Single-cell RNA Sequencing Data Analysis

Mary Yang, PhD

University of Arkansas Little Rock, Information Science
UALR/UAMS Bioinformatics Graduate Program

Why single-cell sequencing?

- Cells are minimal functional units in a biological system.
- The human body is composed of approximately 37.2 trillion single cells that live harmoniously in tissues among their neighbors.
- However, in diseases such as cancer, a single cell can lead to the downfall of the entire organism.
- The single cell analysis is needed to understand the molecular behavior leading to cellular function in organisms.

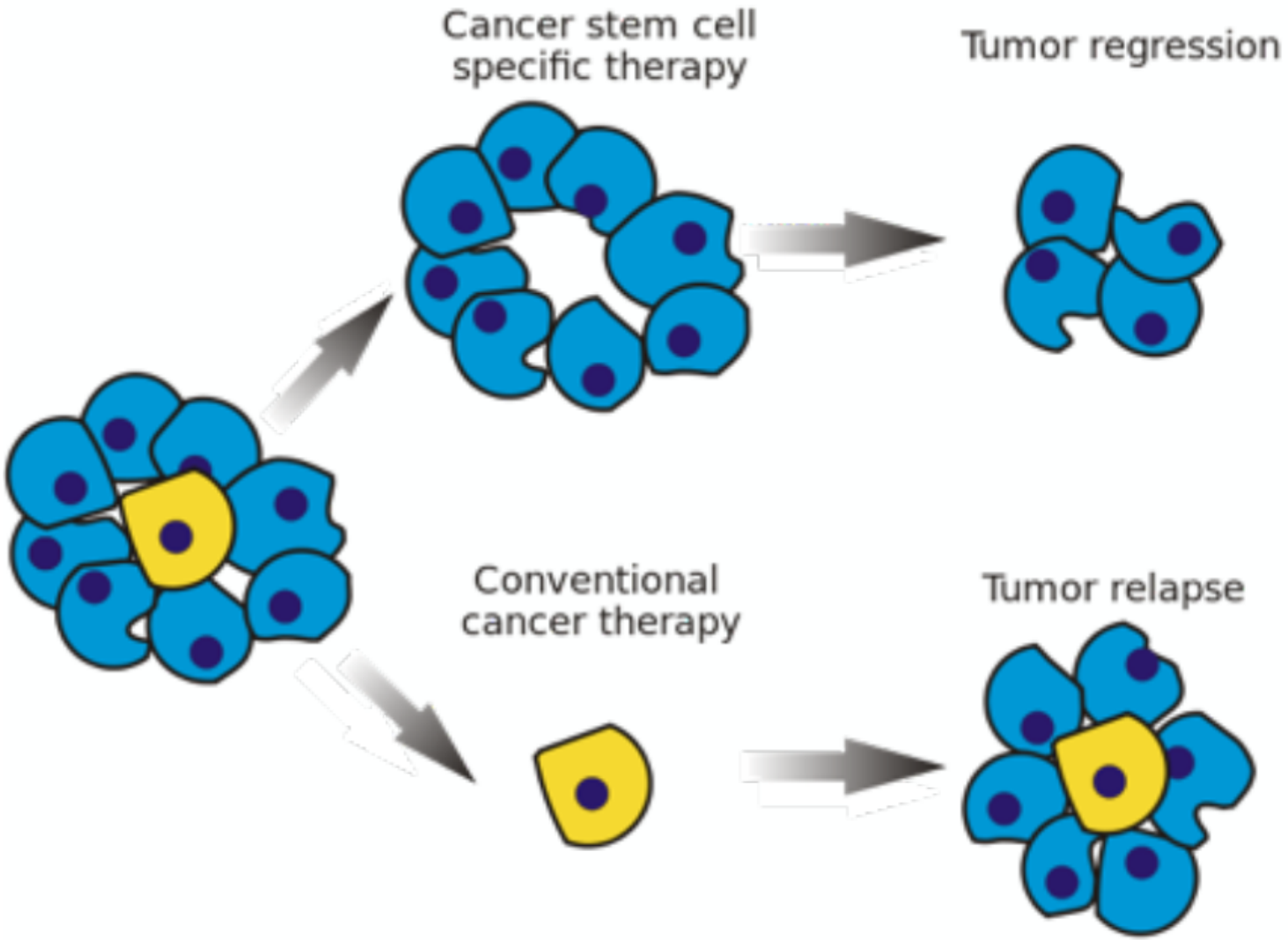
Wang et.al *Advances and Applications of Single Cell Sequencing Technologies*

SIngle-cell vs bulk level study

- Most genomic studies to date have focused on analyzing bulk tissue samples, which are composed of millions of cells.
- Consequently, it is difficult to resolve cell-to-cell variations and identify rare cells that may play an important role in disease progression.
- The single cell technology has enabled dissection of cellular heterogeneity in great detail.

Neil Savage *Eleven grand challenges in single-cell data science*,Genome Biology 2020

Cancer stem cell-specific therapy

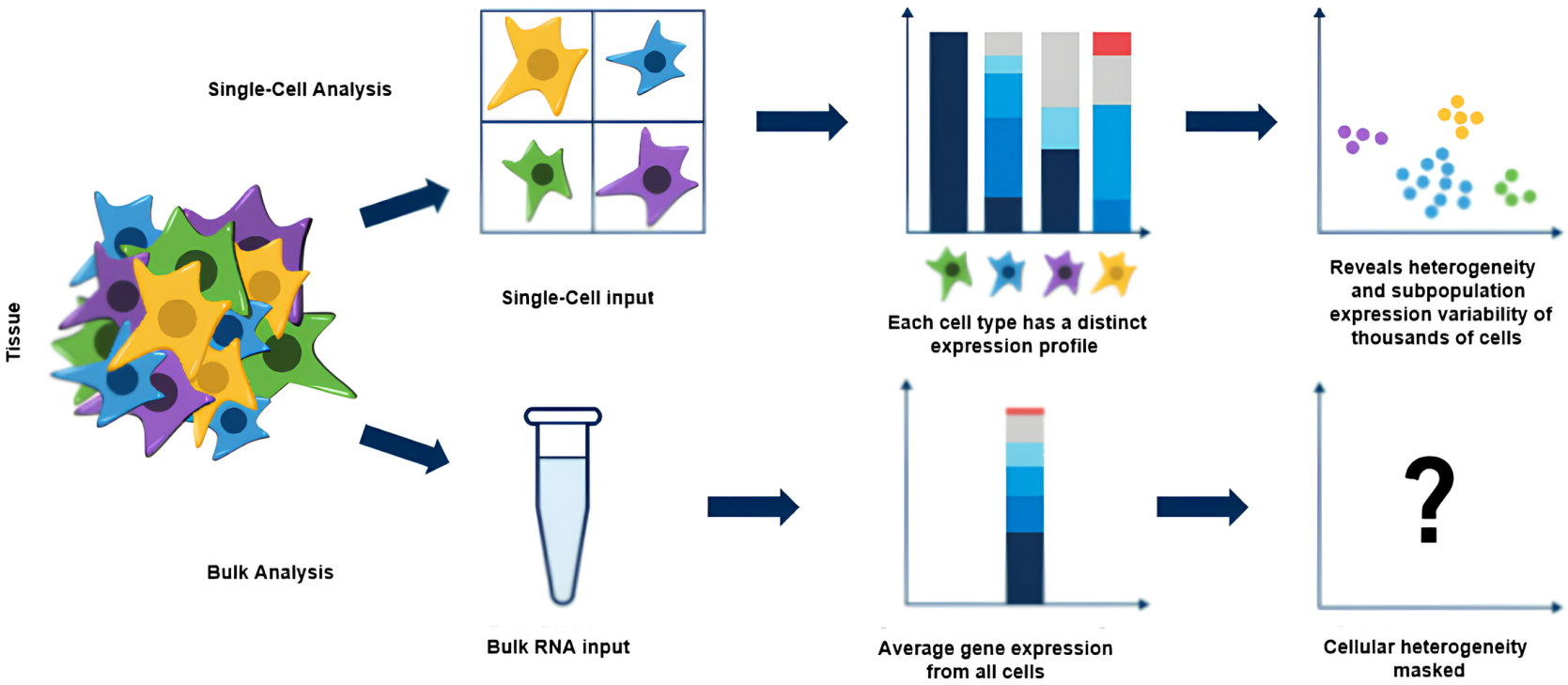


Application of single-cell RNA sequencing technology

- Single-cell RNA sequencing has been employed in different species, humans, animals and plants
- It offers a powerful tool to profile, identify, classify and discover new or rare cell types and subtypes from different human organs and tissues, giving more profound information about health and disease in
  - development
  - immunology
  - diabetes
  - microbiology
  - SARS-CoV-2
  - cancer biology
  - vascular biology
  - neurobiology and
  - clinical diagnosis
  - many other disciplines

Jovic. *Clinical Translational Med., 2021*

Single-cell RNA sequencing vs bulk RNA-sequencing

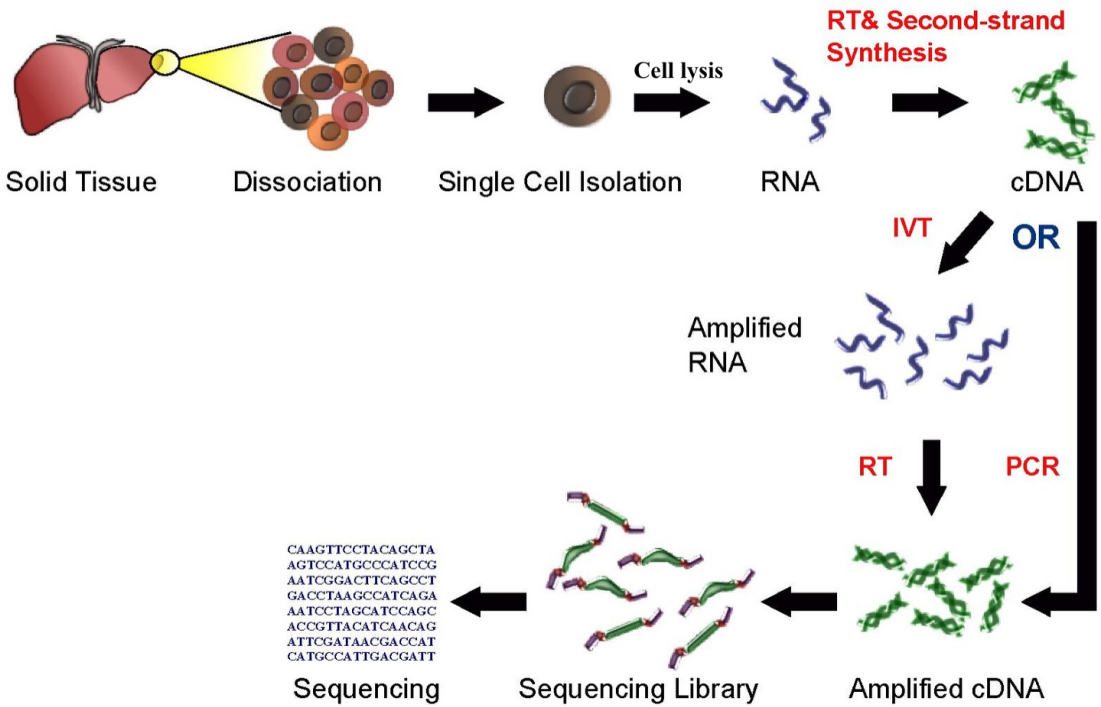


10x Genomics

scRNA-seq library preparation

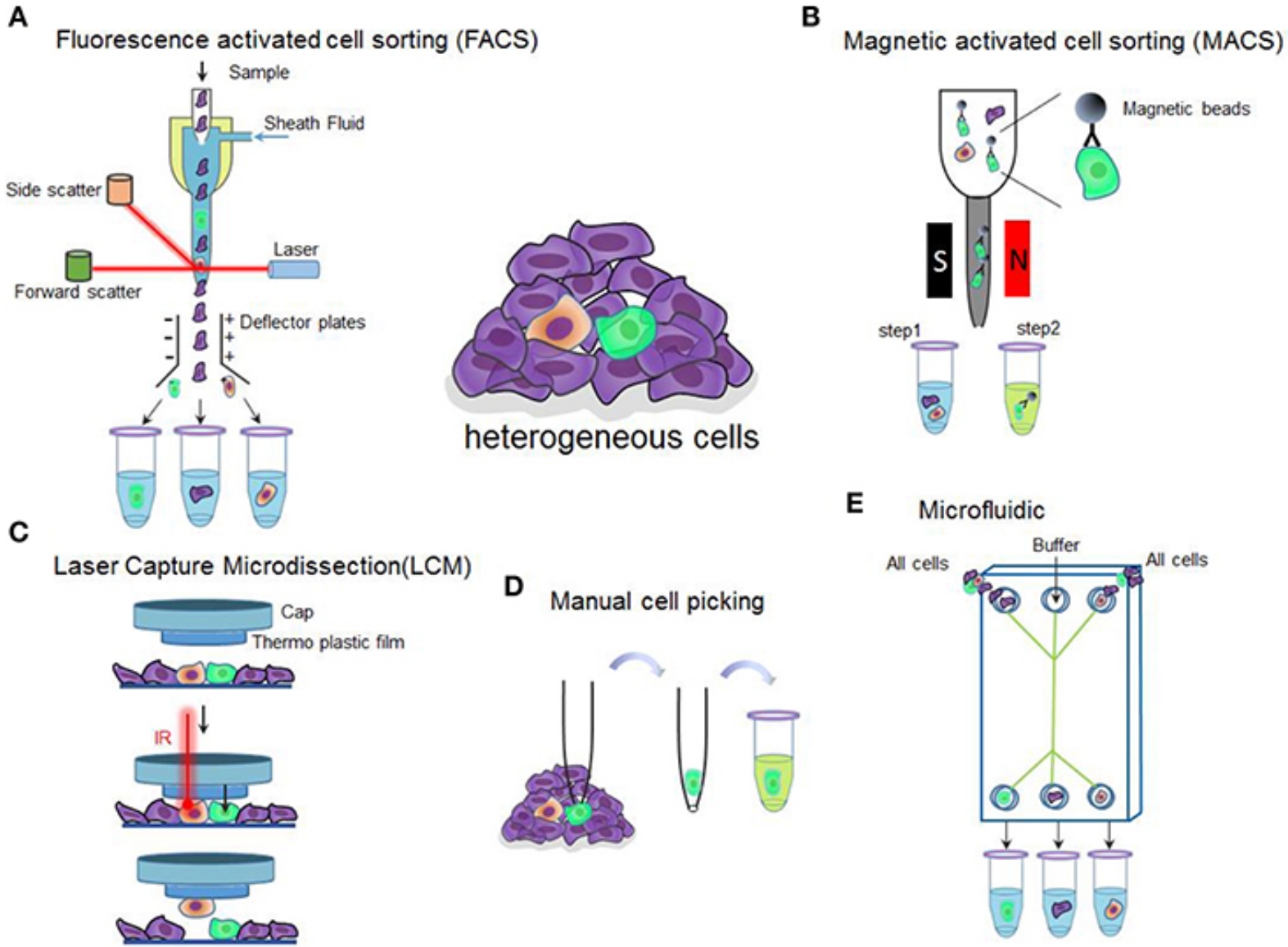
Common steps required for the generation of scRNAseq libraries include

- cell dissociation
- cell lysis
- reverse transcription into first strand cDNA
- second strand synthesis
- cDNA amplification: (PCR or IVT(in vitro transcription))



en.wikipedia.org

Cell isolation protocol



Hu et.al. *Frontiers in Cell and Development Biology 2016*

Cell isolation protocol

Techniques	Throughput	Advantage	Disadvantage	References
Fluorescence-activated cell sorting (FACS)	High	High specificity multiple parameters	Large amount of material, dissociated cells, high skill needed	Gross et al., 2015
Magnetic-activated cell sorting (MACS)	High	High specificity, cost effective	Dissociated cells, non-specific cell capture	Welzel et al., 2015
Laser capture microdissection (LCM)	Low	Intact fixed and live tissue	Contaminated by neighboring cells, high skill needed	Espina et al., 2007; Datta et al., 2015
Manual cell picking	Low	Intact live tissue	High skill needed, low throughput	Citri et al., 2012
Microfluidic	High	Low sample consumption, integrated with amplification	Dissociated cells, high skill needed	Bhagat et al., 2010; Lecault et al., 2012

Hu et.al. *Frontiers in Cell and Development Biology 2016*

- Microfluidics has become popular due to its low sample consumption, precise fluid control, and low operating costs.
- In particular, droplet-based microfluidics (also called microdroplets) is currently the most popular high-throughput platform

scRNA-seq techonogies

- A number of scRNA-seq approaches are available now
- These approaches differ in one or more of the following aspects:
  - cell isolation
  - cell lysis
  - reverse transcription
  - amplification
    - PCR: Polymerase chain reaction
    - IVT: in vitro transcrition
  - transcript coverage
  - strand specificity
  - availability of UMI (unique molecular identifiers, molecular tags that can be applied to detect and quantify the unique transcripts)

Chen et.al. *Frontiers in Genetics, 2019*

Commonly used scRNA-seq technologies

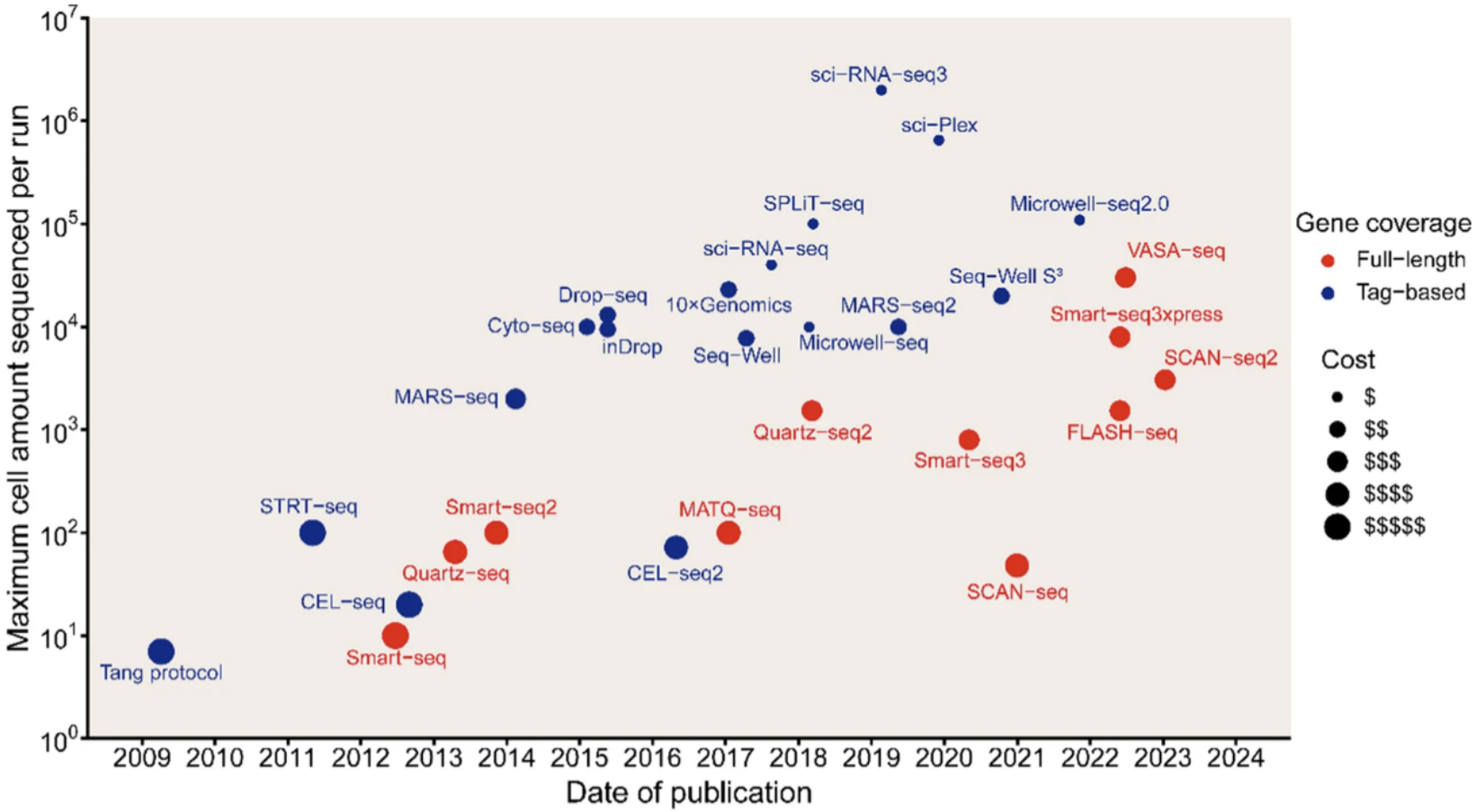
Methods	Transcript coverage	UMI possibility	Strand specific	References
Tang method	Nearly full-length	No	No	Tang et al., 2009
Quartz-Seq	Full-length	No	No	Sasagawa et al., 2013
SUPeR-seq	Full-length	No	No	Fan X. et al., 2015
Smart-seq	Full-length	No	No	Ramskold et al., 2012
Smart-seq2	Full-length	No	No	Picelli et al., 2013
MATQ-seq	Full-length	Yes	Yes	Sheng et al., 2017
STRT-seq and STRT/C1	5′-only	Yes	Yes	Islam et al., 2011, 2012
CEL-seq	3′-only	Yes	Yes	Hashimshony et al., 2012
CEL-seq2	3′-only	Yes	Yes	Hashimshony et al., 2016
MARS-seq	3′-only	Yes	Yes	Jaitin et al., 2014
CytoSeq	3′-only	Yes	Yes	Fan H.C. et al., 2015
Drop-seq	3′-only	Yes	Yes	Macosko et al., 2015
InDrop	3′-only	Yes	Yes	Klein et al., 2015
Chromium	3′-only	Yes	Yes	Zheng et al., 2017
SPLiT-seq	3′-only	Yes	Yes	Rosenberg et al., 2018
sci-RNA-seq	3′-only	Yes	Yes	Cao et al., 2017
Seq-Well	3′-only	Yes	Yes	Gierahn et al., 2017
DroNC-seq	3′-only	Yes	Yes	Habib et al., 2017
Quartz-Seq2	3′-only	Yes	Yes	Sasagawa et al., 2018

Compare commonly used scRNA-seq approaches

- Compared to 3′end and 5′end scRNA-seq technologies, the full-length scRNA-seq methods have incomparable advantages in
  - isoform usage analysis
  - allelic expression detection
  - RNA editing identification
- However, the full-length sequencing methods are limited by lower cellular throughputs and higher costs
- Droplet-based technologies, such as Drop-seq, InDrop, and Chromium 10x genomics, can generally provide a lager throughput of cells and a lower sequencing cost per cell compared to whole-transcript scRNA- seq
  - Suitable for generating huge amounts of cells to identify the cell subpopulations of complex tissues or tumor samples.

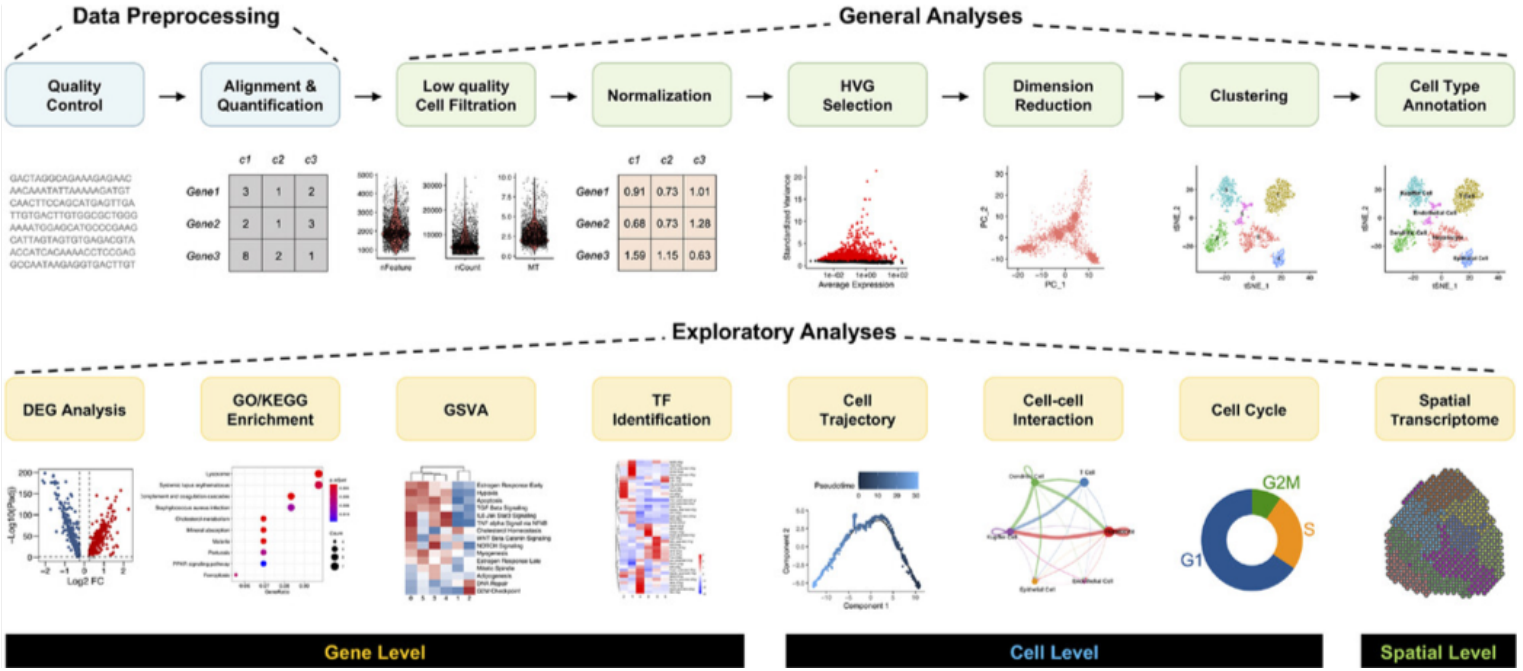
Chen et.al. *Frontiers in Genetics, 2019*

Timeline and throughput of various scRNA-seq methods



Huang et.al. *Journal of Hematology Oncology, 2023*

scRNA-seq data analysis work flow



Jovic. *Clinical Translational Med., 2021*

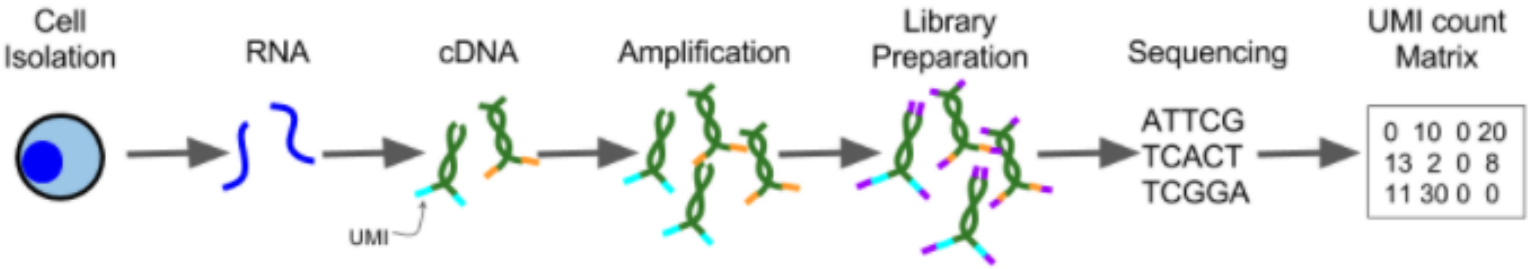
ERCC and UMI

- Sources of technical variability include capture inefficiency, drop out, amplification bias, sequencing depth and coverage, library size, sequencing inefficiency, and batch effects.
- To estimate the technical variances among different cells, spike-ins, such as External RNA Control Consortium (ERCC) controls and Unique molecular identifiers (UMIs) have been widely used in the scRNA-seq methods.
- The RNA spike-ins are RNA transcripts (with known sequences and quantity) that are applied to calibrate the measurements of RNA hybridization assays
- UMIs can theoretically enable the estimation of absolute molecular counts.

Chen et.al. *Frontiers in Genetics, 2019*

ERCC and UMI

- ERCC and UMIs are not applicable to all scRNA-seq technologies due to the inherent protocol differences
- Spike-ins are used in approaches like Smart-seq2 and SUPeR-seq but are not compatible with droplet-based methods
- UMIs are short random bardcodes (4-10 bp), which are added to indivisual transcripts during reverse transcription
- UMIs are typically applied to 3′end sequencing technologies and help to remove the amplication noise and biases from scRNA-seq data.



Chen et.al. *Frontiers in Genetics, 2019*

Preprocessing raw scRNA-seq data

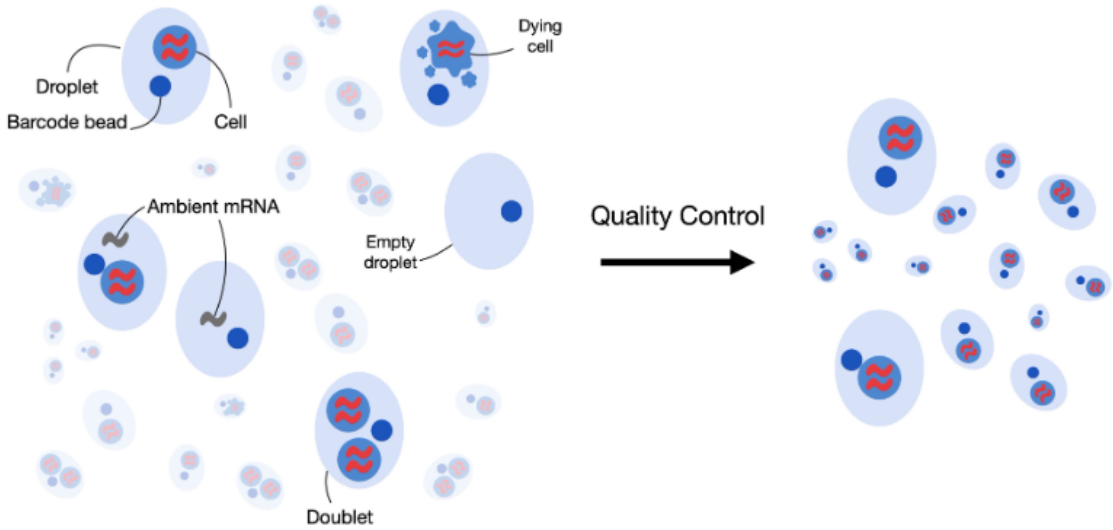
- Assess reads quality: FastQC is a quality control took for both single-cell and bulk sequence data
- Trim sequencing adapter and/or low qualty reads: cutadapt
- Align reads: The alignment tools originally developed for bulk RNA-seq are also applicable to scRNA-seq data.

Tools	Category	URL	References
TopHat2	Read mapping	https://ccb.jhu.edu/software/tophat/index.shtml	Kim et al., 2013
STAR	Read mapping	https://github.com/alexdobin/STAR	Dobin and Gingeras, 2015
HISAT2	Read mapping	https://ccb.jhu.edu/software/hisat2/index.shtml	Kim et al., 2015

Chen et.al. *Frontiers in Genetics, 2019*

Low quality cells

- scRNA-seq experiments can generate a portion of low-quality data from the cells that are broken or dead or mixed with multiple cells
- A series of QC analyses are required to eliminate low-quality cells for downstream analysis



Chen et.al. *Frontiers in Genetics, 2019*

Quality control metrics for filtering out low quality cells

- The samples contain only a few number of reads should be discarded
- The samples with very low mapping ratio should be eliminated
- The cells with an extremely high portion of reads mapped to the spike-ins indicate that they were probably broken during cell capture process and should be removed
- The number of expressed genes/transcripts can be detected in each cell is also suggestive.
- The ratio of reads mapped to mitochondrial genome is also informative for identifying low-quality cells

Chen et.al. *Frontiers in Genetics, 2019*

QC matrics: mitochondrial proportion

- The mtDNA% threshold is used to filter out apoptotic, stressed, low-quality cells in the data.
- mtDNA% threshold depends highly on the tissue type and the questions being investigated
  - A very stringent mtDNA% threshold may cause bias in the recovered cellular composition of the tissue under study and increase the cost of the experiment.
  - Inversely, a relaxed threshold of mtDNA% may allow apoptotic, low-quality cells to remain in the analysis, resulting in the identification of wrong biological patterns.

Osorio et.al . *Bioinformatics, 2020*

Quality control software

Some QC methods for scRNA-seq have been proposed:

- SinQC (Jiang et al., 2016)
- Scater (McCarthy et al., 2017)

Batch effect correction

- Single-cell data is often generated from multiple experiments with differences in capturing times, handling technician, cell dissociation protocols, reagent lots, equipments, and technology platforms.
- These differences lead to large variations or batch effects in the data, and can confound biological variations of interest during data integration

Batch effect correction

- In a published study, fourteen scRNA-seq batch correction algorithms were compared
- The study showed that each batch-effect removal method has its advantages and limitations
- Based on their results, LIGER, Harmony, and Seurat 3 are with top performance

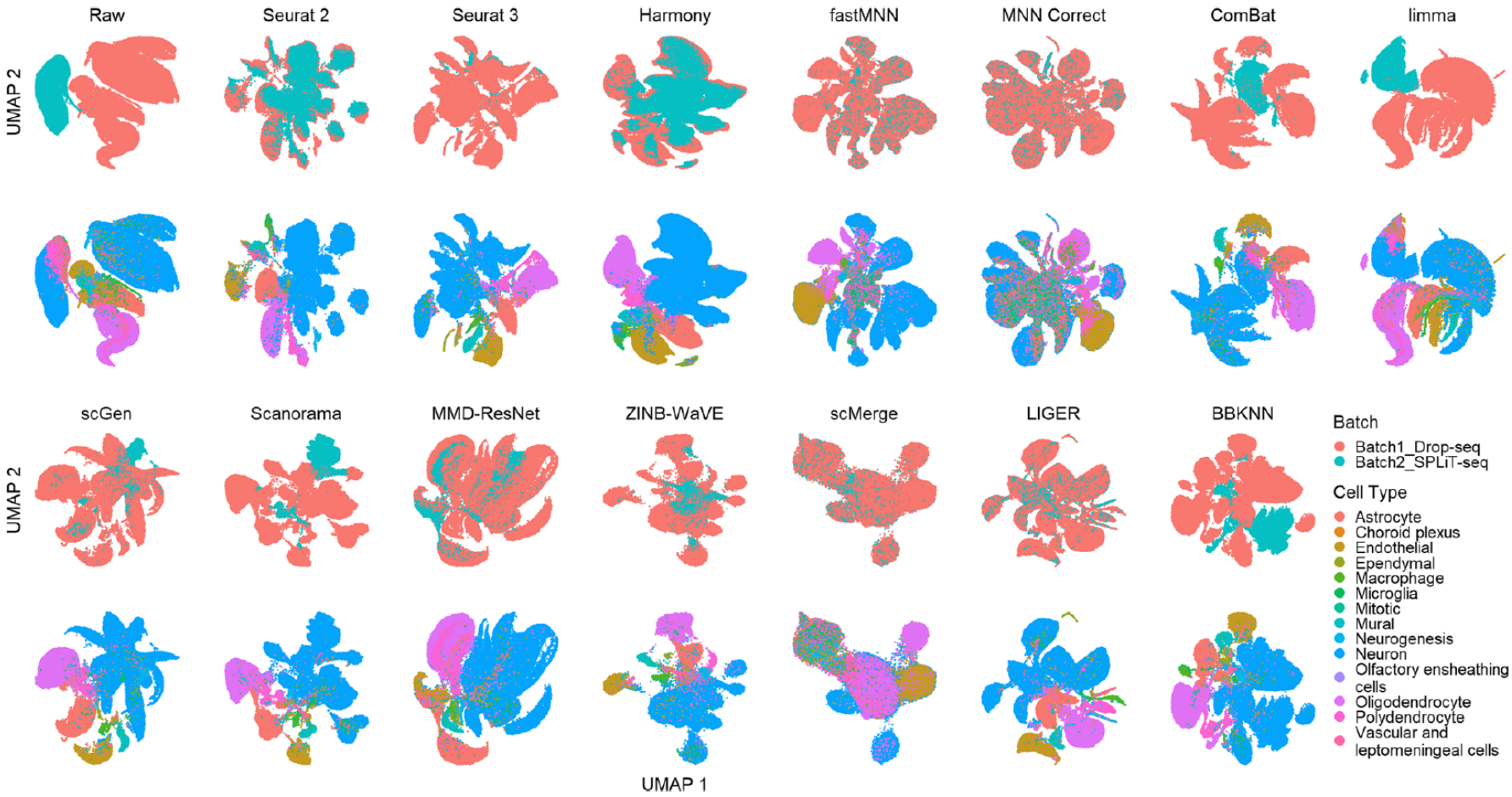
Tran et.al . *Genome Biology, 2020*

Batch effect correction

- Harmony performed well on datasets with common cell types, and also different technologies, as well as comparatively low runtime
- LIGER performed well, especially on datasets with non-identical cell types, longer runtime than Harmony
- Seurat 3 is also able to handle large datasets, however with 20–50% longer runtime than LIGER.
- scMerge is recommended for improving recovery of DEGs in batch-corrected data

Tran et.al . *Genome Biology, 2020*

An example of batch effect correction: mouse brain scRNA-seq data



Tran et.al . *Genome Biology, 2020*

Nomalization

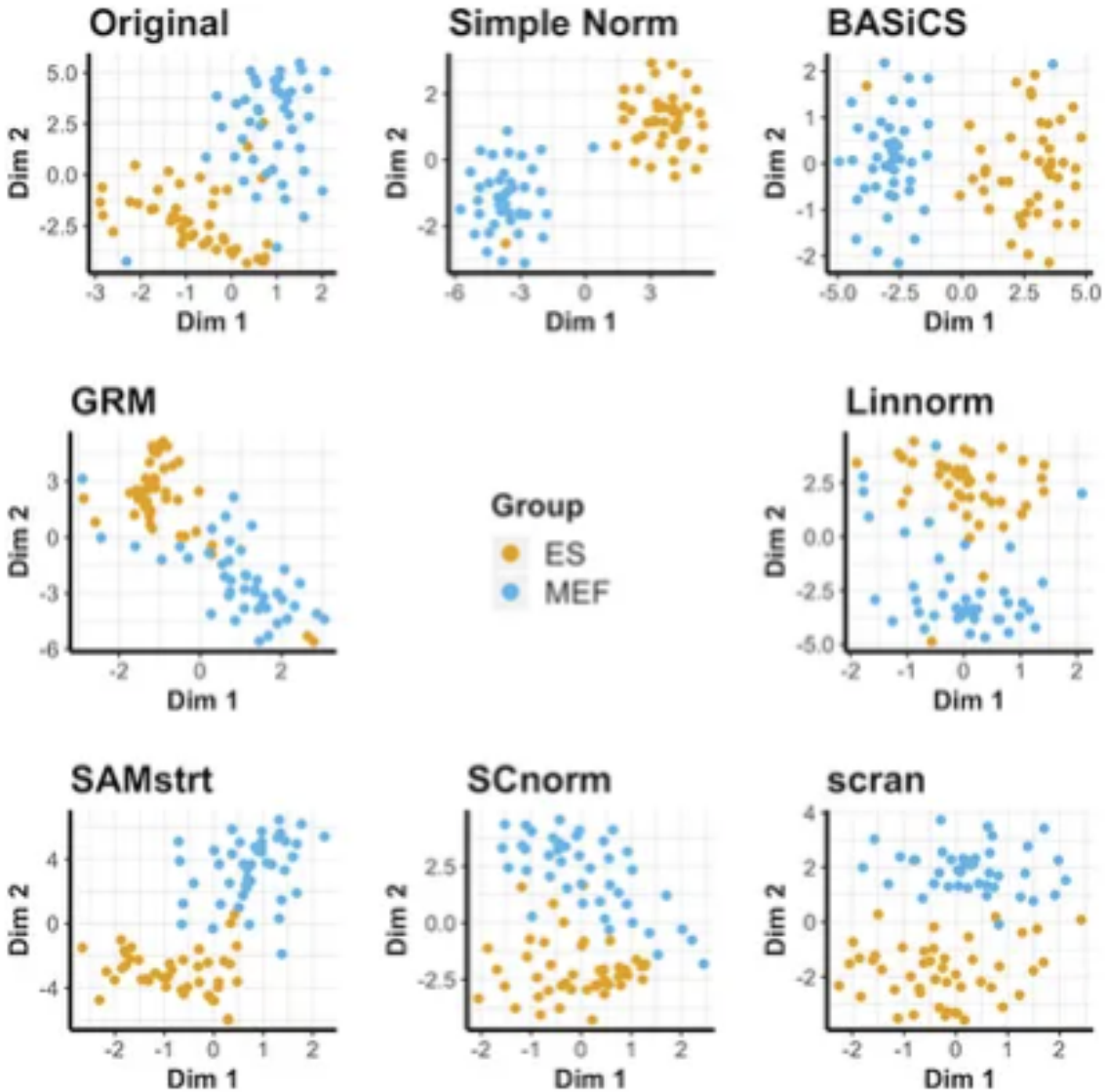
- Normalization is an essential step to reveal true signal by adjusting unwanted biases resulted from technique variations.
- In general, normalization can be divided into two different types: within-sample normalization and between-sample normalization
  - Within- sample normalization aims to remove the gene-specific biases (e.g., GC content and gene length), which makes gene expression comparable within one sample
  - between-sample normalization is to adjust sample- specific differences (e.g., sequencing depth and capture efficiency) to enable the comparison of gene expression between samples.
- If spike-ins or UMIs are used in scRNA-seq protocol, normalization can be refined based on the performance of spike-ins/UMIs

Normlization software

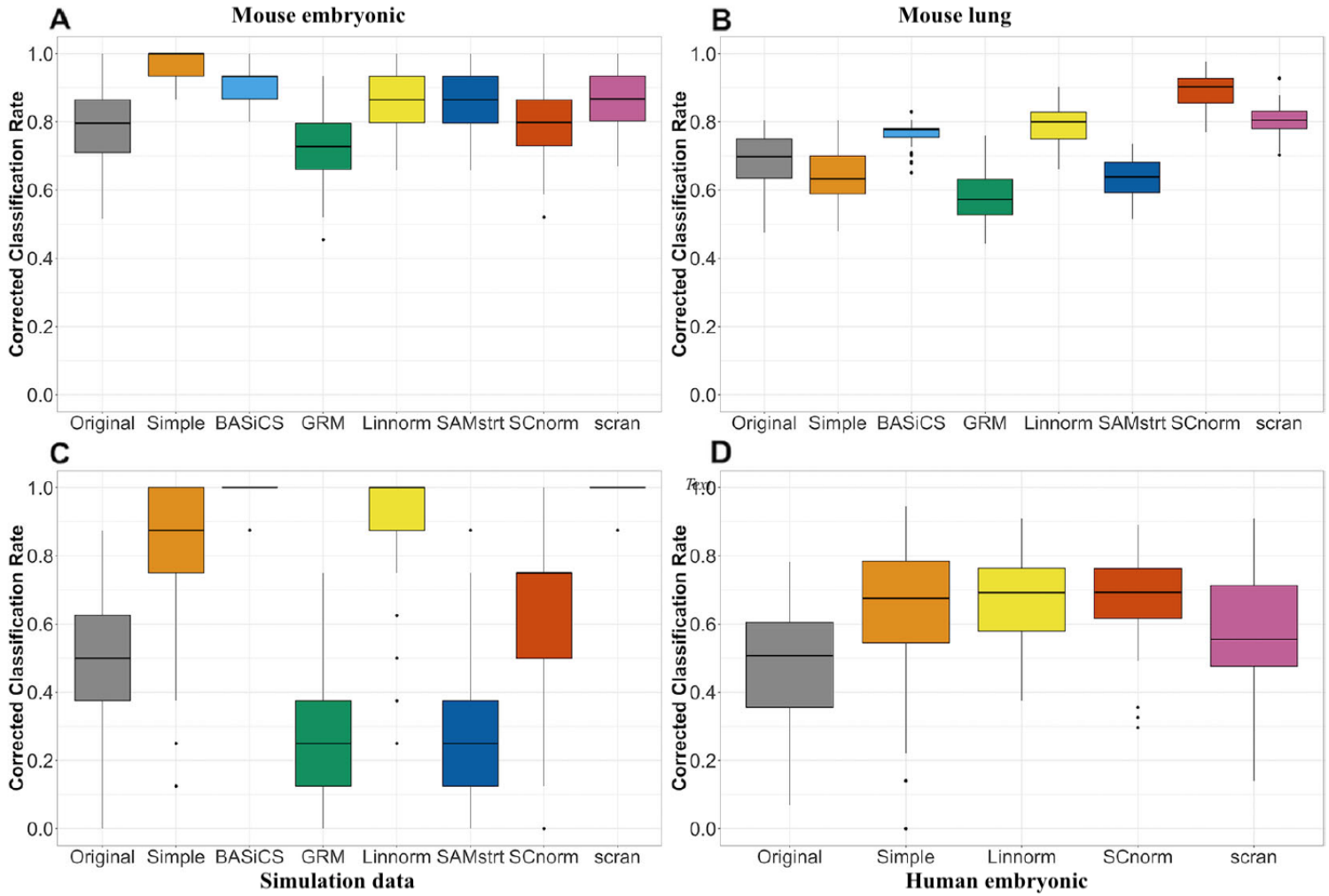
Method	Author	Year	Spike-ins	Model Description
SAMstrt	Katayama et al.	2013	Yes	Poisson resampling and non-parametric statistics
BASiCS	Vallejos et al.	2015	Yes	Use spike-ins for hierarchical Poisson/ Gamma model for technical variability. Expand model to incorporate biological genes with new Poisson model
GRM	Ding et al.	2015	Yes	Gamma regression model from spike-ins
Simple Norm.	Satija et al.	2015	No	Divide gene counts for cells, then multiply by scale factor and apply a log(x +1) transformation to the result (included in the Seurat package as NormalizeData)
scran	Lun et al.	2016	No	Deconvolution of size factors from constructed linear system. Form pools of cells and normalize using summed expression values
SCnorm	Bacher et al.	2017	Optional	Quantile based model for log sequencing depth.
Linnorm	Yip et al.	2017	Optional	Linear models defined with a normalization strength coefficient to update means. Focuses on stable genes to perform normalization

Lytal et.al. *Frontiers in Genetics, 2020*

t-SNE plots of the mouse embryonic data set under various normalization methods



Normalizatin software comparison



Lytal et.al. *Frontiers in Genetics, 2020*

Normalizatin software comparison

Category	BASiCS	GRM	Linnorm	SAMstrt	SCnorm	scran	Simple Norm
Mouse Embryonic Data (sec)	230	35	<5	<5	760	<5	<5
Mouse Lung Data (sec)	510	60	<5	5	1180	<5	<5
Mouse Embryonic Sim Data (sec)	110	15	<5	<5	110	<5	<5
Human Embryonic Data (sec)	_	_	<5	_	370	<5	<5
Classification (Spike-In Genes)	***	*	**	*	***	**	**
Visualization (Spike-In Genes)	**	*	**	*	**	**	**
Classification (Non-Spike-In)	_	_	*	_	**	**	**
Visualization (Non-Spike-In)	_	_	*	_	**	*	**

*A '***' indicates that the method showed exemplary performance overall, a '**' indicates satisfactory performance overall, and a '*' indicates some shortcoming in performance compared to other methods. '_' indicates that the method is not applicable for this type of data. Two variants exist for the Mouse Embryonic Sim Data, but they had negligible differences in execution time.*

Lytal et.al. *Frontiers in Genetics, 2020*

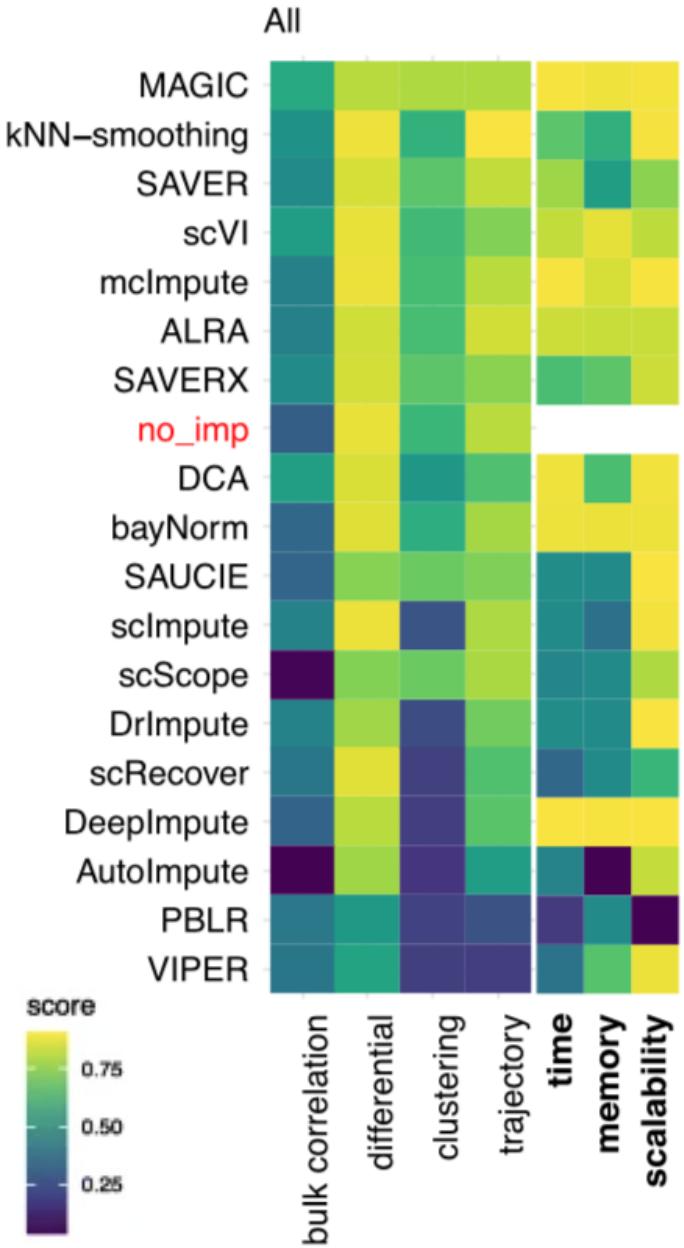
Recover missing values

- scRNA sequencing data often contain many missing values or dropouts that usually caused by the low amounts of mRNA in individual cells.
- Consequently, significant portion of truly expressed transcripts may not be detectable in scRNA-seq.
- Several imputation methods have been recently developed for scRNA-seq
  - Model-based imputation methods: bayNorm, SAVER, SAVER-X, scImpute, scRecover, VIPER;
  - Smooth-based imputation methods: DrImpute, MAGIC, kNN-smoothing
  - Data reconstruction methods using deep-learning methods: AutoImpute, DCA, DeepImpute, SAUCIE, scScope, scVI
  - Low-rank matrix-based methods: ALRA, mcImpute, PBLR.

Hou et.al. *Genome Biology, 2020*

Methods comparison: Recover missing values

- bulk correlation: similarity between imputed single-cell and bulk profiles
- differential: differential expression
- clustering: unsupervised clustering
- trajectory: trajectory inference,
- time
- memory usage
- scalability.
- Performances were all scaled to be in [0,1], a higher score represents a better performance.

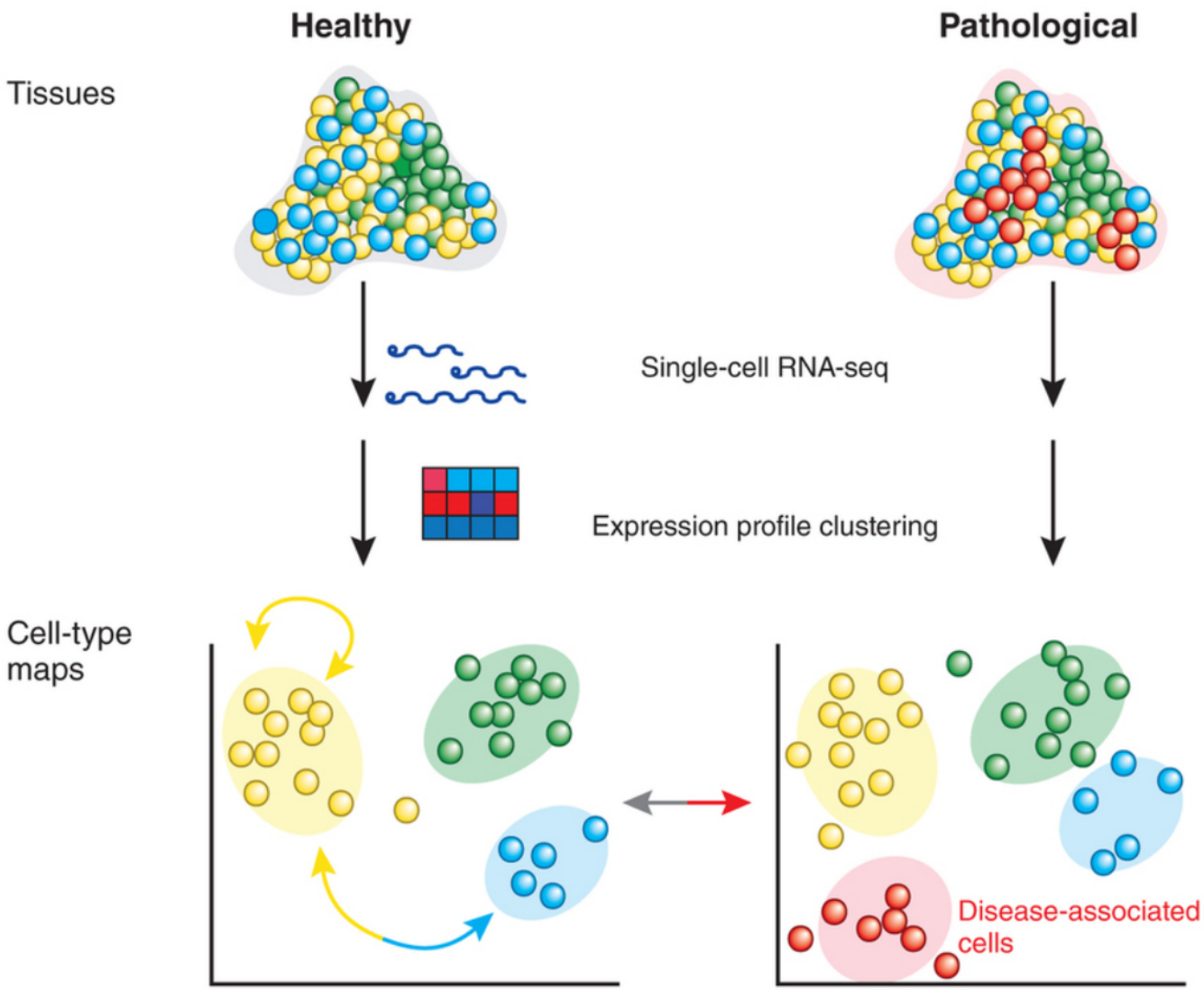


Dimensionality Reduction

- Dimensionality reduction methods that transform the original high-dimensional noisy expression matrix into a low-dimensional subspace with enriched signals
- Commonly used dimensionality reduction methods.
  - PCA is a linear dimensional reduction algorithm, which assumes that the data is approximately normally distributed.
  - T-distributed stochastic neighbor embedding (t-SNE) is a non-linear approach mainly designed for visualizing high dimensional data
  - UMAP (uniform manifold approximation and projection) (Becht et al., 2018), and scvis (Ding et al., 2018) were specially developed for reducing the dimensions of scRNA-seq data

Chen et.al. *Frontiers in Genetics, 2019*

Cell type Identification

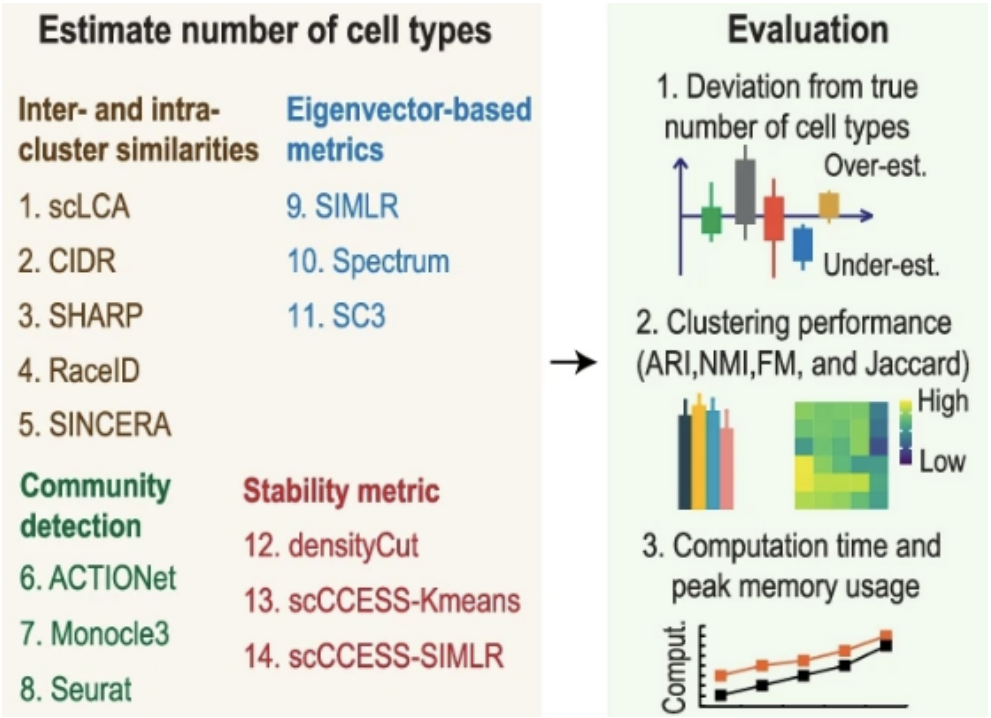


Sandberg *Nature Methods 2014*

Cell type Identification

Approaches forclustering cells can be mainly grouped into two categories based on whether prior information is used.

- If a set of known markers was used in clustering, the methods are prior information based.
  - ScType (Lanevsk et.al 2022), a fully-automated cell type identification using a comprehensive cell marker database as background information.
- Alternatively, unsupervised clustering methods can be used for de novo detection of cell populations with scRNA-seq data.



Yu et al. *Genome Biology*, 2022

Evaluation matricess for clustering

Quantify the concordance of clustering results on each scRNA-seq dataset with respect to their predefined cell-type annotations

- Adjusted Rand index (ARI)
- normalised mutual information (NMI)
- Fowlkes–Mallows index (FM)
- Jaccard index (Jaccard).

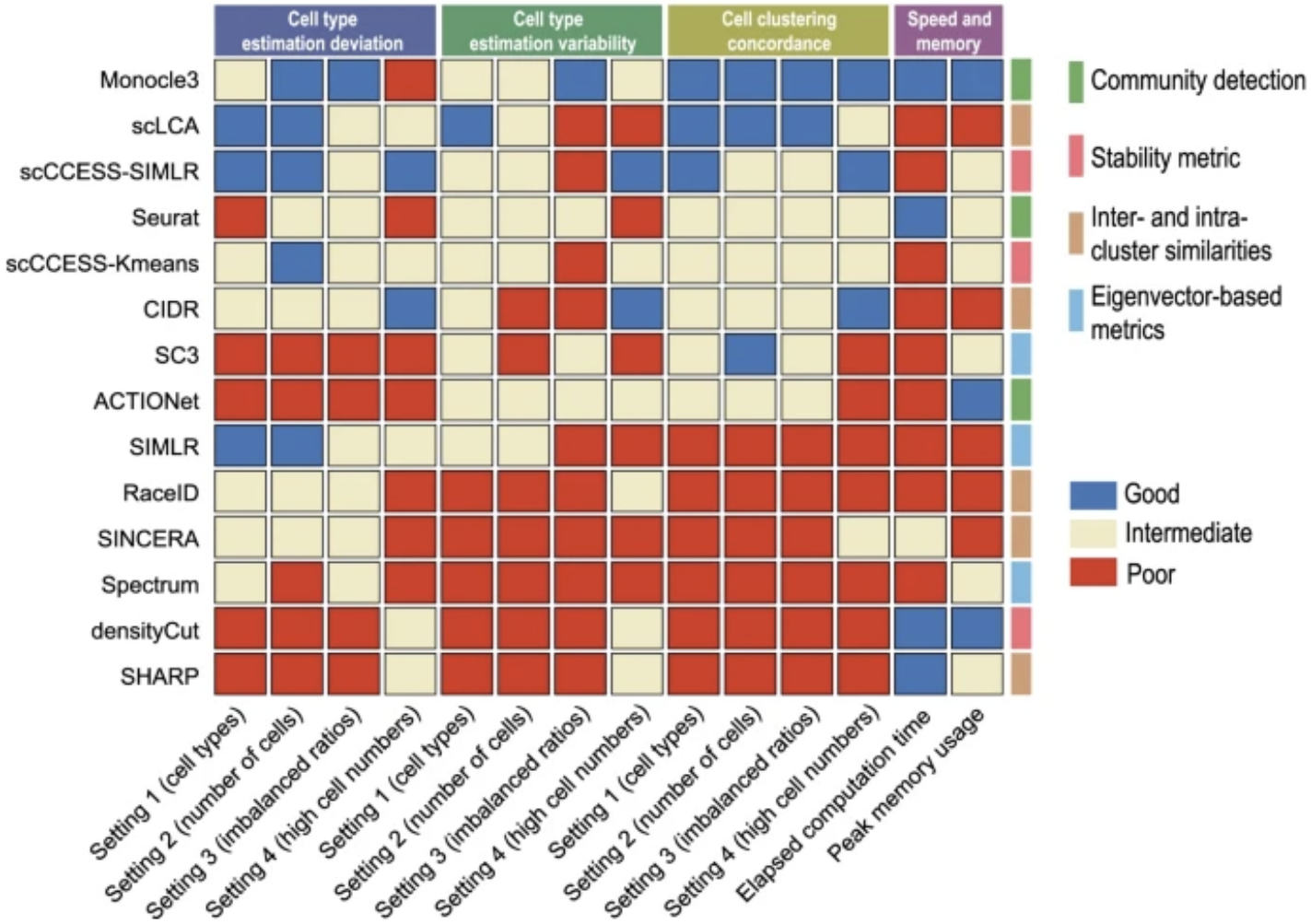
Yu et.al. *Genome Biology, 2022*

Evaluate cell clustering methods

- Benchmark the fourteen clustering methods across a large number of datasets sampled from the Tabula Muris project representing different data characteristics in various settings.
  - Tabula Muris dataset contains 53,760 cells (FACS sorted and sequenced using Smart-Seq2 protocol) from 81 cell types of 20 organs of 7 mice.
- Evalute the accuracy on determining the number of cell types, performance of cell clustering, and computing time and peak memory
- Further cross-compared the performance of clustering algorithms on datasets with a large number of cells using both Tabula Muris and Tabula Sapiens data.

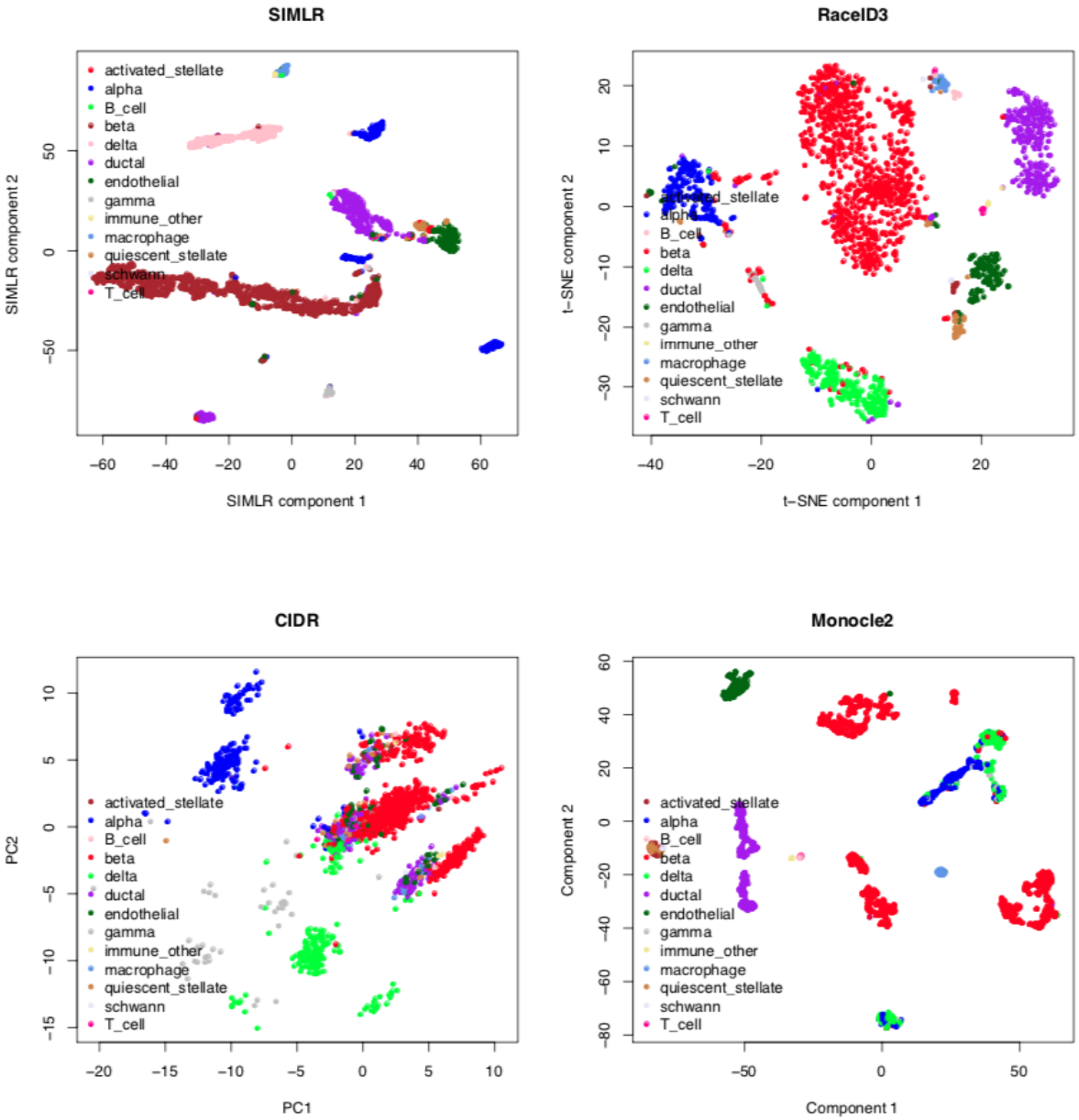
Yu et.al. *Genome Biology, 2022*

clustering method performance across all major evaluation criteria



Yu et.al. *Genome Biology, 2022*

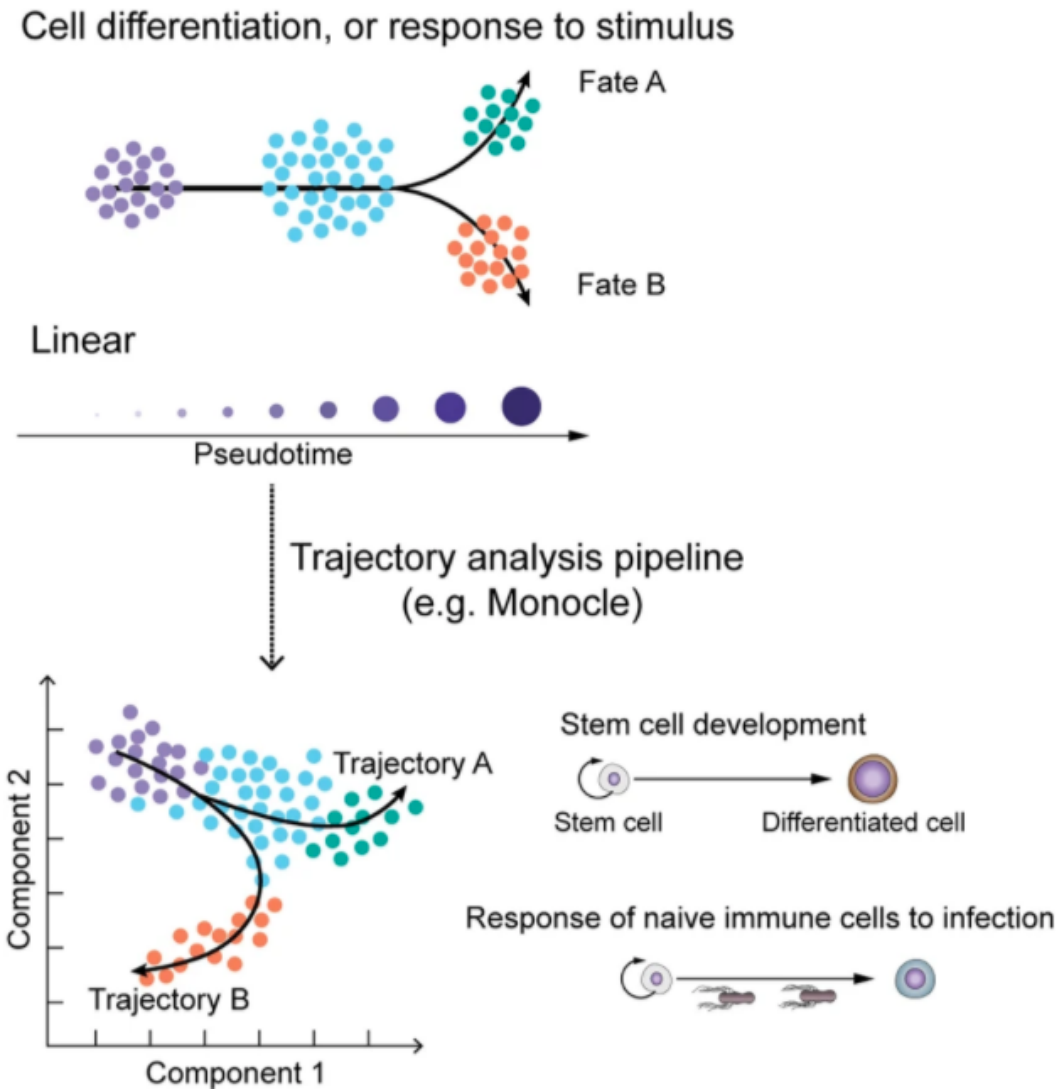
Visulization of clustering results



Cell lineage construction

- The cells in many biological systems exhibit a continuous spectrum of states and involve transitions between different cellular states.
- Such dynamic processes can be computationally modeled by reconstructing the cell trajectory and pseudotime based on scRNA-seq data.
- Pseudotime is an ordering of cells along the trajectory of a continuously developmental process in a system, which allows the identification of the cell types at the beginning, intermediate, and end states of the trajectory

Cell lineage construction

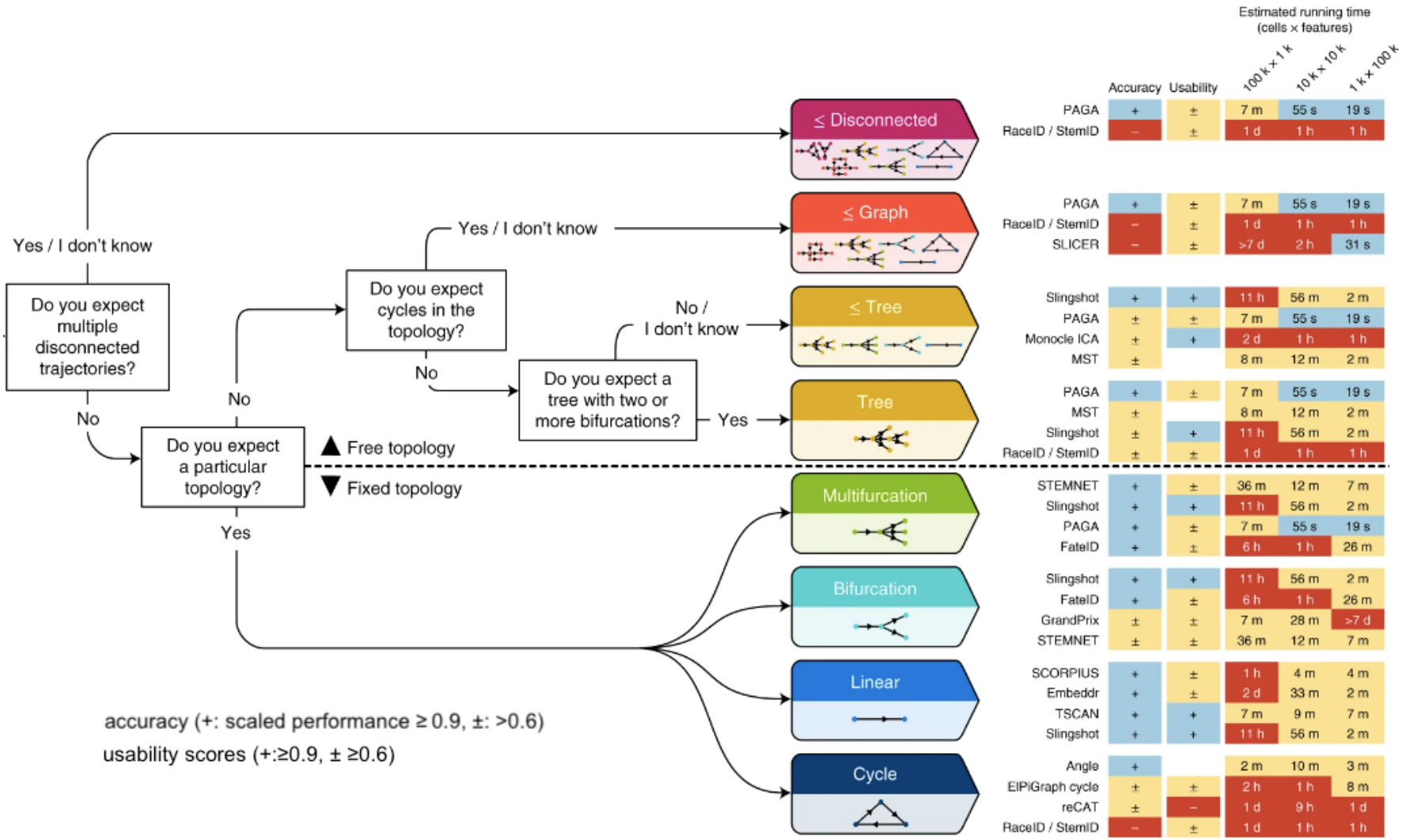


Hwang et.al. *Exp.Mol Med. 2018*

Cell lineage construction methods

Tools	Dimensionality reduction	URL	References
Monocle	ICA	http://cole-trapnell-lab.github.io/monocle-release/	Trapnell et al., 2014
Waterfall	PCA	https://www.cell.com/cms/10.1016/j.stem.2015.07.013/attachment/3e966901-034f-418a-a439-996c50292a11/mmc9.zip	Shin et al., 2015
Wishbone	Diffusion maps	https://github.com/ManuSetty/wishbone	Setty et al., 2016
GrandPrix	Gaussian Process Latent Variable Model	https://github.com/ManchesterBioinference/GrandPrix	Ahmed et al., 2019
SCUBA	t-SNE	https://github.com/gcyuan/SCUBA	Marco et al., 2014
DPT	Diffusion maps	https://media.nature.com/original/nature-assets/nmeth/journal/v13/n10/extref/nmeth.3971-S3.zip	Haghverdi et al., 2016
TSCAN	PCA	https://github.com/zji90/TSCAN	Ji and Ji, 2016
Monocle2	RGE	http://cole-trapnell-lab.github.io/monocle-release/	Qiu et al., 2017
Slingshot	Any	https://github.com/kstreet13/slingshot	Street et al., 2018
CellRouter	Any	https://github.com/edroaldo/cellrouter	Lummertz da Rocha et al., 2018

Compare cell lineage construction methods



Salens et.al. *Nature biotechnoloty, 2019*

Differential expression analysis

- Differentially expressed genes (DEGs) are identified by comparing gene expression levels across one or more conditions, such as diseases, genetic knockouts, or drug treatments.
- The analysis evaluates the magnitude and significance of differences in gene expression patterns between the condition of interest and a reference group.
- In the case of single-cell RNA-Seq is commonly applied on the cell type level.

Differential expression analysis

- The technical variability, high noise (e.g., dropouts) and massive sample size of scRNA-seq data raise challenges in differential expression calling (McDavid et al., 2013).
- Moreover, multiple possible cell states can exist within a population of cells, leading to the multimodality of gene expression in cells (Vallejos et al., 2016).

Differential expression analysis software

	Pseudobulk methods			
	Pseudobulk methods that require built-in normalization		**Pseudobulk methods that can be used with any normalization**	
Method name	DESeq2	edgeR	Limma	ROTS
Normalization	Median of ratios	TMM	TMM+voom	TMM+CPM+log2
Statistical tests	Negative binomial generalized linear model	Negative binomial model + empirical Bayes procedure	Linear model + empirical Bayes procedure	Reproducibility optimized test statistic
R packages (normalization, test)	DESeq2	edgeR	edgeR, Limma	edgeR, ROTS
Filtering	Nonexpressed genes	Nonexpressed genes	Nonexpressed genes	Nonexpressed genes

Junttila et.al. *Briefings in Bioinformatics, 2022*

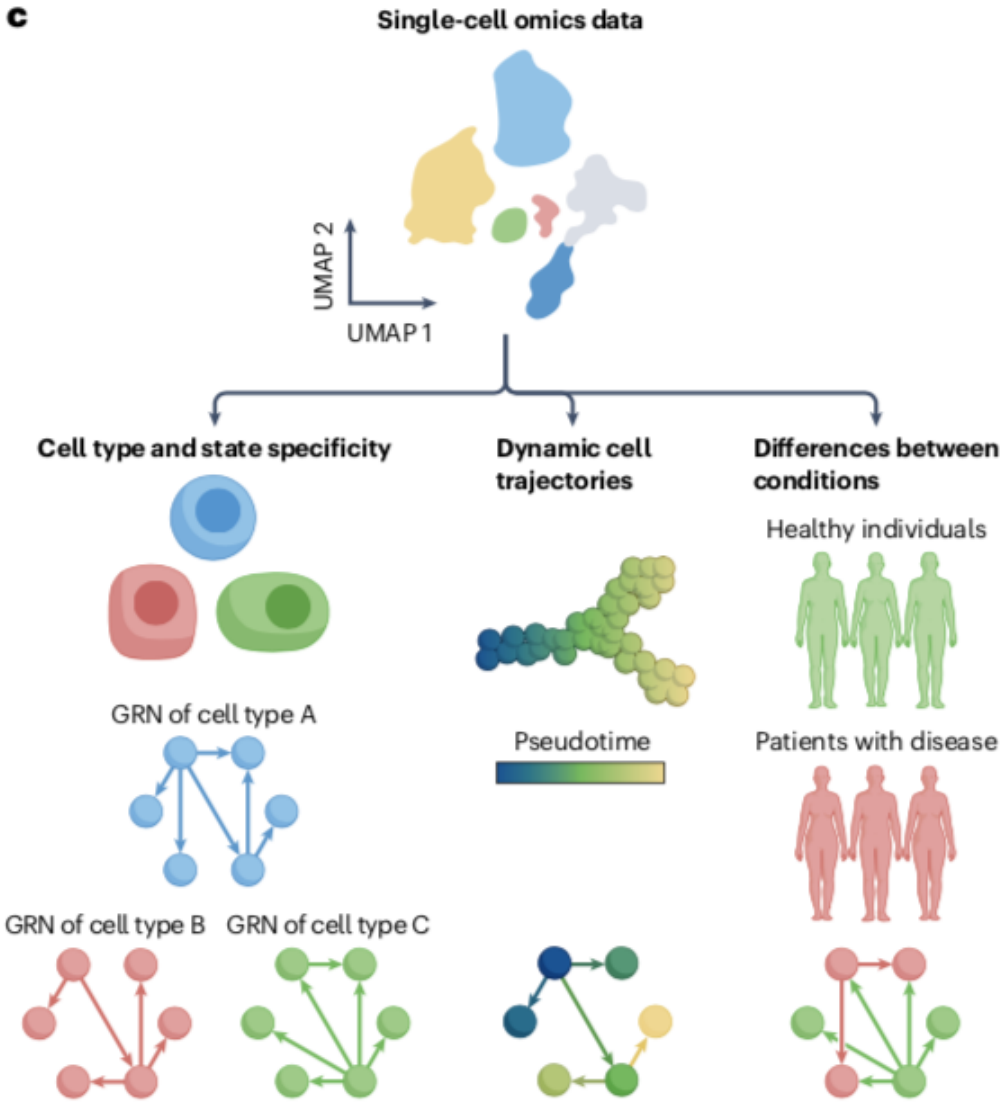
Differential expression analysis software

	Single-cell methods							
	Mixed models accounting for subjects as a random effect			Naïve methods that do not model subjects	Methods that have the option to use latent variables to correct for batches, etc.			
Method name	MAST_RE	muscat_MM	NEBULA-LN	wilcoxon	MAST	LR	negbinom	poisson
Normalization	No default (Log normalize)	Log normalize	Normalization factors from library sizes	Log normalize	Log normalize	Log normalize	Log normalize	Log normalize
Statistical test	Two-part hurdle model with random effect for subject	lme4 linear mixed model with voom weights	Negative binomial mixed model	Wilcoxon rank sum test	Two-part hurdle model	Logistic regression	Negative binomial generalized linear model	Poisson generalized linear model
R packages (normalization test)	MAST	muscat	nebula	Seurat	Seurat, MAST	Seurat	Seurat	Seurat
Filtering	Number cells expressing gene < subjects	Number cells expressing gene <20, Number cells in sample<10	Genes with counts per cell <0.005	Nonexpressed genes	Nonexpressed genes	Nonexpressed genes	Number cells expressing genes <3	Number cells expressing genes <3

Junttila et.al. *Briefings in Bioinformatics, 2022*

Infer gene regulatory network (GRN)

- GRNs models of the regulation of gene expression in the form of networks
- Uncovering the topology and the dynamics of GRNs has important implications for engineering cell fate and for disease prevention
- Gene regulatory networks generated from single-cell omics data allow us to understand
  - cell type and state specificity
  - explain the progression of dynamic trajectories
  - identify differences between condition



Mompel et.al. *Nature Reviews genetics. 2023*

Methods for regulatory network construction

Method	Code	Year
Boolean model		
Boolean Pseudotir	Python	2017
BTR	R	2016
SCNS	F#	2018
Differential equation		
Inference Snapsh	C++/Matlab	2015
SCODE	R/Julia/Ruby	2017
SCOUP	C++	2016
Gene correlation		
Empirical Bayes	Julia	2018
Information	Julia	2017
Measures		
NLNET	R	2016
SINCERA	R	2015
SCENIC	R/Python	2017
Correlation ensemble		
LEAP	R	2016
SINCERITIES	R/Matlab	2017
SCIMITAR	Python	2017
SCINGE	Matlab	2019

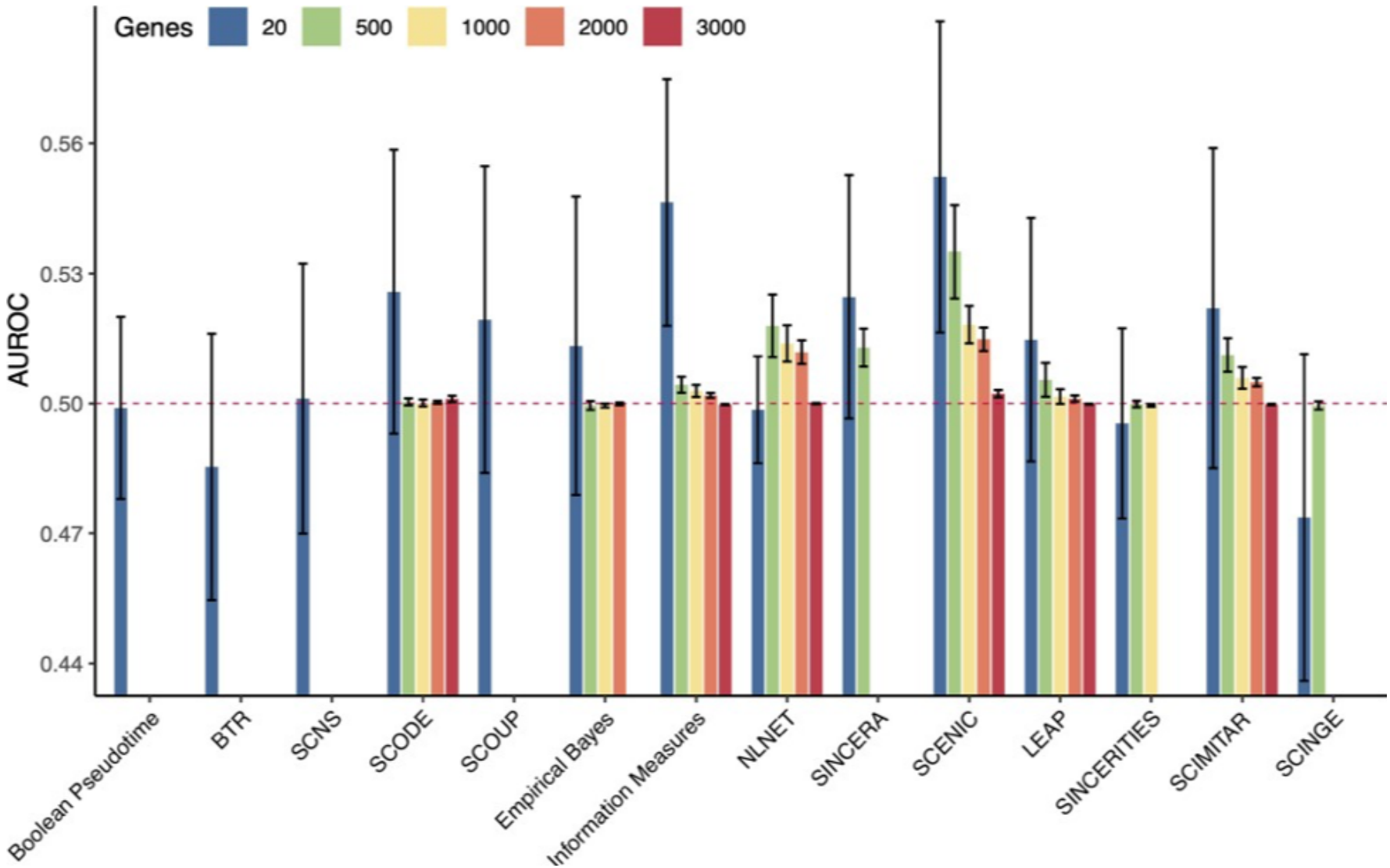
Nguyen et.al.. *Briefings in Bioinformatics, 2021*

Compare GRN methods

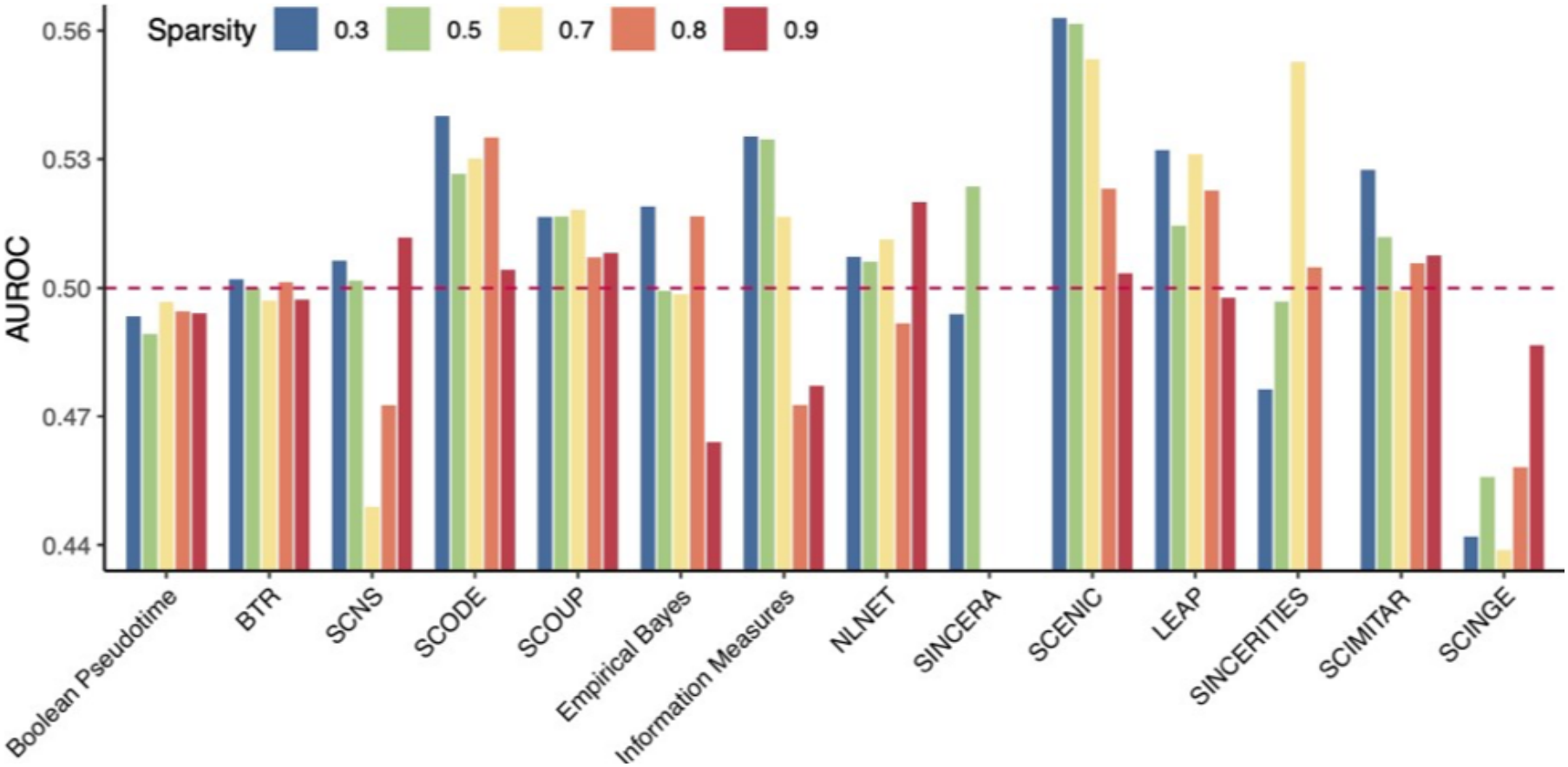
- Peformance with 139 simulated generate scRNA-seq datasets using GeneNetWeaver software based on the curated human network
- Compare the methods based on three essential metrics:
  - accuracy in reconstructing reference networks using scRNA-seq data
  - sensitivity to dropout rate and sparsity
  - time complexity

Nguyen et.al.. *Briefings in Bioinformatics, 2021*

Performance using 100 simulated datasets with 200 samples and varying number of genes

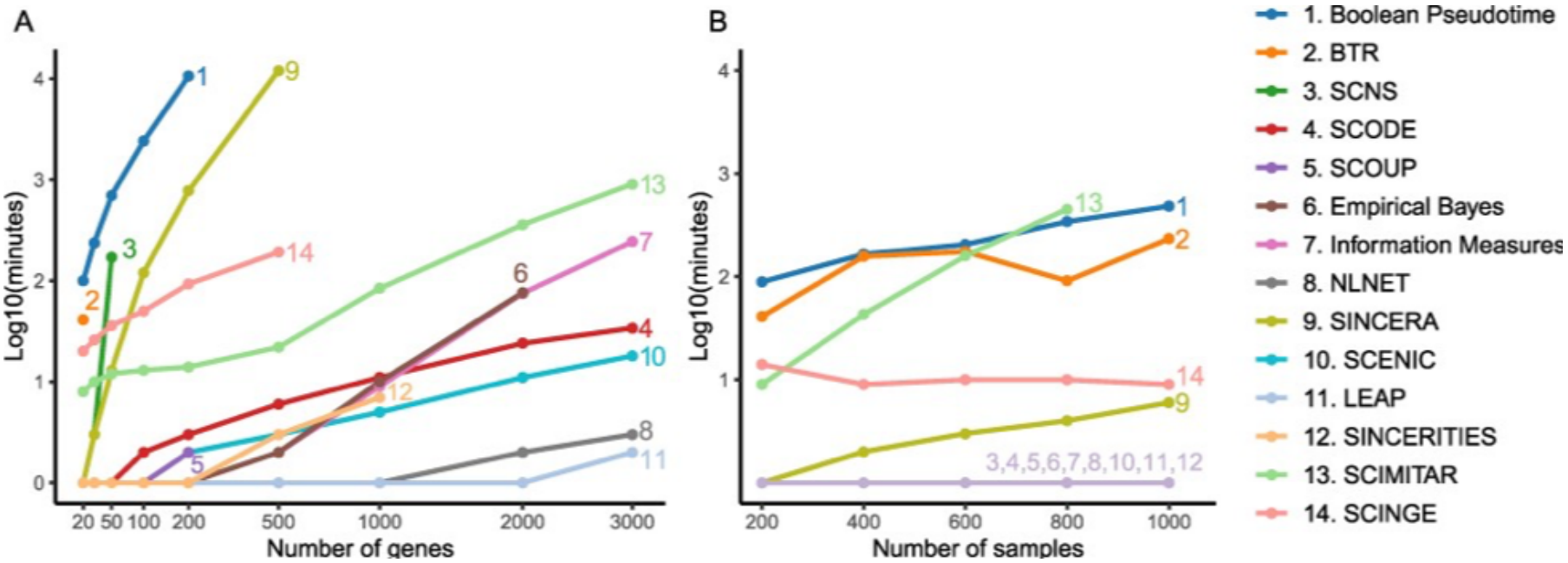


Performance with different levels of sparsity using 25 simulated datasets



Nguyen et.al.. *Briefings in Bioinformatics, 2021*

Running time with varying numbers of genes and samples



Nguyen et.al.. *Briefings in Bioinformatics, 2021*

Cell-cell communication

- Cell-cell interactions are essential to various biological processes
- In the multicellular interacting network, cells can interact and influence each other's behavior through specific signaling molecules, including ligands, receptors, metabolites, ions, and structural or secreted proteins
- Understanding how the cells interact with each other will help to reveal the potential mechanisms behind biological processes, such as organ development and tumor progression

Computation methods to infer cell-cell communication

Table: Statistical-based Tools

Tool	**Subunit**	**Prior Knowledge**	**Language**
CellCall	Single subunit	Ligand-receptor pairs; downstream TF regulation	R
CellChat	Multi-subunit	Ligand-receptor pairs; signaling cofactors and pathways	R
CellPhoneDB	Multi-subunit	Ligand-receptor pairs	Python
ICELLNET	Multi-subunit	Ligand-receptor pairs	R
iTALK	Single subunit	Ligand-receptor pairs	R
SingleCellSignalR	Single subunit	Ligand-receptor pairs	R

Liu et.al. *Genome Biology 2022*

Computation methods to infer cell-cell communication

Table: Network-based Tools

Tool	**Subunit**	**Prior Knowledge**	**Language**
Connectome	Single subunit	Ligand-receptor pairs	R
CytoTalk	Single subunit	Ligand-receptor pairs	R
Domino	Multi-subunit	Ligand-receptor pairs; TF regulation	R
NATMI	Single subunit	Ligand-receptor pairs; ligand-target pairs; receptor-TF pairs	R
NicheNet	Single subunit	Ligand-receptor pairs; ligand-target pairs; receptor-target pairs	R
scMLnet	Single subunit	Ligand-receptor pairs; receptor-TF pairs; TF-target pairs	Python

Liu et.al. *Genome Biology 2022*

Computation methods to infer cell-cell communication

Table: ST-based Tools

Tool	**Subunit**	**Prior Knowledge**	**Language**
CellPhoneDB v3	Multi-subunit	Ligand-receptor pairs; spatial microenvironment	Python
Giotto	Single subunit	Ligand-receptor pairs; cell type colocalization; L-R co-expression	R
stLearn	Single subunit	Ligand-receptor pairs; cell type colocalization; L-R co-expression	Python

Liu et.al. *Genome Biology 2022*

Preformance with 15 simulated and 5 real scRNA-seq and ST datasets

- Evaluate the performance of cell-cell interaction methods by integrating scRNA-seq data with spatial information from spatial transcriptomics (ST) data
  - ST profiles the relative position of different cells.
- CellChat has the best performance in consistencies with spatial information.
- CellChat and CellPhoneDB will generate high-confidence results with scaled computational resources
- SingleCellSignalR showed a good performance but consumes much time and memory.
- CellChat, CellPhoneDB, NicheNet, and ICELLNET show overall better performance than other tools in terms of consistency with spatial tendency and software scalability.

Liu et.al. *Genome Biology 2022*

Summary

- Many methods have been developed for scRNA-seq data analysi. In general, each individual method has its advantage and drawback.
- The selection of the appropriate approaches for analysis depends on characteristics of data and research purpose
- Single-cell sequencing technology has been developed to measure nearly all OMICS,
- One major disadvantage of single cell RNA sequencing is the loss of histological information due to dissociation of tissue samples
- Spatial transcriptomic methods bypass tissue dissociation and retain the spatial information, allowing gene expression assessment across thousands of cells within the context of tissue.
- The continuous development of the technology will broaden its applications in clinical and personalized medicine.