Research statement

Kevin Lin, kevinL1@wharton.upenn.edu

Technological advancements in single-cell sequencing have provided biologists with an influx of increasingly complex data, but few statistically motivated tools exist to analyze it. Recently, my work has focused on problems inspired by the recent revolution in paired multimodal single-cell technology, such as parallel sequencing of the transcriptome and chromatin modifications. This technology enables unprecedented studies of an organ, such as the brain, and its interconnected mechanisms at single-cell resolution. Additionally, multimodal data allows for new data integration frameworks when paired with previously collected data of only one modality. These potential advancements require thoughtful design of new statistical models and computational methods. My background in dimension-reduction and networks makes me suitable for this task – the former uncovers the subtle cross-modal axes of variation, while the latter aids in understanding the regulatory network relating both modalities. In my previous work, I primarily developed these tools to study brain development, but their statistical foundations broadly apply to other systems. For example, I have ongoing collaborations in immunology and oncology, demonstrating my interest in applying my specialty of data integration to several biomedical fields. Such collaborations are becoming evermore crucial, as biomedical research is becoming increasingly interdisciplinary and single-cell sequencing enables studies of more modalities (such as long-read sequencing and spatial transcriptomics). I plan to continue my existing collaborations and initiate new ones, working on projects that overlap with my areas of statistical expertise.

Previous work

I discuss my previous projects, which were primarily motivated by advancing dimension reduction and network tools to enable more powerful biological investigations.

Matrix factorization for over-dispersed and multimodal data. With Kathryn Roeder (CMU, Statistics) and other collaborators, I have developed the exponential-family SVD (eSVD), a new dimension-reduction tool tailored to handle sparse and over-dispersed single-cell RNA-seq (scRNA-seq) data. With this method, I was able to uncover the previously difficult-to-study differentiation among human oligodendrocytes [1] (Figure 1A). My second paper extends the eSVD to test for differentially expressed genes between

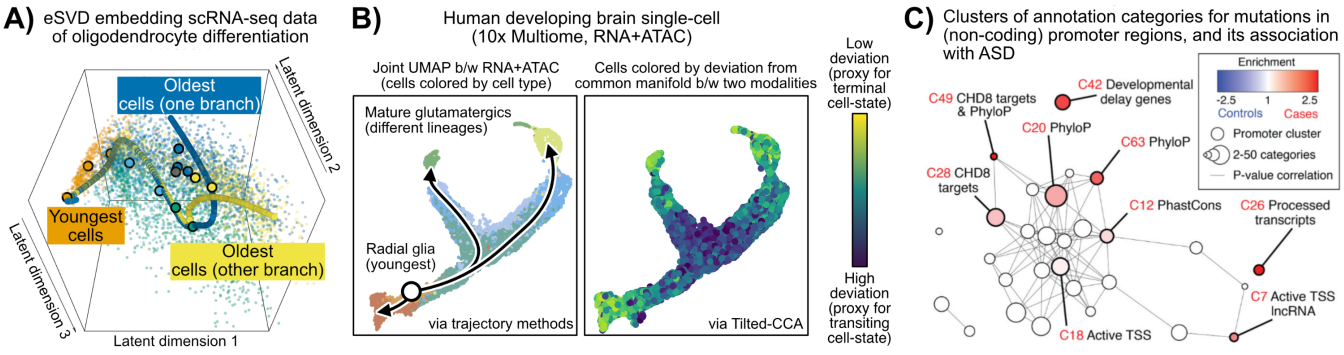

Figure 1: **Summary of previous work**. A) eSVD's 3-dimensional embedding of scRNA-seq data, where the color of cells denotes different stages of oligodendrocyte development. The two estimated lineage trajectories are shown (starting at orange, ending at blue and yellow). B) Tilted-CCA's common embedding of RNA+ATAC single-cell developing brain cells visualized via UMAP, where cells are colored by cell type, and the trajectories estimated using typical lineage reconstruction methods are shown (left). The common embedding captures variation supported by both modalities, and the developmental status of cells can be directly measured by the deviation from this manifold alone without relying on any lineage reconstruction methods (right). C) Network among clusters of annotation categories for non-coding mutations, where we find clusters enriched for association with ASD.

case and control subjects from single-cell data using an empirical Bayes framework [2].

The emergence of paired multimodal single-cell sequencing, such as profiling both gene expression and chromatin accessibility in parallel, has spurred many recent questions on how epigenetic changes coordinate with the transcriptome during differentiation. I developed a novel multimodal dimension-reduction method called Tilted-CCA [3] with Nancy Zhang (UPenn, Statistics) to address this. Tilted-CCA quantifies and separates 1) axes of variation shared between RNA and ATAC (i.e., the "common manifold") capturing the coordination between the transcriptome and chromatin accessibility in developmental systems from 2) the axes of variation unique to either modality. We demonstrate that the deviations from the common manifold alone, which we coin as "asynchrony," provide strong signals of a cell's developmental status during neurogenesis (Figure 1B). Our results suggest that Tilted-CCA can measure the degree of epigenetic priming, whereby enhancer regions become accessible prior to a gene's transcription.

Network analyses of (possibly time-varying) gene co-expression patterns. I have developed new network methods to study the dynamics of gene coordination patterns. In my collaboration with Jing Lei (CMU, Statistics), we posed this as an analysis of multiple networks and started with a simplistic model to formalize the statistical foundations [4]. I then expanded our method to study the time-varying dynamics of gene coordination throughout neurogenesis at single-cell resolution [5]. Our results provided an orthogonal perspective of neurogenesis compared to typical studies of the dynamics in genes' (mean) expression.

In another line of work, I have developed methods with Kathryn Roeder and collaborators to identify mutations associated with autism spectrum disorder (ASD). This was done through a previously developed "guilt-by-association" framework [6], whereby mutations are implicated if they are highly connected to known autism risk mutations in the brain's gene co-expression network. I have bolstered the statistical power of this framework by improving the accuracy of the co-expression network (focusing on mutations in coding regions) [7], and aggregating the mutation categories via a sparse PCA framework so the downstream analysis would be performed on clusters of categories (focusing on non-coding mutations) [8]. In the former, I achieved this by adaptively removing high-dimensional samples to yield a set of homogeneous samples via a two-sample testing framework. In the latter, my aggregation led to novel discoveries of *de novo* mutations in promoter regions that were associated with ASD (Figure 1C).

Advancing statistical methods tailored for studying copy number variation. I have collaborated with Ryan Tibshirani (Berkeley, Statistics) and other statisticians to advance methods for detecting copy number variants (CNV). We study this through the lens of changepoint detection, a broad family of commonly used methods in this field. Specifically, we studied the theoretical requirements to ensure that all copy number variants were found [9], as well as a multiple-testing framework to prune spurious copy number regions to achieve valid Type-I error control [10].

Current research agenda

Paired multimodal single-cell sequencing invites new statistical questions on how to best model the diverse landscape of cross-modal relations. Building off of the statistical insights of my previous works, my current work strives to be among the first to investigate questions that can only be addressed with this technology. I have prepared preliminary results for these projects as part of recent R01 NFS and NIH grant proposals.

Multimodal analysis of acquired resistance and epigenetic priming mechanisms. Cancer mortality has dropped substantially over the last decade due to early-stage diagnoses and multiple therapy options. However, while cancer therapies kill most cancer cells, rare sub-populations of cells could survive and repopulate, potentially leading to future cancer relapse. Is it possible to identify these therapy-resistant cancer sub-populations early on? To study this, Nancy Zhang and I have developed an ongoing collaboration with Sydney Shaffer (UPenn, Bioengineering) to study therapy resistance through the lens of epigenetic priming

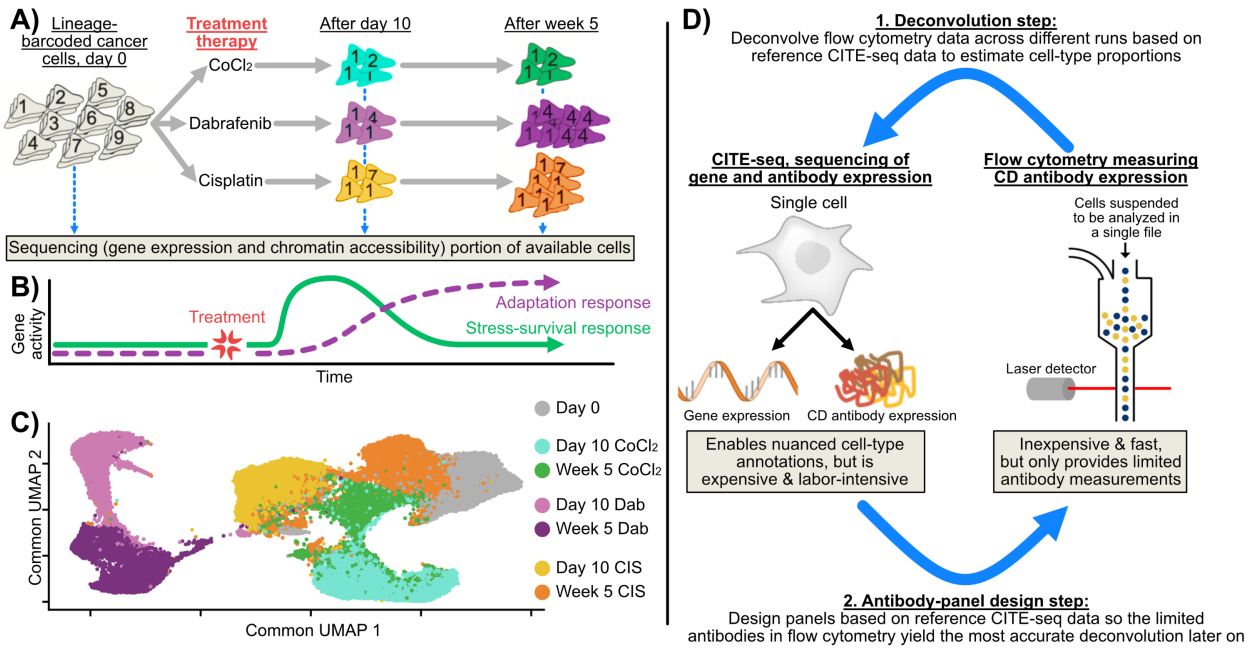

Figure 2: **Summary of current work**. A) Schematic of our *in vitro* cancer therapy experiment. B) Proposed hypotheses of pathways responsible for stress-survival or adaptation after a cancer cell is treated. Both (A) and (B) depict time on the x-axis. C) UMAP of the Tilted-CCA's common embedding of the RNA and ATAC modalities, where cells are colored by their treatment and time-point. D) Schematic of my proposed method to integrate flow cytometry data that is newly collected from a patient with previously-sequenced CITE-seq data for better cell-type proportion estimation.

mechanisms. Dr. Shaffer's expertise in multimodal sequencing and lineage-barcoding protocols for clonal tracing of cells allows me to study this problem through data that is the first of its kind.

We have sequenced a lineage-barcoded melanoma cell line for both RNA and ATAC profiles simultaneously according to a time-course *in vitro* experiment (Figure 2A). Here, cells are exposed to various therapies, causing many cancer cells to die, and the few surviving cells repopulate and are sequenced afterward. The multiple timepoints enable us to investigate acquired resistance mechanisms – perhaps the surviving epigenetically primed cells were able to activate rapid stress-response pathways and gradually acquire resistance through other slowly activated pathways (Figure 2B). My full investigation of the sequenced data (visualized in Figure 2C) involves integrating signals across three modalities: the RNA, ATAC, and lineage barcodes, each contributing vital information about therapy resistance. I have been developing a novel method called *Topic-model CCA* to find K "topics" of cross-modality coordination. Let $X^{(1)} \in \mathbb{R}_+^{p_1 \times n}$ and $X^{(2)} \in \mathbb{R}_+^{p_2 \times n}$ denote the p_1 genes and p_2 chromatin regions measured on the same n cells, where $X^{(\ell)} \geq 0$ and $\|X^{(\ell)}_{\cdot,i}\|_1 = 1$ for each cell $i \in \{1, \ldots, n\}$ and modality $\ell \in \{1, 2\}$. Topic-model CCA then solves

$$\{\widehat{A}, \widehat{B}\} = \underset{\substack{A \in \mathbb{R}_+^{K \times p_1} \\ B \in \mathbb{R}_+^{K \times p_2}}}{\operatorname{argmax}} \left[\operatorname{trace}\left(A X^{(1)} (X^{(2)})^\top B^\top \right) \right], \quad \text{s.t.} \quad \max\left\{ \left\| A X^{(1)}_{\cdot,i} \right\|_1, \left\| B X^{(2)}_{\cdot,i} \right\|_1 \right\} \leq 1 \quad \forall i \in \{1, \ldots, n\}. \tag{1}$$

Here, the objective function is the same as in CCA. As with any non-negative factorization, the additional constraints in (1) aid interpretability. The mixture of topics for each cell is encoded in $X^{(1)}A$ and $X^{(2)}B$, and the associated genes or chromosome regions in A and B. Topic modeling has been prevalently used in single-cell genomics [11], but this has yet to be explored for multimodal data, even in broader fields. Recent promising developments in both topic modeling's computation [12] and underlying geometry [13] make it a ripe time to study Topic-model CCA. I hope our results will be foundational for understanding therapy

resistance, and my method can aid future studies of cross-modal coordination in systems where lineage barcoding is not possible.

Improving flow cytometry's cell-type proportion estimates by leveraging CITE-seq data. Immunology primarily uses flow cytometry (FC) to measure the phenotypes of millions of cells relatively quickly, but it is limited in phenotypic resolution due to the size of the protein panel (i.e., the number of detectable proteins in one run, typically 30). Furthermore, the selection of which proteins to include is primarily based on historical data. In contrast, emerging CITE-seq technology (sequencing single cell's transcriptome and surface antibody markers in parallel) can measure 200+ proteins alongside 10,000+ genes, allowing for high-quality cell-type labels. However, it is hindered by its high cost and low throughput compared to FC. Given each technology's strengths, after an immunologist collects FC data of a new patient, can we design a computational method to better estimate the proportion of granular immune subtypes by leveraging existing CITE-seq data? Furthermore, can we use the full transcriptome and antibodies measured by CITE-seq to guide FC panel design so that immunologists can maximally utilize the limited number of proteins measurable at once? Nancy Zhang and I have initiated collaborations with Andy Minn (UPenn, Oncology) and John Wherry (UPenn, Immunology) to address these pressing questions. My proposed methodology can immediately improve their current research on adaptive resistance to immunotherapy and T-cell exhaustion.

This project involves two tasks (Figure 2D): first, to design M (e.g. $M = 3$) sets $\{\mathcal{A}_m\}_{m=1}^M$ of 30 proteins each based on the existing CITE-seq data. For a new patient, the clinician collects multiple new FC datasets using the prescribed sets $\mathcal{A}_1, \ldots, \mathcal{A}_M$. This leads to the second task, to deconvolve the multiple FC datasets using the CITE-seq data to estimate the cell-type proportions. To solve the second task, I have developed a deconvolution method based on modeling the data as mixtures of Gaussians. This statistical model enables estimation of the cell-type proportions by minimizing the Wasserstein distance between mixtures of Gaussians [14]. Certainly, understanding when my deconvolution accurately estimates cell-type proportions impacts my design of $\{\mathcal{A}_m\}_{m=1}^M$ in the first task. My initial experiments demonstrate this can be recast as an analysis of dependencies among proteins – the proteins within each set $\mathcal{A}_m$ should statistically dependent but not linearly correlated. In contrast, the proteins between two different sets $\mathcal{A}_m$ and $\mathcal{A}'_m$ should be be statistically independent. Altogether, this work will pioneer widely-used methods for clinicians that marry multimodal data with other technologies with markedly different strengths.

Future research agenda

Over the next few years, I seek to develop new data integration tools for multimodal data to address pressing biological questions. As biomedical technologies improve, there will be new modeling and estimation challenges. This will encourage new theoretical questions for unforeseen statistical models. I plan to continue my existing collaboration and initiate new ones, especially with wet-lab biologists and clinicians working in areas that lack targeted statistical tools. As a concrete example, I describe how my statistical expertise can advance the field of developmental biology and cancer research, but my expertise and intended endeavors are not limited to studying these fields.

Investigating cellular differentiation and response through genetic variation and emerging assays. Cell differentiation is regulated in many other ways aside from epigenetics. For example, radial glia cells play an essential role in neuronal migration during brain development [15], which can be studied in more depth via emerging spatial transcriptomic assays. On the other hand, thanks to emerging long-read sequencing assays, differential isoform expression tests have demonstrated how alternative splicing regulates cell development [16]. However, can new hierarchical matrix factorization models accounting for the combinatorial nature of isoforms uncover how genes' isoform diversity (or lack of) regulates differentiation? Additionally, how do we pair our increasing understanding of cell differentiation with genetic variation? Recent work characterizing cell-type specific eQTLs psychiatric and neurological disorders [17] suggest that

SNPs bias how cells commit to particular lineages. Promising frameworks such as dynamic eQTLs [18] have suggested that genetic impact on the transcriptome is not static. These ideas are particularly suitable when studying the neurological impact of stress from a genomic and genetic perspective. Does acute and frequent stress have a genomic impact, and can it disrupt safeguarding mechanisms responsible for restoring normalcy? Recent work on mouse models has demonstrated that frequent stress has temporal effects on the transcriptome [19]. Can this be attributed to epigenetic or genetic variation more broadly? All-in-all, my current work on epigenetic priming will be informative when developing frameworks to integrate across broader biological data.

Formalizing biological mechanisms, illustrated with branching development of regulatory networks. Genomic research has posited many intricate biological concepts that remain statistically elusive. One example is *fate commitment*: what is the last predictable cell-state prior to a cell committing to a specific developmental lineage (i.e., the branchpoint)? Our understanding of fate commitment has a biological impact, as it would suggest how future therapy treatments can control cellular differentiation. While current genomic tools address fate commitment based on cells' mean gene expressions [20], biologists are equally interested in studying fate commitment through cells' epigenetic regulatory network, a critical aspect of cellular identity [21]. This line of questioning can yield tremendous insight for developmental biology, as lineage-tracing experiments suggest that fate commitment occurs much earlier than current transcriptomic analyses suggest [22]. To pose our task abstractly, consider a simplistic example where we observed a collection of five networks $\{G^{(t)}\}_{t=0}^{4}$ that is *unordered* aside from a known "root" $G^{(0)}$ i.e., the start of a *branching* developmental process (Figure 3A). Each network $G^{(t)}$ represents the epigenetic elements' enhancement/suppression of genes for different meta-cells (here, assumed already estimated from multimodal data). Can we estimate the branching development structure and identify when fate commitment occurs? How do we provide a meaningful and rigorous confidence set containing the true branchpoint $G^{(1)}$? Formulating this idea and its more biologically-realistic variants would draw upon many modern statistical ideas, such as graph embeddings, manifold learning, and statistical inference for networks. However, fate commitment is only one of many biological concepts that could benefit from statistical formalization, and doing so would equip biologists with tremendous insight.

Investigating causal impacts, illustrated with translational cancer research. Paired multimodal sequencing has presented many new opportunities to bring biomedical research from "benchside to bedside" by translating biological insight into improved clinical practice. Focusing on cancer research as one of many concrete areas, the diverse heterogeneity of cancer within and across patients has warranted a surge of interest in precision medicine. However, the efficacy of targeted treatments still suggests large room for improvement [23]. Mediation analyses [24] and survival analyses [25] have been critical to explain the diverse treatment efficacies on the patient-level via SNPs or clinical phenotypes. On the other hand, paired multimodal sequencing has enabled high-resolution studies of how cancer cells are impacted by treatment, such as cell-type specific epigenetic modifications [26] or clonal-specific changes in TCR diversity [27]. However, there exist many questions that bridge these two scales: what are robust and predictive cross-modal cancer cell signatures applicable for precision medicine? Which cellular responses yield a desirable clinical outcome? How should we integrate multimodal sequencing alongside clinical data to study the underlying biology dictating a treatment's efficacy? Multimodal sequencing data would tell us a comprehensive picture of which and how cells are impacted by treatment, and clinical data would tell us the causal effect of the cellular changes (Figure 3B). Capitalizing on these opportunities requires substantial interdisciplinary collaborations to develop new multimodal bioinformatics tools and new causal inference techniques. I believe my experiences in multimodal single-cell and statistical analyses make me suitable to initiate such collaborations. The number of opportunities in translational biology will keep increasing as single-cell technology matures, such as in proteomics, metabolomics, and spatial transcriptomics. If successful, such counterfactual causal analyses

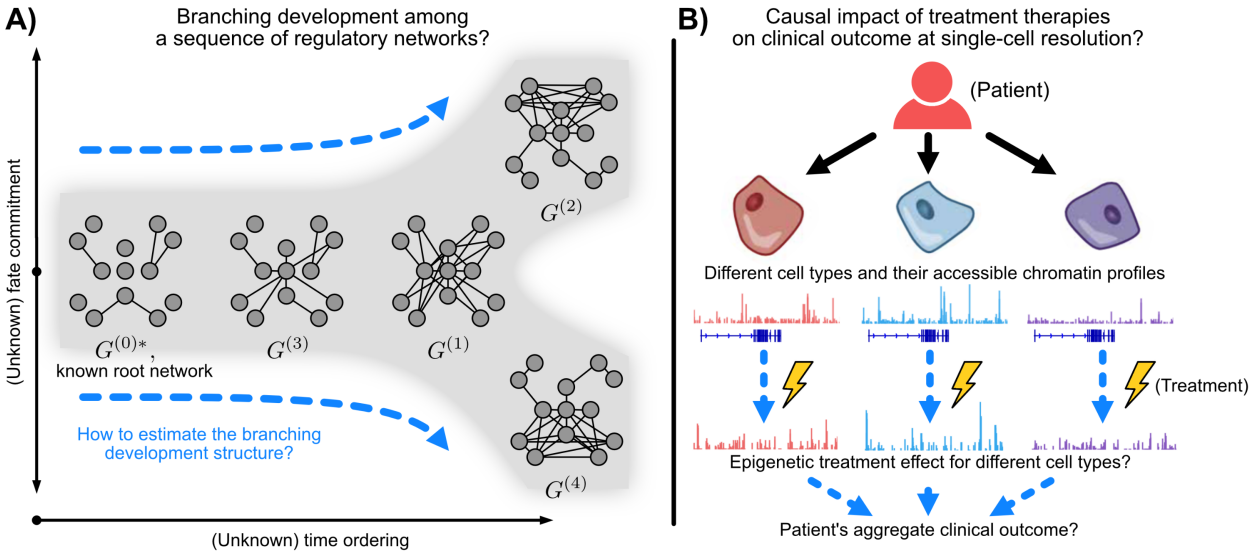

Figure 3: **Schematic of future ideas**. A) An unordered collection of observed networks $\{G^{(t)}\}_{t=0}^{4}$, given a known root $G^{(0)}$. The true (unobserved) branching development we wish to estimate is $G^{(0)} \to G^{(3)} \to G^{(1)}$, followed by a branch into two fates, $G^{(2)}$ and $G^{(4)}$. B) Causal impact of treatment therapies, where paired multimodal sequencing enables investigations of different cell types' (or cancer clones') pre- and post-treatment multiomic landscape, and causal inference enables investigations of how these different cellular responses caused different clinical outcomes.

would be foundational in cancer research (and more broadly, precision medicine) and open up many future research areas.

All-in-all, I wish to address pressing biological questions by expanding upon the latest developments in statistical methods and theory for emerging multimodal sequencing data, and I strive to collaborate with laboratory biologists, clinicians, and statisticians alike to achieve this.

References

[1] **Kevin Lin**, Jing Lei, and Kathryn Roeder. Exponential-family embedding with application to cell developmental trajectories for single-cell RNA-seq data. *Journal of the American Statistical Association*, 116(534):457–470, 2021.

[2] **Kevin Lin**, Yixuan Qiu, and Kathryn Roeder. eSVD: Cohort-level differential expression in single-cell RNA-seq data using exponential-family embeddings. *https://linnykos.github.io/papers/cohort_eSVD.pdf*, 2022.

[3] **Kevin Lin** and Nancy R Zhang. Quantifying common and distinct information in single-cell multimodal data with Tilted-CCA. *bioRxiv preprint (2022.10.07.511320)*, 2022.

[4] Jing Lei and **Kevin Lin**. Bias-adjusted spectral clustering in multi-layer stochastic block models. *Journal of the American Statistical Association*, pages 1–13, 2022.

[5] **Kevin Lin** and Jing Lei. Spectral clustering for heterophilic stochastic block models with time-varying node memberships. *https://linnykos.github.io/papers/dynamicSBM.pdf*, 2022.

[6] Li Liu, Jing Lei, Stephan J Sanders, Arthur Jeremy Willsey, Yan Kou, Abdullah Ercument Cicek, Lambertus Klei, Cong Lu, Xin He, . . ., Matthew W State, Joseph D Buxbaum, Bernie Devlin, and Kathryn Roeder. DAWN: A framework to identify autism genes and subnetworks using gene expression and genetics. *Mol Autism*, 5:22, 2014.

[7] **Kevin Lin**, Han Liu, and Kathryn Roeder. Covariance-based sample selection for heterogeneous data: Applications to gene expression and autism risk gene detection. *Journal of the American Statistical Association*, 116 (533):54–67, 2021.

[8] Joon-Yong An, **Kevin Lin**, Lingxue Zhu, Donna M Werling, Shan Dong, and others. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science*, 362(6420), 2018.

[9] **Kevin Lin**, James Sharpnack, Alessandro Rinaldo, and Ryan J Tibshirani. A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Advances in Neural Information Processing Systems*, pages 6884–6893, 2017.

[10] Sangwon Hyun, **Kevin Lin**, Max G'Sell, and Ryan J Tibshirani. Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics*, 77(3):1037–1049, 2021.

[11] Carmen Bravo González-Blas, Liesbeth Minnoye, Dafni Papasokrati, Sara Aibar, Gert Hulselmans, Valerie Christiaens, Kristofer Davie, Jasper Wouters, and Stein Aerts. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature Methods*, 16(5):397–400, 2019.

[12] Peter Carbonetto, Abhishek Sarkar, Zihao Wang, and Matthew Stephens. Non-negative matrix factorization algorithms greatly improve topic model fits. *arXiv preprint arXiv:2105.13440*, 2021.

[13] Maryam Abdolali and Nicolas Gillis. Simplex-structured matrix factorization: Sparsity-based identifiability and provably correct algorithms. *SIAM Journal on Mathematics of Data Science*, 3(2):593–623, 2021.

[14] Julie Delon and Agnes Desolneux. A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

[15] Cristine R Casingal, Katherine D Descant, and ES Anton. Coordinating cerebral cortical construction and connectivity: Unifying influence of radial progenitors. *Neuron*, 2022.

[16] Xiaochang Zhang, Ming Hui Chen, Xuebing Wu, Andrew Kodani, ..., Douglas L Black, Peter V Kharchenko, Phillip A Sharp, and Christopher A Walsh. Cell-type-specific alternative splicing governs cell fate in the developing cerebral cortex. *Cell*, 166(5):1147–1162, 2016.

[17] Julien Bryois, Daniela Calini, Will Macnair, Lynette Foo, ..., Goncalo Castelo-Branco, Vilas Menon, Philip De Jager, and Dheeraj Malhotra. Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nature Neuroscience*, 25(8):1104–1112, 2022.

[18] Benjamin Strober, Reem Elorbany, Katherine Rhodes, Nirmal Krishnan, Karl Tayeb, Alexis Battle, and Yoav Gilad. Dynamic genetic regulation of gene expression during cellular differentiation. *Science*, 364(6447):1287–1290, 2019.

[19] Lukas M von Ziegler, Amalia Floriou-Servou, Rebecca Waag, Rebecca R Das Gupta, ..., Ferdinand von Meyenn, Hanns U Zeilhofer, Pierre-Luc Germain, and Johannes Bohacek. Multiomic profiling of the acute stress response in the mouse hippocampus. *Nature Communications*, 13(1):1–20, 2022.

[20] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1):1–16, 2018.

[21] María J Barrero, Stephanie Boué, and Juan Carlos Izpisúa Belmonte. Epigenetic mechanisms that regulate cell identity. *Cell Stem Cell*, 7(5):565–570, 2010.

[22] Caleb Weinreb, Alejo Rodriguez-Fraticelli, Fernando D Camargo, and Allon M Klein. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479), 2020.

[23] Ryann Quinn, Rajvi Patel, Cristina Sison, Amandeep Singh, and Xin-Hua Zhu. Impact of precision medicine on clinical outcomes: A single-institution retrospective study. *Frontiers in Oncology*, 11, 2021.

[24] Anna MM Boers, Ivo GH Jansen, Scott Brown, Hester F Lingsma, ..., Michael D Hill, Mayank Goyal, Henk A Marquering, and Charles BLM Majoie. Mediation of the relationship between endovascular therapy and functional outcome by follow-up infarct volume in patients with acute ischemic stroke. *JAMA Neurology*, 76(2):194–202, 2019.

[25] Atefeh Talebi, Afsaneh Mohammadnejad, Abolfazl Akbari, Mohamad Amin Pourhoseingholi, Hassan Doosti, Bijan Moghimi-Dehkordi, Shahram Agah, and Mansour Bahardoust. Survival analysis in gastric cancer: A multi-center study among Iranian patients. *BMC Surgery*, 20(1):1–8, 2020.

[26] Yuanyuan Zhang, Hongyan Chen, Hongnan Mo, Xueda Hu, ..., Binghe Xu, Fei Ma, Zemin Zhang, and Zhuihua Liu. Single-cell analyses reveal key immune cell subsets associated with response to PD-L1 blockade in triple-negative breast cancer. *Cancer Cell*, 39(12):1578–1593, 2021.

[27] Hussein A Abbas, Dapeng Hao, Katarzyna Tomczak, Praveen Barrodia, ..., Kunal Rai, Linghua Wang, Naval Daver, and Andrew Futreal. Single cell T cell landscape and T cell receptor repertoire profiling of AML in context of PD-1 blockade therapy. *Nature Communications*, 12(1):1–13, 2021.