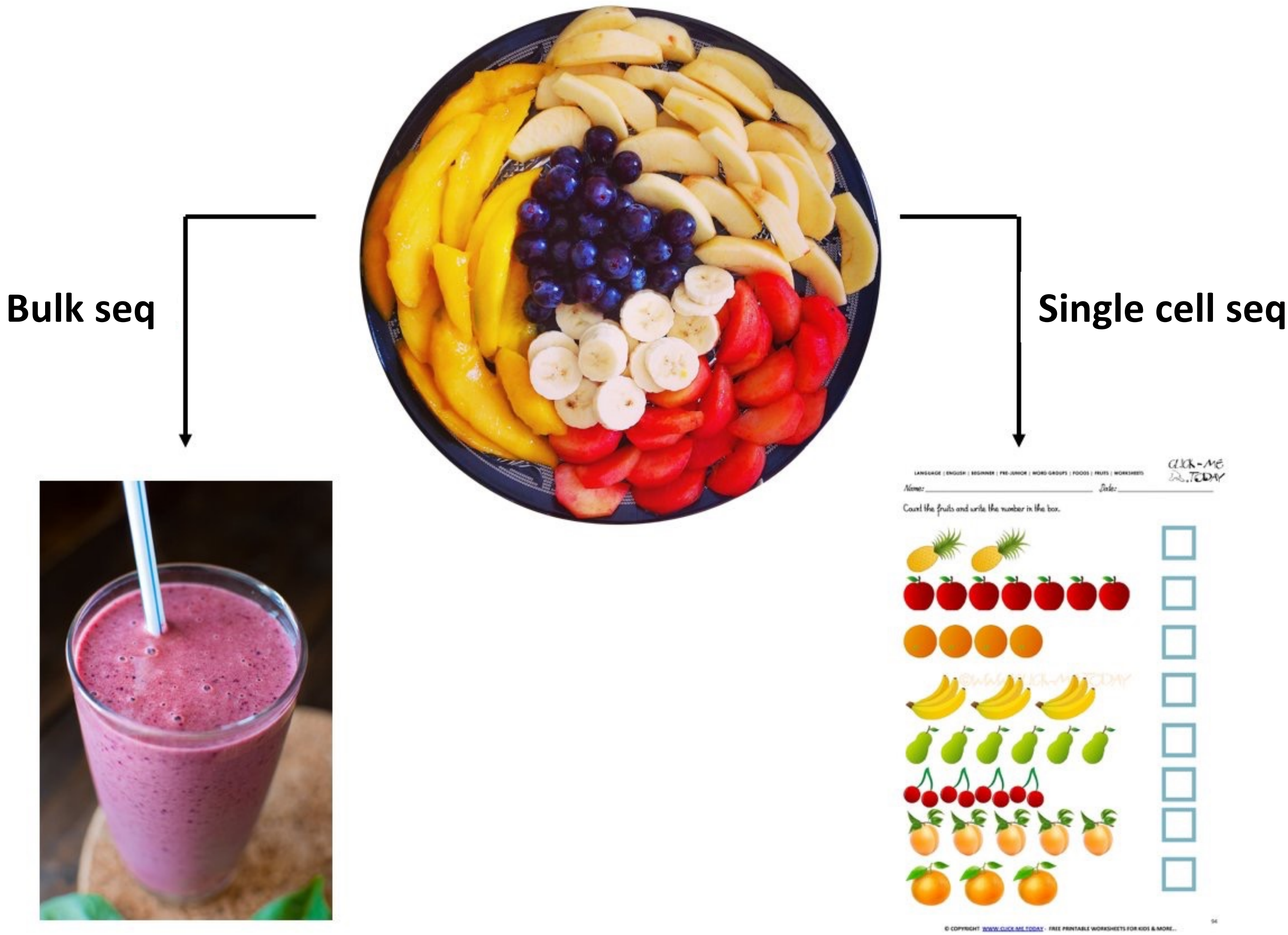
Analysis of single-cell sequencing data

“Bulk” sequencing

- Performed on “bulk” samples, which contains a large number of cells (millions).
- The “bulk” data measure the average signals (gene expression, TF binding, methylation, etc.) of many cells, and ignore the inter-cellular heterogeneities:
  - Different cell types.
  - Variation among the same cell type.

Single cell sequencing

- The cells are isolated from multi-cellular organism.
- Experiment is performed for each cell individually.
- Different types of sequencing at the single-cell level:
  - DNA-seq
  - ATAC-seq, ChIP-seq
  - BS-seq
  - RNA-seq
- Very active research field in the past several years.


Bulk seq
Single cell seq

Basic experimental procedure

- Isolation of single cell. Techniques include
  - Laser-capture microdissection (LCM)
  - Fluorescence-activated cell sorting (FACS)
  - Microfluidics
- Open the cell and obtain DNA/mRNA/etc.
- PCR amplification to get enough materials.
- Perform sequencing.
- Note that single cell sequencing usually has higher error rates than bulk data.

Single cell DNA-seq (scDNA-seq)

- For a comprehensive review, read Gawad *et al.* (2016) NRG.
- Examples of biological applications:
  - Identify and assemble the genome of unculturable microorganisms.
  - Determine the contribution of intra-tumor genetic heterogeneity in cancer development of treatment response.

Single cell BS-seq (scBS-seq)

- Similar to scDNA-seq, but with bisulfite treatment before sequencing.
- There's scWGBS and scRRBS.
- The methylation levels from scBS-seq should be 0/1, with some exceptions caused by technical artifacts.
- Data is very sparse.

Smallwood et al. 2014, NM

Single cell ChIP/ATAC-seq

- ATAC-seq: profile the active genomic regions.
- Data look like ChIP-seq, but very sparse

Cell1	1	1	2	1	0	2	2	1	0	2	1	0	1	0	0	0	2
Cell2	0	1	0	1	0	0	1	0	0	0	0	3	1	0	1	1	0
Cell3	2	0	0	1	1	0	1	0	1	0	1	1	1	2	1	0	0
.	1	1	2	1	0	2	2	1	0	2	1	0	1	0	0	0	2
.	0	1	0	1	0	0	1	0	0	0	0	3	1	0	1	1	0
.	2	0	0	1	1	0	1	0	1	0	1	1	1	2	1	0	0
	1	1	2	1	0	2	2	1	0	2	1	0	1	0	0	0	2
	Peak1	Peak2	Peak3	. . .													

Buenrostro et al. 2015 Nature

Single cell RNA-seq (scRNA-seq)

- The most active in the single cell field.
- Scientific goals:
  - Composition of different cell types in complex tissues.
  - New/rare cell type discovery.
  - Gene expression, alternative splicing, allele specific expression at the level of individual cells.
  - Transcriptional dynamics (pseudotime construction).
  - Many others.

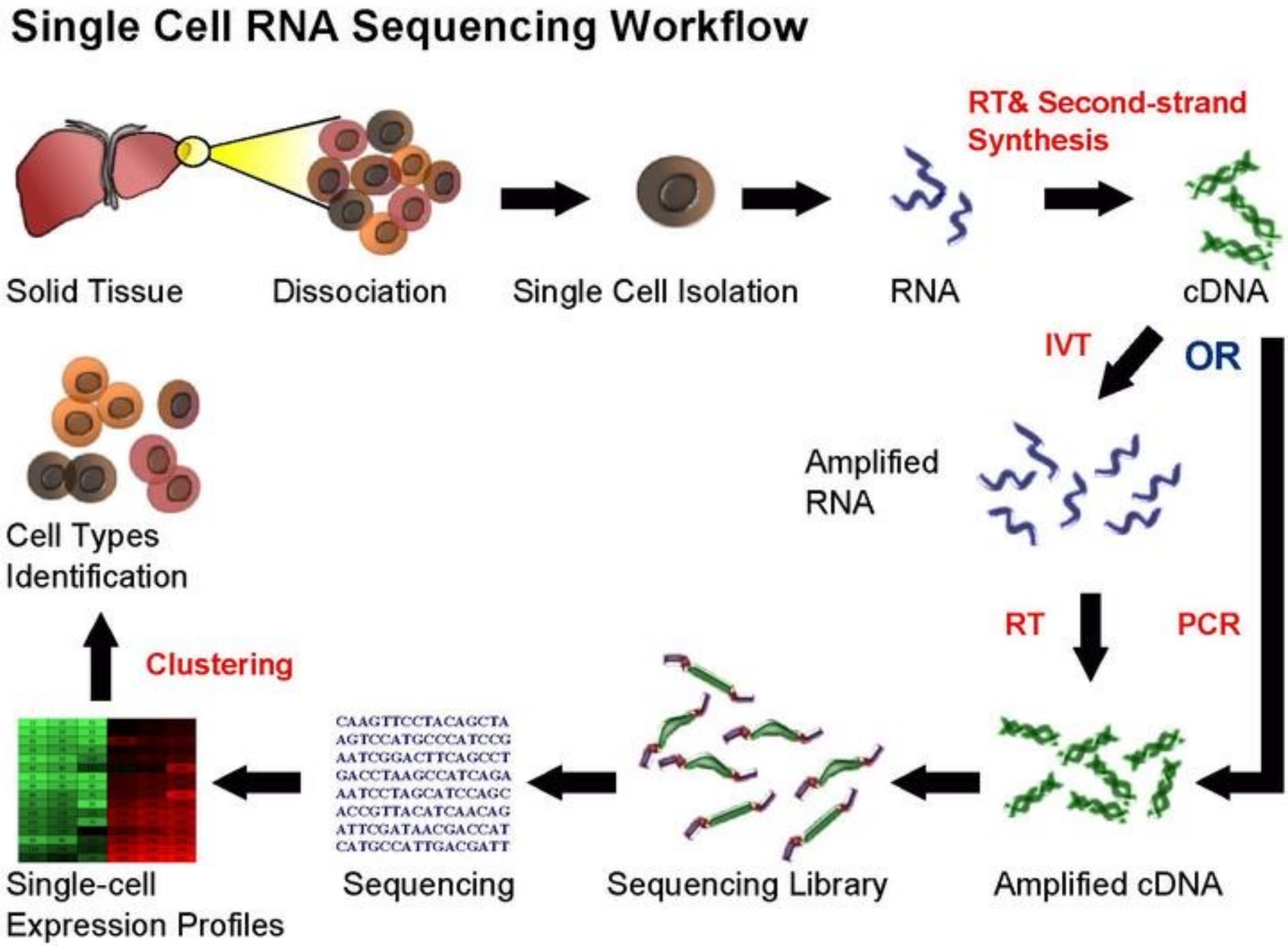


Figure source: Wikipedia

Technologies by cell capturing method

- **Plate-based methods**: Smart-Seq/Smart-Seq2, CEL-seq:
  - Sort cells into the wells on a multi-well plate.
  - Lower throughput (in terms of number of cells).
  - High sequencing depth
  - Can be combined with FACS for cell sorting.
  - Better at detecting low expression genes
  - Good for isoform analysis, allele specific expression

Microwell plates

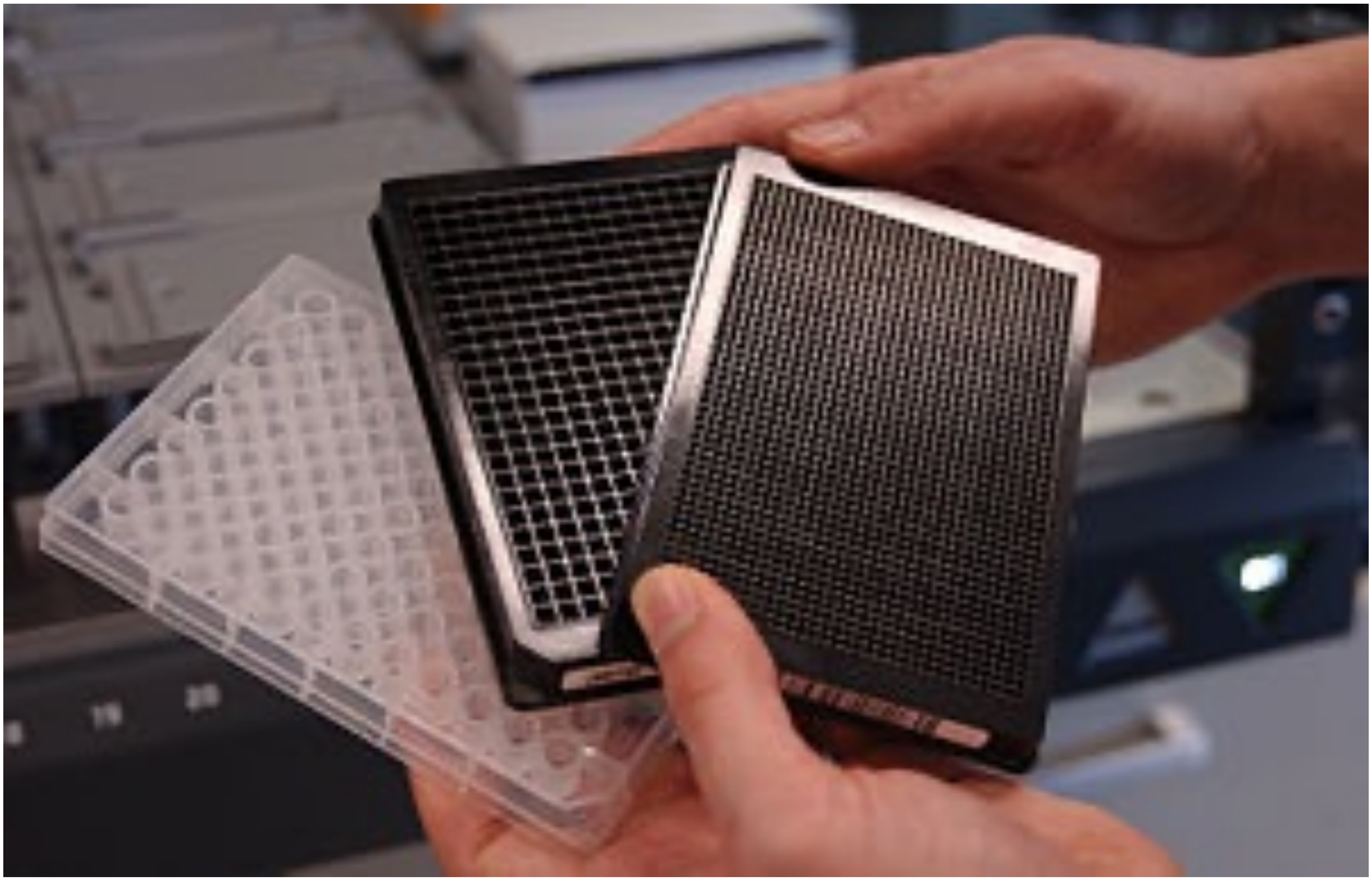
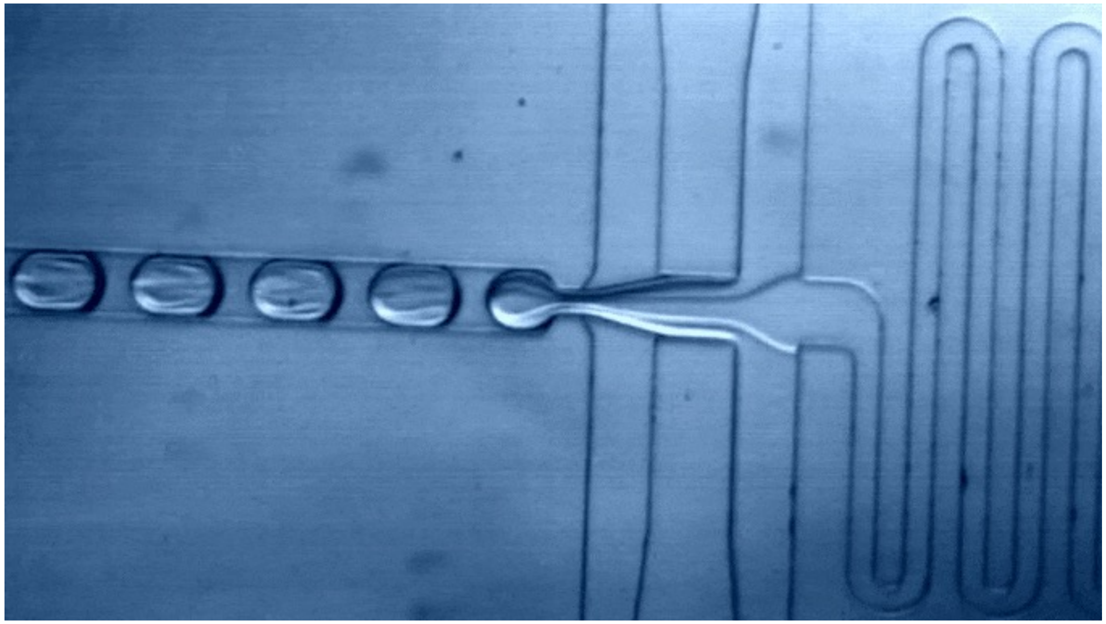
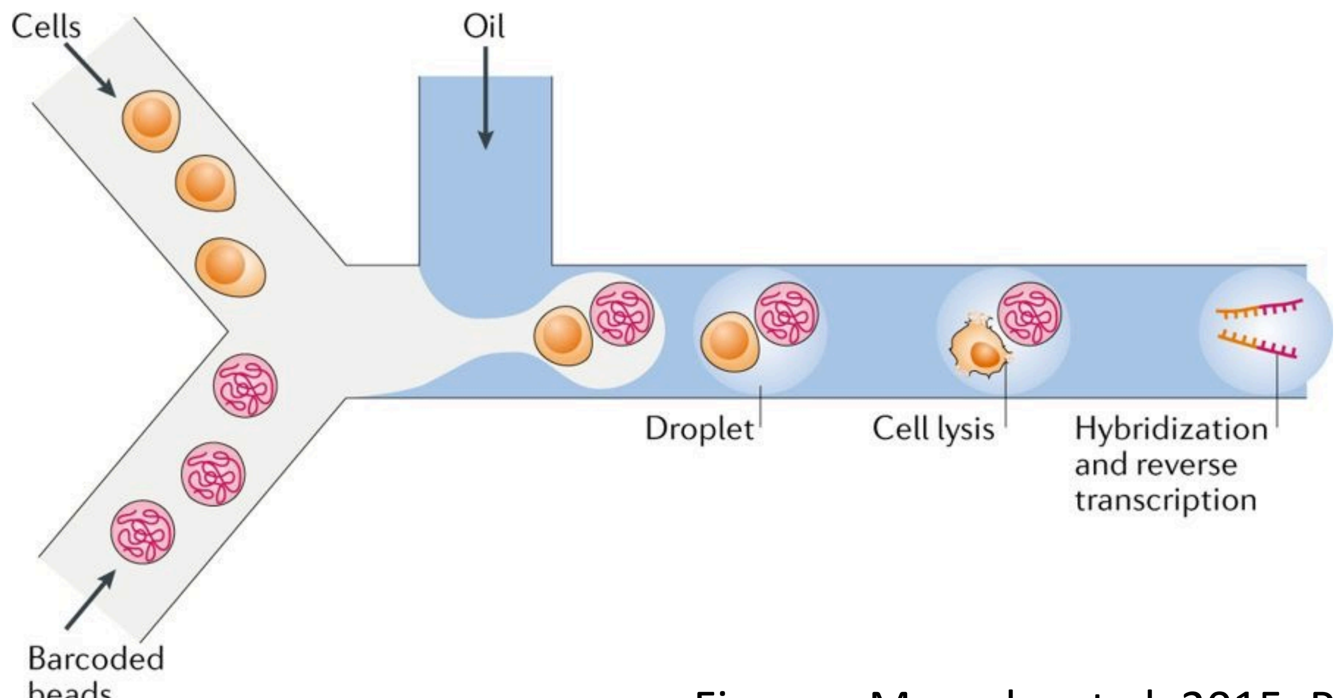


Figure source: wikipedia

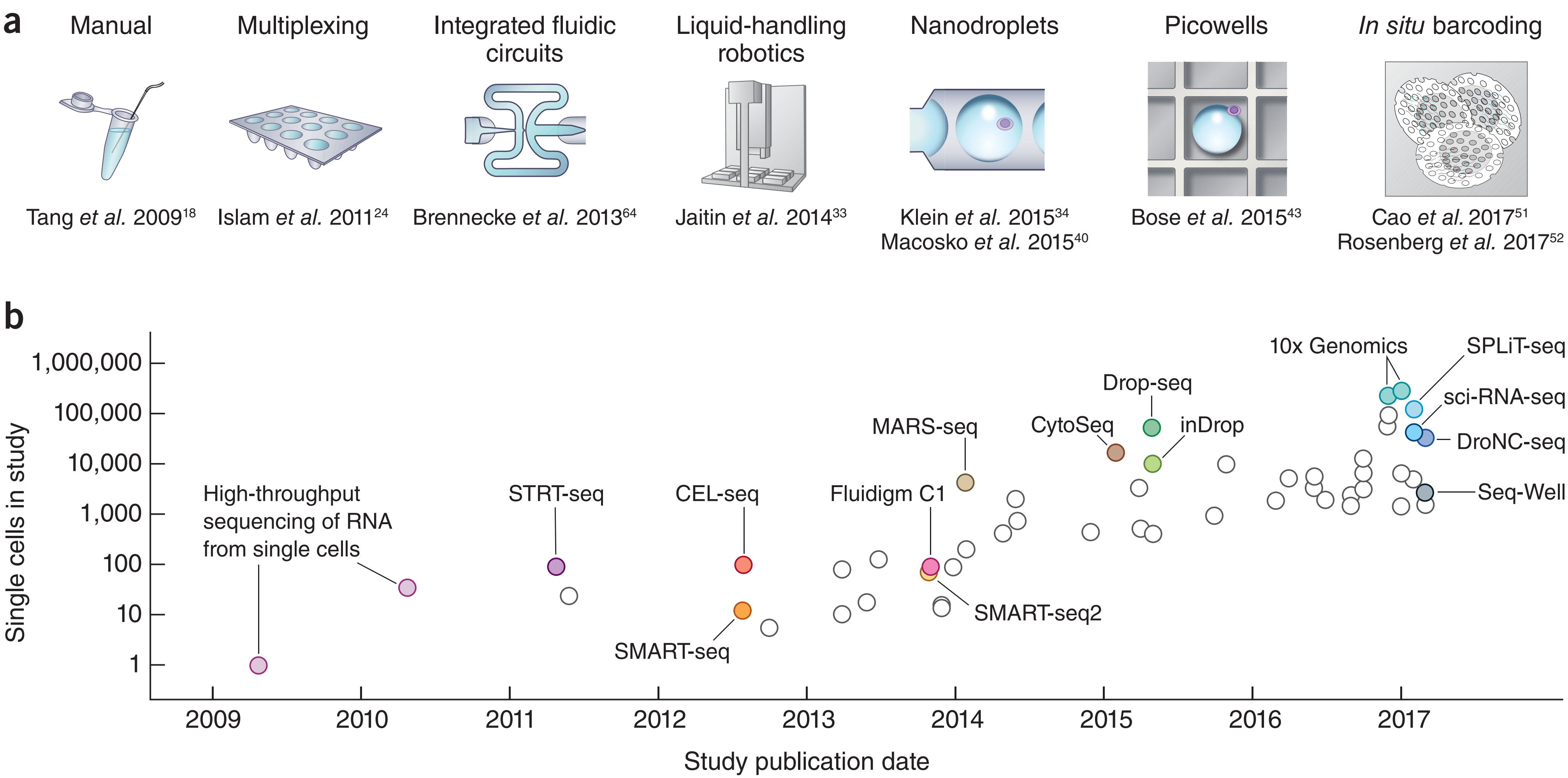
- **Droplet-based methods:** Drop-seq, inDrop, 10x genomics
  - Put each cell in a nanoliter droplet with a bead.
  - Each droplet is a reactor for PCR.
  - Much higher throughput (number of cells).
  - Lower sequencing depth.
  - Good for identifying cell subpopulations.





Figures: Macosko et al. 2015, Potter SS. 2018

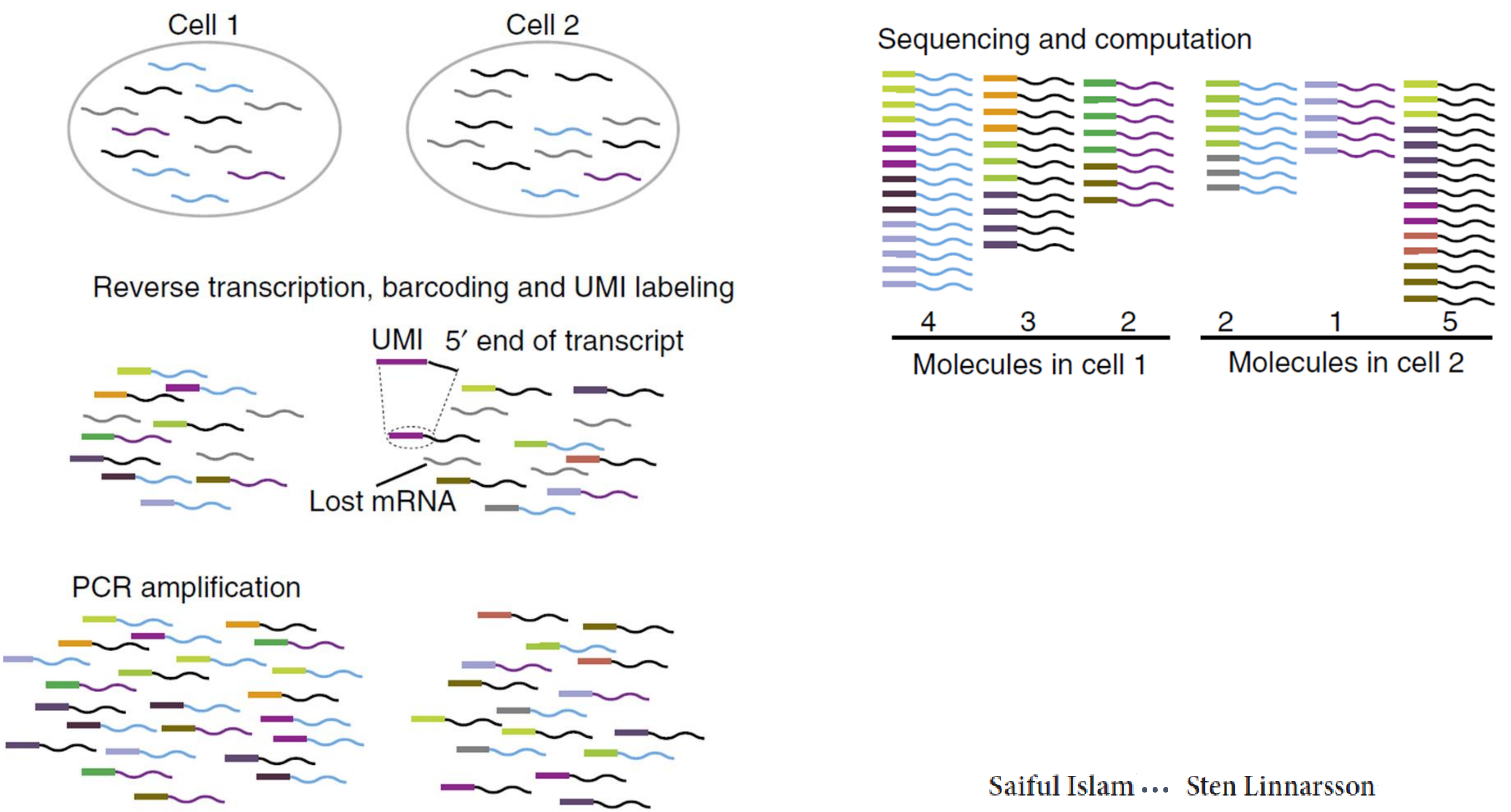
Technologies over the years



Svensson et al. (2017) Nature Protocols

Universal molecular identifier (UMI)

- Short sequence tag added to the mRNA molecular before PCR, for reducing PCR bias.



Saiful Islam ... Sten Linnarsson

NATURE METHODS | VOL.11 NO.2 | FEBRUARY 2014 | 163

Multi-omics single cell assays

- CITE-seq (**C**ellular **I**ndexing of **T**ranscriptomes and **E**pitopes by **Seq**uencing)
  - Jointly profile transcriptome and proteome.
- scNMT-seq (single-cell **N**ucleosome, **M**ethylation and **T**ranscription sequencing)
  - Jointly profile chromatin accessibility, DNA methylation, and transcription

scRNA-seq data analyses

- **Data preprocessing**
  - Normalization
  - Batch effect correction
  - Imputation
- **Data analyses**
  - Cell clustering
  - Pseudo-time construction
  - Cell type identification
  - Differential expression
  - Rare cell type discovery; alternative splicing; allele specific expression; RNA velocity
- **Visualization**
  - TSNE and UMAP

Data processing

- Preprocessing
- Data characteristics
- Normalization
- Batch effect correction
- Imputation

scRNA-seq data preprocessing

- Sequence alignment and expression quantification
  - RNA-seq alignment software (Tophat, STAR, HISAT, etc.) can be used
  - Some commercial software, such as Cell Ranger for 10x genomics data.
- Result: a matrix of read counts. Rows are genes and columns are cells

Some data characteristics

- Data is very sparse (many zeros), especially for Drop-seq data.
- Number of transcripts detected is much lower compared to bulk RNA-seq under the same sequencing depth.

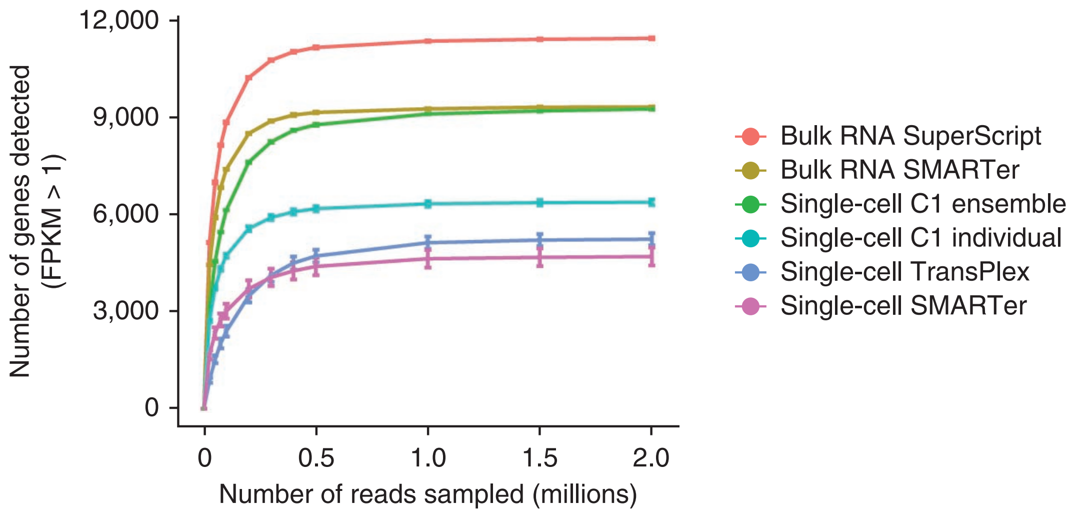
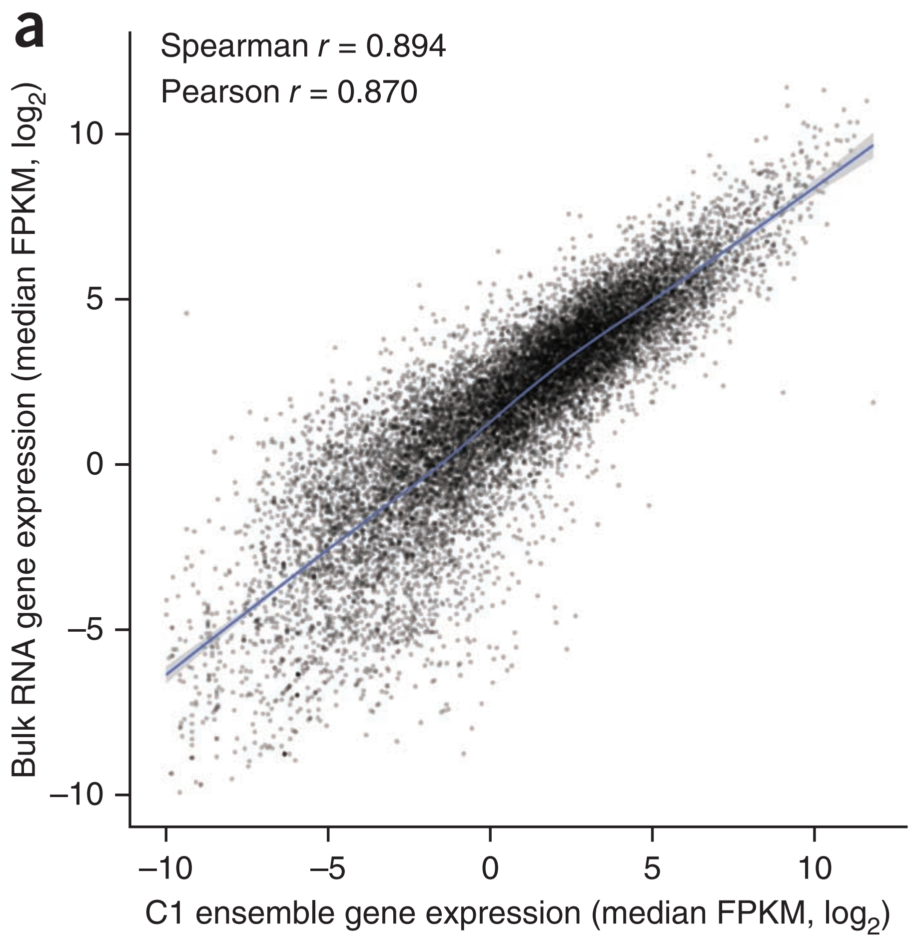


Figure 5 | Saturation curves for the different sample preparation methods. Each point on the curve was generated by randomly selecting a number of raw reads from each sample library and then using the same alignment pipeline to call genes with mean FPKM >1. Each point represents four replicate subsamplings. Error bars, standard error.

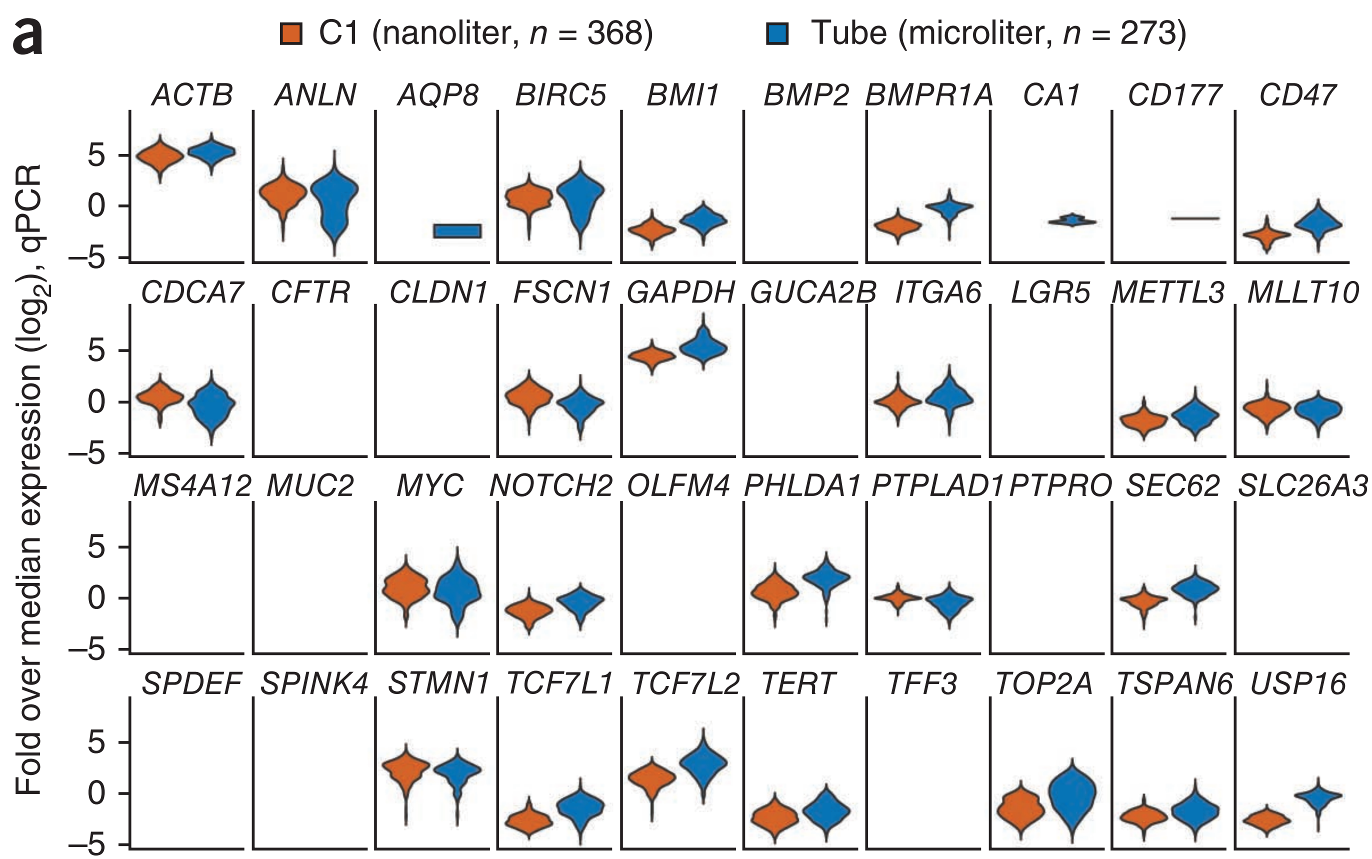
Wu et al. 2013 Nature Method

- Bulk and aggregated single cell expressions have good correlation.



Wu et al. 2013 Nature Method

- Expression levels for a gene in different cells sometimes show bimodal distribution.



Wu et al. 2013 Nature Method

Data normalization

- scRNA-seq is very noisy.
- Spike-in data is usually available.
  - Spike-ins from the external RNA Control Consortium (ERCC) panel contains 92 synthetic spikes based on bacterial genome with known expression level.
- UMI is helpful for removing amplification noise.
- A combination of spike-in and UMI can potentially be used for data normalization.
- Simple normalization (such as by sequencing depth) for bulk RNA-seq can be applied, e.g., TPM or FPKM.

Lun et al. Genome Biology (2016) 17:75
DOI 10.1186/s13059-016-0947-7

Genome Biology

METHOD Open Access

CrossMark

Pooling across cells to normalize single-cell RNA sequencing data with many zero counts

Aaron T. L. Lun[1*], Karsten Bach[2] and John C. Marioni[1,2,3*]

- Works for data without spike-in.
- The goal is to estimate a size factor for each cell.
- The idea is to normalize on summed expression values from pools of cells – it's more stable than using individual cell.
- Bioconductor package **scran.**

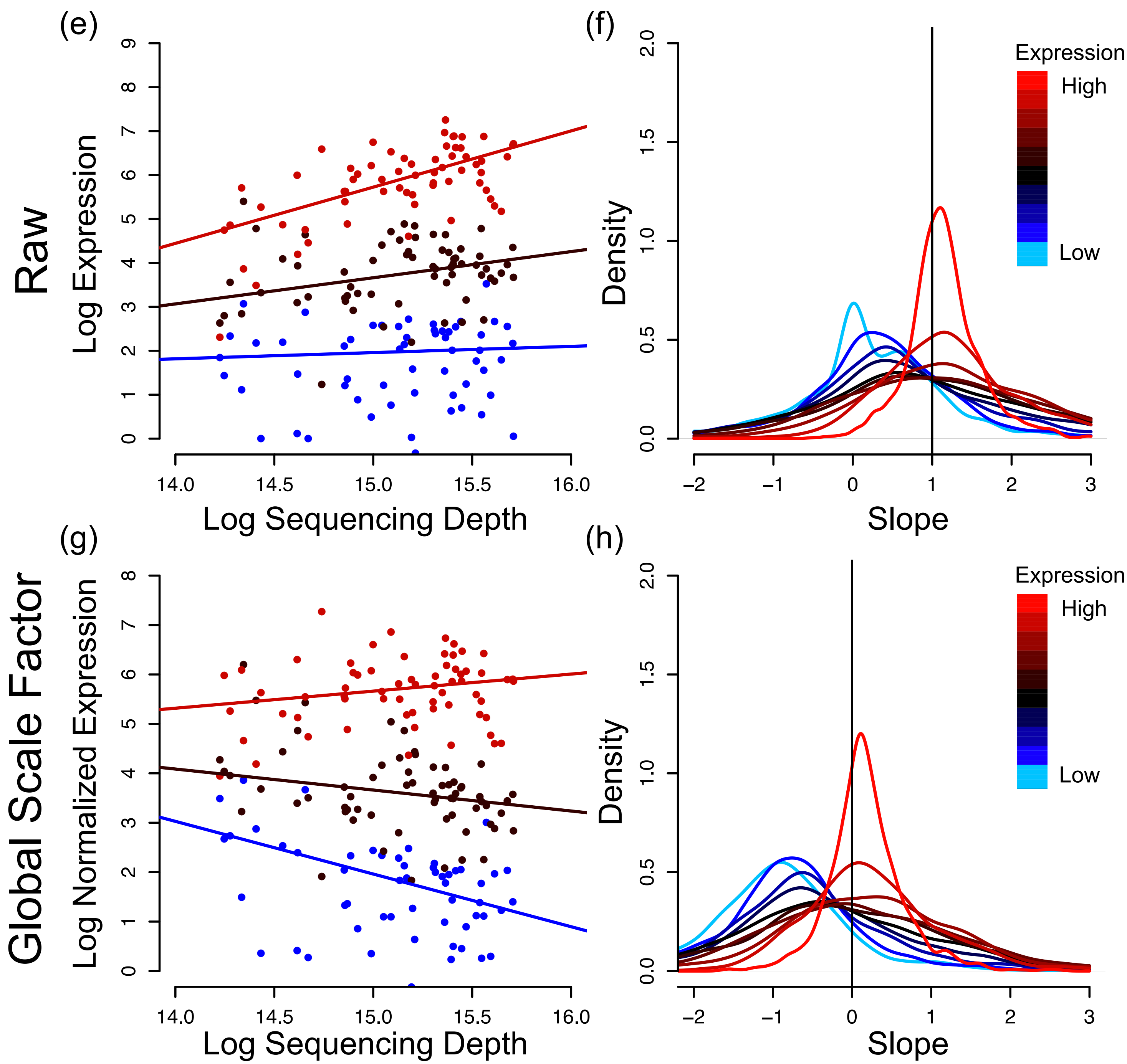
SCnorm: robust normalization of single-cell RNA-seq data

584 | VOL.14 NO.6 | JUNE 2017 | NATURE METHODS

Rhonda Bacher[1,5], Li-Fang Chu[2,5], Ning Leng[2], Audrey P Gasch[3], James A Thomson[2], Ron M Stewart[2], Michael Newton[1,4] & Christina Kendziorski[4]

- Basic idea: one normalization factor per cell doesn’t fit all genes.
- Relationships of read counts and sequencing depths vary and depend on the expression levels.

Single cell



(e)
Raw
Log Expression
Log Sequencing Depth
(f)
Density
Slope
Expression
High
Low
(g)
Global Scale Factor
Log Normalized Expression
Log Sequencing Depth
(h)
Density
Slope
Expression
High
Low

SCnorm Solution

- Uses quantile regression to estimate the dependence of read counts on sequencing depth for every gene.
- Genes with similar dependence are then grouped, and a second quantile regression is used to estimate scale factors within each group.
- Bioconductor package **SCnorm**.

Batch effect correction

- Batch effect in scRNA-seq can be severe.
- It’s difficult to randomize the design, i.e., batch is often confounded with individual, so it causes trouble for analyzing data from multiple individuals (more on this later).
- Bulk data methods such as Combat/SVA don’t work well
- There are several methods designed for scRNA-seq:
  - MNN (Haghverdi et al. 2018. Nat. Biotech.)
  - ZINB-WaVE (Risso et al. 2018 Nat. comm.)
  - LIGER (Welch et al. 2019. Cell)
  - Harmony (Korsunsky et al. 2019 Nat. Method)
  - BUSseq (Song et al. 2020. Nat. Comm.)

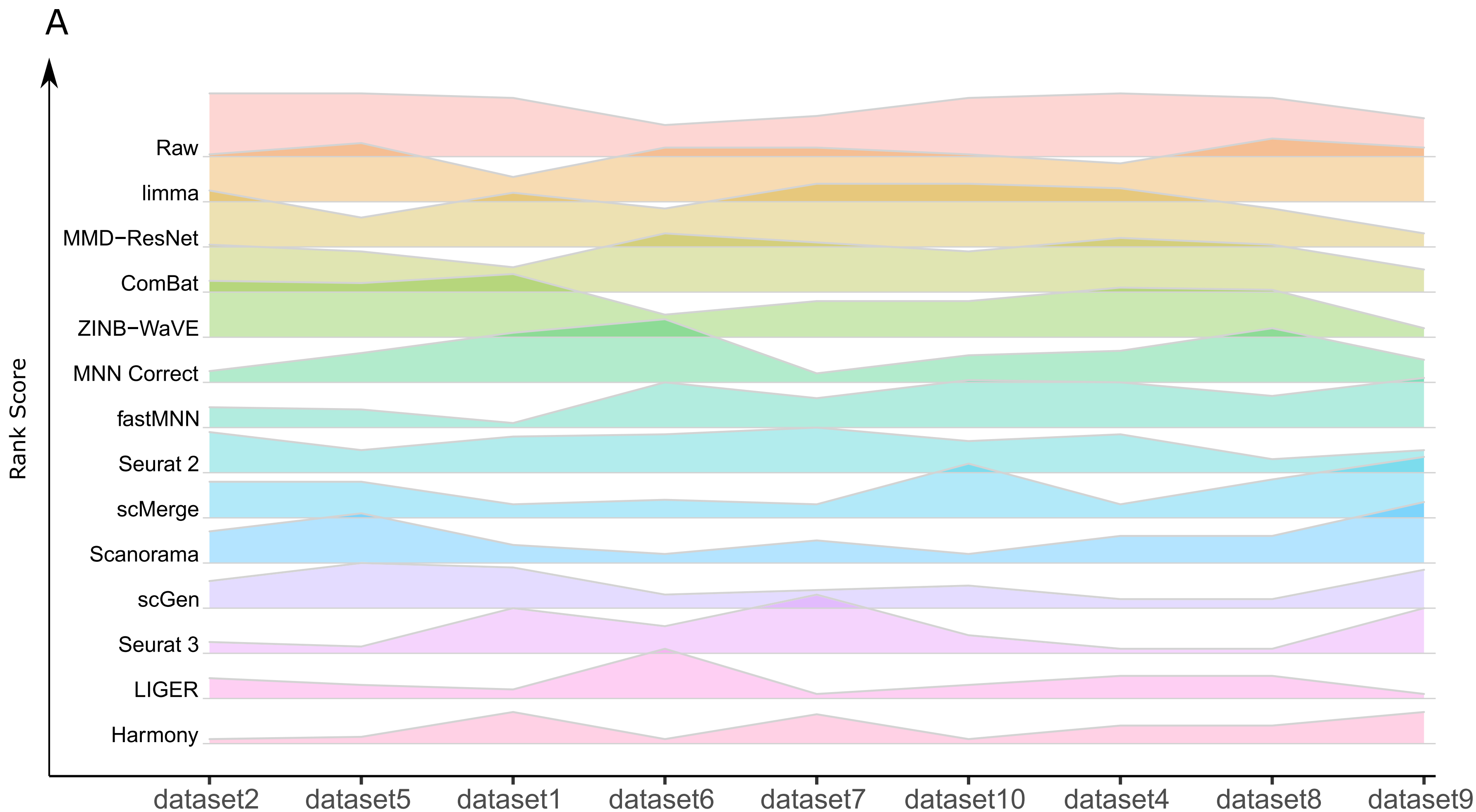
Tran et al. Genome Biology (2020) 21:12
https://doi.org/10.1186/s13059-019-1850-9

Genome Biology

RESEARCH Open Access

A benchmark of batch-effect correction methods for single-cell RNA sequencing data

Hoa Thi Nhu Tran[†], Kok Siong Ang[†], Marion Chevrier[†], Xiaomeng Zhang[†], Nicole Yee Shin Lee, Michelle Goh and Jinmiao Chen[*]



Data imputation

- scRNA-seq has lots of missing data (dropout).
- Imputing the missing data help the downstream analyses.
- There are a number of methods:
  - SAVER (Huang et al. 2018 Nat. Methods)
  - ScImpute (Li et al. 2018 Nat. Comm.)
  - MAGIC (van Dijk et al. 2018 Cell)
  - SCRABBLE (Peng et al. 2019 GB)

General strategy for imputation

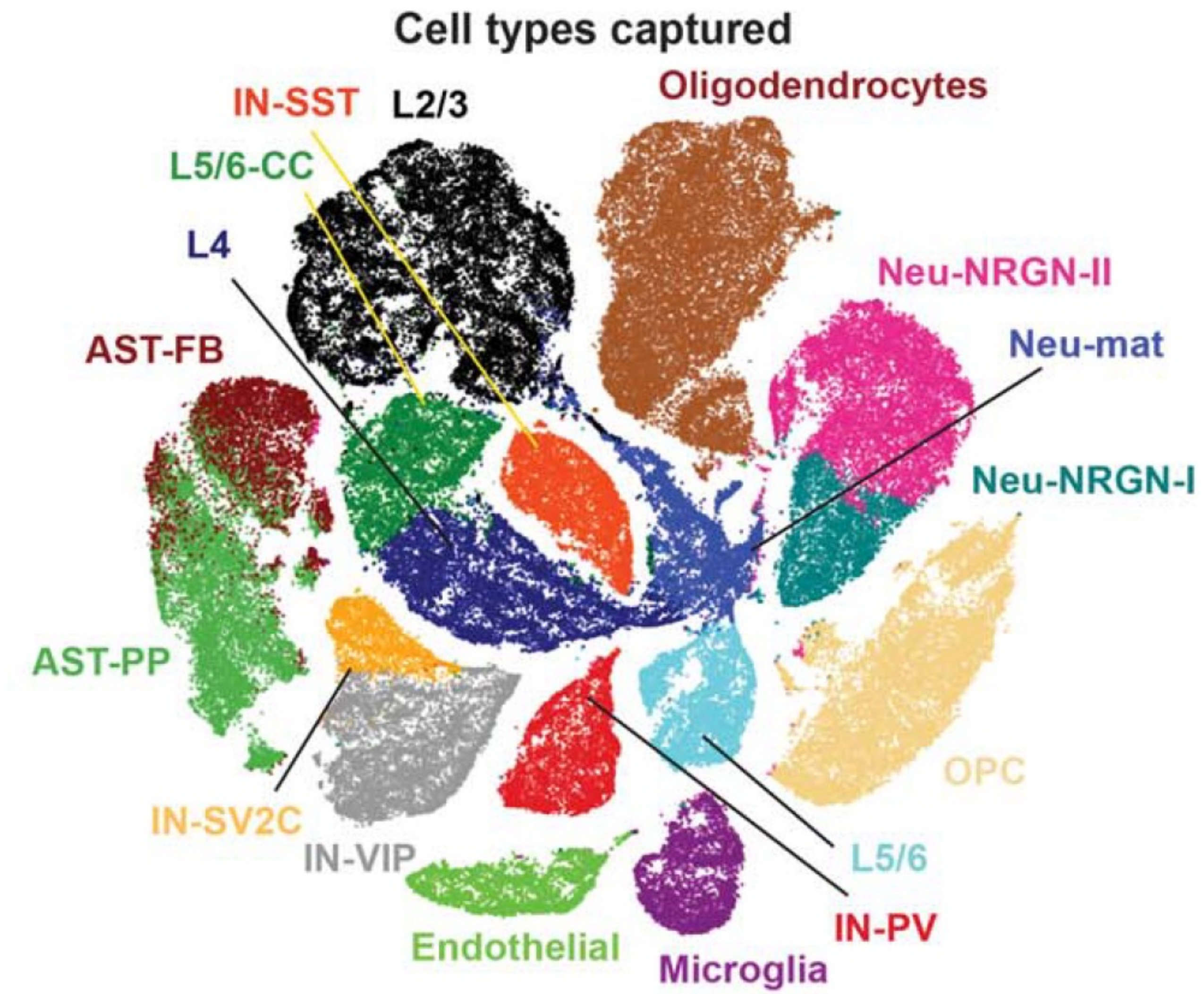
- The problem is similar to a “recommendation system”.
  - First compute the similarities among genes and cells.
  - To impute one element, borrow information from similar gene/cell.

Data analyses tasks

- Cell clustering
- Pseudotime construction
- Cell type identification
- Differential expression
- Rare cell type discovery
- Alternative splicing
- Allele specific expression
- RNA velocity

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.



AST-FB	Fibrous astrocytes	**L2/3**	Layer 2/3 excitatory neurons
AST-PP	Protoplasmic astrocytes	**L4**	Layer 4 excitatory neurons
OPC	Oligodendrocyte precursor cells	**L5/6**	Layer 5/6 corticofugal projection neurons
IN-PV	Parvalbumin interneurons	**L5/6-CC**	Layer 5/6 cortico-cortical projection neurons
IN-SST	Somatostatin interneurons	**Neu-mat**	Maturing neurons
IN-SV2C	SV2C interneurons	**Neu-NRGN-I**	NRGN-expressing neurons
IN-VIP	VIP interneurons	**Neu-NRGN-II**	NRGN-expressing neurons

Velmeshev et al., (2019) Science

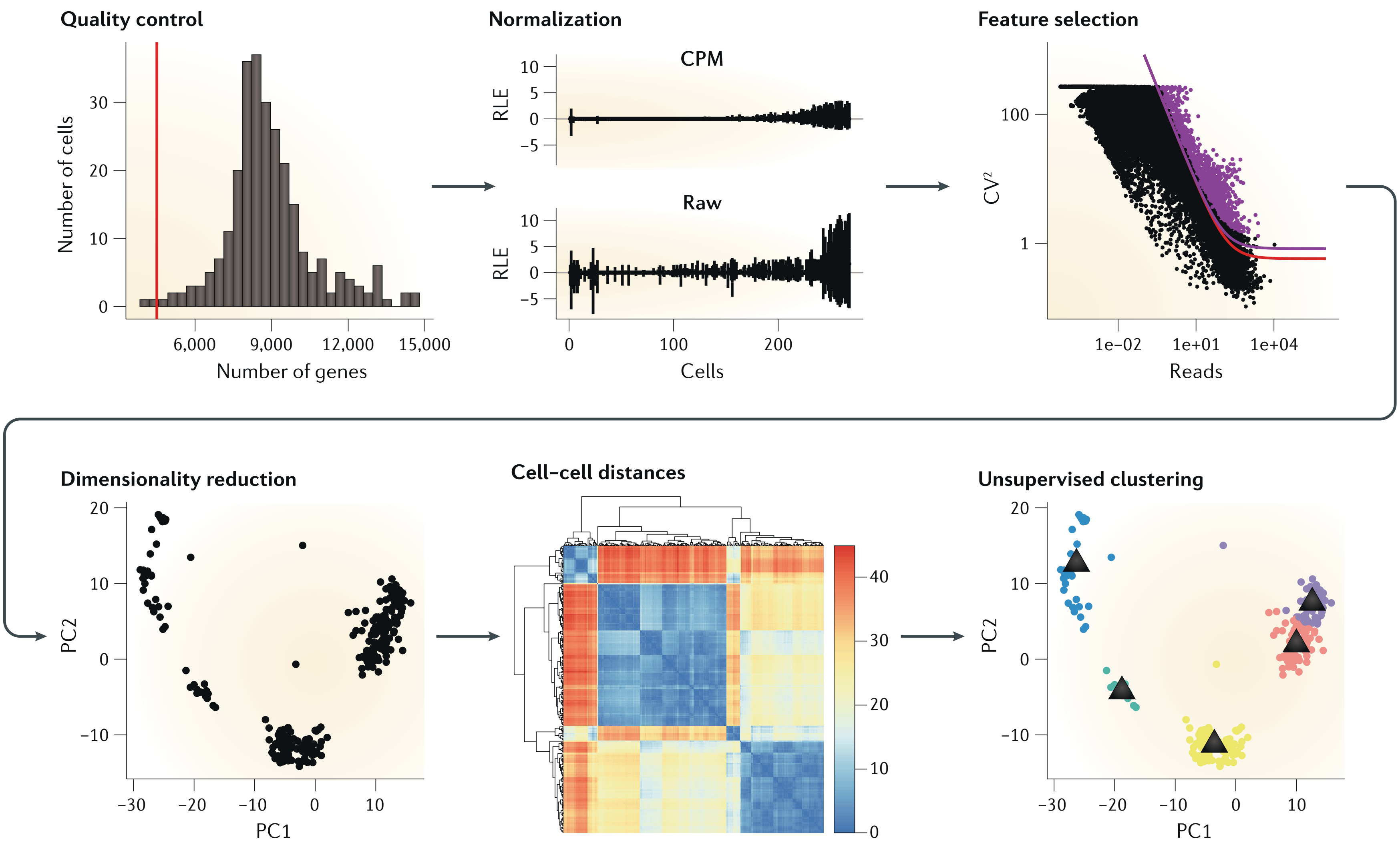
Cell clustering methods

- Many methods available
  - SC3, Seurat, TSCAN, Monocle, CIDR, …
  - Comprehensively compared in Duo et. al (2018) F1000 Research.

and robust [73]. Due to the heavy time consuming nature of consensus clustering, a rule of thumb for unsupervised single cell clustering is to use single-cell consensus clustering (SC3, integrated in Scater [52]) when the number of cells is < 5000 but use Seurat instead when there are more than 5000 cells.

Mu et al. Genomics Proteomics Bioinformatics (2019)

Principal of the clustering methods



Kiselev et al. (2019) Nat. Rev. Genet.

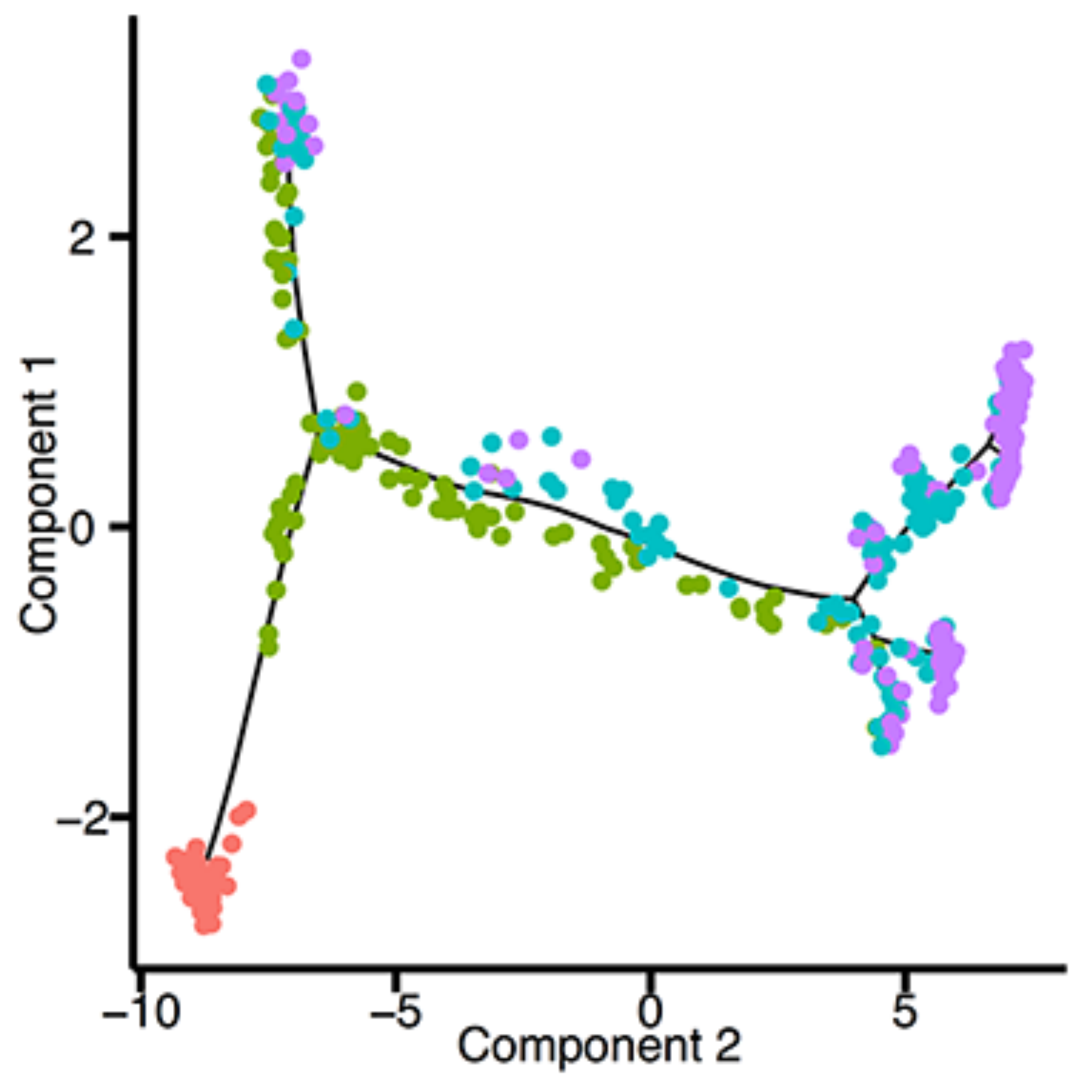
Pseudotime construction

- This belongs to the “clustering” category.
- Instead of putting cells into independent, exchangeable groups, it orders the cells by underlying temporal stage (estimated).
- Methods/tools:
  - Monocle/monocle2: Trapnell et al. (2014) Nat. Biotechnol; Qiu et al. (2017) Nat. Methods.
  - Waterfall: Shin et al. (2015) Cell Stem Cell
  - Wanderlust: Bendall et al. (2014) Cell
  - TSCAN: Ji et al. (2016) NAR

Pseudotime construction method

General steps:

1. Select informative genes.
2. Dimension reduction of GE.
3. Cluster the cells based on reduced data. Often want to over-cluster them to have many groups.
4. Construct a MST (miminum spanning tree) from the clustering results.
5. Map cells to the MST.



Component 1
2
0
−2
−10
−5
0
5
Component 2

Clustering vs. pseudotime

- Clustering: assumes the cells belong to discrete groups.
- If assumption not hold, clustering still partition the data, and thus mistake random noise for true structure.
- Pseudotime construction:
  - place cells on a continuum connecting two or more end states
  - useful for understanding development or disease progression
- When assumptions are not clear, explore both.

Cell clustering for multiple samples

- When scRNA-seq data are from multiple samples, batch effects could have significant impact on the results.
- Cells from the same sample, instead of the same cell type form different sample, can cluster together.
- Possible solution:
  – Remove batch effect then cluster: MNN + SC3
  – Jointly model cell type and sample effect: BAMM-SC (Sun et al. 2019, Nat. Comm)

Cell type identification

- Another paradigm to identify cell type.
- Cell clustering (**unsupervised**):
  - Cluster cells to multiple clusters (unsupervised). then assign cell type for each cluster.
- Cell type identification (**supervised**):
  - Requires reference, or training data.
  - Directly assign each cell to a cell type.
  - In general works better.
  - Cannot identify new cell types (restricted to the known cell types in the reference).

Cell type identification methods

- Pre-train a classifier using training set first, predict labels by kNN/correlation/RF etc.
  - scmap (Kiselev et al. 2018 Nat. Methods)
  - CaSTLe (Lieberman et al. 2018 Plos One)
  - Garnett (Pliner et al. 2019 Nat. Methods)
  - CHETAH (Kanter et al. 2019 Nucleic Acids Research)
- Marker-based classifier
  - CellAssign (Zhang et al. 2019 Nat. Methods)
- Other generic machine learning methods: SVM, LDA, RF, kNN, RF
- Semi-supervised method (transfer learning): ItClust, MARS.
- Comprehensively compared in Abdelaal et al. Genome Biology 2019

Comparison of the methods

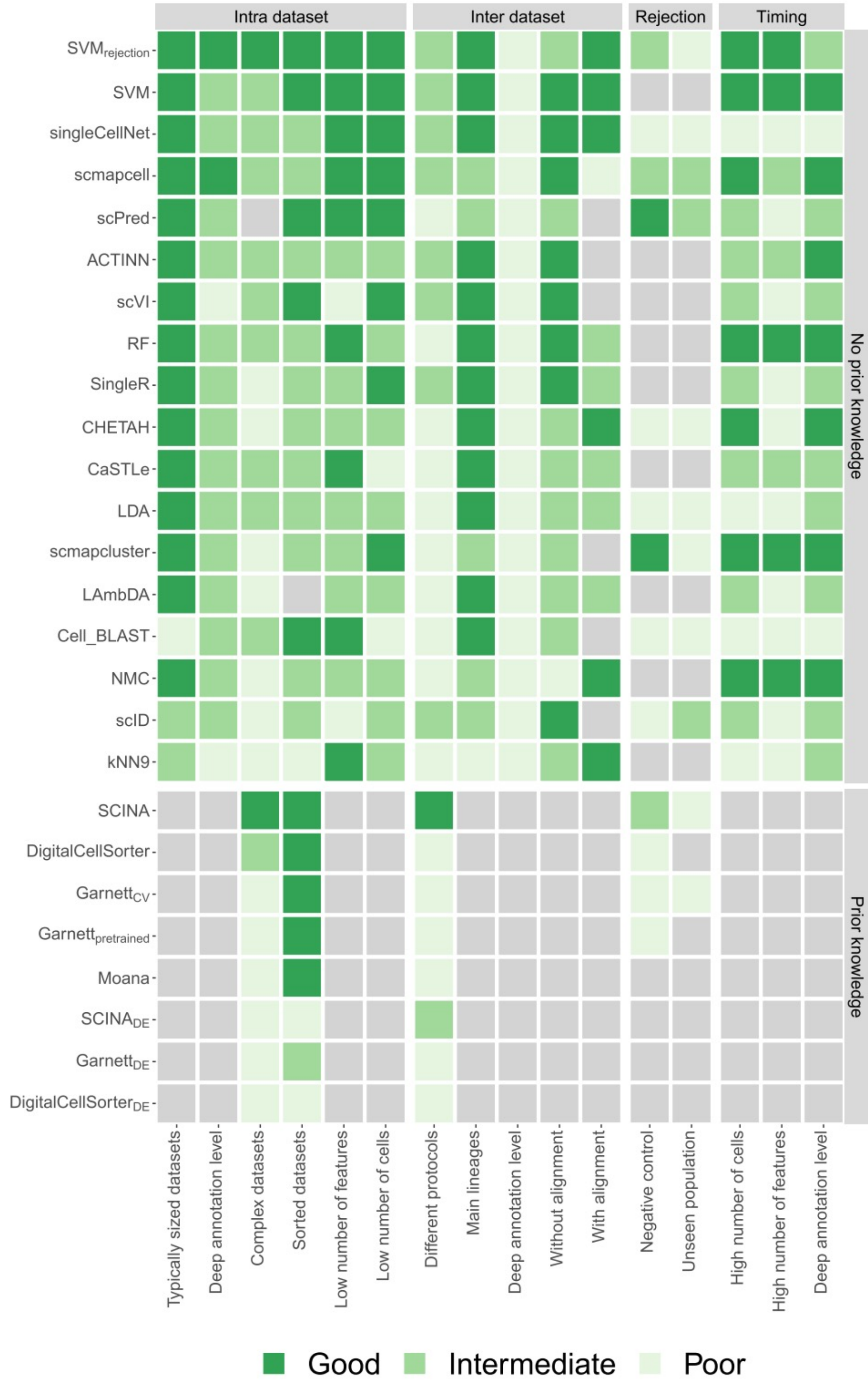
Abdelaal *et al. Genome Biology* (2019) 20:194
https://doi.org/10.1186/s13059-019-1795-z

Genome Biology

RESEARCH Open Access

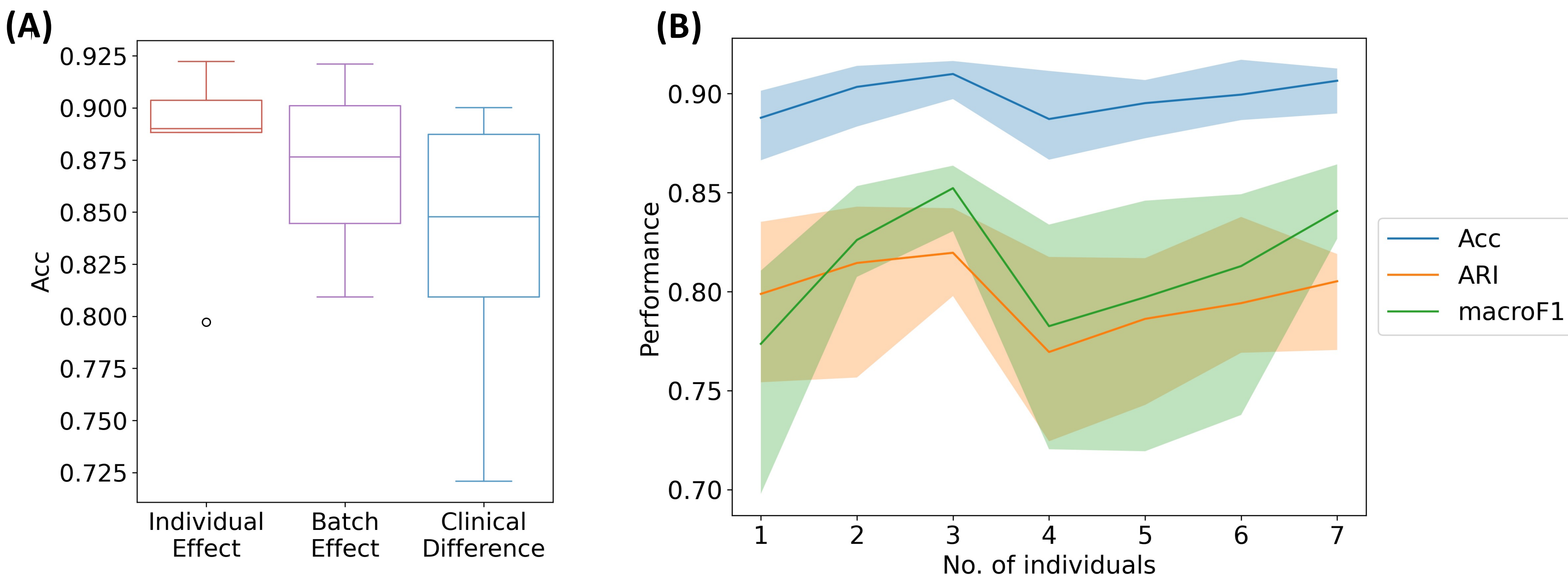
A comparison of automatic cell identification methods for single-cell RNA sequencing data

Tamim Abdelaal[1,2†], Lieke Michielsen[1,2†], Davy Cats[3], Dylan Hoogduin[3], Hailiang Mei[3], Marcel J. T. Reinders[1,2] and Ahmed Mahfouz[1,2*]



Intra dataset
Inter dataset
Rejection
Timing
SVM$_{rejection}$
SVM
singleCellNet
scmapcell
scPred
ACTINN
scVI
RF
SingleR
CHETAH
CaSTLe
LDA
scmapcluster
LAmbDA
Cell_BLAST
NMC
scID
kNN9
SCINA
DigitalCellSorter
Garnett$_{CV}$
Garnett$_{pretrained}$
Moana
SCINA$_{DE}$
Garnett$_{DE}$
DigitalCellSorter$_{DE}$
No prior knowledge
Prior knowledge
Typically sized datasets
Deep annotation level
Complex datasets
Sorted datasets
Low number of features
Low number of cells
Different protocols
Main lineages
Deep annotation level
Without alignment
With alignment
Negative control
Unseen population
High number of cells
High number of features
Deep annotation level
Good
Intermediate
Poor

Choice of reference is important



Ma et al (2021) GB

Differential expression (DE)

- DE analysis is the most important task for bulk expression data (microarray or RNA-seq).
- DE in scRNA-seq is a little different:
  - Traditional methods test mean changes, while the consideration and modeling of “drop-out” event (non-expressed) is important in sc data.
  - Considering cell types: can compare cross cell types or compare the same cell type cross biological conditions.

DE methods

- SCDE (Kharchenko et al. 2014 Nat. Methods)
- MAST (Finik et al. 2015 GB)
- SC2P (Wu et al. 2018 Bioinformatics)
- Seurat and monocle also provides DE functions.
- Bulk methods (DESeq, edgeR) are sometimes used.
- A comparison paper: Soneson and Robinson (2018) Nat. Methods

Finak *et al. Genome Biology* (2015) 16:278
DOI 10.1186/s13059-015-0844-5

Genome Biology

METHOD Open Access

MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

Greg Finak[1†], Andrew McDavid[1†], Masanao Yajima[1†], Jingyuan Deng[1], Vivian Gersuk[2], Alex K. Shalek[3,4,5,6], Chloe K. Slichter[1], Hannah W. Miller[1], M. Juliana McElrath[1], Martin Prlic[1], Peter S. Linsley[2] and Raphael Gottardo[1,7*]

- MAST: “Model-based Analysis of Single- cell Transcriptomics.”
- Bioconductor package **MAST**.

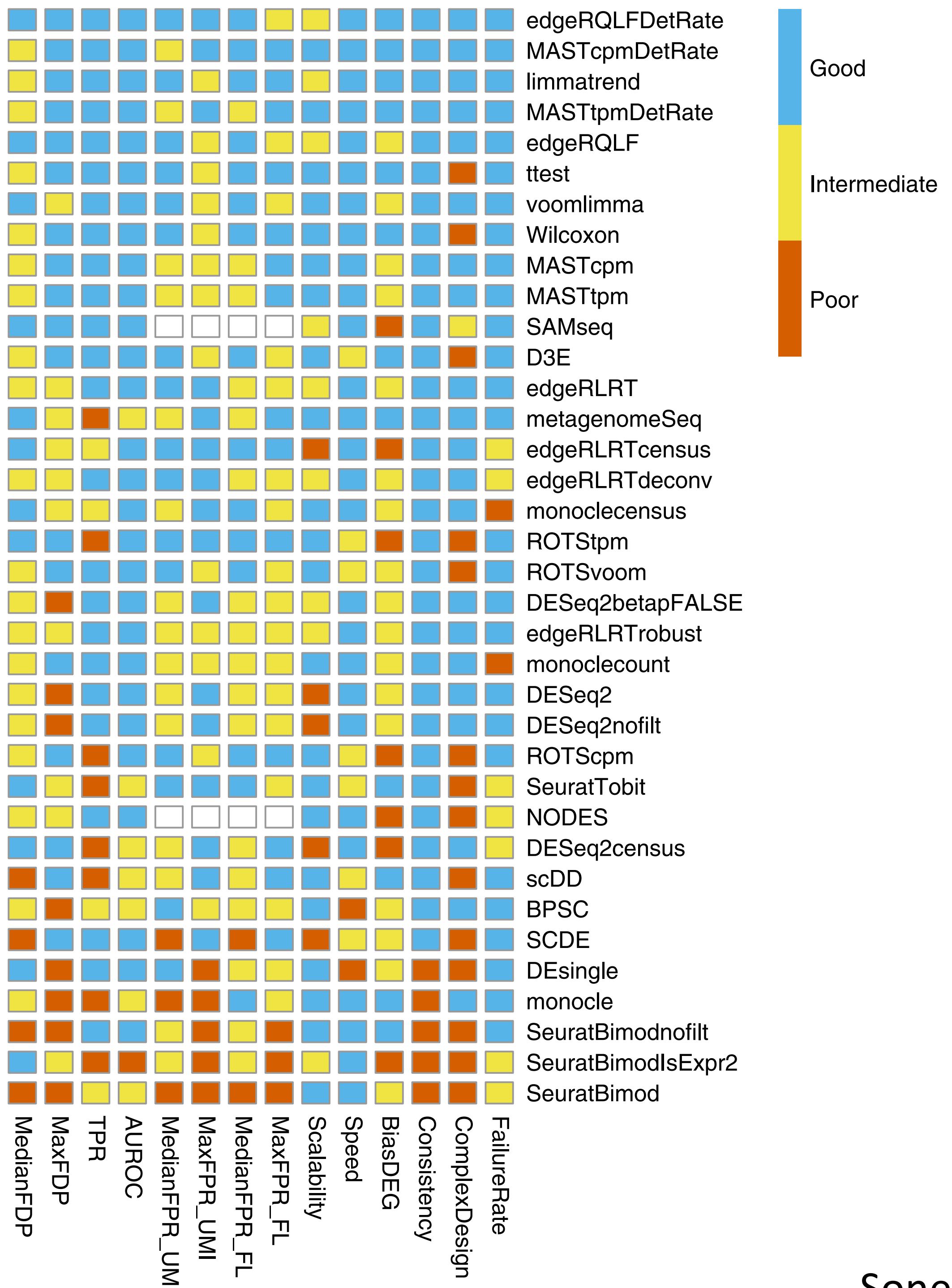
MAST for DE

- Main ideas:
  - Use log2(TPM+1) as input data
  - Both dropout probability and expression level depends on experimental conditions.

$$logit\left(Pr(Z_{ig} = 1)\right) = \mathrm{X_i}\ \upbeta_\mathrm{g}^\mathrm{D}$$

$$\Pr\left(Y_{ig} = y | Z_{ig} = 1\right) = \mathrm{N}\left(\mathrm{X_i}\upbeta_\mathrm{g}^\mathrm{C},\ \sigma_g^2\right)$$

  - Model fitting with some regularization.
  - DE is based on chi-square or Wald test.



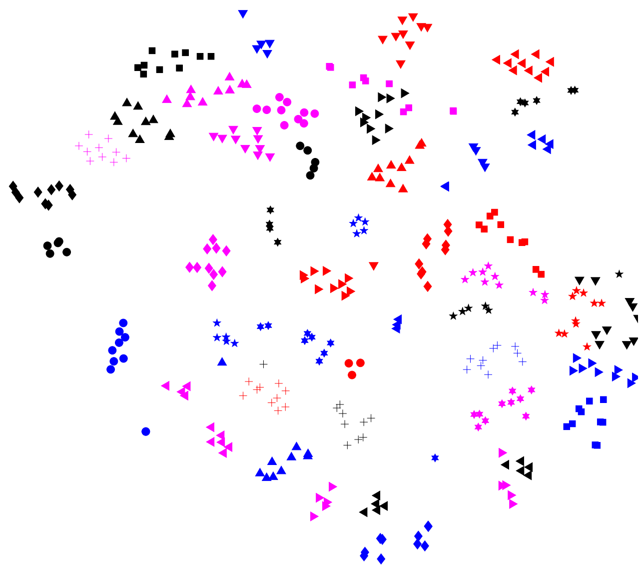
Soneson and Robinson (2018) Nat. Methods

Visualization

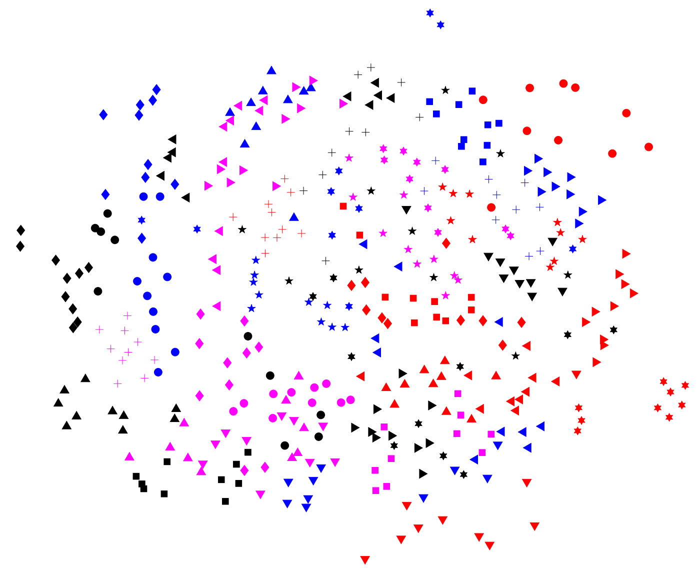
- TSNE
- UMAP

t-SNE: a useful visualization tool

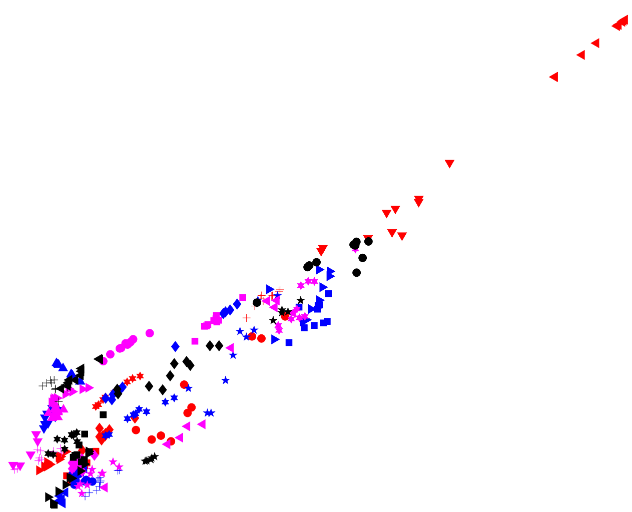
- t-SNE (t-distributed stochastic neighbor embedding): visualize high-dimensional data on 2-/3-D map.
- When project high-dimensional data into lower dimensional space, preserve the distances among data points.
  - This alleviate the problem that many clusters overlap on low dimensional space.
- Try to make the pairwise distances of points similar in high and low dimension.
- This is used in almost all scRNA-seq data visualization.
- Has “Rtsne” package on CRAN.



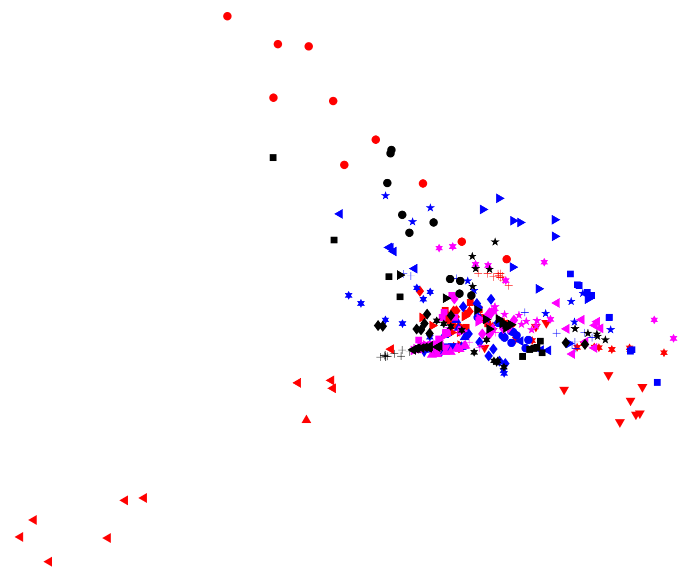
(a) Visualization by t-SNE.



(b) Visualization by Sammon mapping.



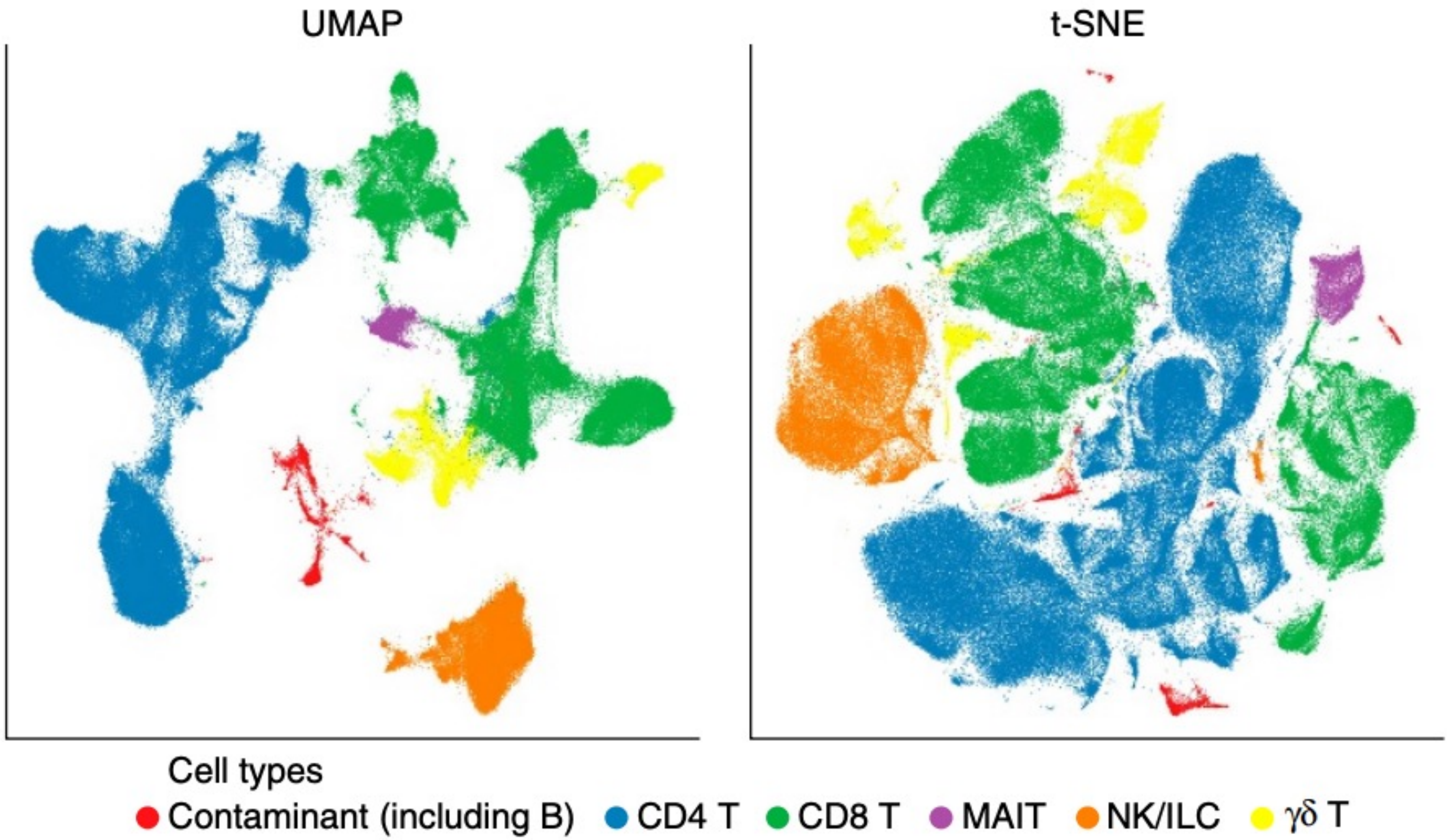
(c) Visualization by Isomap.



(d) Visualization by LLE.

UMAP: a newer (and better?) visualization tool

- UMAP (uniform manifold approximation and projection): a recently developed dimension reduction tool
- "*Comparing the performance of UMAP with five other tools, we find that UMAP provides the fastest run times, highest reproducibility and the most meaningful organization of cell clusters.* " ---- Betcht et al. 2018 Nat Biotech
- "*UMAP, which is based on theories in Riemannian geometry and algebraic topology, has been developed, and soon demonstrated arguably better performance than t-SNE due to its higher efficiency and better preservation of continuum.*" ---- Mu et al. 2018 GBP
- Has "umap" package on CRAN.



Betcht et al. 2018 Nat Biotech

Summary

- The main interests are inter-cellular heterogeneity, expression dynamics, cell type discovery, etc.
- Many statistical methods and computational tools for different biological questions.
  - Data pre-processing: normalization, batch effect, imputation
  - Cell clustering and cell type identification
  - Differential expression

Other useful resources

- My ENAR short course (with lab practices): http://www.haowulab.org//teaching/ENAR2021/scRNAseq.html
- https://github.com/theislab/single-cell-tutorial/
- https://scrnaseq-course.cog.sanger.ac.uk/website/index.html
- https://broadinstitute.github.io/2019_scWorkshop/