Supplemental Materials

Title: How to get started with single cell-RNA sequencing data analysis

Running title: Single cell RNA-seq data analysis guide

Name of Authors:

Michael S. Balzer[1,2,3], Ziyuan Ma[1,2,3], Jianfu Zhou[1,2,3], Amin Abedini[1,2,3], and Katalin Susztak[1,2,3]

Affiliations:

[1]Renal Electrolyte and Hypertension Division, Department of Medicine, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA 19104
[2]Department of Genetics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA 19104
[3]Institute for Diabetes, Obesity and Metabolism, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA 19104

Correspondence:
Katalin Susztak, MD, PhD, MSc
12-123 Smilow Translational Research Center
3400 Civic Center Blvd
Philadelphia, PA 19104
+1 (215) 898-2009
ksusztak@pennmedicine.upenn.edu

Table of contents

1. Spatially resolved single cell datasets

Spatial transcriptomics was voted as the method of the year in 2020. Experimental and computational method development for spatially resolved single cell profiling is probably the fastest growing area in single cell biology.[1] At present there are a large number of experimental methods available to generate spatial transcriptomics data. The most ambitious methods employ direct in situ sequencing. Other methods use multiplexed fluorescent in situ hybridization such as SeqFISH,[2] SeqFISH+[3] or MerFISH[4] to read out cell type gene expression. In addition, barcoding and bead-based methods are improving significantly such as 10x Visium spatial transcriptomics,[5] Slide-Seq,[6] and sci-Space (currently under development by the Trapnell lab), which are spatially resolved single nucleus RNA-seq techniques that use an array of oligonucleotides arranged in a grid on a slide. This field is rapidly developing and several additional methods, such as e.g., DBiT-seq[7] for co-mapping of mRNAs and proteins have recently been published.

Data integration for spatial transcriptomics is also developing rapidly. The Yuan lab has recently developed Giotto, an open-source pipeline for spatial transcriptomic data analysis and visualization.[8] The data analysis module implements tasks from pre-processing to cell-cell interaction characterization. The data visualization module allows interactive visualization, exploration, and comparison of multiple layers of information. Giotto can resolve tissue spatial organization and allows the interactive exploration of multi-layer information in spatial transcriptomic and imaging data.[9] Giotto provides a single cell resolution spatial information for the investigation of ligand-receptor mediated cell-cell interactions. This is superior to previous methods described Skelly et al.[10] and Ramilowski et al.[11] or CellPhoneDB.[12] Seurat[13, 14] is also capable of analyzing spatial transcriptomics data. Employing such analyses will help to identify altered cellular interactions under a disease condition and reveal signals that are missed in univariate-based differential gene expression analysis.

2. Integration of multi-omics datasets: epigenome, protein expression and beyond

Multi-omics profiles can recover the missing values lost in single modality analysis. For instance, dropout issues are prevalent in scRNA-seq but they are likely recoverable by employing snATAC-seq, either in a separate experiment or in parallel such as with the 10X Single Cell Multiome ATAC+Gene Expression kit. However, analyzing multi-omics data can be challenging because people must harmonize different modalities and correct the underlying batch effects between them.[15] To address this issue, several theoretical models have been developed for multi-omics integration. Some models co-cluster[16] data from different experiments. Canonical correlation vectorization (CCV)[16] hypothesizes that cells originating from identical biological state, even though coming from different data sets, should correspond to each other. Through maximizing the pairwise correspondences, CCV is able to establish mapping between data sets. The advantage of these co-clustering approaches is that they can upregulate features distinguishing cell types while depressing batch-specific noise, which helps link different modalities to each other. Other methods tackle individual cells and multilayer data types at the same time, where they capture the innate heterogeneity via different regression models. For example, Hidden Markov random field (HMRF) performs spatial transcription analysis by connecting gene expression patterns with cell spatial coordinates.[16] Other models depend on decomposing the data into such matrices as one for identifying gene co-expression patterns and another for clustering cells.

Multiple integration methods can also be used for comprehensive data analysis. We have found Seurat particularly powerful for co-embedding scRNA-seq and unmatched snATAC-seq data to reveal cell type-specific regulatory loci, such that joint analysis can improve cluster predition.[17] It can also be used for spatial transcriptomic data to predict spatial gene expression patterns and classify subpopulations. MATCHER[18] also uses co-clustering and a manifold alignment and has the advantage that it can provide a trajectory path and gene expression changes along the path. LIGER[19] uses matrix factorization method to understand relationship between the epigenome and gene expression.

Genome-wide association studies (GWAS) have identified close to 300 loci where nucleotide variants are associated with kidney function.[20] More than 90% of these signals are in the non-coding region of the genome and more often than not the closest gene is the GWAS target gene.[21] snATAC-seq can provide critical information to GWAS signal annotation by prioritizing the causal variants, causal cell types and even imply the causal gene. Our group has successfully used human kidney snATAC-seq data for GWAS SNP, gene and cell type prioritization.[22] Interestingly, epigenetic signals appear to be conserved for many loci and even the mouse kidney snATAC-seq data can be successfully used.[17] Leveraging cis-coaccessibility network analyses such as Cicero,[23] single cell open chromatin information enables to infer not only the implication of affected cell type and variant, but also the target gene. Analysis of allelic imbalance in snATAC-seq data is another important step. Allelic imbalance is defined as the unequal contribution of paternal and maternal DNA sequences to chromatin openness or gene expression. Allelic expression in scRNA-seq data could also be detected by e.g. SCALE[24] and scBASE[25] (ASE). Furthermore, when imputing transcription factor (TF) binding sites, one needs to be aware of the limitations of motif enrichment analyses as implemented by packages HOMER,[26] SCENIC,[27] and chromVAR,[28] as these are mostly not able to exactly distinguish between TFs with similar binding sites. Generation of promoter-enhancer pairs and looking for TF binding are also important downstream analyses. These are exciting new prioritization methods, however at present they will still need to be combined with experimental validation.

3. Supplemental References

1. Asp, M, Bergenstrahle, J, Lundeberg, J: Spatially Resolved Transcriptomes-Next Generation Tools for Tissue Exploration. *Bioessays,* 42**:** e1900221, 2020.
2. Shah, S, Lubeck, E, Zhou, W, Cai, L: In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron,* 92**:** 342-357, 2016.
3. Eng, C-HL, Lawson, M, Zhu, Q, Dries, R, Koulena, N, Takei, Y, Yun, J, Cronin, C, Karp, C, Yuan, G-C, Cai, L: Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature,* 568**:** 235-239, 2019.
4. Chen, KH, Boettiger, AN, Moffitt, JR, Wang, S, Zhuang, X: RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science,* 348**:** aaa6090, 2015.
5. Ståhl, PL, Salmén, F, Vickovic, S, Lundmark, A, Navarro, JF, Magnusson, J, Giacomello, S, Asp, M, Westholm, JO, Huss, M, Mollbrink, A, Linnarsson, S, Codeluppi, S, Borg, Å, Pontén, F, Costea, PI, Sahlén, P, Mulder, J, Bergmann, O, Lundeberg, J, Frisén, J: Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science,* 353**:** 78-82, 2016.
6. Rodriques, SG, Stickels, RR, Goeva, A, Martin, CA, Murray, E, Vanderburg, CR, Welch, J, Chen, LM, Chen, F, Macosko, EZ: Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science,* 363**:** 1463-1467, 2019.
7. Liu, Y, Yang, M, Deng, Y, Su, G, Enninful, A, Guo, CC, Tebaldi, T, Zhang, D, Kim, D, Bai, Z, Norris, E, Pan, A, Li, J, Xiao, Y, Halene, S, Fan, R: High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell,* 183**:** 1665-1681.e1618, 2020.
8. Ruben Dries, QZ, Chee-Huat Linus Eng, Arpan Sarkar, Feng Bao, Rani E George, Nico Pierson, Long Cai, Guo-Cheng Yuan: Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data. *BioRx,* doi: https://doi.org/10.1101/701680, 2019.
9. Dries, R, Zhu, Q, Eng, C-HL, Sarkar, A, Bao, F, George, RE, Pierson, N, Cai, L, Yuan, G-C: Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data. *bioRxiv***:** 701680, 2019.
10. Skelly, DA, Squiers, GT, McLellan, MA, Bolisetty, MT, Robson, P, Rosenthal, NA, Pinto, AR: Single-Cell Transcriptional Profiling Reveals Cellular Diversity and Intercommunication in the Mouse Heart. *Cell Rep,* 22**:** 600-610, 2018.
11. Ramilowski, JA, Goldberg, T, Harshbarger, J, Kloppmann, E, Lizio, M, Satagopam, VP, Itoh, M, Kawaji, H, Carninci, P, Rost, B, Forrest, AR: A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat Commun,* 6**:** 7866, 2015.
12. Efremova, M, Vento-Tormo, M, Teichmann, SA, Vento-Tormo, R: CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc,* 15**:** 1484-1506, 2020.
13. Satija, R: Strength in numbers from integrated single-cell neuroscience. *Nat Biotechnol,* 36**:** 41-42, 2018.
14. Satija, R, Farrell, JA, Gennert, D, Schier, AF, Regev, A: Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol,* 33**:** 495-502, 2015.

15. Cao, J, Cusanovich, DA, Ramani, V, Aghamirzaie, D, Pliner, HA, Hill, AJ, Daza, RM, McFaline-Figueroa, JL, Packer, JS, Christiansen, L, Steemers, FJ, Adey, AC, Trapnell, C, Shendure, J: Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science,* 361**:** 1380-1385, 2018.
16. Zhu, Q: A Hidden Markov Random Field Model for Detecting Domain Organizations from Spatial Transcriptomic Data. *Methods Mol Biol,* 1935**:** 251-268, 2019.
17. Miao, Z, Balzer, MS, Ma, Z, Liu, H, Wu, J, Shrestha, R, Aranyi, T, Kwan, A, Kondo, A, Pontoglio, M, Kim, J, Li, M, Kaestner, KH, Susztak, K: Single cell resolution regulatory landscape of the mouse kidney highlights cellular differentiation programs and renal disease targets. *bioRxiv***:** 2020.2005.2024.113910, 2020.
18. Welch, JD, Hartemink, AJ, Prins, JF: MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol,* 18**:** 138, 2017.
19. Liu, J, Gao, C, Sodicoff, J, Kozareva, V, Macosko, EZ, Welch, JD: Jointly Defining Cell Types from Multiple Single-Cell Datasets Using LIGER. *bioRxiv***:** 2020.2004.2007.029546, 2020.
20. Hellwege, JN, Velez Edwards, DR, Giri, A, Qiu, C, Park, J, Torstenson, ES, Keaton, JM, Wilson, OD, Robinson-Cohen, C, Chung, CP, Roumie, CL, Klarin, D, Damrauer, SM, DuVall, SL, Siew, E, Akwo, EA, Wuttke, M, Gorski, M, Li, M, Li, Y, Gaziano, JM, Wilson, PWF, Tsao, PS, O'Donnell, CJ, Kovesdy, CP, Pattaro, C, Kottgen, A, Susztak, K, Edwards, TL, Hung, AM: Mapping eGFR loci to the renal transcriptome and phenome in the VA Million Veteran Program. *Nature communications,* 10**:** 3842, 2019.
21. Sullivan, KM, Susztak, K: Unravelling the complex genetics of common kidney diseases: from variants to mechanisms. *Nature reviews Nephrology*, 2020.
22. Sheng, X, Ma, Z, Wu, J, Liu, H, Qiu, C, Miao, Z, Seasock, MJ, Palmer, M, Shin, MK, Duffin, KL, Pullen, SS, Edwards, TL, Hellwege, JN, Hung, AM, Li, M, Voight, B, Coffman, T, Brown, CD, Susztak, K: Mapping the genetic architecture of human traits to cell types in the kidney identifies mechanisms of disease and potential treatments. *bioRxiv***:** 2020.2011.2009.375592, 2020.
23. Pliner, HA, Packer, JS, McFaline-Figueroa, JL, Cusanovich, DA, Daza, RM, Aghamirzaie, D, Srivatsan, S, Qiu, X, Jackson, D, Minkina, A, Adey, AC, Steemers, FJ, Shendure, J, Trapnell, C: Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell,* 71**:** 858-871 e858, 2018.
24. Xiong, L, Xu, K, Tian, K, Shao, Y, Tang, L, Gao, G, Zhang, M, Jiang, T, Zhang, QC: SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nature communications,* 10**:** 4576, 2019.
25. Choi, K, Raghupathy, N, Churchill, GA: A Bayesian mixture model for the analysis of allelic expression in single cells. *Nature communications,* 10**:** 5188, 2019.
26. Heinz, S, Benner, C, Spann, N, Bertolino, E, Lin, YC, Laslo, P, Cheng, JX, Murre, C, Singh, H, Glass, CK: Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell,* 38**:** 576-589, 2010.
27. Aibar, S, Gonzalez-Blas, CB, Moerman, T, Huynh-Thu, VA, Imrichova, H, Hulselmans, G, Rambow, F, Marine, JC, Geurts, P, Aerts, J, van den Oord, J, Atak, ZK, Wouters, J, Aerts, S: SCENIC: single-cell regulatory network inference and clustering. *Nat Methods,* 14**:** 1083-1086, 2017.

28. Schep, AN, Wu, B, Buenrostro, JD, Greenleaf, WJ: chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods,* 14**:** 975-978, 2017.