Single-cell RNA-seq

Mikhail Dozmorov

2021-04-21

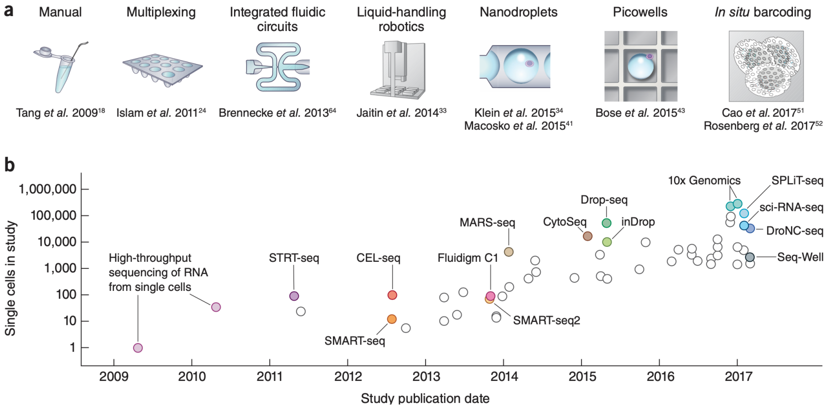
Background

- Most of the biological experiments are performed on “bulk” samples, which contain a large number of cells (millions).
- The high-throughput data we discussed so far are all “bulk” data, which measures the average (gene expression, TF binding, methylation, etc.) of many cells.
- The bulk measurement ignores the cellular heterogeneity:
  - Different cell types.
  - Biological variation among the same type of cell.

Single cell sequencing applications

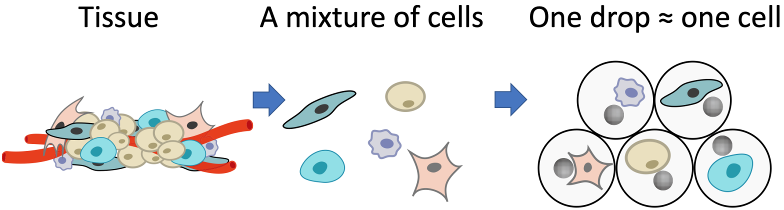
- Sequencing of genetic material from individual cells provides more detailed, higher resolution information.
- Infer cell lineages
- Identify subpopulations
- Outline temporal evolution
- Define subpopulation-specific biological characteristics, e.g., differentially expressed genes

Single cell timeline



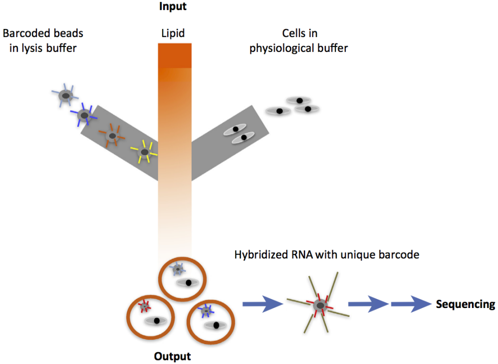
Svensson, Valentine, Roser Vento-Tormo, and Sarah A. Teichmann. "Exponential Scaling of Single-Cell RNA-Seq in the Past Decade." Nature Protocols 13, no. 4 (April 2018): 599–604. https://doi.org/10.1038/nprot.2017.149.

Drop-Seq: We measure single-cell expression by counting cells and genes (mRNA)



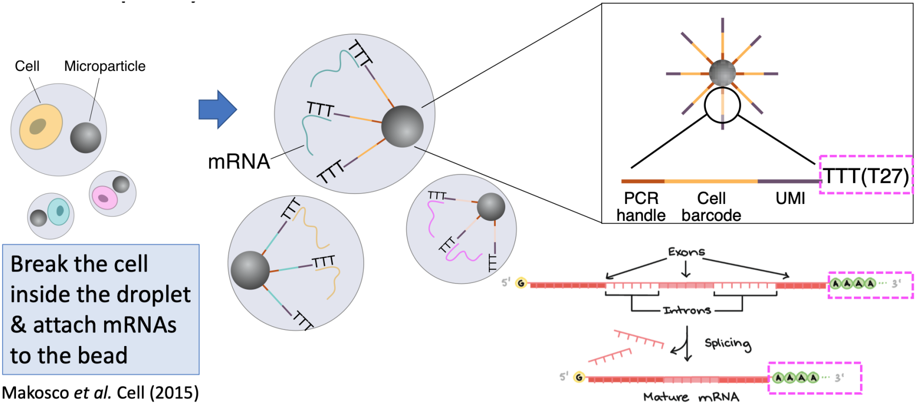
Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets." Cell 161, no. 5 (May 21, 2015): 1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

Single-cell Sequencing Technology


Input
Barcoded beads in lysis buffer
Lipid
Cells in physiological buffer
Hybridized RNA with unique barcode
Sequencing
Output

A single device has three input ports (oil, barcoded beads in lysis buffer, and cells of interest) and a single output port used for collecting bead–cell-containing lipid droplets. Then each cell (or RNA in the cell) is marked by the unique barcode and processed on the bead for sequencing

Cells are barcoded and sorted & genes are uniquely counted

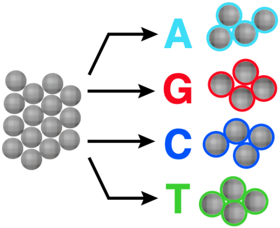


Break the cell inside the droplet & attach mRNAs to the bead

Makosco *et al.* Cell (2015)

Image: www.khanacademy.

Create unique barcode sequence patterns with combinatorics on our side (bead-specific indexes)

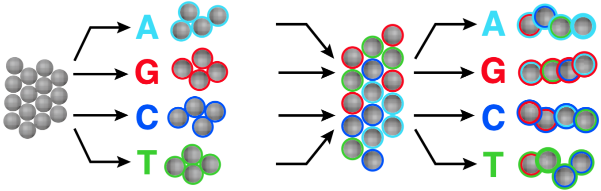


Attach the same base in each set (A, G, C, T)

Randomly split microparticles into 4 sets

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets." Cell 161, no. 5 (May 21, 2015): 1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

Create unique barcode sequence patterns with combinatorics on our side (bead-specific indexes)



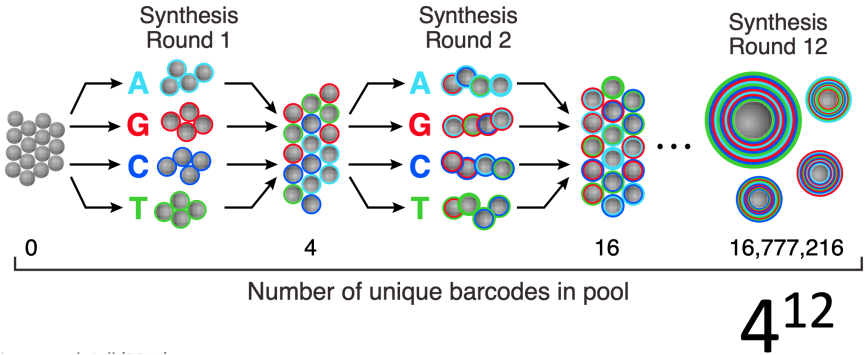
Attach the same base in each set (A, G, C, T)

Pool them all

Random split into 4 sets

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets." Cell 161, no. 5 (May 21, 2015): 1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

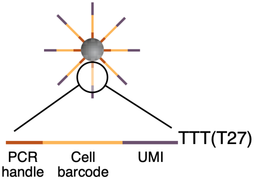
Create unique barcode sequence patterns with combinatorics on our side (bead-specific indexes)

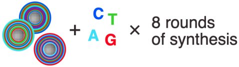


Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets." Cell 161, no. 5 (May 21, 2015): 1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

Unique Molecular Identifier (UMI) to quantify number of genes within each cell

- Random 8-mers incorporated into mRNA fragments before amplification. This is in addition to cell barcodes and primers
- For each cell barcode, count the number of UMIs for every transcript and aggregate this number across all transcripts derived from the same gene locus
- Significantly reduce noise, improve gene expression quantification



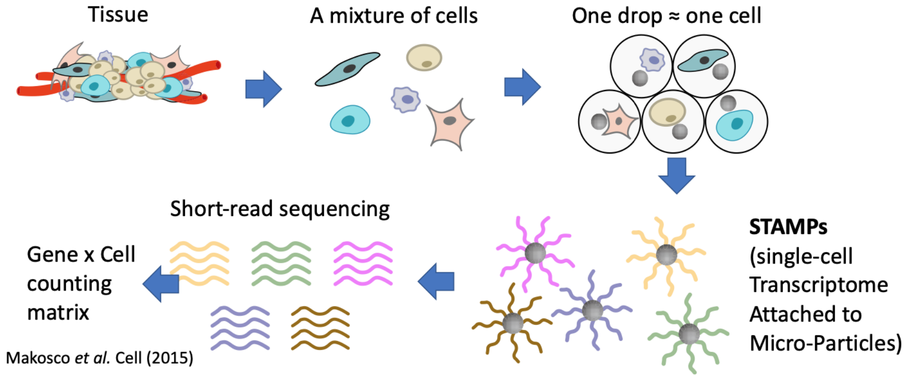


- Millions of the same **cell barcode** per bead
- 4^8 different **molecular barcodes** (UMIs) per bead

4^{12} cells $\quad$ 4^8 molecules

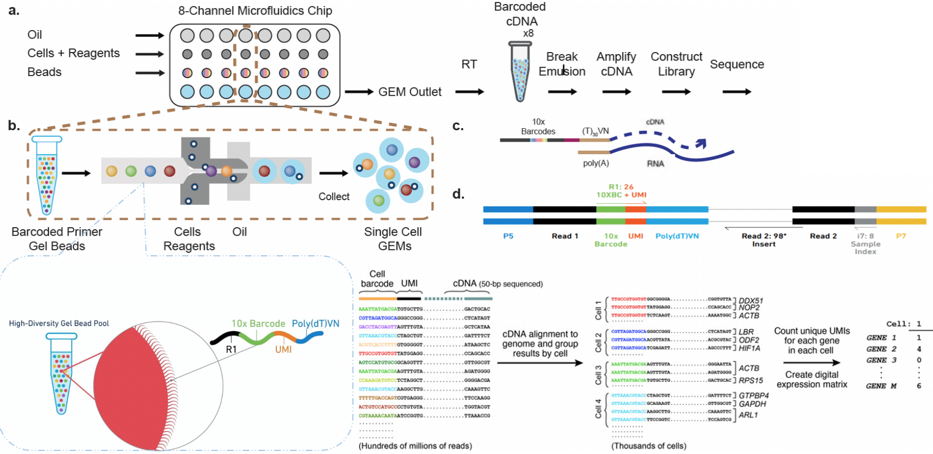
l. Cell (2015)

Drop-seq: highly parallelized counting of cells and genes using cell- and UMI indexes



Makosco *et al.* Cell (2015)

3'-end Sequencing with UMIs (10X Genomics)


a.
8-Channel Microfluidics Chip
Oil
Cells + Reagents
Beads
GEM Outlet
RT
Barcoded
cDNA
x8
Break
Emulsion
Amplify
cDNA
Construct
Library
Sequence
b.
Collect
Barcoded Primer
Gel Beads
Cells
Reagents
Oil
Single Cell
GEMs
High-Diversity Gel Bead Pool
10x Barcode
Poly(dT)VN
R1
UMI
c.
10x
Barcodes
(T)30VN
cDNA
poly(A)
RNA
d.
R1: 26
10XBC + UMI
P5
Read 1
10x
Barcode
UMI
Poly(dT)VN
Read 2: 98*
Insert
Read 2
i7: 8
Sample
Index
P7
Cell
barcode
UMI
cDNA (50-bp sequenced)
AAATTATGACGATGTGCTTG............GACTGCAC
CGTTAGATGGCAGGGCCGGG............CTCATAGT
GACCTACGAGTTAGTTTGTA............GCTCATAA
GTTAAACGTACCCTAGCTGT............GATTTTCT
ACGTCACCTTTTGTGGGGGT............ATAAGCTC
TTGCCGTGGTGTTATGGAGG............CCAGCACC
AGTCCATGTGCGGCAGGTTT............GTTGGCGT
AAATTATGACGAAGTTTGTA............AGATGGGG
CCAAAGATGTCCTCTAGGCT............GGGGACGA
GTTAAACGTACCAAGGCTTG............CAAAGTTC
TTTTTGACCAGTCGTGAGGG............TTCCAAGG
ACTGTCCATGCCCCTGTGTA............TGGTACGT
CGTAAAACAATAATCCGGTG............TTAAACCG
(Hundreds of millions of reads)
cDNA alignment to
genome and group
results by cell
Cell 1
TTGCCGTGGTGTGGCGGGGA............CGGTGTTA DDX51
TTGCCGTGGTGTTATGGAGG............CCAGCACC NOP2
TTGCCGTGGTGTTCTCAAGT............AAAATGGC ACTB
Cell 2
CGTTAGATGGCAGGGCCGGG............CTCATAGT LBR
CGTTAGATGGCAACGTTATA............ACGCGTAC ODF2
CGTTAGATGGCATCGAGATT............AGCCCTTT HIF1A
Cell 3
AAATTATGACGAAGTTTGTA............GGGAATTA
AAATTATGACGAAGTTTGTA............AGATGGGG ACTB
AAATTATGACGATGTGCTTG............GACTGCAC RPS15
Cell 4
GTTAAACGTACCCTAGCTGT............GATTTTCT GTPBP4
GTTAAACGTACCGCAGAAGT............GTTGGCGT GAPDH
GTTAAACGTACCAAGGCTTG............CAAAGTTC
GTTAAACGTACCTTCCGGTC............TCCAGTCG ARL1
(Thousands of cells)
Count unique UMIs
for each gene
in each cell
Create digital
expression matrix
Cell: 1
GENE 1 1
GENE 2 4
GENE 3 0
GENE M 6

scRNA-seq design considerations

Each step is vulnerable to experimental/technical noise and human errors:

- Randomize batch effects
- Spike-ins (debatable), or unique molecular identifiers (UMIs)
- Record all sources of variability, check for confounding with the main effect

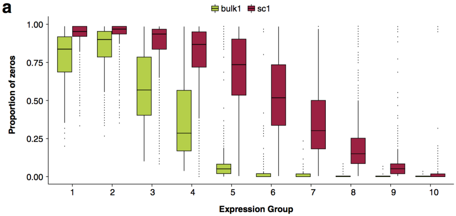
Low amount of starting material

- Approximately $10^5 - 10^6$ mRNA molecules are present in a typical single mammalian cell, and up to 10,000 different genes may be expressed
- $\sim$500,000 to 1M reads per cell (sometimes less ($\sim$50,000) reads is sufficient for cell classification [Pollen AA et.al. Nat. Biotechnol. 2014]) vs. 20-30M reads in bulk RNA-seq

Computational analysis of scRNA-seq data

How does single-cell data differ from bulk RNA-seq

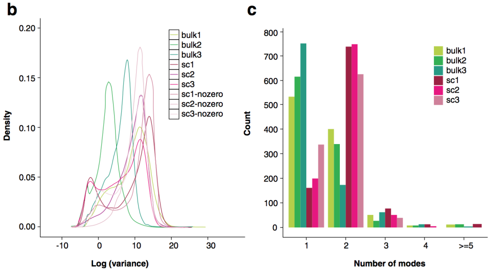
- **Abundance of zeros:** Even with the most sensitive platforms, the data are relatively sparse owing to a high frequency of dropout events (lack of detection of specific transcripts)
- The numbers of expressed genes detected from single cells are typically lower compared with population-level ensemble measurements



Bacher, Rhonda, and Christina Kendziorski. "Design and Computational Analysis of Single-Cell RNA-Sequencing Experiments." Genome Biology 17 (April 7, 2016): 63. https://doi.org/10.1186/s13059-016-0927-y.

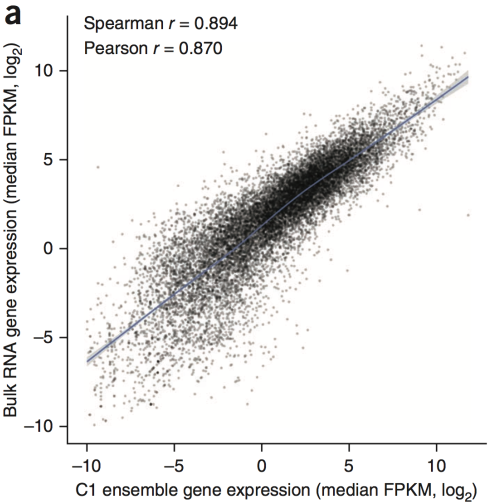
How does single-cell data differ from bulk RNA-seq

- **Multimodal distribution of variance:** scRNA-seq data, in general, are much more variable than bulk data
- Distributions of transcript quantities are often more complex in single-cell datasets than in bulk RNA-seq - negative binomial or multimodal distributions



Bacher, Rhonda, and Christina Kendziorski. "Design and Computational Analysis of Single-Cell RNA-Sequencing Experiments." Genome Biology 17 (April 7, 2016): 63. https://doi.org/10.1186/s13059-016-0927-y.

Correlation with regular RNA-seq data

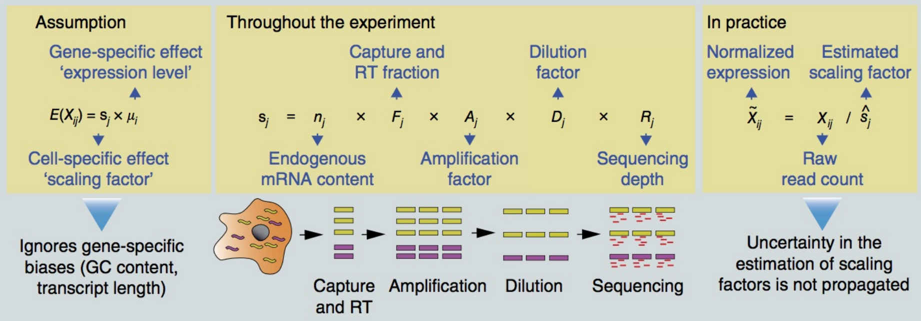


https://www.nature.com/nmeth/journal/v11/n1/full/nmeth.2694.html

Filtering

- Filter cells and/or genes
- No single consensus, frequently used criteria include:
    - relative library size (cells with insufficient number of reads should be removed)
    - number of detected genes ($<$ 10,000)
    - fraction of reads mapping to mitochondria-encoded genes (indicator of dead/broken cells)

Global-scaling normalization


Assumption
Gene-specific effect 'expression level'
$E(X_{ij}) = s_j \times \mu_i$
Cell-specific effect 'scaling factor'
Ignores gene-specific biases (GC content, transcript length)
Throughout the experiment
Capture and RT fraction
Dilution factor
$s_j = n_j \times F_j \times A_j \times D_j \times R_j$
Endogenous mRNA content
Amplification factor
Sequencing depth
Capture and RT
Amplification
Dilution
Sequencing
In practice
Normalized expression
Estimated scaling factor
$\tilde{X}_{ij} = X_{ij} / \hat{s}_j$
Raw read count
Uncertainty in the estimation of scaling factors is not propagated

Vallejos, Catalina A, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni. "Normalizing Single-Cell RNA Sequencing Data: Challenges and Opportunities." Nature Methods 14, no. 6 (May 15, 2017): 565–71. https://doi.org/10.1038/nmeth.4292.

Between-sample normalization

TPM or RPKM/FPKM (within-cell normalization) is insufficient - between-sample normalization is needed

- **Median normalization** - identify relatively stable genes to calculate global scaling factors (one for each cell, common across genes in the cell)
- **Spike-in based normalization** - estimate global rescaling factors from known spike-in concentration

Spike-in sequences and normalization

- External RNA Control Consortium (ERCC) spike-in controls can be used for normalization in the context of a global expression shift
  - Count the number of cells in each sample
  - Add the ERCC spike-in sequences to each sample in proportion to the number of cells
  - Normalize read counts based on cyclic loess robust local regression on the spike-in counts

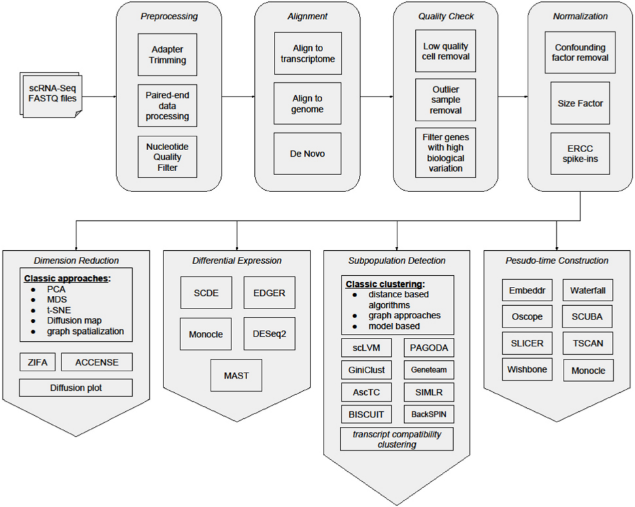
Baker, S.C. et al. The external RNA controls consortium: a progress report. Nat. Methods 2, 731–734 (2005).

Jiang, L. et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 21, 1543–1551 (2011).

Loven, J. et al. Revisiting global gene expression analysis. Cell 151, 476–482 (2012).

Tools and methods

The computational workflow for single cell experiments



scRNA-Seq FASTQ files

Preprocessing

Adapter Trimming

Paired-end data processing

Nucleotide Quality Filter

Alignment

Align to transcriptome

Align to genome

De Novo

Quality Check

Low quality cell removal

Outlier sample removal

Filter genes with high biological variation

Normalization

Confounding factor removal

Size Factor

ERCC spike-ins

Dimension Reduction

Classic approaches:

- PCA
- MDS
- t-SNE
- Diffusion map
- graph spatialization

ZIFA

ACCENSE

Diffusion plot

Differential Expression

SCDE

EDGER

Monocle

DESeq2

MAST

Subpopulation Detection

Classic clustering:

- distance based algorithms
- graph approaches
- model based

scLVM

PAGODA

GiniClust

Geneteam

AscTC

SIMLR

BISCUIT

BackSPIN

transcript compatibility clustering

Pesudo-time Construction

Embeddr

Waterfall

Oscope

SCUBA

SLICER

TSCAN

Wishbone

Monocle

From FASTQ files to the matrix of single-cell gene expression

CellRanger - a set of tools that process 10X Genomics Chromium single-cell data, generate feature-barcode matrices, perform clustering, and more.

https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger

Alevin - end-to-end droplet-based scRNA-seq (10X Genomics) processing pipeline performing cell barcode detection (two-step whitelisting procedure), read mapping, UMI deduplication (parsimonious UMI graphs, PUGs), resolving multimapped reads (EM method to resolve UMI collisions), gene count estimation. Intelligently handles UMI deduplication and multimapped reads, resulting in more accurate gene abundance estimation.

https://salmon.readthedocs.io/en/latest/alevin.html

https://github.com/mdozmorov/scRNA-seq_notes#preprocessing-pipelines

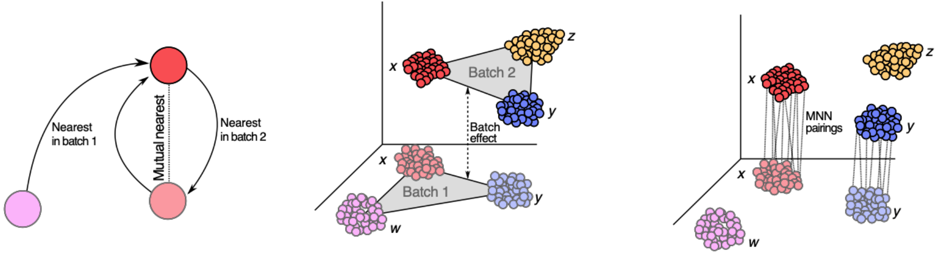
Quality control

- Mapping statistics (% uniquely mapping)
- Mismatch rate
- Fraction of exon-mapping reads
- 3' bias (degraded RNA)
- mRNA-mapping reads
- Number of detected genes
- Empty droplets, droplets with barcode-swapped pseudo-cells
- Doublets
- Mitochondrial gene expression
- Outliers (genes, cells)

https://github.com/mdozmorov/scRNA-seq_notes#quality-control

Batch correction

- Particularly important when combining multiple datasets.
- Many methods for data integration across batches and technologies.
- Example: MNN - mutual nearest neighbors method for single-cell batch correction.



Haghverdi, Laleh, Aaron T L Lun, Michael D Morgan, and John C Marioni. "Batch Effects in Single-Cell RNA-Sequencing Data Are Corrected by Matching Mutual Nearest Neighbors." Nature Biotechnology, April 2, 2018. https://doi.org/10.1038/nbt.4091.
https://github.com/mdozmorov/scRNA-seq_notes#batch-effect-merging

Imputation

- To account for the high dropout level, many methods for imputing scRNA-seq data were developed.
- Goal - to improve clustering, differential expression.
- Issues - unknown origin of dropouts, may be technical or biological.
- Methods range from k-nearest neighborhood, graph imputation, matrix decomposition, to (autoencoder-based) neural networks

https://github.com/mdozmorov/scRNA-seq_notes#imputation

Filtering

- In most cases, all genes are not used.
- Filtering based on:
  - Biologically variable genes (Brenneke method based on spike-in data) or top variable genes if no spike-in data.
  - Genes expressed in X cells.
    – Filter out genes with correlation to few other genes
  - Prior knowledge / annotation
  - DE genes from bulk experiments

Dimensionality reduction

- PCA (principal component analysis)
- ICA (independent component analysis)
- MDS (multidimensional scaling)
- Non-linear PCA – t-SNE (t-distributed stochastic neighbor embedding)
- Diffusion maps
- Neural network-based methods

https://github.com/mdozmorov/scRNA-seq_notes#dimensionality-reduction

Cell clustering

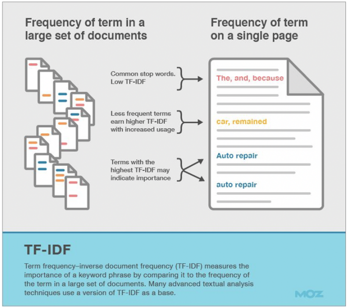
- Perhaps the most active topic in scRNA-seq
- The goals include:
  - Cluster cells into subgroups
  - Model temporal transcriptomic dynamics: reconstruct "pseudo-time" for cells. This is useful for understanding development or disease progression
- Traditional method like k-means or hierarchical clustering need to be used with caution due to dropout events
- In single-cell world, clustering is typically performed on reduced dimensionality data

https://github.com/mdozmorov/scRNA-seq_notes#clustering

TF-IDF transformation

- **Term Frequency x Inverse Document Frequency**
  - Successfully employed in *information retrieval*
- Two parts:
  - How many times a term occurs in a document? (Considers *term frequency*)
  - How 'important' the term is?(Considers *document/collection frequency*)

  (Intuition: rare terms in a collection are more *informative* than frequent terms; Think stop-words!)

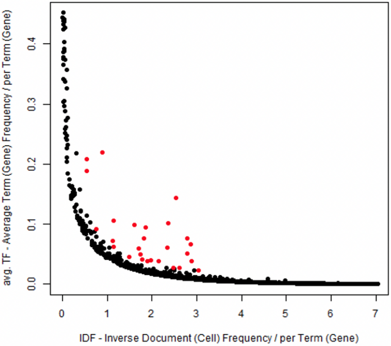

Frequency of term in a large set of documents
Frequency of term on a single page
Common stop words. Low TF-IDF
Less frequent terms earn higher TF-IDF with increased usage
Terms with the highest TF-IDF may indicate importance
The, and, because
car, remained
Auto repair
auto repair
MOZ

TF-IDF

Term frequency–inverse document frequency (TF-IDF) measures the importance of a keyword phrase by comparing it to the frequency of the term in a large set of documents. Many advanced textual analysis techniques use a version of TF-IDF as a base.

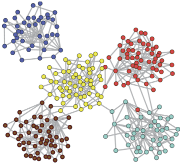
TF-IDF transformation

- Term Frequency x Inverse Document Frequency for scRNA-Seq data:
  - For gene i in cell j with count f :
    $$TF_{ij} = f_{ij} / \max_k f_{kj}$$
  - If gene i is detected in n_i out of N cells:
    $$IDF_i = \log_2(N/n_i)$$
  - TF-IDF score:
    $$TF_{ij} \ * IDF_i$$



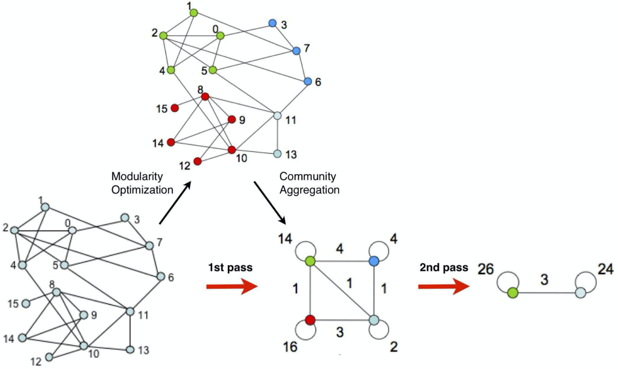
Graph-based clustering

- Undirected graph
  - Cells = vertices
  - Edges = connecting pairs of cells for which the binarized TF-IDF transformed expression signature vectors have Euclidean, Pearson, Cosine, or Jaccard similarity above a user-defined threshold
  - Weights = edges weighted by the corresponding pairwise similarity measures
- Clustering by greedy/Louvain modularity optimization (igraph R).
- Keep on partitioning based on silhouette score for homogeneity and to force a minimum number of clusters (or "optimal number of clusters") when required.



Louvain Method

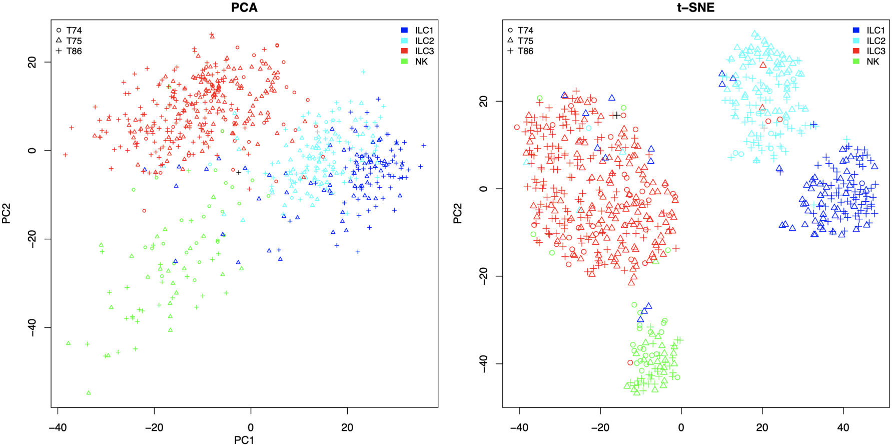
- First, small communities are found by optimizing modularity locally on all nodes (evaluates the change of modularity by removing i from its community and then by moving it into a neighboring community).
- Then each small community is grouped into one node and the first step is repeated.



t-SNE: a useful visualization tool

- t-SNE (t-distributed stochastic neighbor embedding): visualize high-dimensional data on 2-/3-D map
- Method often used in single cell data.
  - Step 1 - probability distribution for all pairs in PCA space with N principal components
  - Step 2 – dimensionality reduction with similar probability distribution and minimization of divergence between distributions
- Implementations in R:
  – tsne – Rtsne (Barnes-Hut t-SNE)
- For other languages (python, java, matlab, C++ etc.) - http://lvdmaaten.github.io/tsne/

t-SNE vs PCA dimensionality reduction



Visualization

- Many visualization tools are parts of pipelines for performing other analyses (filtering, normalization, dimensionality reduction, etc.)
- Many are web-based or Shiny apps
- **Loupe Browser** - a desctop application (Windows, Mac) that provides interactive visualization for 10X Genomics data

https://support.10xgenomics.com/single-cell-gene-expression/software/visualization/latest/what-is-loupe-cell-browser

https://github.com/mdozmorov/scRNA-seq_notes#visualization-pipelines

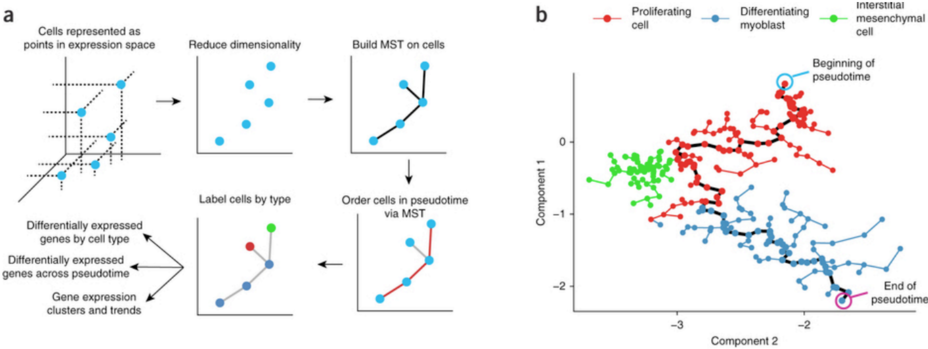
Pseudotemporal ordering

- Idea - cells at different differentiation (or other biological process) stage are presented with different expression profiles
- Dynamics of cellular processes can be reconstructed from expression profiles
- Key assumption: genes do not change direction very often, thus samples with similar transcriptional profiles should be close in order
- Most approaches are dimensionality reduction-based, and apply graph theory designed to traverse nodes in a graph efficiently
- **Monocle** - Independent component analysis, then a minimum spanning three through the dimension-reduced data

https://github.com/mdozmorov/scRNA-seq_notes#time-trajectory-inference

Monocle, An analysis toolkit for single-cell RNA-seq

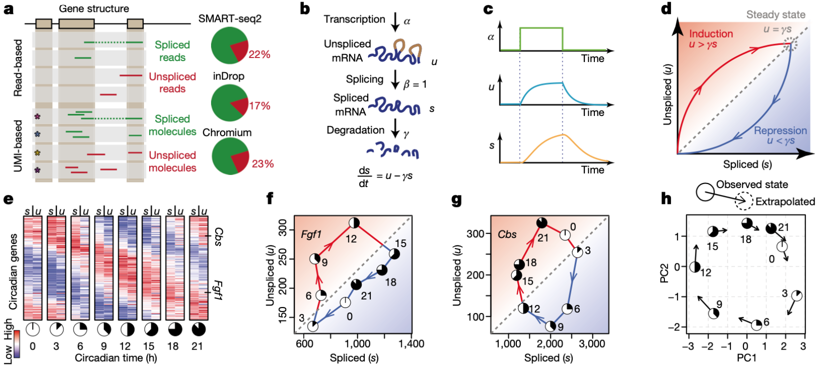
Single-cell trajectories, clustering, visualization, differential expression



https://cole-trapnell-lab.github.io/monocle-release/

Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. "The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells." Nature Biotechnology 32, no. 4 (April 2014): 381–86. https://doi.org/10.1038/nbt.2859.

RNA velocity



La Manno, Gioele, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, et al. "RNA Velocity of Single Cells." Nature 560, no. 7719 (August 2018): 494–98. https://doi.org/10.1038/s41586-018-0414-6.

https://github.com/mdozmorov/scRNA-seq_notes#rna-velocity

Differentially expressed genes

- Goal is to identify cluster/condition-specific genes
- Many scRNA-seq-specific methods were developed
- Methods for bulk RNA-seq (edgeR, DESeq2) perform well

https://github.com/mdozmorov/scRNA-seq_notes#differential-expression

Annotation, subpopulation identification

- Grand question - what cell types are in diffefent clusters?
- Main idea - match cluster/condition-specific gene expression profiles to reference profiles/gene markers
- Many methods, from clustering, joint embedding to graph networks, machine learning
- Need cell markers databases

https://github.com/mdozmorov/scRNA-seq_notes#annotation-subpopulation-identification

https://github.com/mdozmorov/scRNA-seq_notes#cell-markers

scRNA-seq databases

- Numerous scRNA-seq experiments were performed - databases collect processed data
- Each data frequently has meta-data - which cells are which (what cluster they belong to)
- Interesting datasets include immune cells, brain cells, cancer cells
- Human Cell Atlas - all cells in human body https://www.humancellatlas.org/

https://github.com/mdozmorov/scRNA-seq_notes#data

Other analyses

- Functional enrichment analysis - what functions are shared by cluster/condition-specific genes?
- Gene networks - infer gene regulatory networks, https://github.com/mdozmorov/scRNA-seq_notes#networks
- Single Nubleotide Polymorphisms, Copy Number Variants, https://github.com/mdozmorov/scRNA-seq_notes#cnv
- Immuno-analyses, T/B cell receptor sequencing, https://github.com/mdozmorov/scRNA-seq_notes#immuno-analysis
- Power analysis - how many cells to sequence to identify subpopulations. https://github.com/mdozmorov/scRNA-seq_notes#power

scRNA-seq analysis pipelines

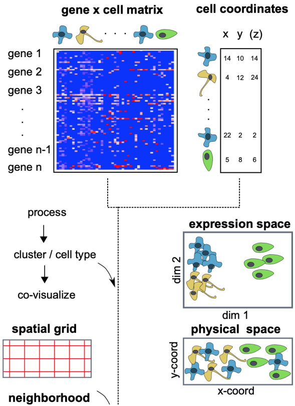
- **Seurat** - the comprehensive collection of tools for single-cell multi-omics analysis. Single ecosystem.
- **Bioconductor** - SingleCellExperiment object, wide variety of packages described in the “Orchestrating Single-Cell Analysis with Bioconductor” (OSCA) book

https://satijalab.org/seurat/

https://bioconductor.org/books/release/OSCA/

https://github.com/mdozmorov/scRNA-seq_notes#courses

Spatial transcriptomics

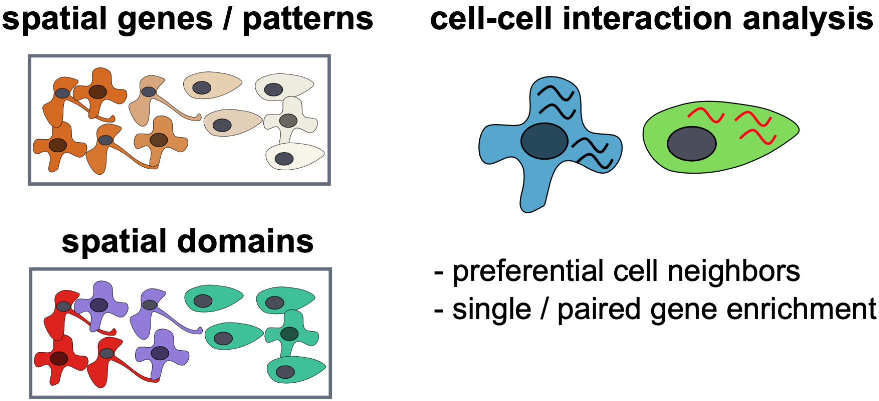


https://github.com/mdozmorov/scRNA-seq_notes#spatial-transcriptomics

https://www.10xgenomics.com/spatial-transcriptomics

Dries, Ruben, Qian Zhu, Chee-Huat Linus Eng, Arpan Sarkar, Feng Bao, Rani E George, Nico Pierson, Long Cai, and Guo-Cheng Yuan. "Giotto, a Pipeline for Integrative Analysis and Visualization of Single-Cell Spatial Transcriptomic Data." Preprint. Bioinformatics, July 13, 2019. https://doi.org/10.1101/701680.

Spatial transcriptomics



spatial genes / patterns

spatial domains

cell-cell interaction analysis

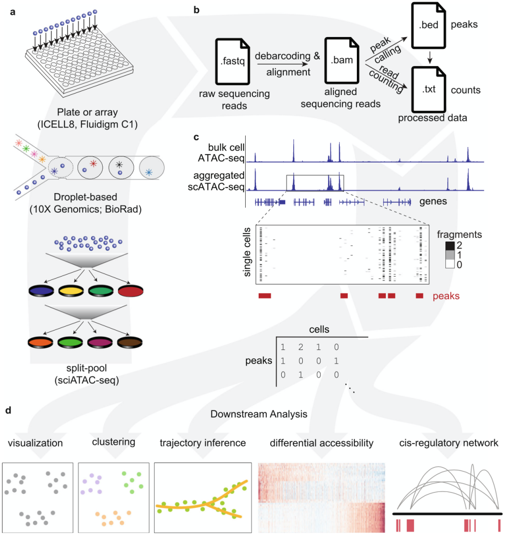
- preferential cell neighbors
- single / paired gene enrichment

https://github.com/mdozmorov/scRNA-seq_notes#spatial-transcriptomics

https://www.10xgenomics.com/spatial-transcriptomics

Dries, Ruben, Qian Zhu, Chee-Huat Linus Eng, Arpan Sarkar, Feng Bao, Rani E George, Nico Pierson, Long Cai, and Guo-Cheng Yuan. "Giotto, a Pipeline for Integrative Analysis and Visualization of Single-Cell Spatial Transcriptomic Data." Preprint. Bioinformatics, July 13, 2019. https://doi.org/10.1101/701680.

single-cell ATAC-seq



	cells			
peaks	1	2	1	0
	1	0	0	1
	0	1	0	0

Chen et al., "Assessment of Computational Methods for the Analysis of Single-Cell ATAC-Seq Data."

single-cell ATAC-seq challenges

- While for an expressed gene several RNA molecules are present in a single cell, scATAC-seq assays profile DNA, a molecule which is present in only two copies per cell (for a diploid organism)
- The low copy number results in an inherent per-cell data sparsity, where only 1– 10% of expected accessible peaks are detected in single cells from scATAC-seq data, compared to 10–45% of expressed genes detected in single cells from scRNA-seq data

Chen et al., "Assessment of Computational Methods for the Analysis of Single-Cell ATAC-Seq Data."

single-cell ATAC-seq challenges

- The potential feature set in scATAC-seq, which includes genome-wide regions of accessible chromatin, is typically 10–20X the size of the feature set in scRNA-seq experiments (which is defined and limited by the number of genes expressed).

Chen et al., "Assessment of Computational Methods for the Analysis of Single-Cell ATAC-Seq Data."

References

https://github.com/mdozmorov/scRNA-seq_notes

https://github.com/mdozmorov/scATAC-seq_notes