Exciting recent developments in single cell genomics: overview and resources

Compiled by the Satija Lab
Single Cell Genomics Day, 2018

SINGLE NUCLEUS SEQUENCING

While widely applicable to any biological tissue in principle, in practice single cell RNA-seq requires a sample to be first dissociated into a clean single cell suspension. This is easy for some tissues (ie. blood, a liquid suspension of cells) but remarkably difficult for others (i.e. the brain, cells can project long distances and are intertwined). This is a particularly pressing challenge for profiling many human tissues, where samples are often obtained post-mortem. By contrast, suspensions of individual nuclei can be easily generated, even from preserved tissue. Single nucleus RNA-sequencing (sNuc-seq) therefore presents an attractive alternative, not only to study differences in nuclear vs. cytoplasmic expression, but also to extend single cell profiling to challenging tissues.

Remarkably, numerous studies have demonstrated that libraries for single nuclear sequencing can be generated with only minor modifications to standard scRNA-seq protocols, and with only slightly reduced data quality. Recently, multiple groups have developed technologies to rapidly multiplex and scale these experiments, enabling a similar scale to scRNA-seq. We expect to see a significant increase in sNuc-seq datasets, in particular for a variety of human tissues, alongside comprehensive technical benchmarks and experimental power analyses for this exciting new method.

1. Lacar B. *et al.* Nuclear RNA-seq of single neurons reveals molecular signatures of activation.
   Nat Commun. 2016.

   One of the first studies to apply sNuc-seq to individual neurons.

2. Habib N. *et al.* Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons.
   Science. 2016.

   Applies sNuc-seq to study transcriptional dynamics of the neurogenic niche in adult mice.

3. Lake B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA-seq of the human brain.
   Science. 2016.

   Reports an initial taxonomy of cell states in the adult human brain based on sNuc-Seq on the Fluidigm-C1.

4. Habib N. *et al.* DroNc-Seq: Deciphering cell types in human archived brain tissues by massively-parallel single nucleus RNA-seq.
   Nat Methods. 2017.

   Modifies the Drop-seq technique to examine ~40,000 nuclei from mouse and archived human brain tissue.

5. Lake B. *et al.* Integrative Single-Cell Analysis By Transcriptional And Epigenetic States In Human Adult Brain.
   Nat Biotechnol. 2018.

   Reports massively parallel technologies to separately profile RNA and chromatin state from single nuclei.

6. Tasic B. *et al.* Equivalent high-resolution identification of neuronal cell types with single-nucleus and single-cell RNA-sequencing.
   bioRxiv. December 2017.

   An exciting report that datasets based on scRNA-seq or sNuc-seq are equally powered to discover cell states in the mouse brain.

DATASET INTEGRATION, BATCH EFFECT CORRECTION

'Batch effects' have been demonstrated to pose significant challenge for scRNA-seq analyses. In particular, when combining datasets across different experiments, cells of the same type often do not cluster together, but instead cluster together by experiment. This can significantly confound downstream analyses, both for individual labs and large consortia. Unfortunately, techniques for batch effect correction that have been developed for bulk RNA-seq do not work well for single cell data, necessitating new methods.

Recently, there has been interest around solving this by identifying (or 'matching-up') cell types that are present in multiple datasets. Not only can these methods help correct for batch effects, but they can effectively help to 'co-cluster' datasets that are generated from multiple conditions. By learning a consistent set of cell types across datasets, we believe these types of approaches are essential to compare scRNA-seq experiments produced across genetic perturbations or disease states. They may also be potentially useful for 'cross-species' comparisons, to identify evolutionarily conserved and novel cell states. We therefore expect continued development of exciting new methods to address these challenges.

1. Hicks SC. *et al.* Missing data and technical variability in single-cell RNA-seq experiments.
   Biostatistics. November 2017.
   https://github.com/stephaniehicks/scBatchPaper
   Comprehensively describes the challenge of batch effects, and their potential technical sources.

2. Shaham U. *et al.* Removal of batch effects using distribution-matching residual networks.
   Bioinformatics. 2017.
   https://github.com/ushaham/BatchEffectRemoval.git
   Uses residual neural networks to correct for batch effects in mass cytometry and scRNA-seq data.

3. Haghverdi L. *et al.* Correcting batch effects in scRNA-seq data by matching mutual nearest neighbours.
   bioRxiv. July 2017.
   https://github.com/MarioniLab/scran
   Identifies 'matching' cells with similar expression between datasets to quantify and correct for batch effects.

4. Butler. *et al.* Integrated analysis of scRNA-seq data across conditions, technologies, and species
   bioRxiv. July 2017.
   https://github.com/satijalab/seurat
   'Aligns' cells into a shared low-dimensional space to find shared cell types across diverse datasets.

MULTI-MODAL SINGLE CELL ANALYSIS

While transcriptome-wide measurements represent a detailed characterization of a single cell, this view is incomplete as variation in DNA, epigenetic state, and protein levels also are fundamental to cellular state and function. As a result, there is enormous interest in technologies enabling 'multi-modal' measurements of cell state, where multiple molecular phenotypes are simultaneously measured in a single cell. This enables illuminating analyses of how variability in one phenotype affects another, for example, how variation in epigenetic state affects downstream gene expression. Methods initially focused on the simultaneous measurement of genomic, epigenetic, and transcriptomic information in single cells. As these methods often require the physical separation of DNA from RNA into different tubes, they are challenging to scale to profile thousands of cells.

Recently, there has been a rapid development of methods using DNA oligo-barcoded antibodies to measure protein levels in single cells. Multiple groups have explored designing these oligos to represent 'synthetic transcripts', so that their presence can be read out alongside the transcriptome. Importantly, these methods are compatible with most if not all scRNA-seq technologies, enabling massively parallel multi-modal measurements of RNA and surface protein levels. We anticipate that large, optimized panels will be developed for a multitude of human tissues, and that these technologies will be further extended to profile intracellular proteins. More generally, we expect that the potential of novel and scalable multimodal analysis strategies will be an exciting area of future development, enabling not only a taxonomic classification of cell states but a detailed understanding of their molecular regulation and function.

1. Macaulay IC. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods.* 2015.

   Pioneering work to simultaneous sequence DNA and RNA in single cells, after physical separation.

2. Angermueller C. *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods*. 2016.

   An important modification of G&T-seq to measure bisulfite modifications in parallel with RNA in single cells.

3. Pott S. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *eLife*. 2017.

   A clever extension of existing methods to generate trimodal single cell data.

4. Stoeckius M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods.* 2017. http://www.cite-seq.com

5. Peterson V. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol.* 2017.

   These two manuscripts utilize DNA-barcoded antibodies to convert protein levels into a sequenceable readout.

MULTIPLEXING SINGLE CELL EXPERIMENTS

Sample 'multiplexing', i.e. pooling cells from different samples together and running a single experiment, has significant potential benefits for single cell experiments. Multiplexed designs require that cells from each sample contain a unique fingerprint (or barcode), enabling each cell to be assigned to its sample of origin. Such a design is possible when samples originate from different individuals, enabling sample-specific genetic differences (i.e. single nucleotide polymorphisms) to act as barcodes. A recently introduced barcoding strategy using antibodies against ubiquitously expressed surface proteins represents a complementary approach for samples with the same genotype.

Multiplexed experimental designs eliminate sample-specific technical 'batch effects' that may confound downstream analysis, but they also facilitate the identification of across-sample 'doublets', as these cells will exhibit multiple sample barcodes. Doublet identification not only removes artifacts from expression data, but it also enables the 'super-loading' of commercial droplet-based single cell platforms, which can greatly reduce costs. We therefore expect that multiplexing will become a widely used strategy for single

cell experiments, with further technology development that will enable this strategy to be used for any experiment.

1. Kang HM. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation.
   Nat. Biotechnol. 2017.
   https://github.com/statgen/demuxlet

   Introduces the 'demuxlet' algorithm, which enables genetic demultiplexing, doublet detection, and super-loading for droplet-based scRNA-seq. Recommended approach when samples have distinct genotypes.

2. Stoeckius M*, Zheng S*. *et al.* Cell "hashing" with barcoded antibodies enables multiplexing and doublet detection for single cell genomics.
   bioRxiv. July 2017.

   A complementary approach to multiplex samples, based on barcoded antibodies against ubiquitously expressed surface proteins. Recommended approach when samples have the same genotype.

RNA VELOCITY

There has been significant interest in computational methods to 'order' cells along developmental trajectories, an approach commonly referred to as 'pseudotime' estimation. However, these approaches are based primarily on a cell's quantified RNA expression, often do not include an element of real time, and cannot infer a cell's direction. The Kharchenko and Linnarson labs recently introduced a new computational approach to measure RNA 'velocity'. Here, they not only quantify a cell's RNA expression, but *predict* what it will be in the future. They do this by quantifying spliced and unspliced reads for each gene. Genes with an abundance of unspliced reads (precursor mRNA) are in the process of upregulating expression, suggesting that more transcript production is expected in the future.

Remarkably, the authors demonstrate that this computational prediction can be applied across a wide range of technologies and existing datasets, including human tissue. We anticipate exiting further developments, including new technological modifications to improve the quantification of RNA 'velocity', and its integration with developmental trajectory reconstruction algorithms.

1. La Manno, G. *et al.* RNA velocity in single cells.
   bioRxiv. October 2017.

   Introduces the concept of expression velocity, and demonstrates its utility on multiple datasets across tissues and technologies.

SPLIT-POOL COMBINATORIAL BARCODING

Single cell analysis typically begins by placing individual cells in isolated compartments, so that they can be uniquely barcoded or profiled. Innovative methods based on 'Split-pool' workflows modify this workflow, splitting a sample into groups (~100) of cells, and placing a unique barcode on each group. Next, all cells are combined into a single pool, re-split into new groupings, and barcoded again. With the right experimental parameter, each cell will have a unique combination of two barcodes! These workflows require little to no specificied equipment, and are particularly compatible with single cell chromatin profiling techniques.

Recently, two groups introduced exciting methods to extend split-pool strategies to single cell RNA-seq. Importantly, these strategies are compatible with fixed samples and single nuclear profiling, and can be

have been used to profile entire organisms in a single experiment! Most importantly, the potential number of cells profiled in a single experiment grows *exponentially* with each round of barcoding. Therefore, adding a third round of split-pool barcoding could enable the sequencing of >1,000,000 cells in a single experiment. While the sensitivity of these approaches is still being optimized, we expect that the unique scale of combinatorial barcoding will lead to the generation of transformative and exciting datasets in the near future.

1. Cusanovich DA. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science.* 2015.

   Pioneering development of transposase-mediated split-pool combinatorial barcoding for single cell ATAC-seq.

2. Lake B. *et al.* Integrative Single-Cell Analysis By Transcriptional And Epigenetic States In Human Adult Brain. *Nat Biotechnol*. 2018.

   A similar method to profile open chromatin in single cells, using linear amplification, applied to human neurons.

3. Cao J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. 2017.

4. Rosenberg AB. *et al.* Scaling single cell transcriptomics through split pool barcoding. *bioRxiv. February 2017.* https://sites.google.com/uw.edu/splitseq

   These two manuscripts extend initial reports, which focused on epigenetic profiling, to perform single cell RNA-seq using split-pool workflows.

IMPUTATION

Due to miniscule quantities of mRNA in single cells, the resulting gene expression data contain numerous false-negatives ('drop-outs'), particularly for lowly expressed genes. Recently, there has been significant interest around using imputation or smoothing to counter the extensive noise and sparsity in single cell data. These methods leverage the fact that groups of genes have highly correlated expression, facilitating the identification and correction of technical outliers. We expect continued interest around the development of these methods, alongside the development of potential best practices to avoid over-smoothing or loss of signal.

1. Satija R. *et al.* Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*. 2015 https://github.com/satijalab/seurat

   Imputation of scRNA-seq values to integrate with FISH, using LASSO.

2. van Dijk D. *et al.* MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *bioRxiv*. February 2017. https://github.com/pkathail/magic

   Powerful diffusion-based approach to recover regulatory networks from scRNA-seq.

3. Wagner F. *et al.* K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv*. November 2017. https://github.com/yanailab/knn-smoothing

   Innovative approach to pool expression measurements across similar cells to reduce technical noise.

4. Ronen J. *et al.* netSmooth: Network-smoothing based imputation for single cell RNA-seq.. *bioRxiv*. November 2017. https://github.com/BIMSBbioinfo/netSmooth

   Incorporates prior knowledge (i.e. protein-protein interaction networks) for imputation.

5. Zhang L. *et al.* Comparison of computational methods for imputing single-cell RNA-sequencing data. *bioRxiv*. December 2017.

   Initial benchmarking of scRNA-seq imputation methods.