang *et al.* 2020).

**Gerstein lab experience with integrating allele-specific expression**

We have extensive experience in conducting analyses related to allele-specific expression and binding. We developed a computational pipeline AlleleSeq (Rozowsky *et al.* 2011) which was originally used for identifying and quantifying ASE and allele-specific binding (ASB) in GM12878. We have applied our tool to a broad spectrum of personal and functional genomics data. We have used it in multiple large Consortium projects and publications, including ENCODE and the 1000 Genomes Project (Djebali *et al.* 2012; Gerstein *et al.* 2012) (Khurana *et al.* 2013). We annotated variants associated with ASE and ASB in a large pool of individuals from the 1000 Genomes Project. For this analysis, we integrated matching functional datasets (955 RNA-seq and 165 ChIP-seq), including ChIP-seq datasets from 14 lymphoblastoid cell lines in ENCODE (Chen *et al.* 2016). We detected more than 6K and 63K SNVs associated with ASB and ASE, respectively. These results were made available as an online resource, AlleleDB (alleledb.gersteinlab.org). Furthermore, using the extensive Roadmap dataset, we constructed a high-resolution map that reveals allelic imbalances in DNA methylation, histone marks, and transcription across 71 epigenomes from 36 distinct cell and tissue types from 13 donors (Onuchic *et al.* 2018).

As part of these and other projects, we continued to develop our pipeline for ASE and ASB analysis. Firstly, to address the inherent variability in functional genomics readcount data, we have implemented the beta-binomial test to determine the significance of allelic imbalances. Additionally, we added supplementary filters to mitigate potential biases stemming from ambiguous mapping (Chen *et al.* 2016). And we expanded the tool to call allele-specific genomic elements, such as genes or regulatory regions, giving rise to our current tool, AlleleSeq2 (Onuchic *et al.* 2018). Most recently, we applied AlleleSeq2 to the EN-TEx resource (Rozowsky *et al.* 2023) encompassing ~1.6K datasets from four donors (~30 tissues x 15 assays) (Figure 2).



**Figure 2. An overview of the EN-TEx project, including ASE and transferQTL.**

We observed a significant increase in the number of detected allele-specific events when aggregating genomic reads from multiple tissues (as opposed to a simple union of sample-specific analyses). We generated a catalog of >1M allele-specific loci (Rozowsky *et al.* 2023). Combining EN-TEx with existing genome annotations revealed strong associations between ASE and GWAS loci.

We have previously used bulk RNA-seq (and ChIP-seq) to show good correspondence between the allele-specific read counts and the corresponding eQTL size and direction (Rozowsky *et al.* 2023). We studied the association of an eQTL with the ASE of its target gene and found a positive correlation with eQTL effect size (Figure 3).



**Figure 3. Correspondence between AS gene expression, AS binding in an upstream regulatory element, and the effect size of the eQTL.**

**References**

Chen, J., J. Rozowsky, T. R. Galeev, A. Harmanci, R. Kitchen *et al.*, 2016a A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat Commun 7**:** 11101.

Djebali, S., C. A. Davis, A. Merkel, A. Dobin, T. Lassmann *et al.*, 2012 Landscape of transcription in human cells. Nature 489**:** 101-108.

Du, J., J. Leng, L. Habegger, A. Sboner, D. McDermott *et al.*, 2012 IQSeq: integrated isoform quantification analysis based on next-generation sequencing. PLoS One 7**:** e29175.

Garrido-Martin, D., B. Borsari, M. Calvo, F. Reverter and R. Guigo, 2021 Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. Nat Commun 12**:** 727.

Gerstein, M. B., A. Kundaje, M. Hariharan, S. G. Landt, K. K. Yan *et al.*, 2012 Architecture of the human regulatory network derived from ENCODE data. Nature 489**:** 91-100.

Habegger, L., S. Balasubramanian, D. Z. Chen, E. Khurana, A. Sboner *et al.*, 2012 VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. Bioinformatics 28**:** 2267-2269.

Khurana, E., Y. Fu, V. Colonna, X. J. Mu, H. M. Kang *et al.*, 2013 Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342**:** 1235587.

Lee, D., J. Zhang, J. Liu and M. Gerstein, 2020 Epigenome-based splicing prediction using a recurrent neural network. PLoS Comput Biol 16**:** e1008006.

Onuchic, V., E. Lurie, I. Carrero, P. Pawliczek, R. Y. Patel *et al.*, 2018 Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. Science 361.

Pan, L., H. Q. Dinh, Y. Pawitan and T. N. Vu, 2022 Isoform-level quantification for single-cell RNA sequencing. Bioinformatics 38**:** 1287-1294.

Psych, E. C., S. Akbarian, C. Liu, J. A. Knowles, F. M. Vaccarino *et al.*, 2015 The PsychENCODE project. Nat Neurosci 18**:** 1707-1712.

Rozowsky, J., A. Abyzov, J. Wang, P. Alves, D. Raha *et al.*, 2011 AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol 7**:** 522.

Rozowsky, J., J. Gao, B. Borsari, Y. T. Yang, T. Galeev *et al.*, 2023 The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models. Cell 186**:** 1493-1511 e1440.

Wang, D., S. Liu, J. Warrell, H. Won, X. Shi *et al.*, 2018 Comprehensive functional genomic resource and integrative model for the human brain. Science 362.

Ye, J., G. Zou, R. Zhu, C. Kong, C. Miao *et al.*, 2021 Structural basis of GABARAP-mediated GABA(A) receptor trafficking and functions on GABAergic synaptic transmission. Nat Commun 12**:** 297.

Zhang, J., J. Liu, D. Lee, J. J. Feng, L. Lochovsky *et al.*, 2020 RADAR: annotation and prioritization of variants in the post-transcriptional regulome of RNA-binding proteins. Genome Biol 21**:** 151.