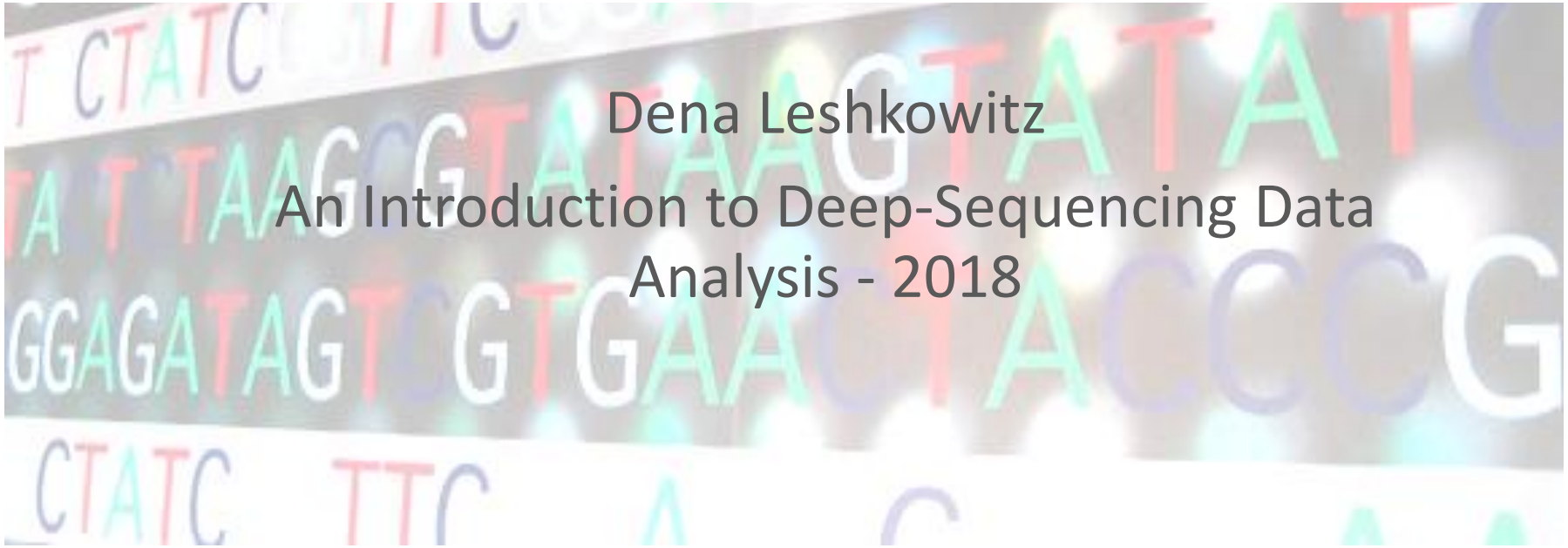
מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Introduction to Single-cell RNA-sequencing (scRNA-seq) Analysis



Dena Leshkowitz

An Introduction to Deep-Sequencing Data Analysis - 2018

RNA Biology ›
Volume 14, 2017 - Issue 5: RNA in disease and development

Review

Single-cell RNA-sequencing: The future of genome biology is now

Simone Picelli

Pages 637-650 | Received 30 Mar 2016, Accepted 09 Jun 2016, Published online: 21 Jul 2016

Goal :

Study gene expression profiling at single cell resolution

- Make Every Cell Count!

Plan

- Introduction to scRNA-Seq
- Cell Capture Methods
- Experimental Design
- Research examples
- Bioinformatics Analysis- Cell Ranger

Bulk RNA-Seq Versus Single Cell RNA-Seq

Bulk RNA-Seq

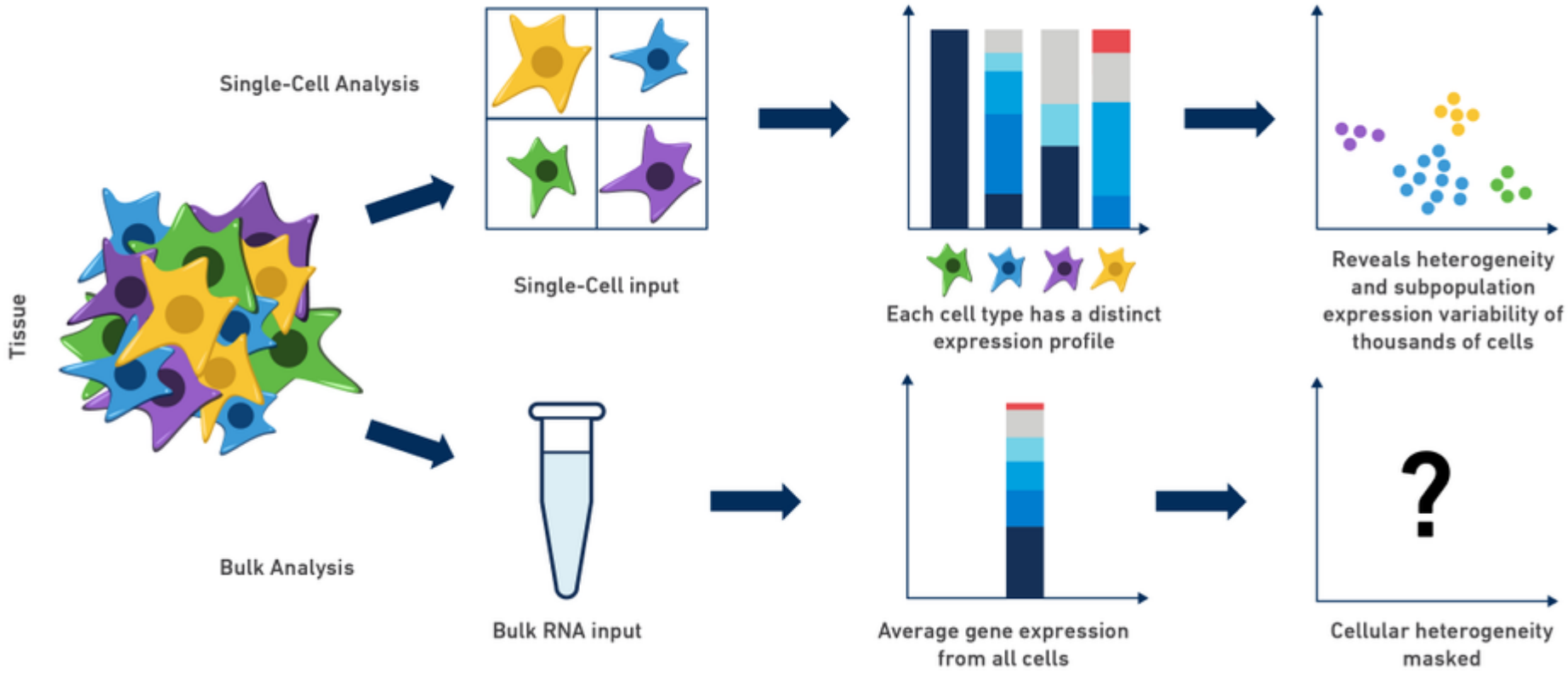
Provides gene measurements that are **averaged** over thousands of cells

scRNA-Seq

Provides gene measurements for individual cells

Categorize Cell types

Bulk RNA Versus scRNA-Seq



community.10xgenomics.com

Bulk RNA-Seq

- A major breakthrough (replacing microarrays) been widely used in the last decade
- Measures the average expression level for each gene across a large population of input cells
- Useful for comparative transcriptomics (different conditions, tissues, genotypes…)
- Insufficient for studying heterogeneous systems, such as early development studies and complex tissues (brain)
- Does not provide insights into the stochastic nature of gene expression

scRNA-seq

- A new technology, first publication by (Tang et al. 2009)
- Gained popularity from ~2014 when new protocols and lower sequencing costs made it more accessible
- Measures the distribution of expression levels for each gene across a population of cells
- Allows to study new biological questions in which cell-specific transcriptome are important, e.g. identification of cell types, discovering heterogeneity in cell responses…
- Data sets range from hundreds to thousands of cells and increase in size every year
- Currently there are several different protocols in use, e.g. SMART-seq2 (Picelli et al. 2013), CELL-seq (Hashimshony et al. 2012) and Drop-seq (Macosko et al. 2015)
- There are commercial platforms available supporting the technology, including the Fluidigm C1, Wafergen ICELL8 and the 10X Genomics Chromium
- Some computational analysis methods from bulk RNA-seq can be applied
- **Computational analysis requires alterations of the existing methods or development of new ones**

Plan

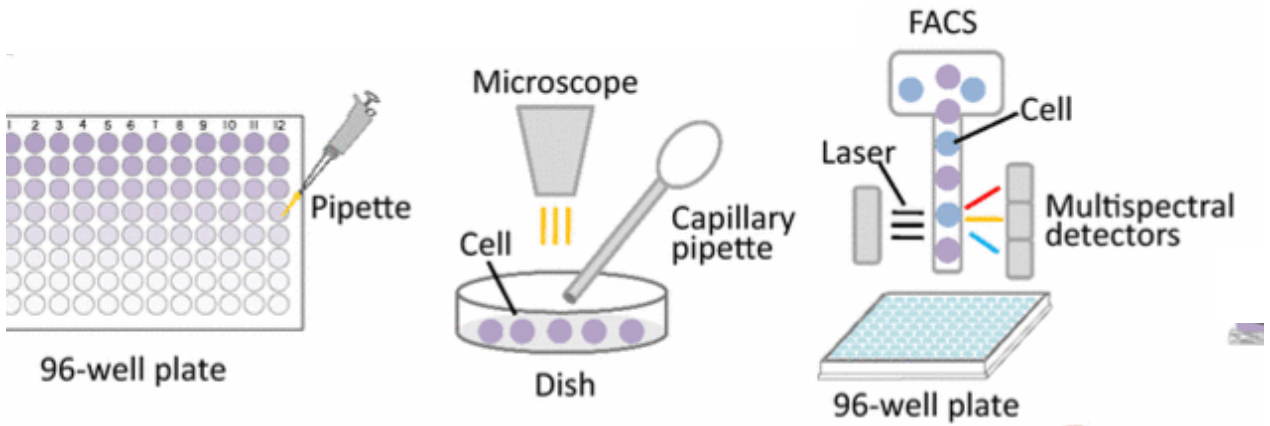
- Introduction to scRNA-Seq
- Cell Capture Methods & Library creation
- Experimental Design
- Research examples
- Bioinformatics Analysis

Methods for Cell Capture

Low-throughput

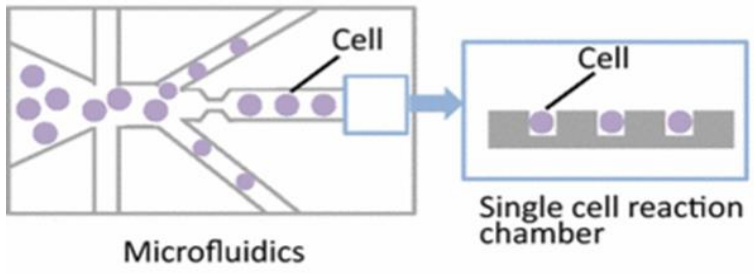
Well-based platform

Cells are isolated using for example pipette or laser capture or FACS and placed in microfluidic wells

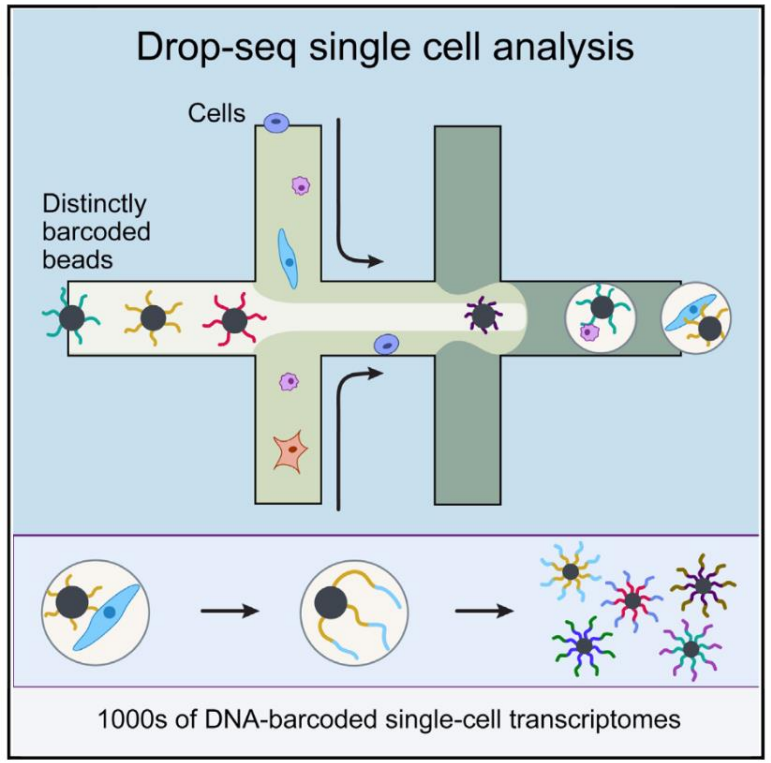


Microfluidic platforms

Such as Fluidigm's C1, advantage small volume

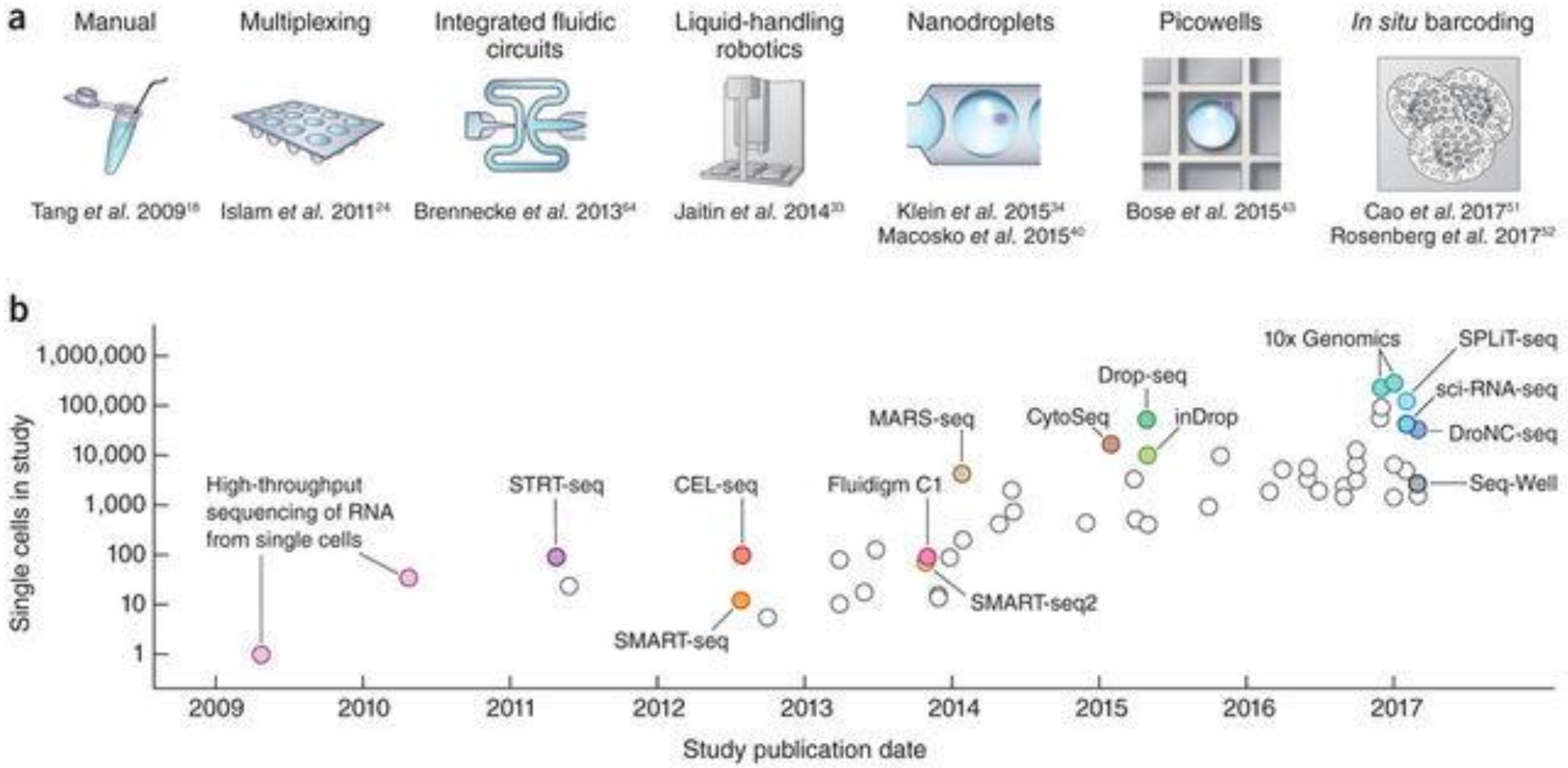


High-throughput Droplet based platform



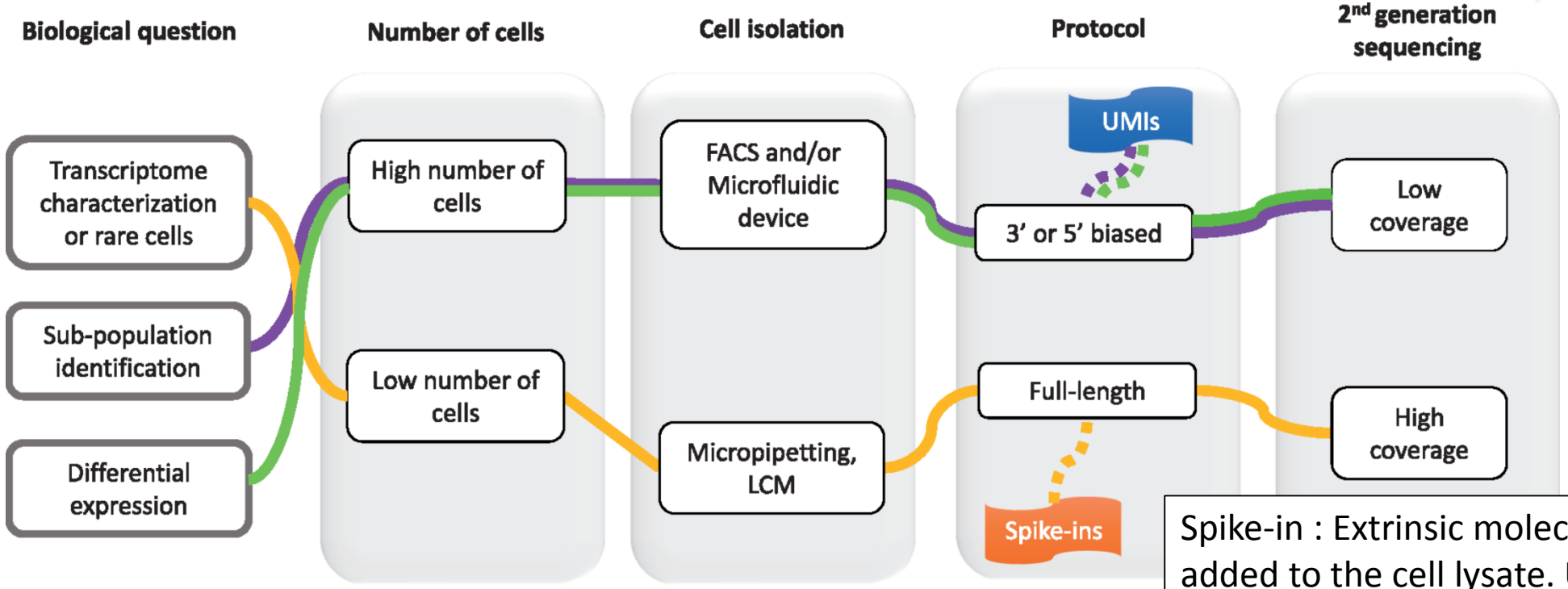
- Each individual cell is captured inside a nanoliter droplet together with a bead.
- The bead is loaded with the enzymes required to construct the library and unique barcode
- All of the droplets can be pooled and sequenced together

Exponential Growth of Scale in Number of Cells per scRNA-seq experiment



Svensson et al. Nature Protocols volume 13, pages 599–604 (2018)

Major Differences in Single Cell Platforms



Spike-in : Extrinsic molecules added to the cell lysate. Used to filter empty wells & normalize

From: How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives

Brief Bioinform. Published online January 31, 2018. doi:10.1093/bib/bby007

Platforms available at Weizmann

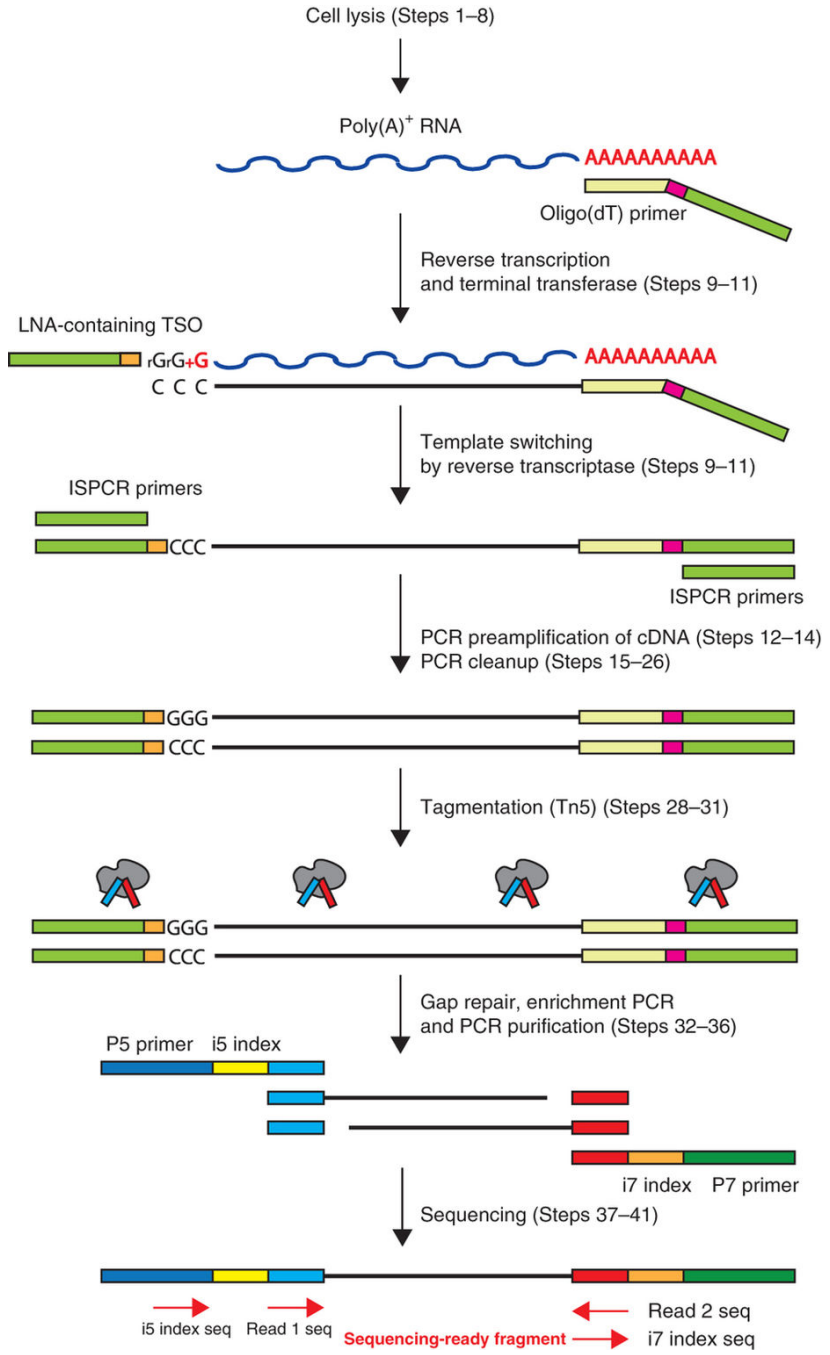
Read more - Experimental design for single-cell RNA sequencing
Baran-Gale et al. Briefings in Functional Genomics, Published: 08 November 2017

Protocol	Number of cells	Sequenc-ing method	Number of reads per per cell	Sequencing machine	Advantages	Disadvan-tages	Usage
10x Genomics Chromium **Includes Cell Isolation + library construction**	Up to 100000, usually 100 to thousands	3' tag method	50,000 reads/cell for RNA-rich cells; 30,000 reads/cell for small primary cells	**LSCF Sandbox: Next-Seq**	**UMI** – ability to remove PCR duplicates	Doublets Dropouts Empty drops	Assess large numbers of cells
SmartSeq2 Nextera **Construct Library -** mRNA capture, RT and amplification	Up to 384	Full length	Minimum 1M reads/cell	**INCPM: HiSeq or Next-Seq**	Can add **spike-ins** more genes per cell	Batch effect; 3'- bias	Alleles, isoforms

Smart-seq2 protocol

Full-length amplification of the transcripts :

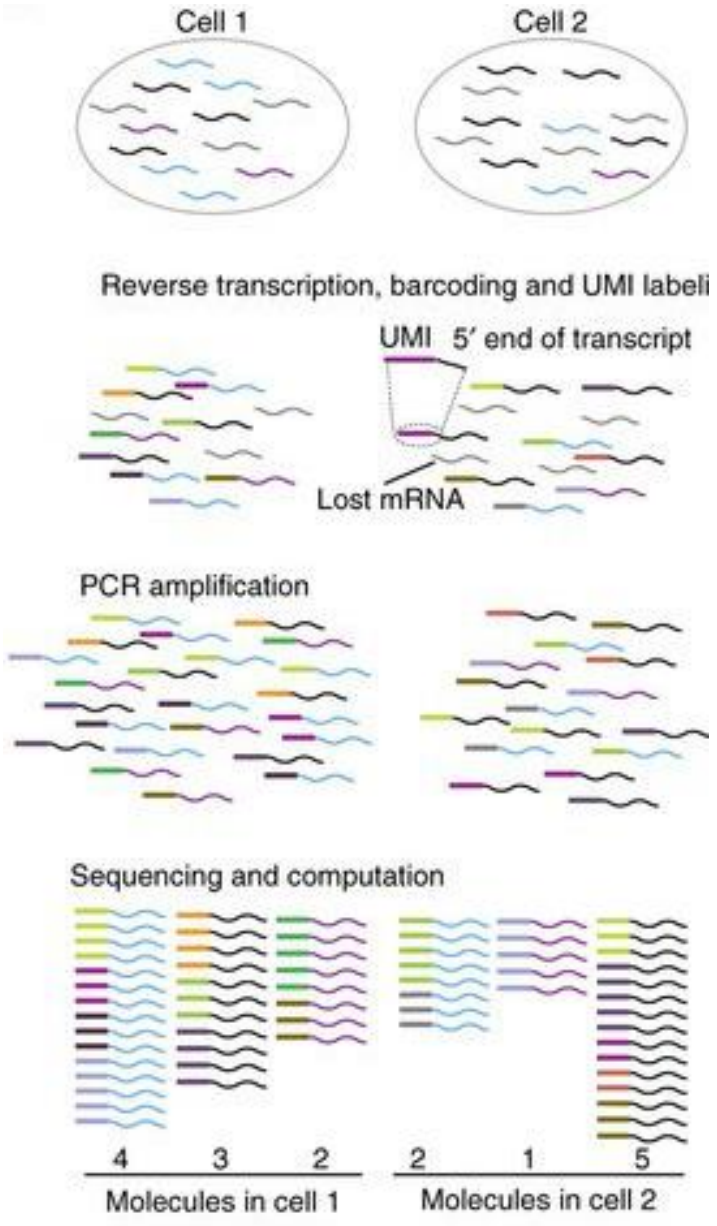
- MMLV reverse transcriptase is able to add at the 5’-end of the RNA template, which corresponds to the 3’-end of the new cDNA strand, a few non-templated cytosines.
- These cytosines serve as an extended template for a helper oligonucleotide (called Template Switching Oligo - Locked Nucleic Acid) that allows the reverse transcriptase to ‘switch’ the template and synthesize the new cDNA strand



Using UMI to Remove PCR Amplification

Reads are considered duplicated, if they map to the same gene and have the same UMI

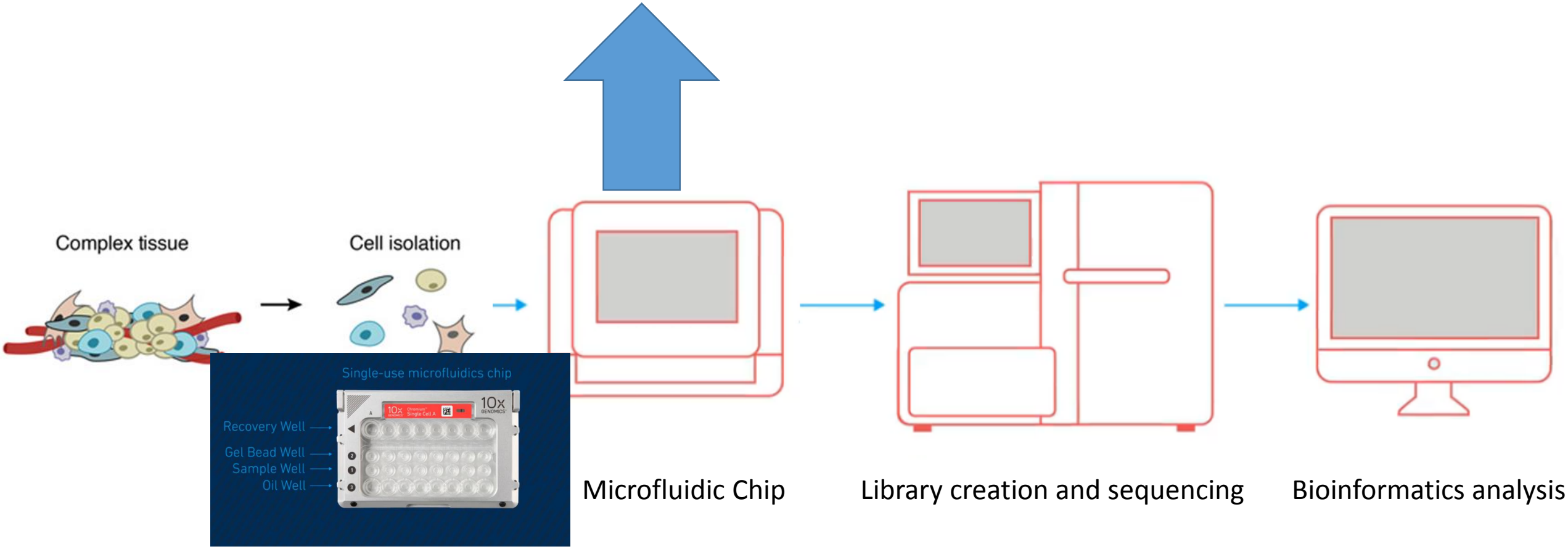
Instead of counting reads we will count number of unique UMIs per gene.



This figure is adapted from [Islam et al (2014)]

10X Genomics - Chromium™ Single Cell 3’ Solution (LSCF)

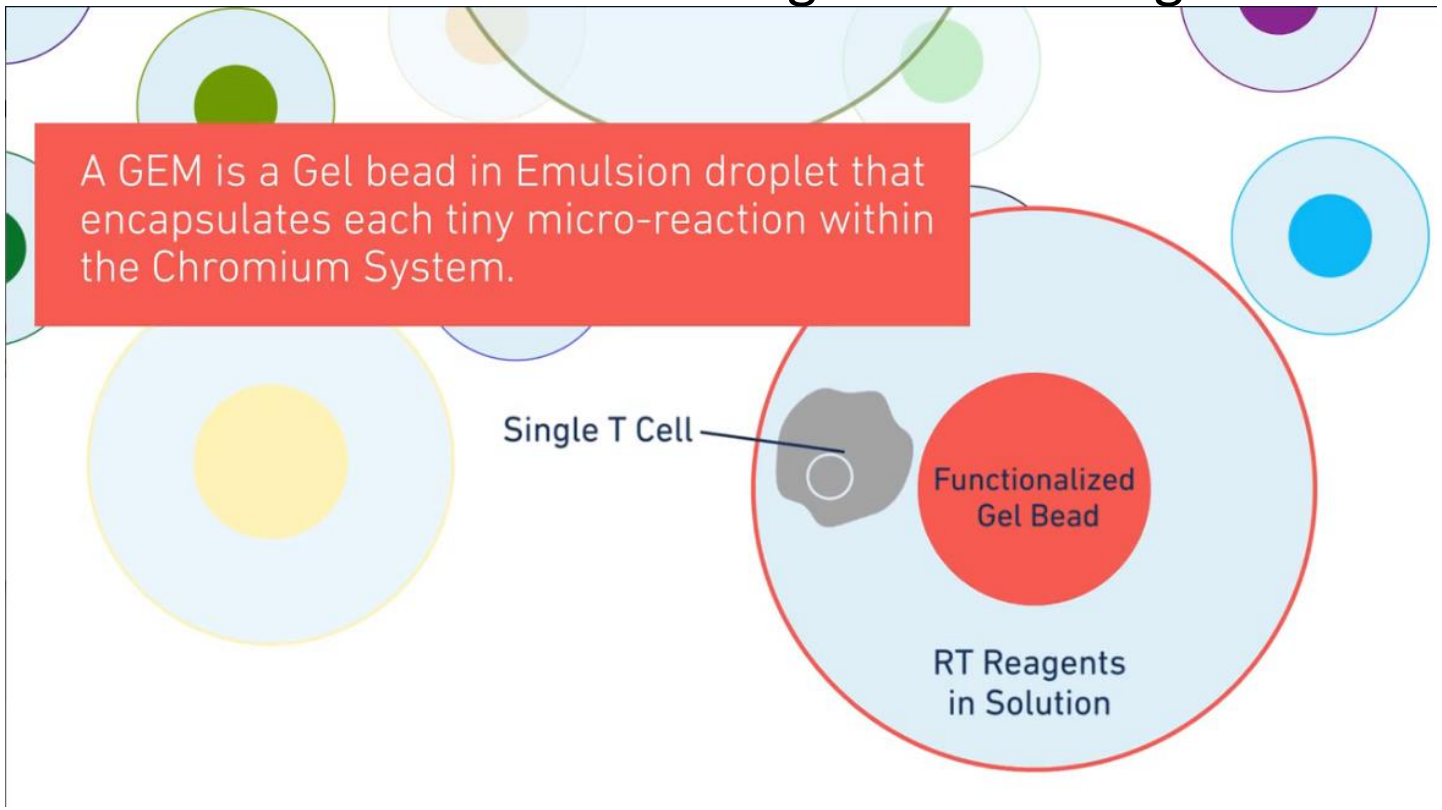
The microfluidic chip enables us to capture single cells and prepare cell-barcoded cDNA libraries to sequence with Illumina machines

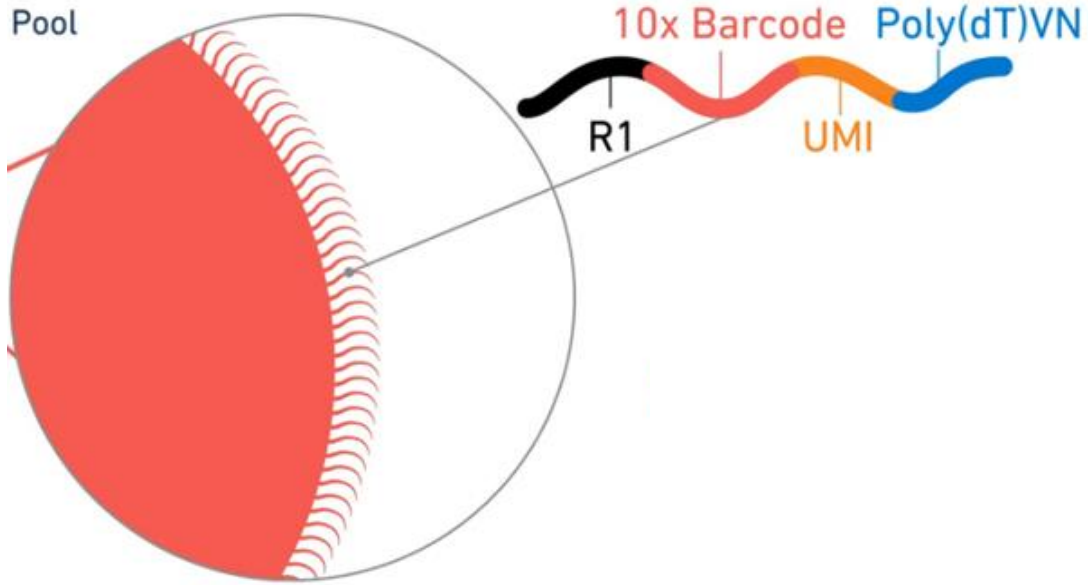


Within the Chip - Gel Bead Emulsion (GEM) Droplets Are Formed

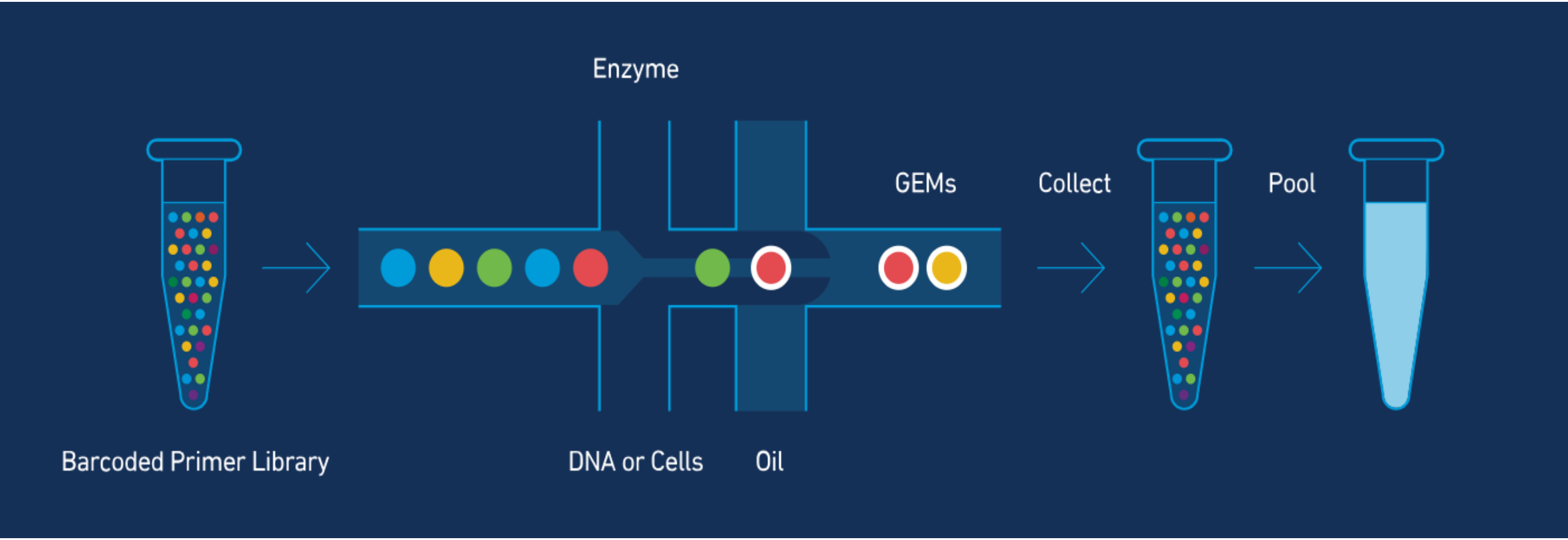
These droplets are tiny micro-reactions containing:

- A single cell
- Reverse transcription (RT)
- Reagents
- Gel Beads containing barcoded oligonucleotides

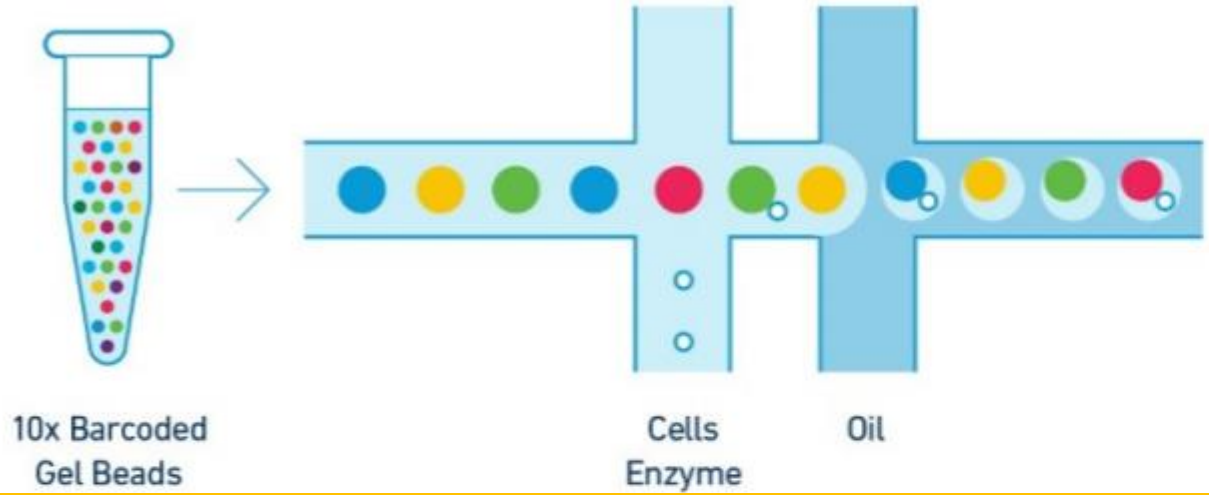


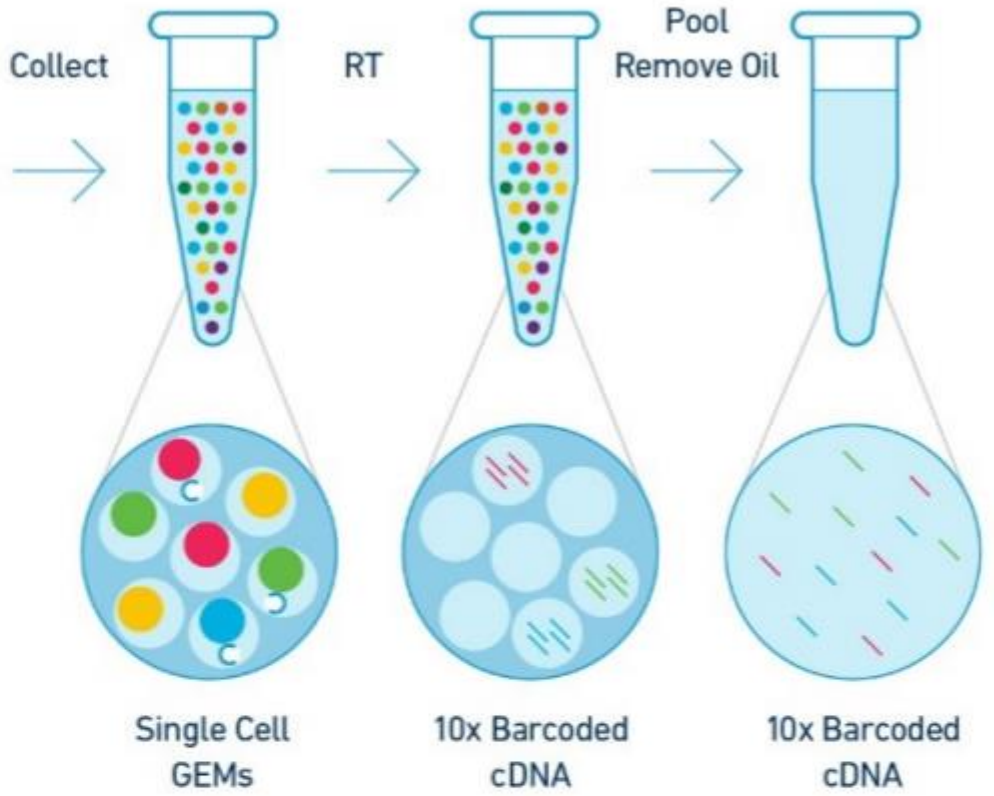


Within each microfluidic channel, thousands of GEMs are formed per ~6-min run, encapsulating thousands of cells in GEMs.



Gel beads loaded with primers and barcoded oligonucleotides are first mixed with cells and reagents, and subsequently mixed with oil-surfactant solution at a microfluidic junction.

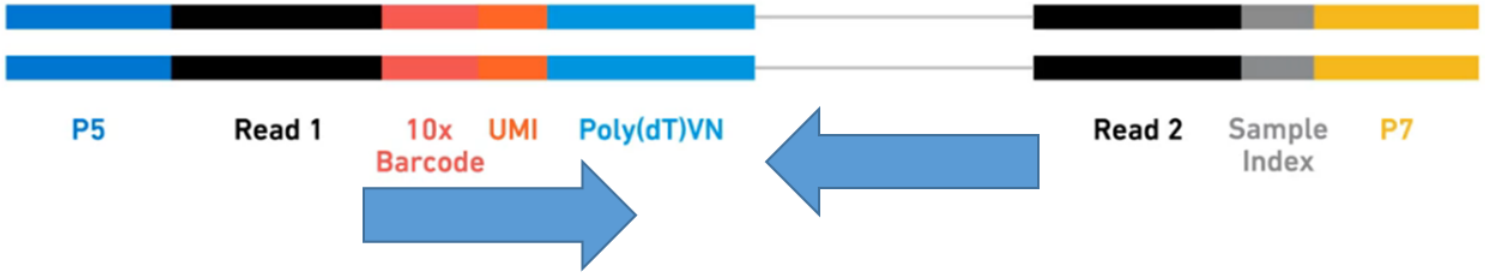




- Within each GEM reaction vesicle a single cell is lysed
- The Gel Bead is dissolved to free the identically barcoded RT oligonucleotides into solution, and reverse transcription of poly-adenylated mRNA occurs.
- As a result, all cDNAs from a single cell will have the same barcode, allowing the sequencing reads to be mapped back to their original single cells of origin.

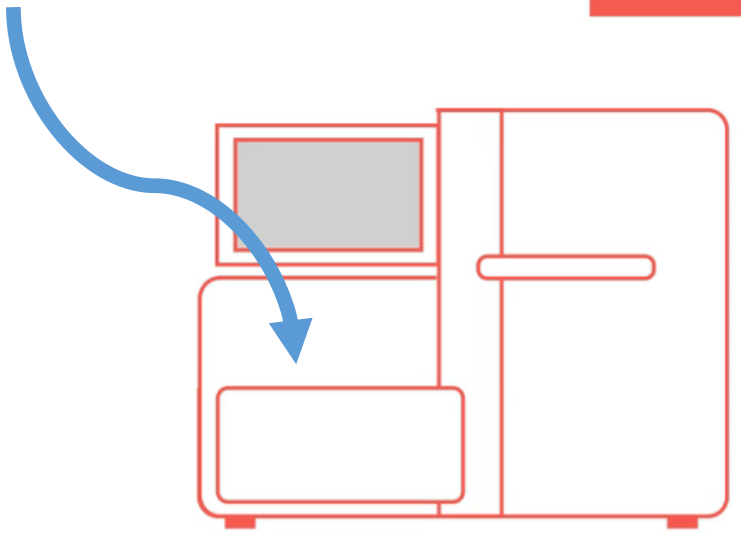
From Library construction to Sequencing and Data analysis

After additional steps we have a library compatible with Illumina sequencing

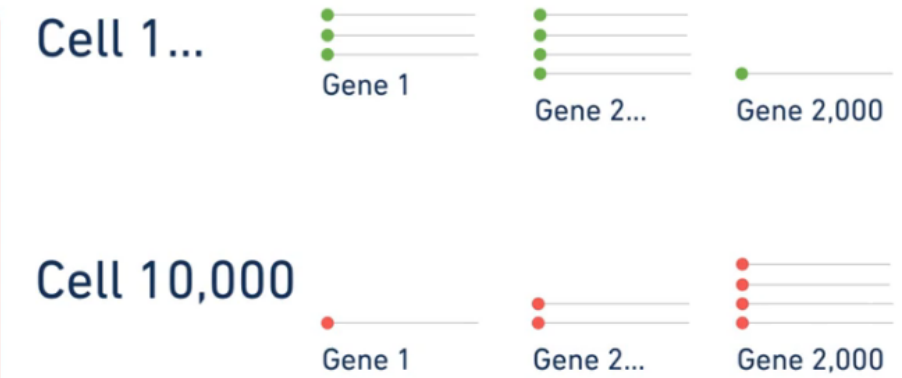


READ 1 contains cell barcode and UMI (26bases)

READ 2 contains insert sequence



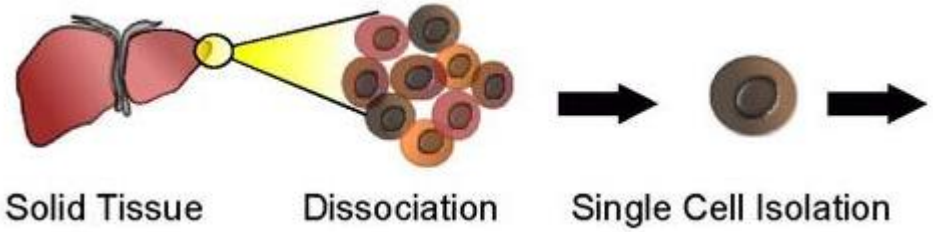
...enables massive transcription profiling of thousands of individual cells...



Plan

- Introduction to scRNA-Seq
- Capture Methods
- Experimental Design
- Research examples
- Bioinformatics Analysis

Experiment Design



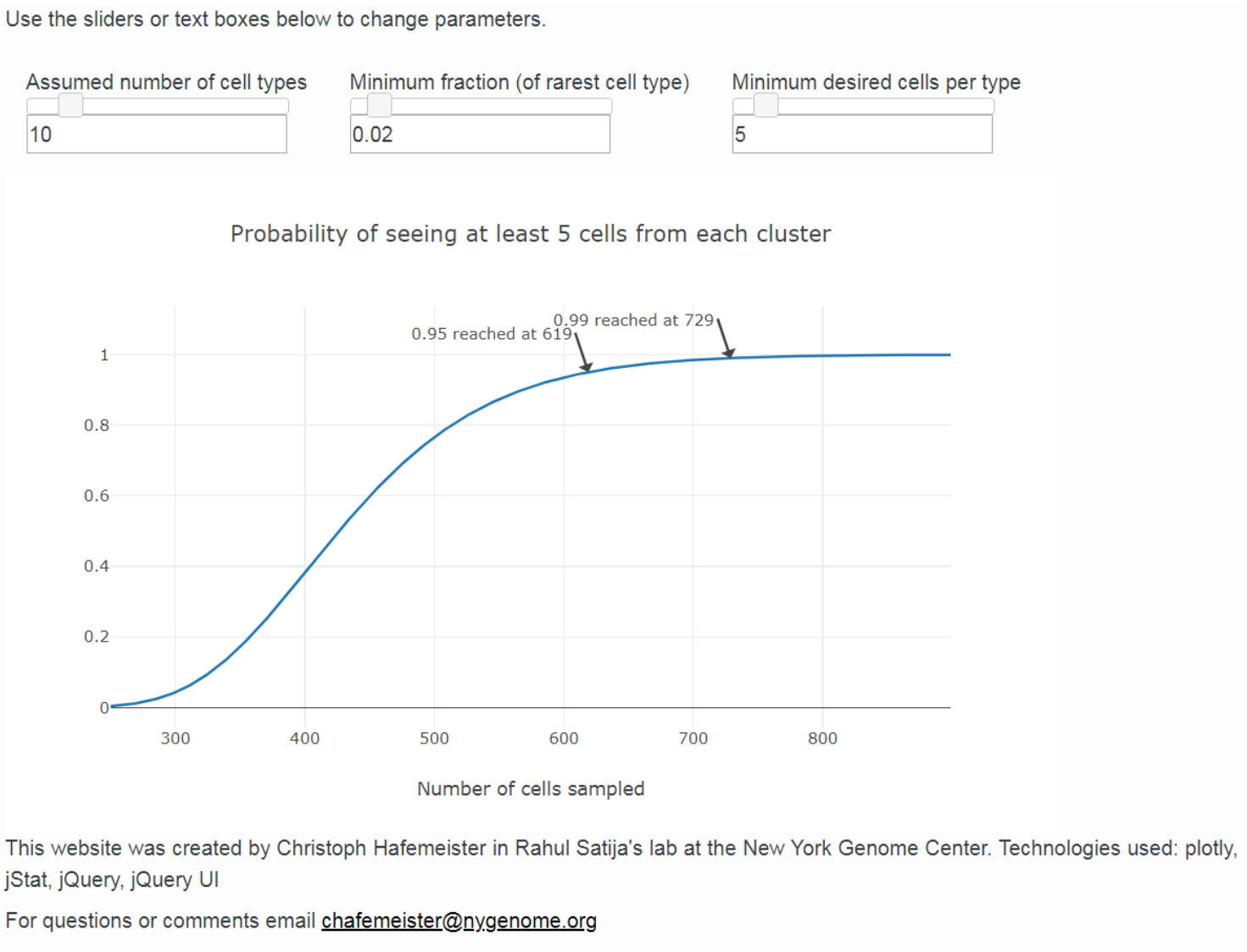
- Before conducting a single-cell experiment, trial single-cell separation methods and assess cell viability
- Cell isolation should be performed as fast as possible with all downstream work being carried out on ice
- There are cell-type-specific differences in recovery after FACS sorting, possibly because of cell size, with larger cells resulting in fewer cells recovered
- Avoid batch effects between experiments

Deciding on Appropriate Cell Number

The required number of cells can be estimated based on the expected heterogeneity of all cells in a sample using a negative binomial distribution.

- Interactive tool

https://satijalab.org/howmanycells



Plan

- Introduction to scRNA-Seq
- Capture Methods
- Experimental Design
- Research examples
- Bioinformatics Analysis

scRNA-Seq in Research

- In the human body there are $4*10^{13}$ cells, that are classified to ~200 cell types based on morphology
  - Discover new and rare cell types
    - For example, Campbell et al., 2017 sequenced 20,921 cells from mouse hypothalamus and they were able to identify neuronal subpopulations comprised of fewer than 50 cells (<0.2%).
- Characterize differences between similar cell-types
  - For instance, dissecting differences among hematopoietic stem-cells requires detection of relatively lowly expressed transcription factors
- Single-cell RNA-seq could play a key role in personalized medicine by facilitating characterization of cells, pathways, and genes associated with human diseases such as cancer

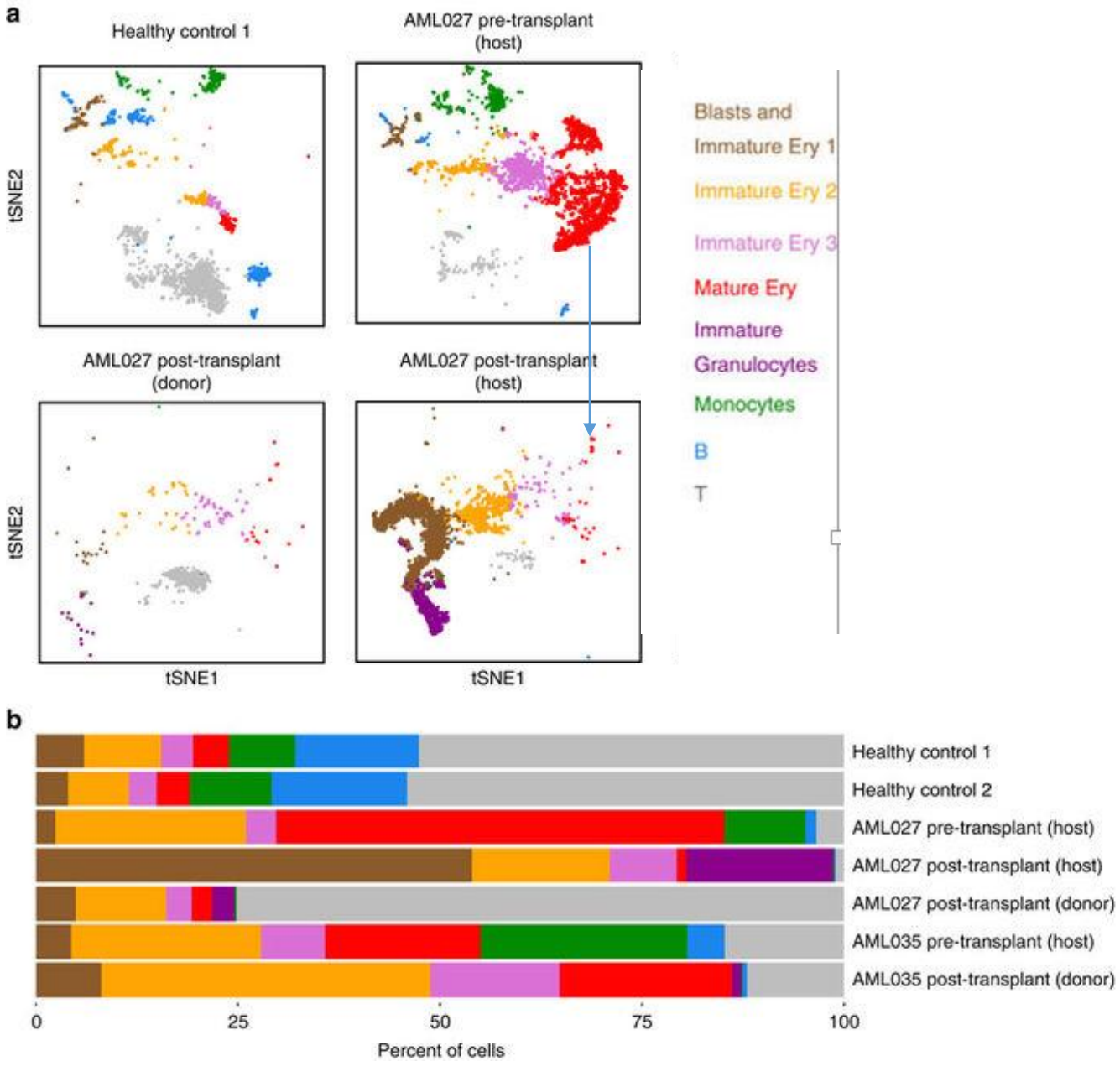
Study Example

Massively parallel digital transcriptional profiling of single cells Zheng et al. Nature Communications volume 8, Article number: 14049 (2017)

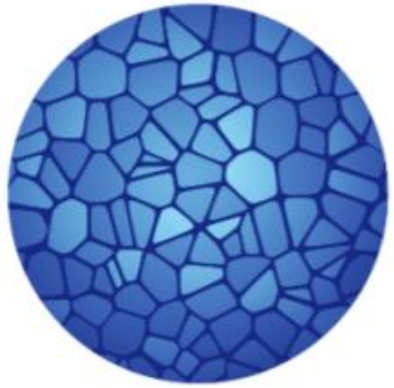
Single-cell RNA-seq libraries were generated from cryopreserved bone marrow mononuclear cell (BMMC) samples obtained from two AML patients before and after undergoing hematopoietic stem cell transplantation.

AML027 showed the highest level of erythroid cells (>80%, consist of mostly mature erythroids) before transplant, consistent with the erythroleukaemia diagnosis of AML027.

Third, monocytes are abundant in both AML patients before transplant (10% and 25% in AML027 and AML035 respectively), but are not detectable after transplant.



Where are we heading? The Single Cell Future

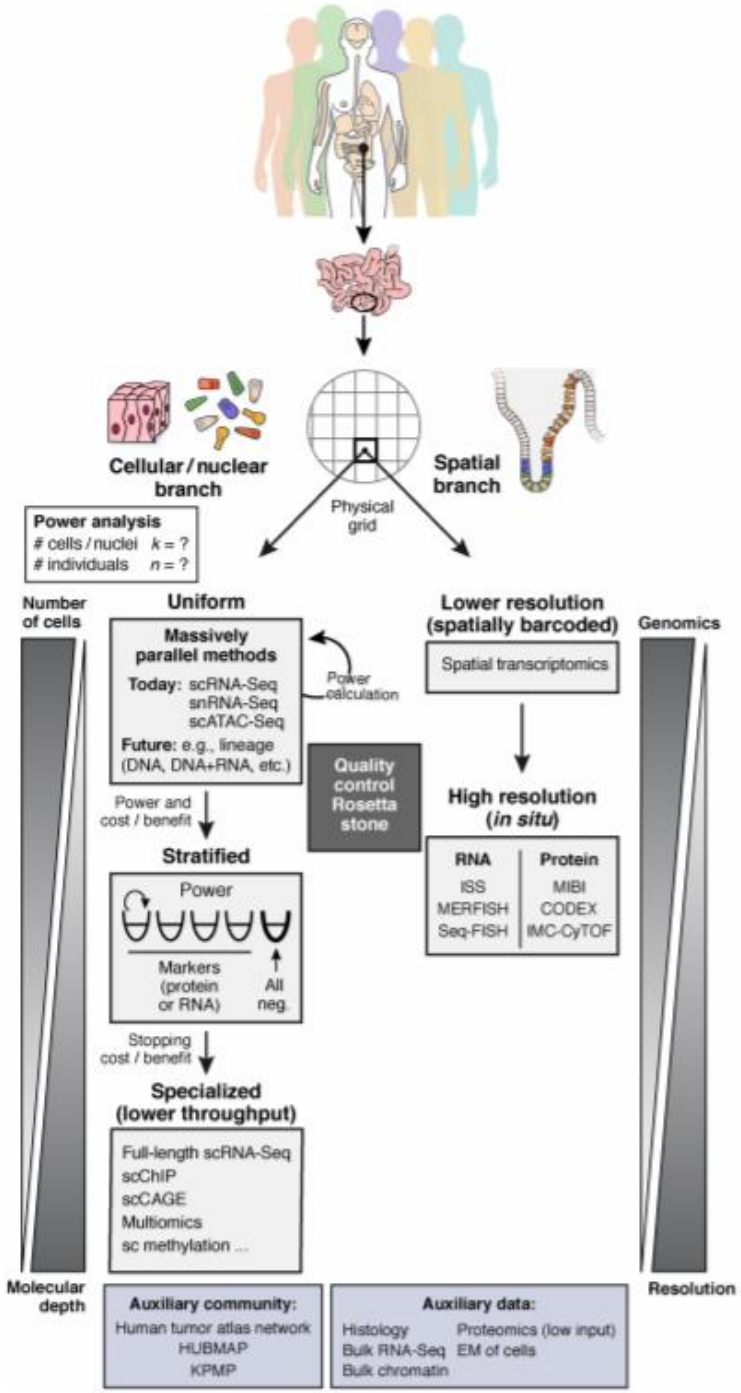


THE HUMAN CELL ATLAS

White Paper

The HCA Consortium
October 18, 2017

The Human Cell Atlas (HCA) will be made up of comprehensive reference maps of all human cells — the fundamental units of life — as a basis for understanding fundamental human biological processes and diagnosing, monitoring, and treating disease.



Plan

- Introduction to scRNA-Seq
- Capture Methods
- Experimental Design
- Research examples
- Bioinformatics challenges and Cell Ranger analysis pipeline

Challenges

Despite its power and high resolution, scRNA-seq has some open challenges related to the higher level of technical noise and data complexity with respect to bulk RNA-seq :

- Amplification (up to 1 million fold) : The amount of RNA present in a single cell is limited, and ranges from 1–to 50 pg depending on cell type.
  - Solution - UMIs
- Gene ‘dropouts’ : a gene is observed at a moderate expression level in one cell but is not detected in another cell.
  - An estimated 10–20% of transcripts are sampled.
- Doublets : some droplets/wells contain more than a single cell (two cells may be physically captured together)
  - Few percent of the droplets
- Empty droplets: need to distinguish cells from empty droplets
  - Majority of the droplets are empty

10X Genomics analysis

- Cell Ranger analysis pipeline performs the following:

1. Identifies the read cell origin using the cell-barcode (correct cell barcode allowing 1 mismatch - hamming distance of 1)
2. Aligns reads to genome using STAR (uniquely aligned to gene exons)
3. Count UMIs per cell per gene (allowing hamming distance of 1 in UMI)
4. Filter real cells (barcodes) from empty droplets

Cell Ranger Report

Cell Ranger · CD3_TCRb ·

SUMMARY ANALYSIS

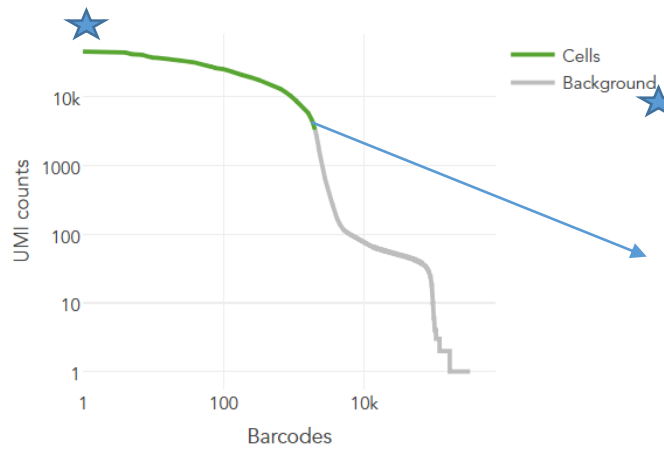
Estimated Number of Cells
2,017

Mean Reads per Cell	Median Genes per Cell
232,047	2,528

Sequencing

Number of Reads	468,040,079
Valid Barcodes	96.8%
Reads Mapped Confidently to Transcriptome	50.1%
Reads Mapped Confidently to Exonic Regions	51.7%
Reads Mapped Confidently to Intronic Regions	13.3%
Reads Mapped Confidently to Intergenic Regions	3.6%
Reads Mapped Antisense to Gene	3.7%
Sequencing Saturation	87.5%
Q30 Bases in Barcode	96.0%
Q30 Bases in RNA Read	83.0%
Q30 Bases in Sample Index	94.1%
Q30 Bases in UMI	95.0%

Cells



Estimated Number of Cells	2,017
Fraction Reads in Cells	84.3%
Mean Reads per Cell	232,047
Median Genes per Cell	2,528
Total Genes Detected	18,631
Median UMI Counts per Cell	9,198

Sample

Name	CD3_TCRb

M = Cell with maximum total UMI counts (99th percentile of the top N)

A barcode is called a cell if -
Total UMI count > M/10 is called a cell

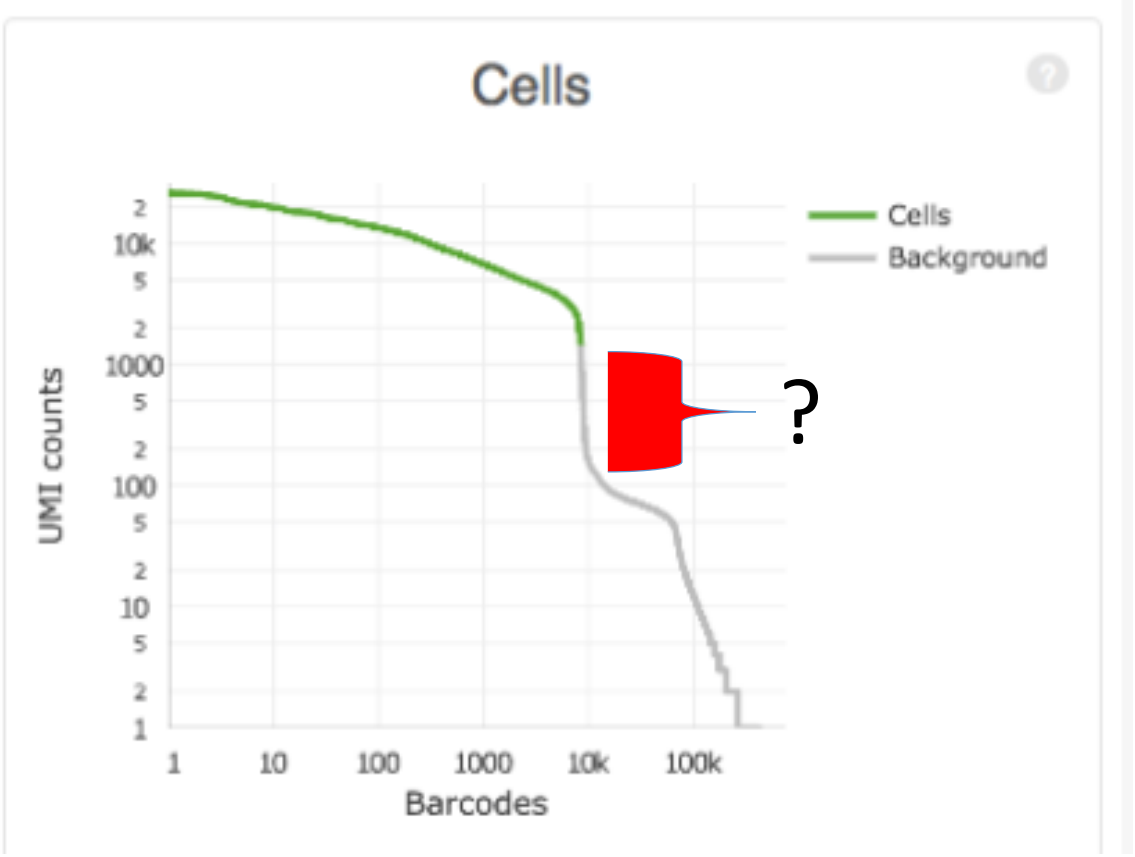
Why do we have Barcodes with Counts that are not Called Cells? Ambient RNA

An empty droplet does not contain a cell but will still contain “ambient” RNA i.e. cell-free transcripts in the solution in which the cells are suspended.

Ambient RNA can be actively secreted by cells or released upon cell lysis (possibly induced by the stresses of dissociation and microfluidics).

The presence of ambient RNA means that many empty droplets will contain material for reverse transcription and library preparation, resulting in non-zero total UMI counts for the corresponding barcodes.

However, the expression profiles for these barcodes do not originate from any individual cell and need to be removed prior to further analysis to avoid misleading biological conclusions



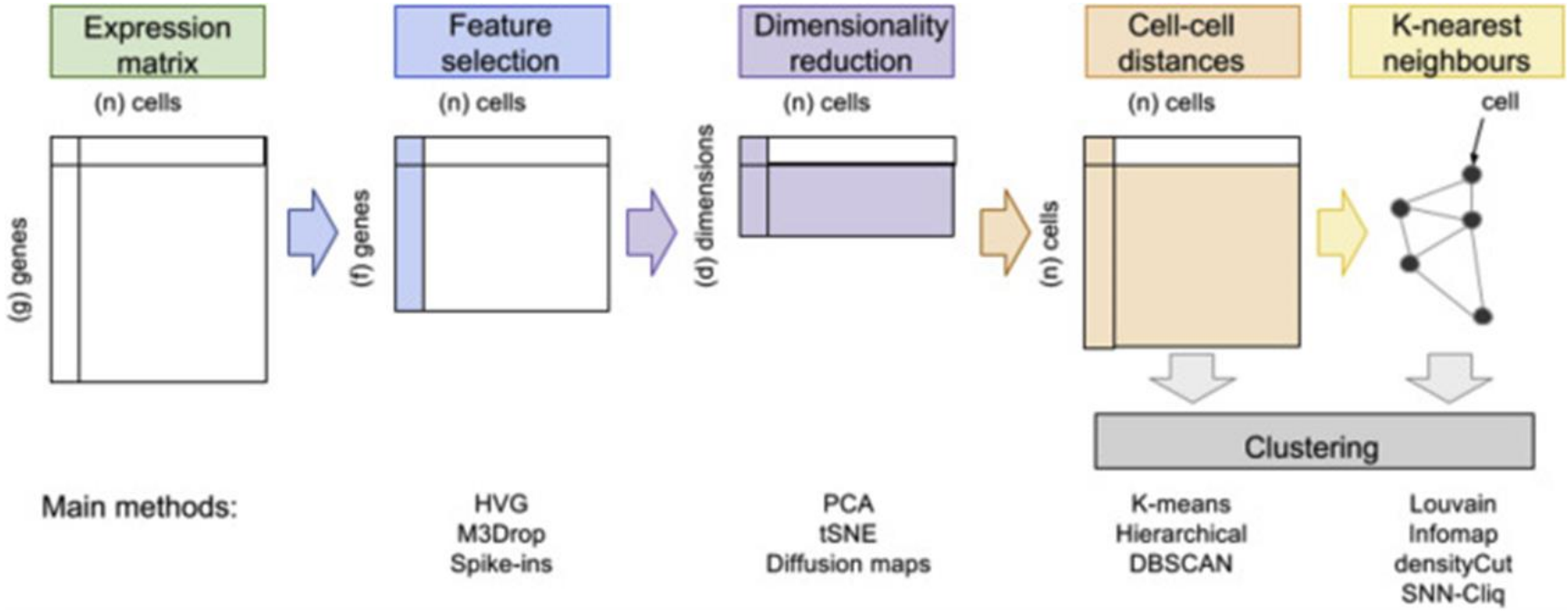
Lun et al. doi: https://doi.org/10.1101/234872

Cell Ranger Analysis Steps

6. Generates gene-cell (barcodes) matrices
7. PCA - Dimensionality Reduction

- In order to reduce the gene expression matrix to its most important features, Cell Ranger uses Principal Components Analysis (PCA)
- This changes the dimensionality of the dataset: (cells x genes) -> (cells x d)

d is a user-selectable number of principal components

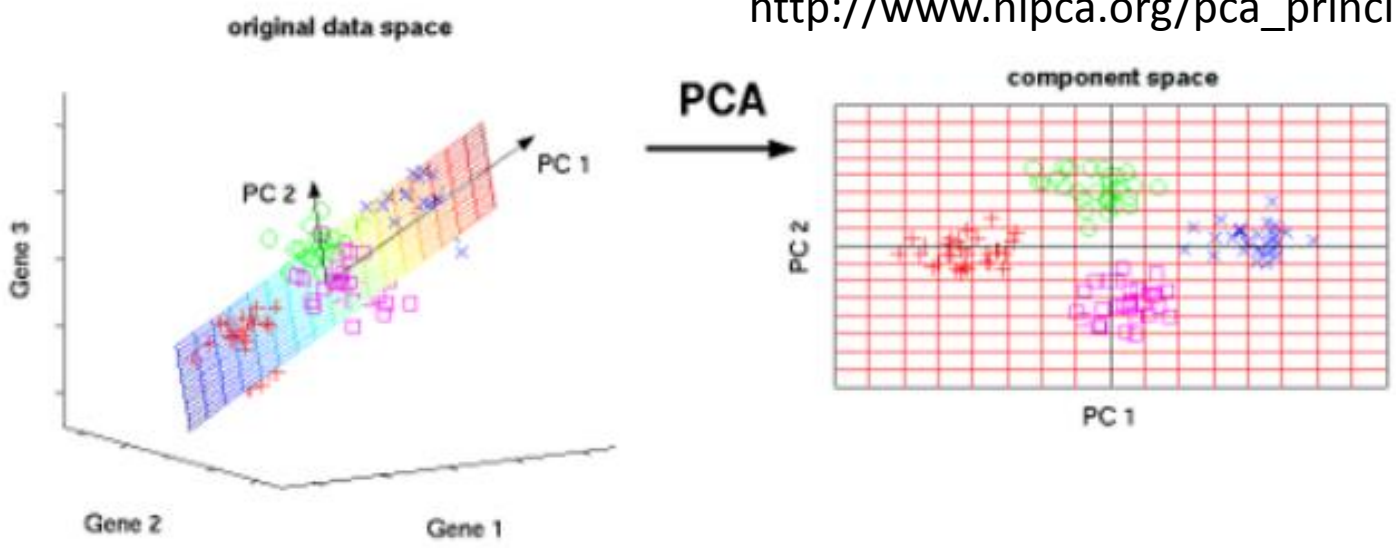


Andrews et al. Molecular Aspects of Medicine Volume 59, February 2018, Pages 114-122

PCA

Finds a linear projection of high dimensional data so that the variance is maximized (and reconstruction error is minimized)

http://www.nlpca.org/pca_principal_component_analysis.html



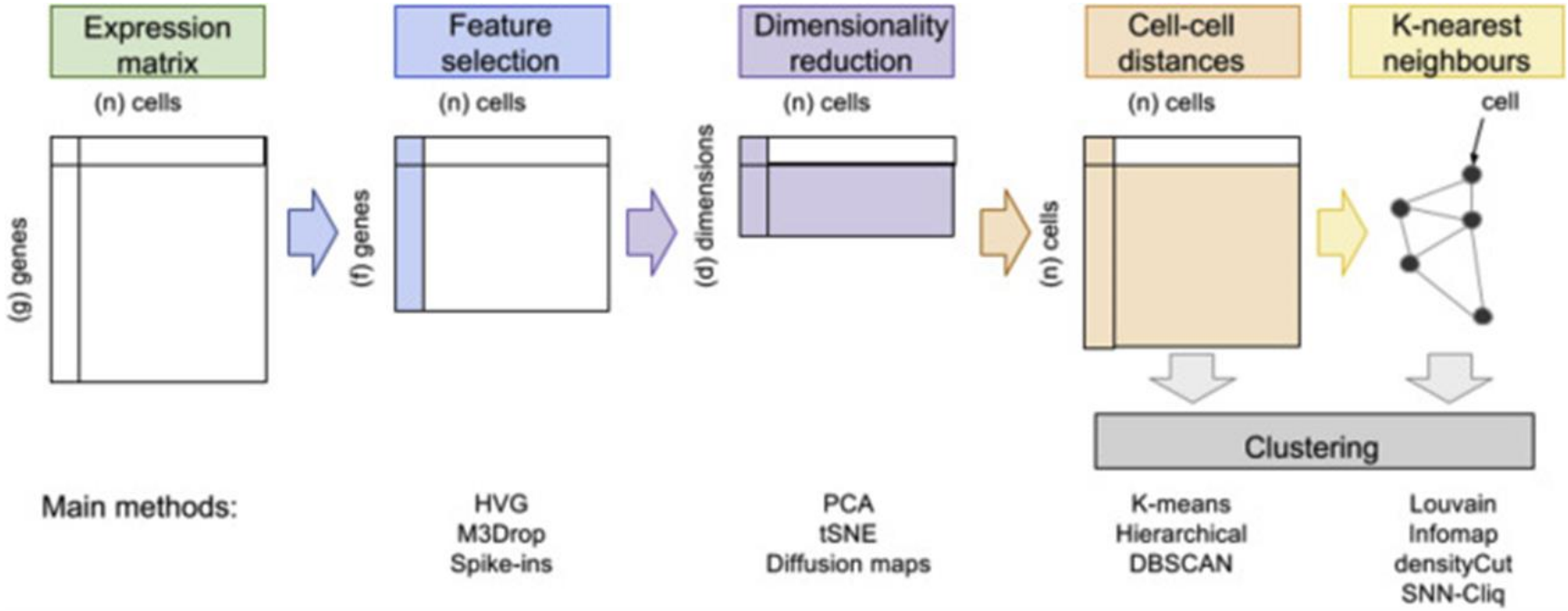
Read more : https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigen

Cell Ranger Analysis Steps

6. Generates gene-cell (barcodes) matrices
7. PCA - Dimensionality Reduction

- In order to reduce the gene expression matrix to its most important features, Cell Ranger uses Principal Components Analysis (PCA)
- This changes the dimensionality of the dataset: (cells x genes) -> (cells x d)

d is a user-selectable number of principal components



Andrews et al. Molecular Aspects of Medicine Volume 59, February 2018, Pages 114-122

Cell Ranger Analysis Steps

8. Clustering Cells : K-means & Graph-based, both of which **operate in the PCA space**

K-means

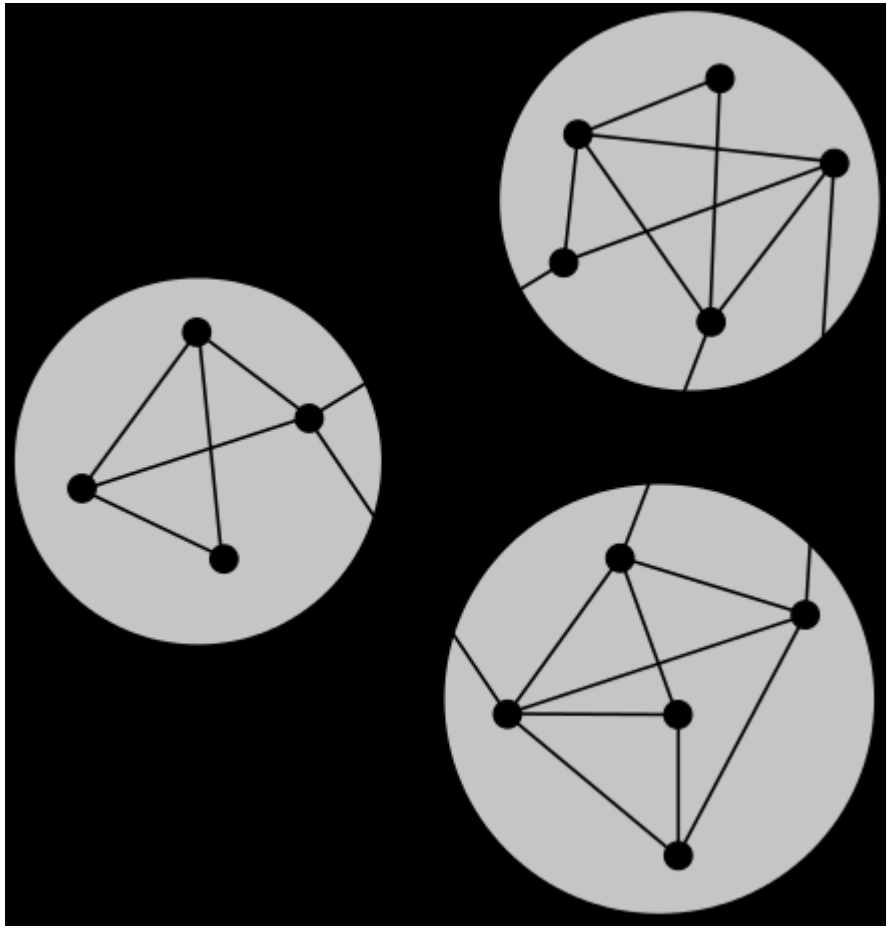
Clustering across a range of K values, where K is the preset number of clusters.

Graph Based

The graph-based clustering algorithm consists of building a sparse nearest-neighbor graph - where cells are linked if they are among the k nearest Euclidean neighbors of one another.

This is followed by Louvain Modularity Optimization : Communities are groups of nodes within a network that are more densely connected to one another than to other nodes.

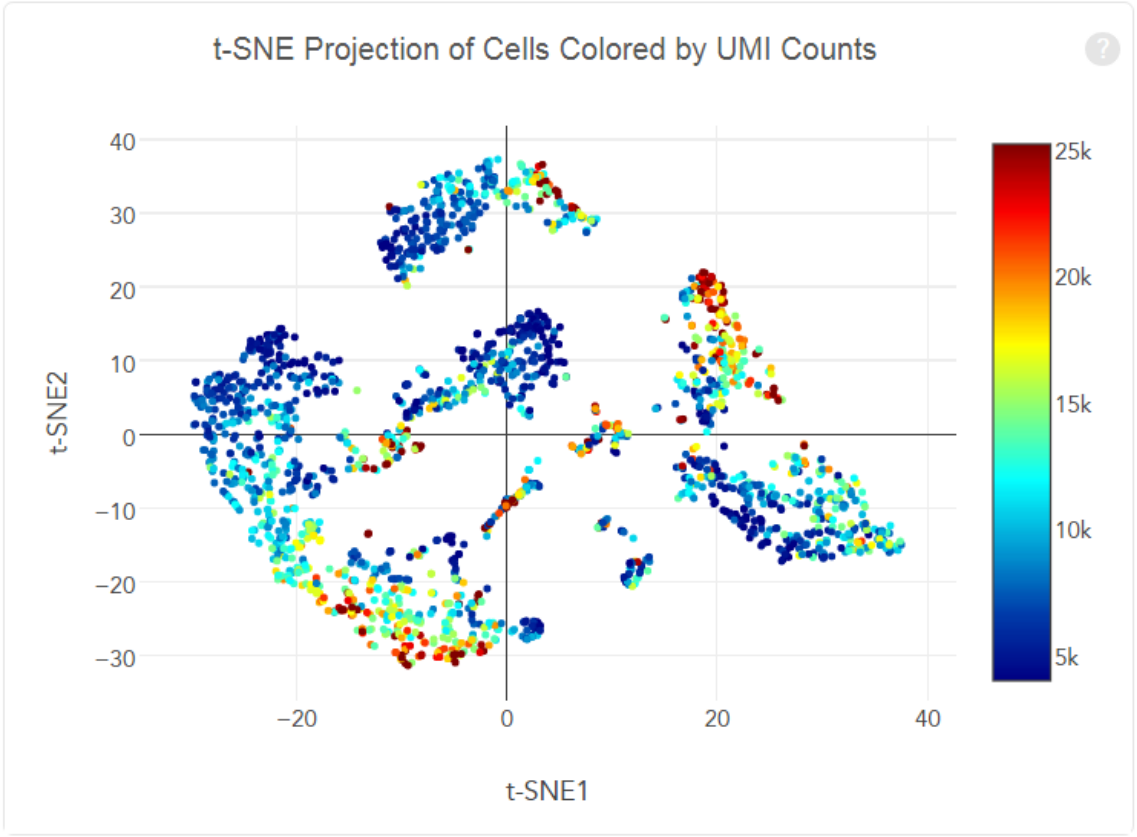
An additional cluster-merging step is done: Perform hierarchical clustering on the cluster-medoids in PCA space and merge pairs of sibling clusters if there are no genes differentially expressed between them.

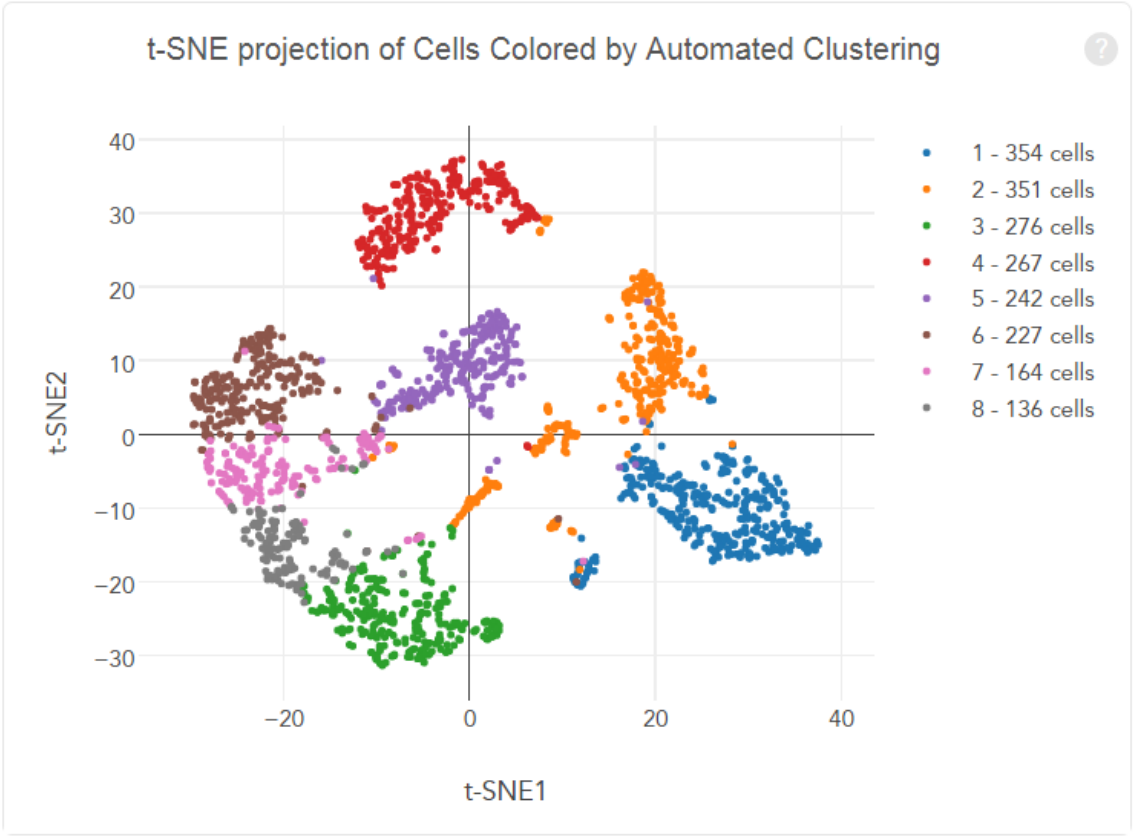


https://en.wikipedia.org/wiki/Community_structure

Cell Ranger Analysis Report : Viewing Clustering results with tSNE Plots

Clustering Type: Graph-based





Why tSNE?

Allows to project 50 PCs in two dimensions.
A better view - separation of cell populations in two dimensions.

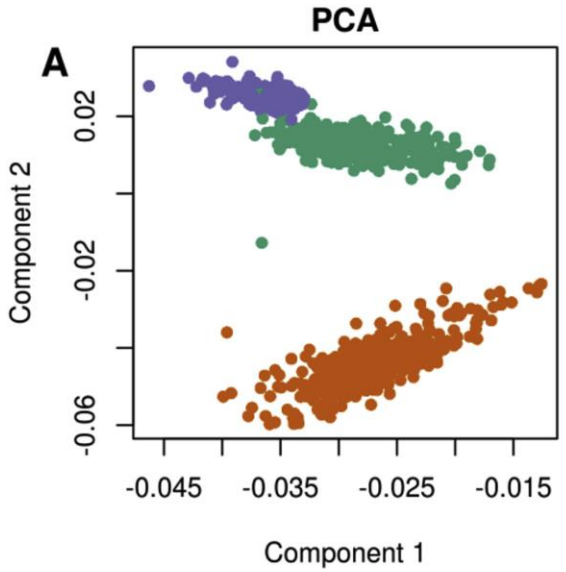
tSNE

T-distributed stochastic neighbor embedding (tSNE) is a non-linear dimensionality reduction method.

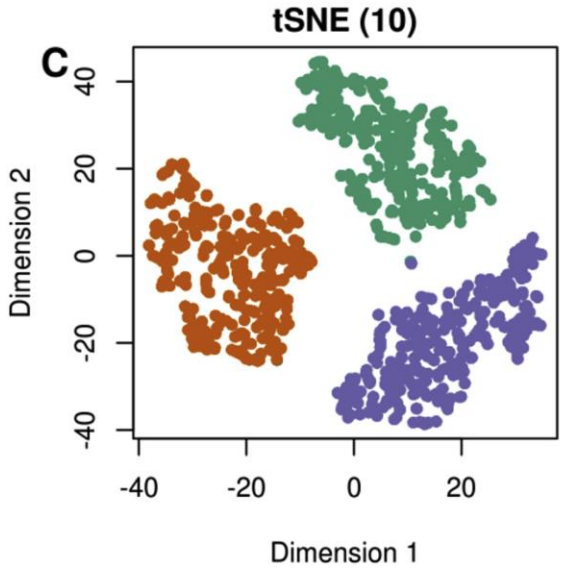
- Preserves local structure amongst cells
- Long-range information is lost
- Cluster sizes in a t-SNE plot mean nothing

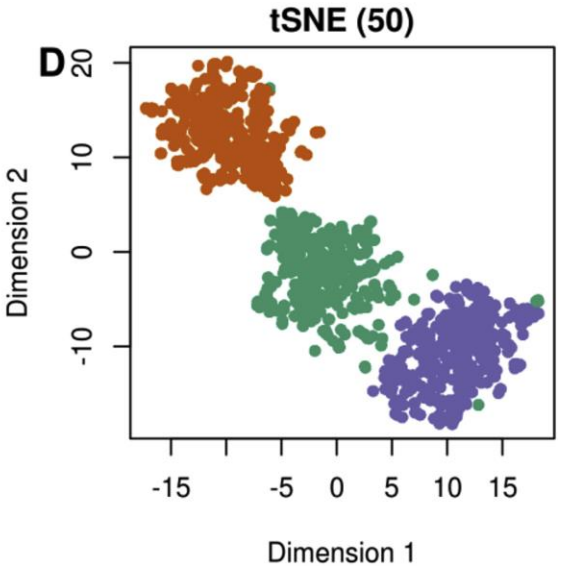
Notice: Orange cluster is not separated from blue and green as in PCA

The authors of the method recommend using tSNE for visualization purposes (Maaten et al., 2008).



Andrews et al. Molecular Aspects of Medicine Volume 59, February 2018, Pages 114-122





More on tSNE https://distill.pub/2016/misread-tsne/

Cell Ranger Analysis Steps

8. Differentially expressed Genes

Identify genes whose expression is specific to each cluster, by testing for each gene and each cluster, whether the in-cluster mean differs from the out-of-cluster mean.

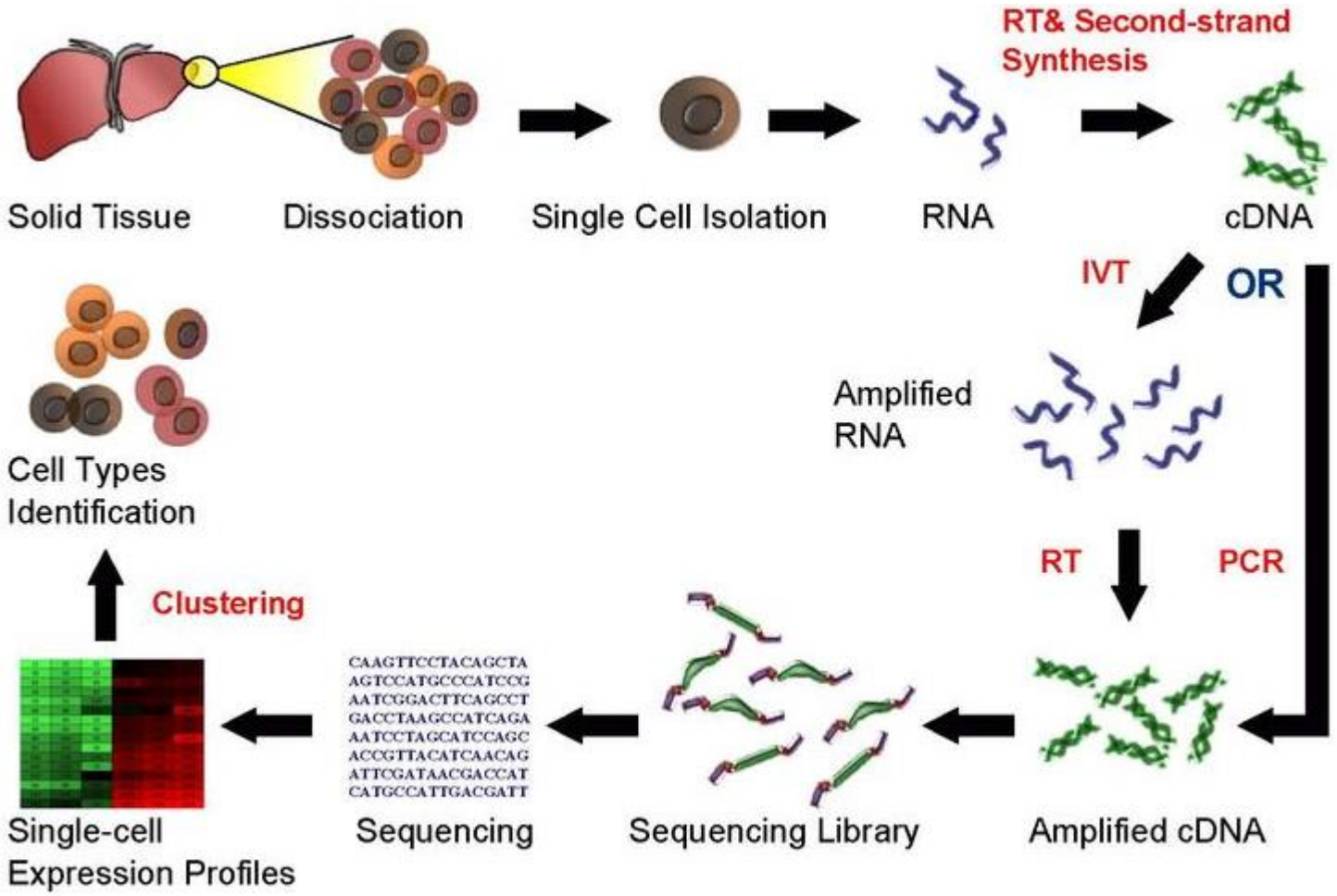
Top Genes By Cluster (Log2 fold-change, p-value)

		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Clus
Gene ID	Gene name	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC
ENSMUSG00000000983	Wfdc18	6.87	4e-09	-3.77	3e-01	-6.24	2e-01	-3.97	1e+00	-3.79
ENSMUSG00000028100	Nudt17	5.85	2e-44	-2.52	3e-05	-4.24	3e-08	-4.34	2e-05	-2.72
ENSMUSG00000094686	Ccl21a	5.79	9e-06	-3.05	6e-01	-5.22	5e-01	-2.64	1e+00	-3.03
ENSMUSG00000029075	Tnfrsf4	5.75	1e-38	-2.75	2e-05	-3.29	2e-05	-4.16	1e-04	-2.69
ENSMUSG00000025938	Slco5a1	5.68	5e-38	-2.12	1e-03	-5.81	8e-10	-4.39	6e-05	-2.62
ENSMUSG00000044309	Apol7c	5.36	5e-28	-1.55	7e-02	-5.17	6e-08	-4.82	1e-04	-3.40
ENSMUSG00000041782	Lad1	5.33	1e-35	-1.82	5e-03	-4.69	2e-08	-3.81	2e-04	-2.72
ENSMUSG00000035042	Ccl5	5.31	4e-14	-2.50	9e-05	-3.72	2e-06	-4.01	2e-04	-2.41
ENSMUSG00000029581	Fscn1	5.23	4e-14	-2.26	9e-05	-4.30	1e-09	-3.05	8e-04	-1.92
ENSMUSG00000003352	Cacnb3	4.96	2e-34	-1.63	9e-03	-4.15	2e-08	-3.05	1e-03	-2.64
ENSMUSG00000049382	Krt8	4.86	3e-09	-2.39	2e-01	-3.96	5e-02	-3.80	4e-01	-4.27

Normalization: computes relative library size as the total UMI counts for each cell divided by the median UMI counts per cell.

Cell Ranger uses method sSeq, (Yu, Huber, & Vitek, 2013) which employs a negative binomial exact test. When the counts become large, Cell Ranger switches to the fast asymptotic beta test used in edgeR.

Summarizing scRNA-Seq Workflow



https://hemberg-lab.github.io/scRNA.seq.course/introduction-to-single-cell-rna-seq.html

Closing Remarks

Molecular Aspects of Medicine
Volume 59, February 2018, Pages 114-122

Identifying cell populations with scRNASeq

Tallulah S. Andrews, Martin Hemberg

- Identifying novel or known cell populations is likely to remain a key goal of scRNA-Seq experiments in the future.
- Due to the trade-offs between cell number and sensitivity, it is likely there will never be a single optimal platform for scRNA-Seq experiments.
- No computational methods for dimensionality reduction, feature selection and unsupervised clustering will be optimal in all situations.
- Although novel cell populations can be readily identified using existing methods, these findings must be validated using external data or experiments to ensure they are not technical artifacts.

References

- https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/algorithms/overview#header
- https://hemberg-lab.github.io/scRNA.seq.course/introduction-to-single-cell-rna-seq.html
- How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives; Molin & Camillo ; Briefings in Bioinformatics, 31 January 2018
- Bioinformatics workflows:
  - https://bioconductor.org/packages/release/workflows/html/simpleSingleCell.html Lun et al.
  - https://satijalab.org/seurat/ Seurat v2.0 Butler et al., Nature Biotechnology 2018.

The End

Questions??

In the Exercise we will review a cell ranger report and analyse clusters with Loupe