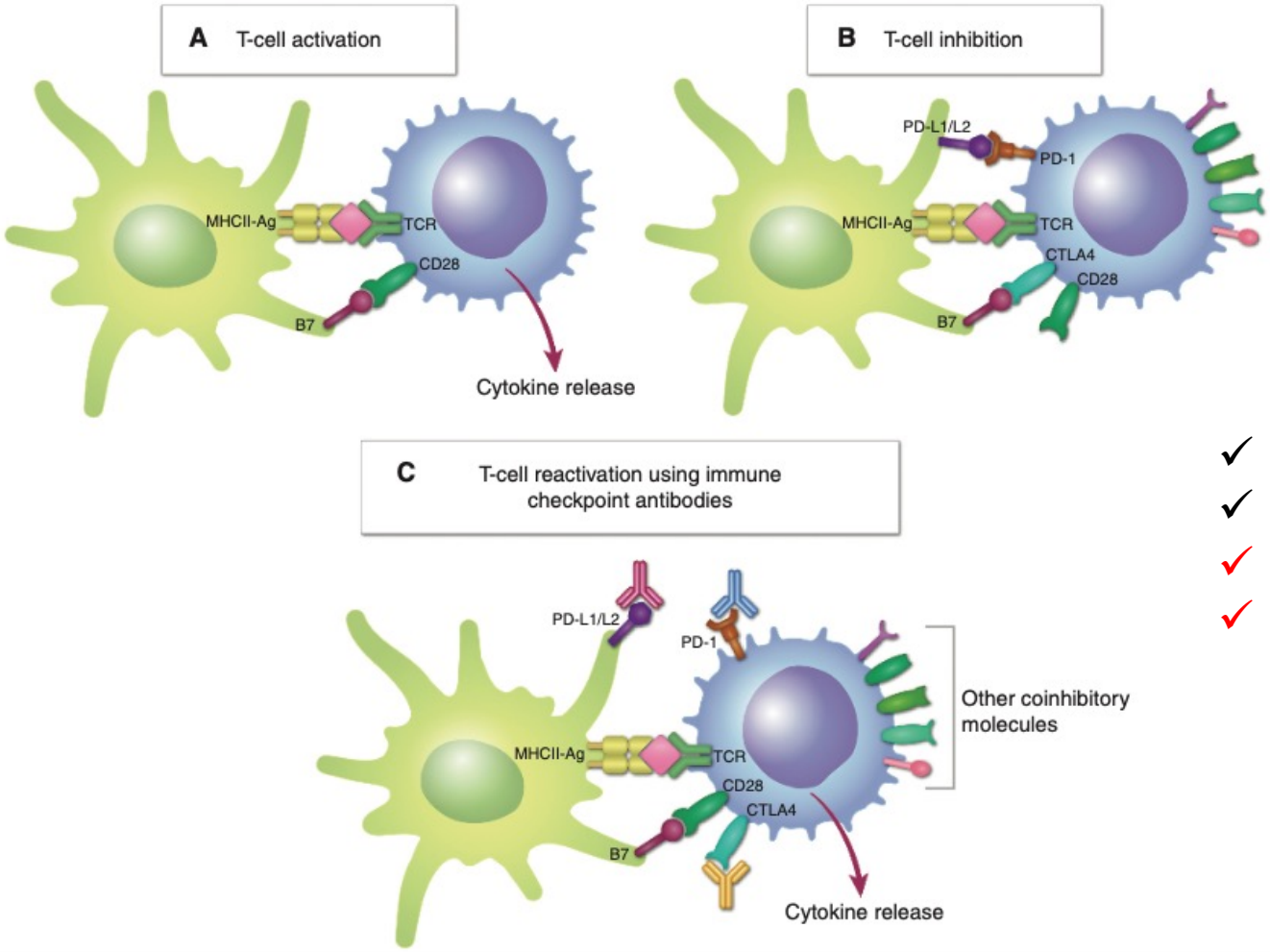
How to Handle Big Data and High Throughput Data

Single Cell Sequencing Data Analysis and Data Handling Tools

Dinesh Kumar, Ph.D.
Director, Informatics
Parker Institute for Cancer Immunotherapy
SITC Winter School 2022

PARKER INSTITUTE FOR CANCER IMMUNOTHERAPY

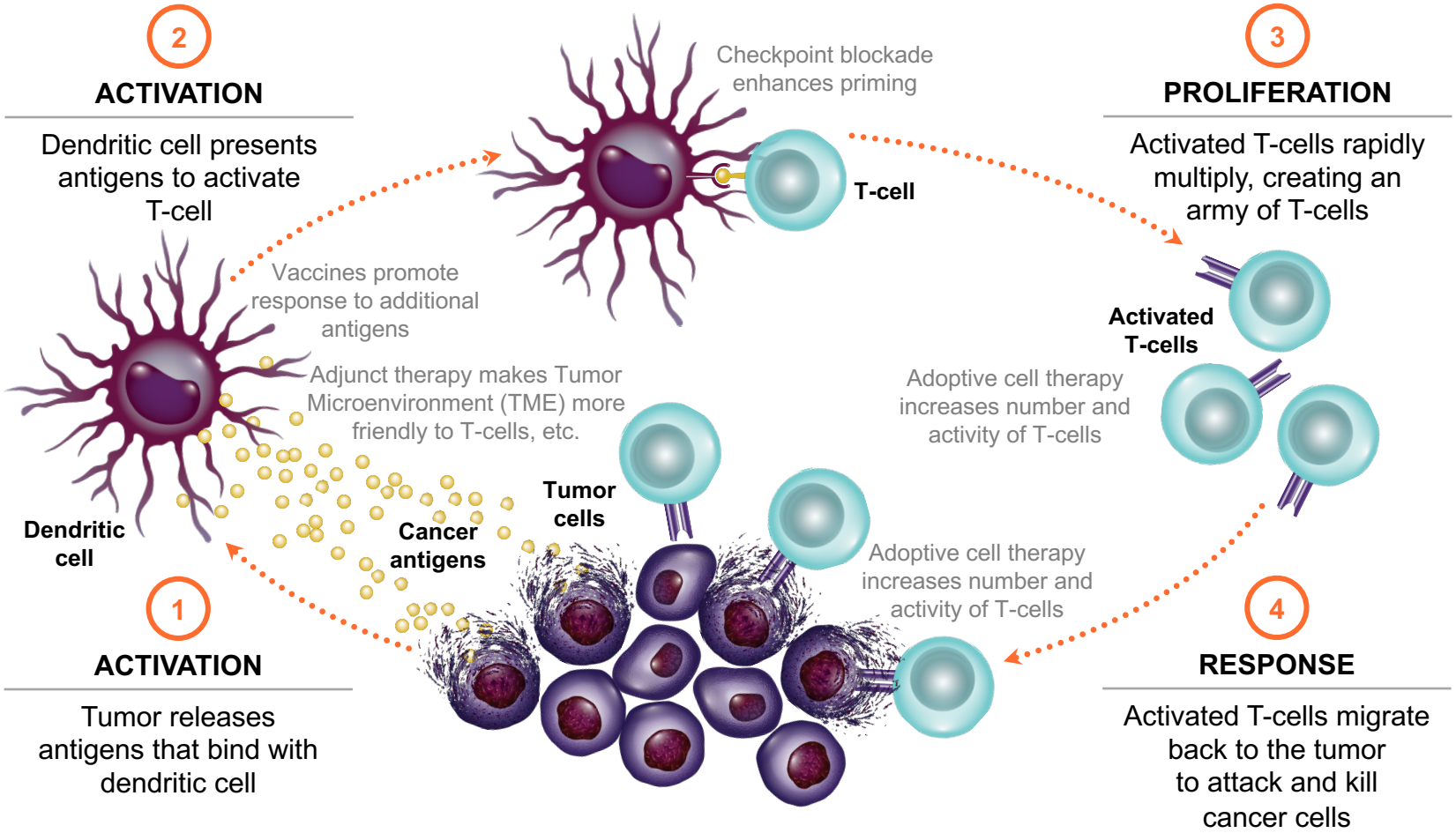
Central Dogma in ICT



- ✓ ICT can provide durable clinical responses
- ✓ Improve overall survival
- ✓ Works only in few patients
- ✓ Patient develop resistance

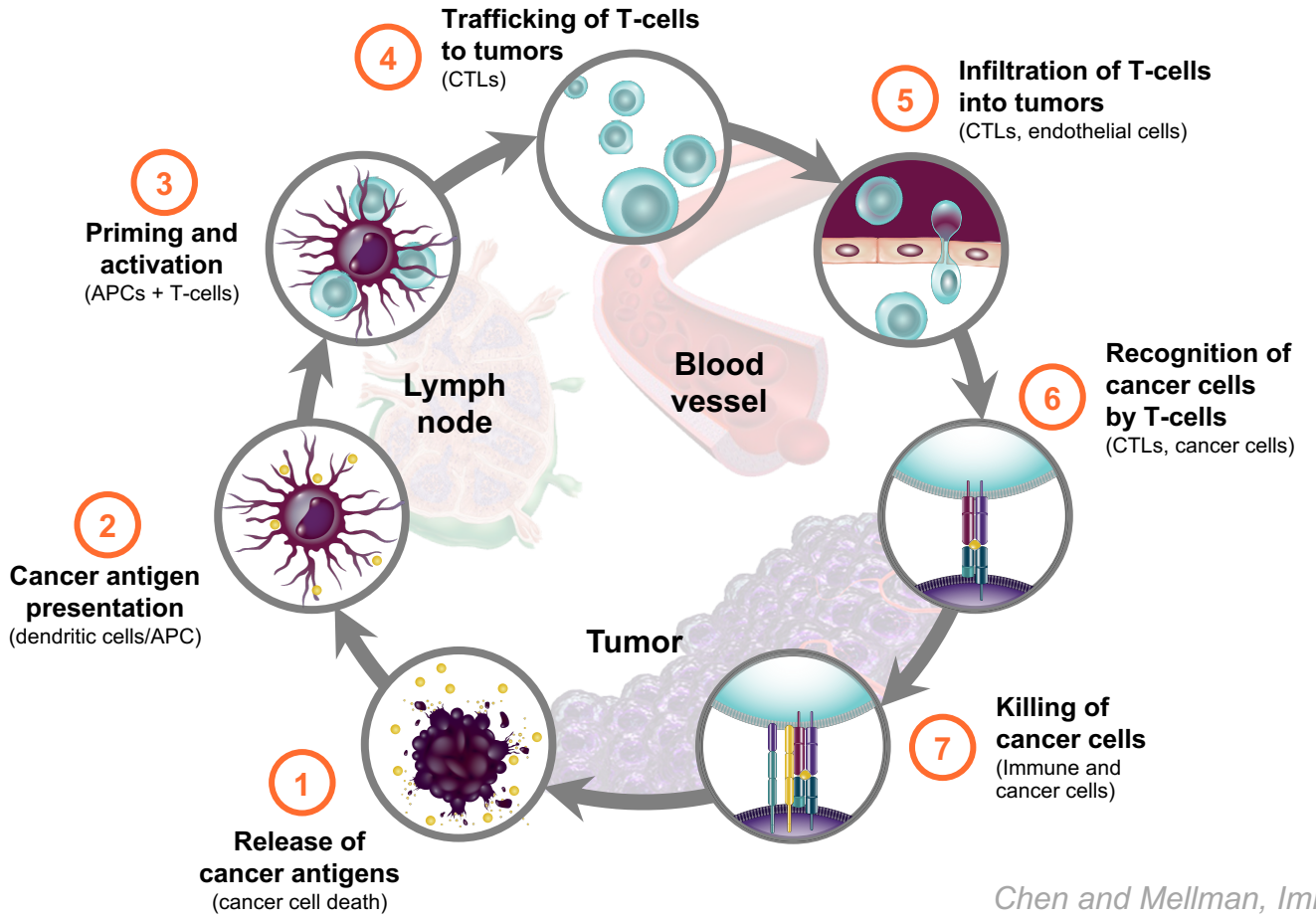
Sharma et. al., Cancer Discov. (Special Anniversary Issue) 2021 Apr;11(4):838-857

T-cell Activation



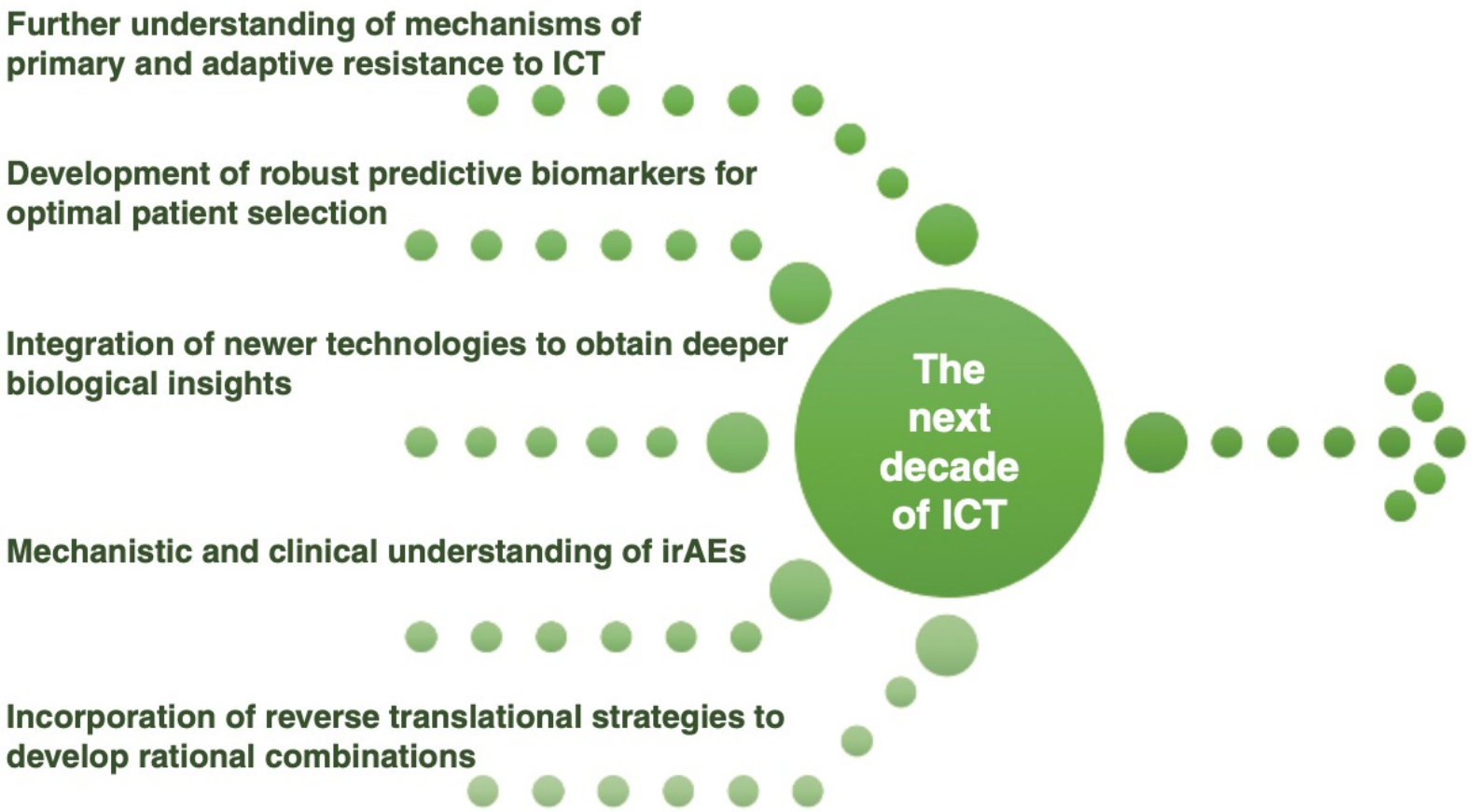
Cancer Immunity Cycle

A complex set of tumor, host and environmental factors govern strength and timing of anti-cancer immune responses.



Chen and Mellman, Immunity 2013

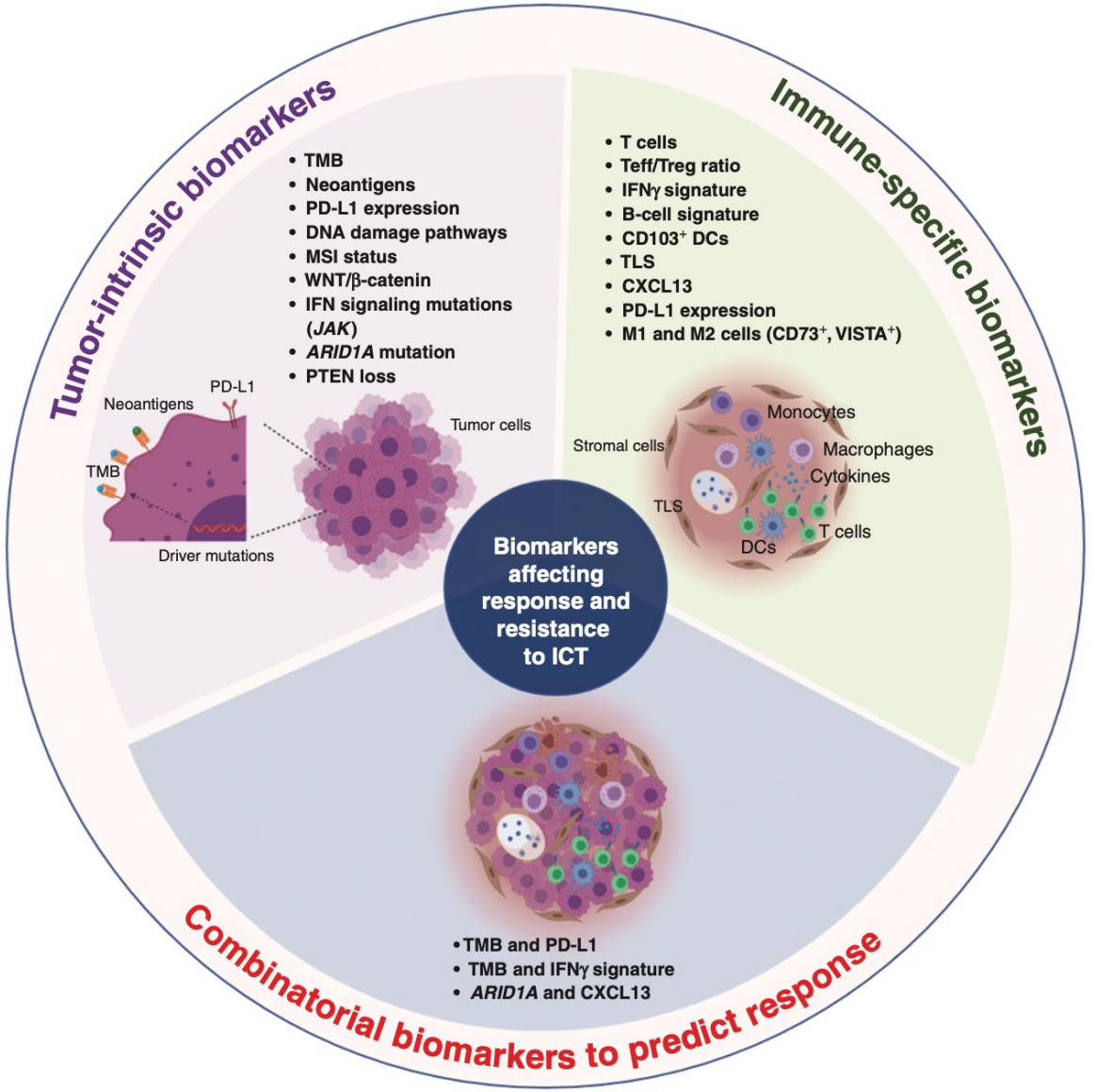
Focus Areas in the Coming Decade in the Field of ICT



Sharma et. al., Cancer Discov. (Special Anniversary Issue) 2021 Apr;11(4):838-857

REVIEW

Biomarkers of Response and Resistance to ICT



Sharma et. al., Cancer Discov. (Special Anniversary Issue) 2021 Apr;11(4):838-857. doi: 10.1158/2159-8290.CD-20-1680.

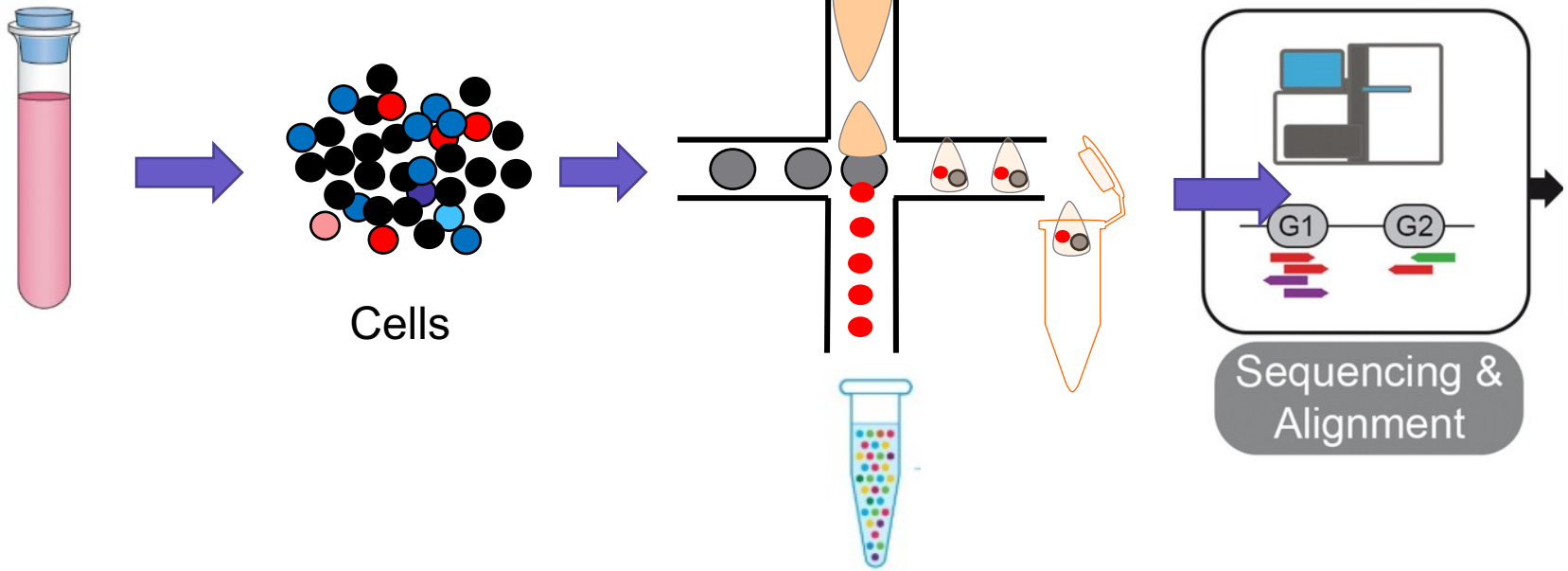
Introduction: bulk vs single cell

bulk analysis—is the most common way to start with for genomics analysis

Bulk

Single Cell

Introduction: Workflow Overview

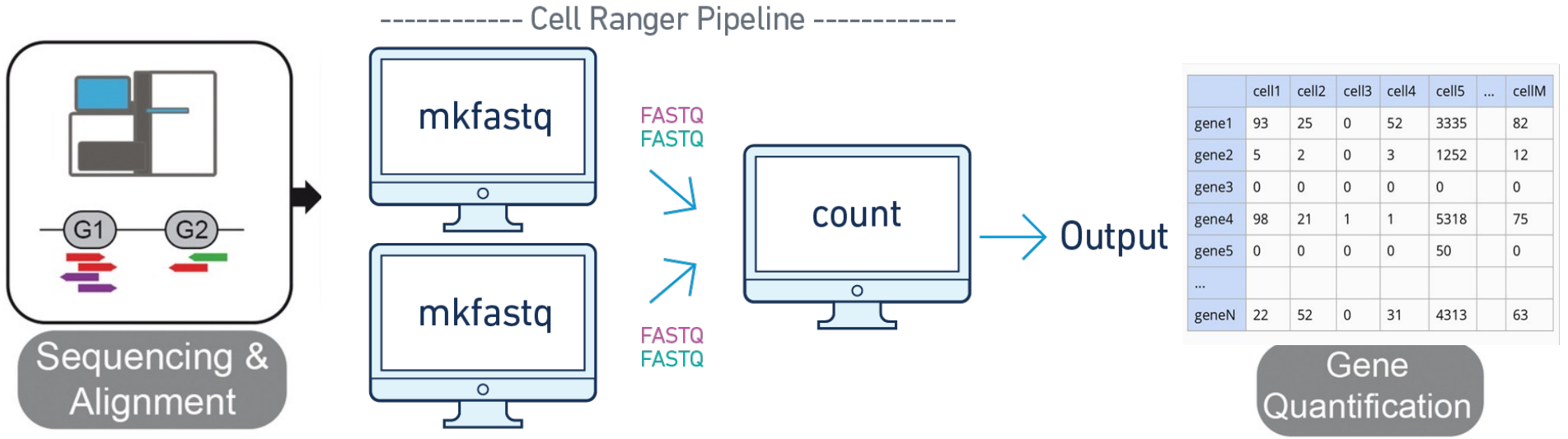


	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

(Gene x Cell) Count Matrix

Modified from Stewart et al. Nature reviews nephrology (2019)

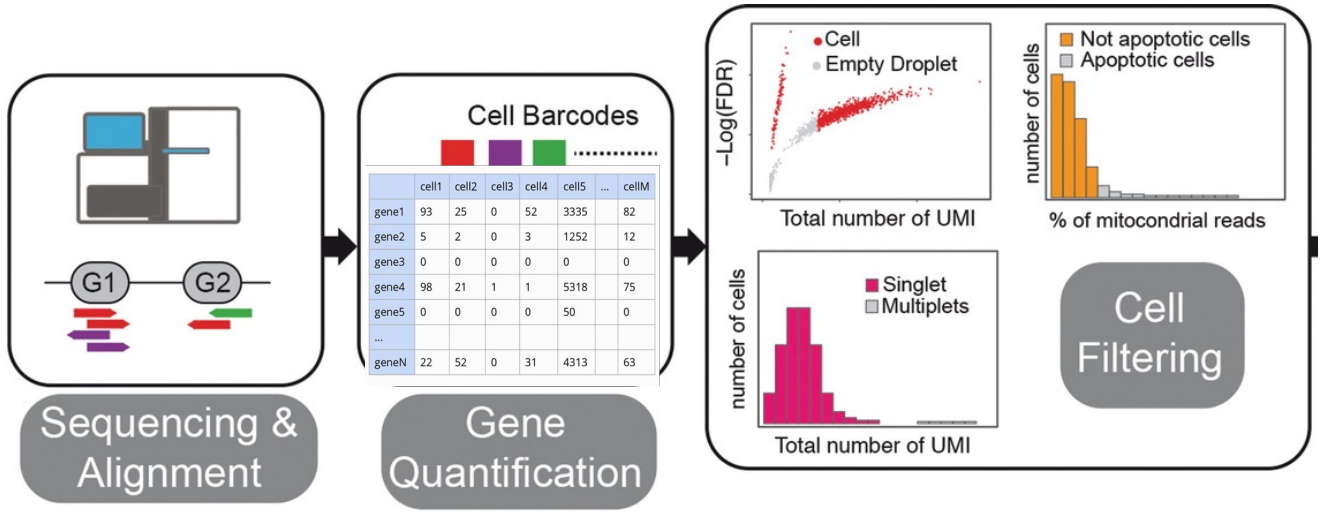
Introduction: Analysis overview



	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

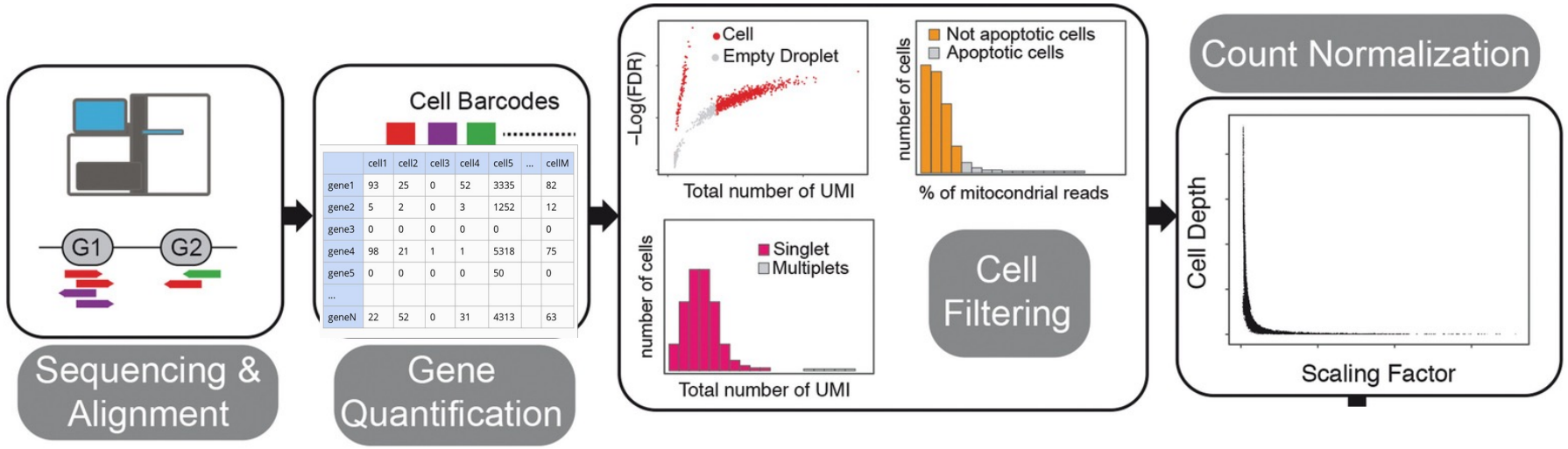
Slovin S. et al. (2021) Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview. In: Picardi E. (eds) RNA Bioinformatics. Methods in Molecular Biology, vol 2284. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-1307-8_19

Introduction: Analysis overview



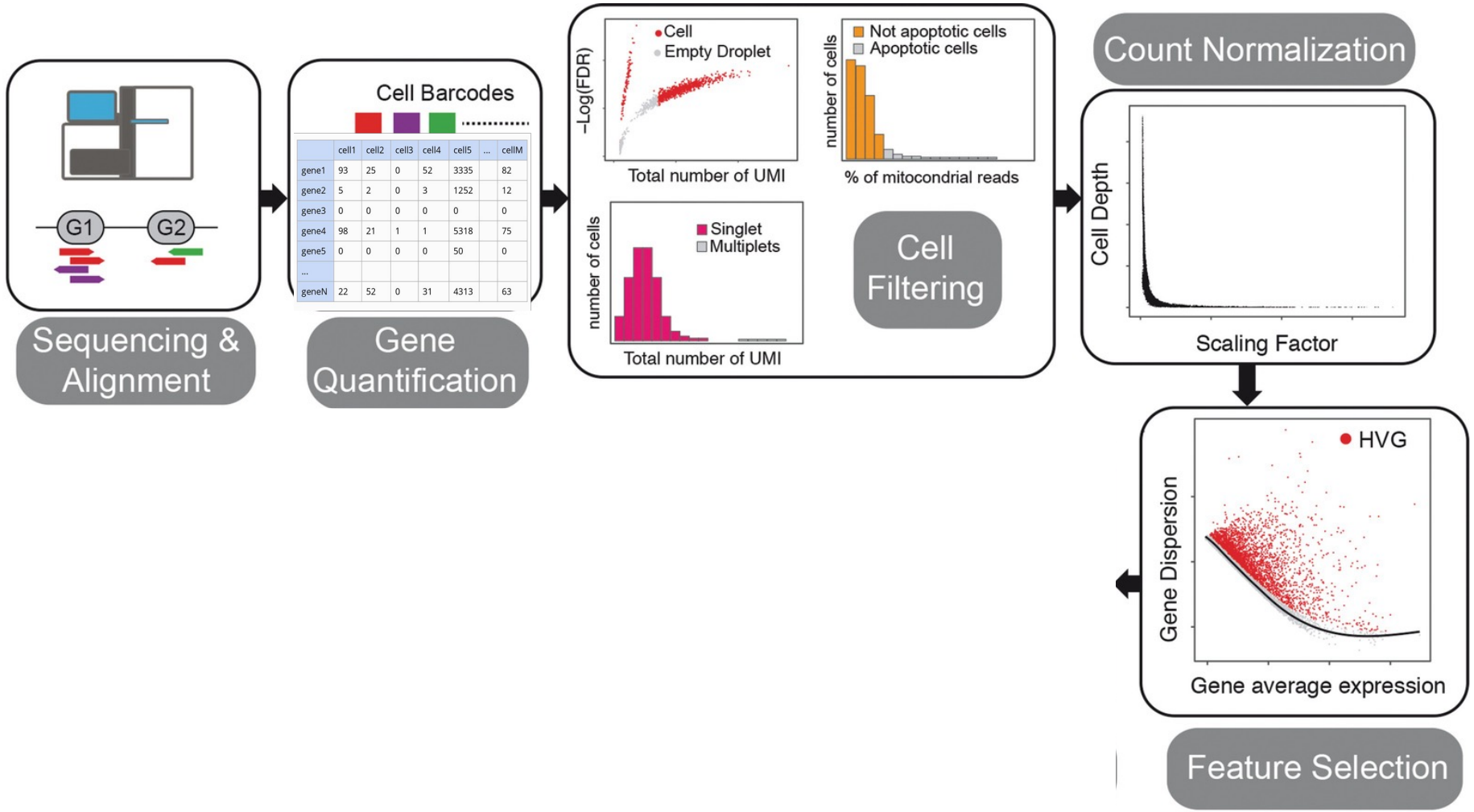
Slovin S. et al. (2021) Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview. In: Picardi E. (eds) RNA Bioinformatics. Methods in Molecular Biology, vol 2284. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-1307-8_19

Introduction: Analysis overview



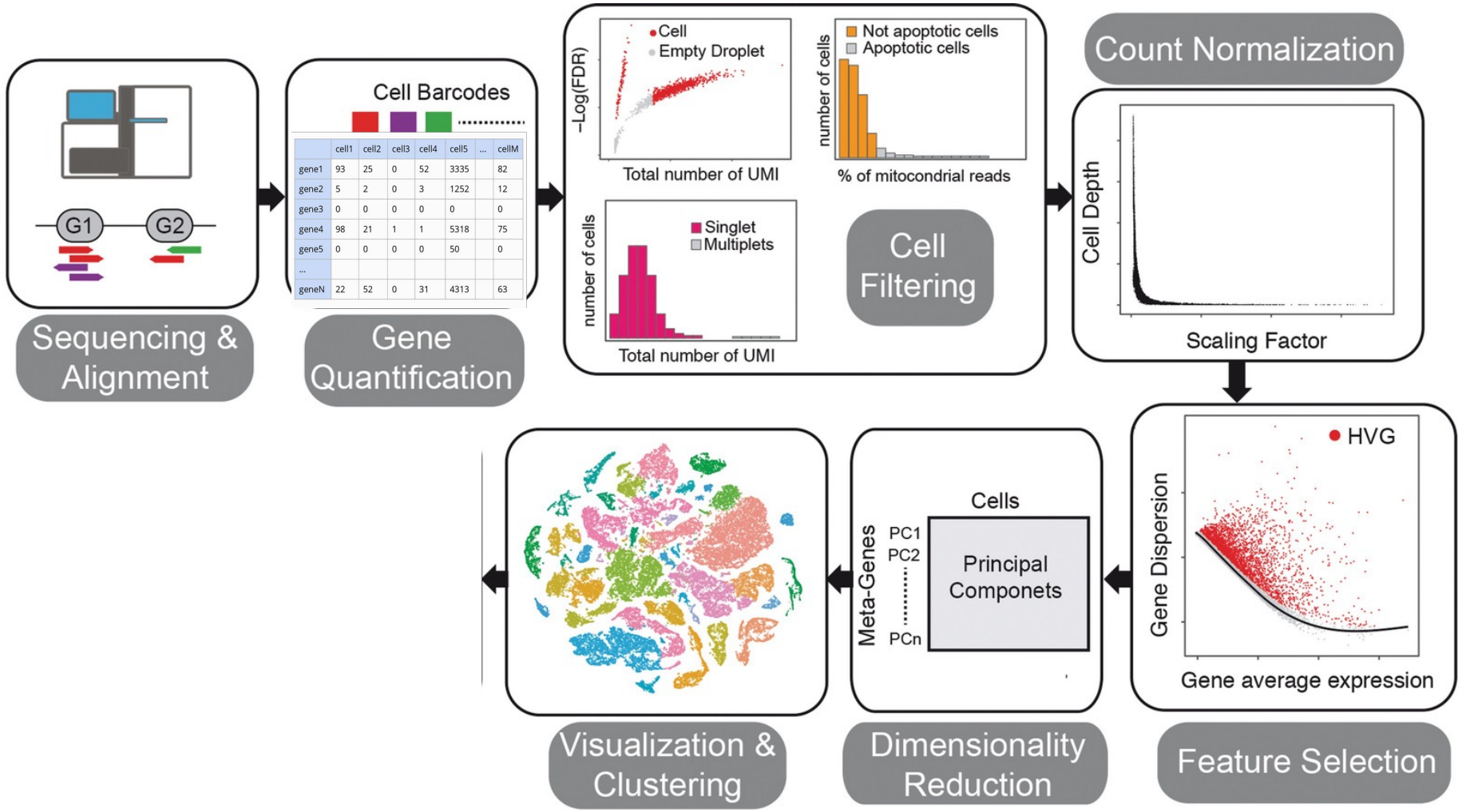
Slovin S. et al. (2021) Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview. In: Picardi E. (eds) RNA Bioinformatics. Methods in Molecular Biology, vol 2284. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-1307-8_19

Introduction: Analysis overview



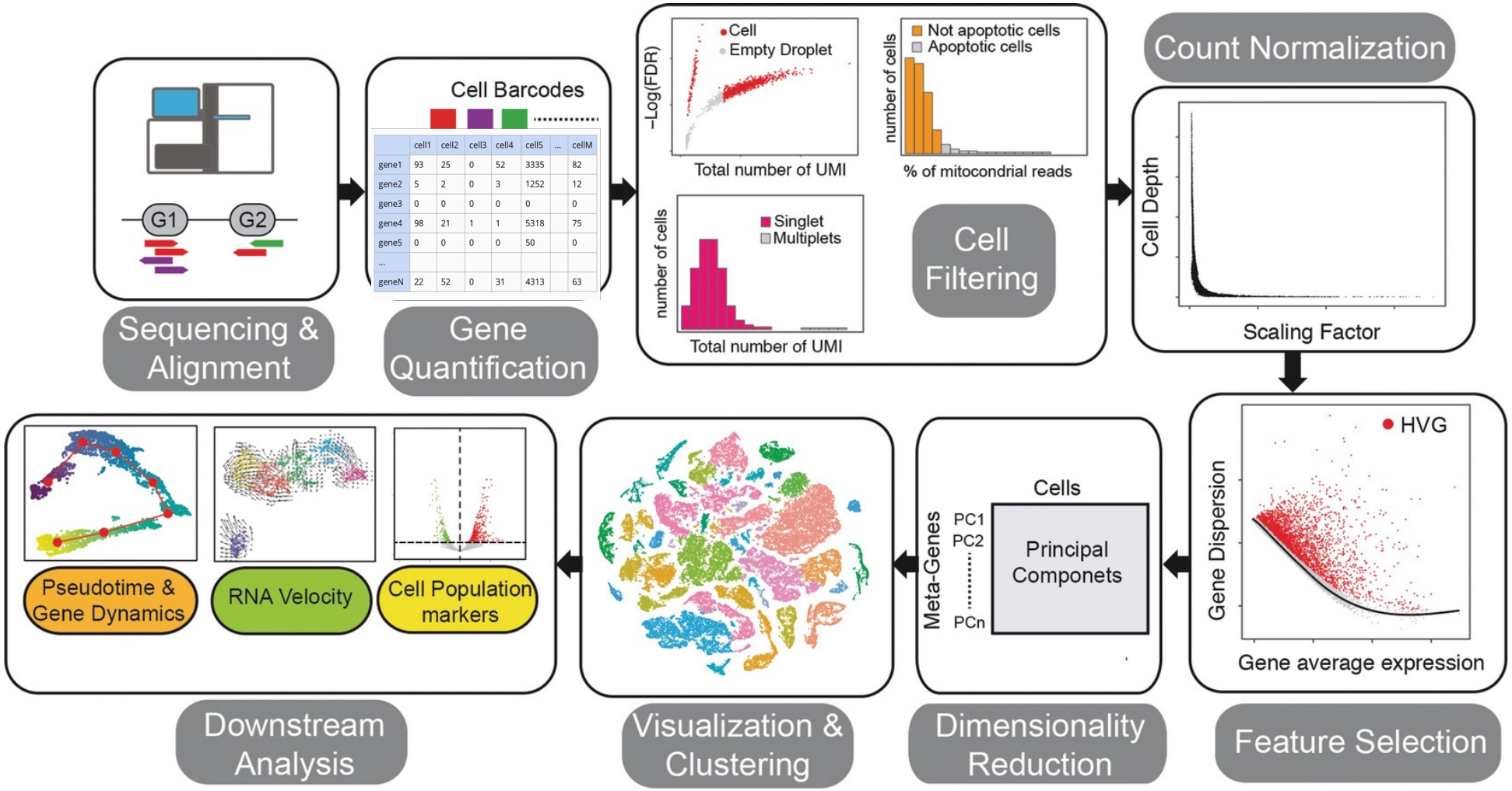
Slovin S. et al. (2021) Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview. In: Picardi E. (eds) RNA Bioinformatics. Methods in Molecular Biology, vol 2284. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-1307-8_19

Introduction: Analysis overview



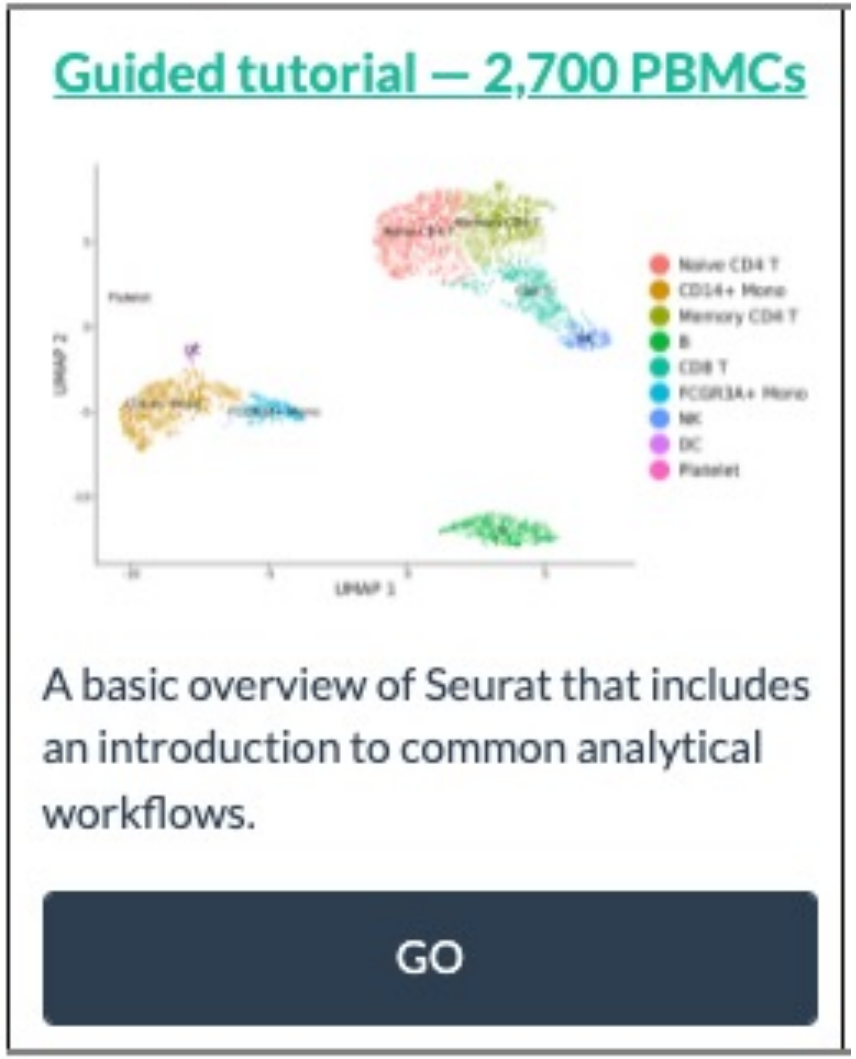
Slovin S. et al. (2021) Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview. In: Picardi E. (eds) RNA Bioinformatics. Methods in Molecular Biology, vol 2284. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-1307-8_19

Introduction: Analysis overview

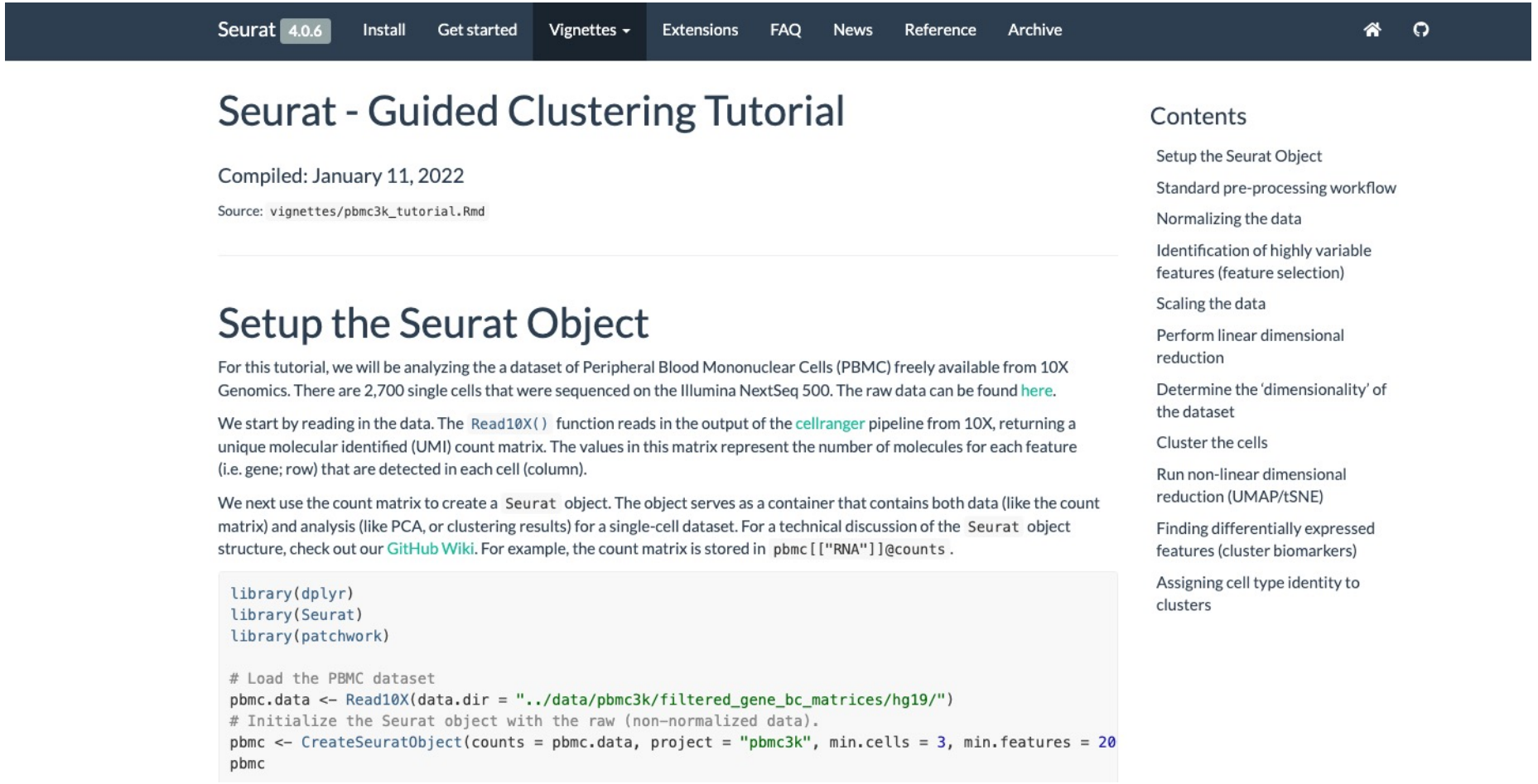


Slovin S. et al. (2021) Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview. In: Picardi E. (eds) RNA Bioinformatics. Methods in Molecular Biology, vol 2284. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-1307-8_19

Seurat: Great Place to Start for Single Cell Analysis



Seurat: Great Place to Start for Single Cell Analysis



Seurat 4.0.6 | Install | Get started | Vignettes | Extensions | FAQ | News | Reference | Archive

Seurat - Guided Clustering Tutorial

Compiled: January 11, 2022

Source: `vignettes/pbmc3k_tutorial.Rmd`

Setup the Seurat Object

For this tutorial, we will be analyzing the a dataset of Peripheral Blood Mononuclear Cells (PBMC) freely available from 10X Genomics. There are 2,700 single cells that were sequenced on the Illumina NextSeq 500. The raw data can be found here.

We start by reading in the data. The `Read10X()` function reads in the output of the cellranger pipeline from 10X, returning a unique molecular identified (UMI) count matrix. The values in this matrix represent the number of molecules for each feature (i.e. gene; row) that are detected in each cell (column).

We next use the count matrix to create a `Seurat` object. The object serves as a container that contains both data (like the count matrix) and analysis (like PCA, or clustering results) for a single-cell dataset. For a technical discussion of the `Seurat` object structure, check out our GitHub Wiki. For example, the count matrix is stored in `pbmc[["RNA"]]@counts`.

```
library(dplyr)
library(Seurat)
library(patchwork)

# Load the PBMC dataset
pbmc.data <- Read10X(data.dir = "../data/pbmc3k/filtered_gene_bc_matrices/hg19/")
# Initialize the Seurat object with the raw (non-normalized data).
pbmc <- CreateSeuratObject(counts = pbmc.data, project = "pbmc3k", min.cells = 3, min.features = 20
pbmc
```

Contents

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

Single Cell Multi-Omics: The Data Challenges

=

	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

Transcriptome

+

	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

Methylome

+

	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

Proteome

+

	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

Spatial expression

+

	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

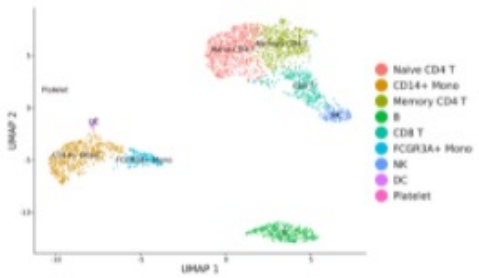
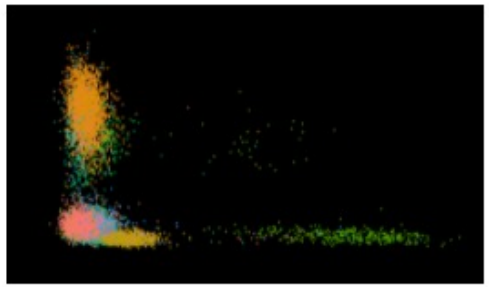
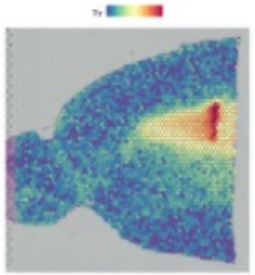
Chromatin (ATAC)

Single Cell Analysis: Seurat is a Great Place to Start

Introductory Vignettes

For new users of Seurat, we suggest starting with a guided walk through of a dataset of 2,700 Peripheral Blood Mononuclear Cells (PBMCs) made publicly available by 10X Genomics. This tutorial implements the major components of a standard unsupervised clustering workflow including QC and data filtration, calculation of high-variance genes, dimensional reduction, graph-based clustering, and the identification of cluster markers.

We provide additional introductory vignettes for users who are interested in analyzing multimodal single-cell datasets (e.g. from CITE-seq, or the 10x mulitome kit), or spatial datasets (e.g. from 10x visium or SLIDE-seq).

Guided tutorial — 2,700 PBMCs	Multimodal analysis	Analysis of spatial datasets
		
A basic overview of Seurat that includes an introduction to common analytical workflows.	An introduction to working with multi-modal datasets in Seurat.	Learn to explore spatially-resolved transcriptomic data with examples from 10x Visium and Slide-seq v2.
GO	GO	GO

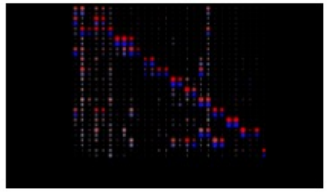
https://satijalab.org/seurat

Single Cell Analysis: Seurat is a Great Place to Start

Data Integration

Recently, we have developed computational methods for integrated analysis of single-cell datasets generated across different conditions, technologies, or species. As an example, we provide a guided walk through for integrating and comparing PBMC datasets generated under different stimulation conditions. We provide additional vignettes demonstrating how to leverage an annotated scRNA-seq reference to map and label cells from a query, and to efficiently integrate large datasets.

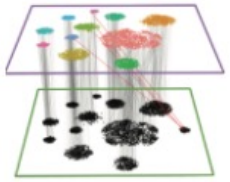
Introduction to scRNA-seq integration



An introduction to integrating scRNA-seq datasets in order to identify and compare shared cell types across experiments.

GO

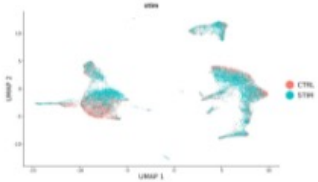
Mapping and annotating query datasets



Learn how to map a query scRNA-seq dataset onto a reference in order to automate the annotation and visualization of query cells.

GO

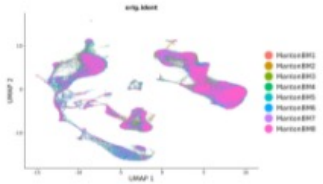
Fast integration using reciprocal PCA (RPCA)



Identify anchors using the reciprocal PCA (rPCA) workflow, which performs a faster and more conservative integration.

GO

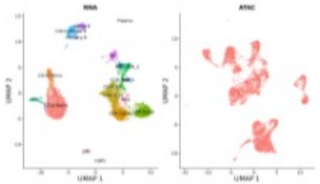
Tips for integrating large datasets



Tips and examples for integrating very large scRNA-seq datasets (including >200,000 cells).

GO

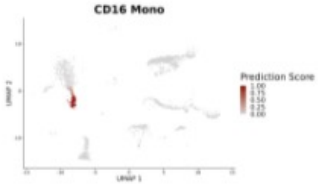
Integrating scRNA-seq and scATAC-seq data



Annotate, visualize, and interpret an scATAC-seq experiment using scRNA-seq data from the same biological system.

GO

Multimodal Reference Mapping



Analyze query data in the context of multimodal reference atlases.

GO

https://satijalab.org/seurat

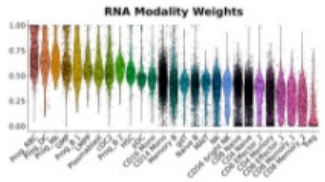
Single Cell Analysis: Seurat is a Great Place to Start

Additional New Methods

Seurat also offers additional novel statistical methods for analyzing single-cell data. These include:

- Weighted-nearest neighbor (WNN) analysis: to define cell state based on multiple modalities [paper]
- Mixscape: to analyze data from pooled single-cell CRISPR screens [paper]
- SCTransform: Improved normalization for single-cell RNA-seq data [paper]]
- SCTransform, v2 regularization [paper]]

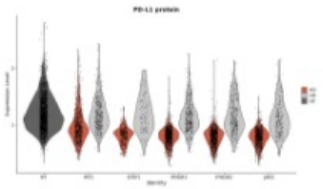
Weighted Nearest Neighbor Analysis



Analyze multimodal single-cell data with weighted nearest neighbor analysis in Seurat v4.

GO

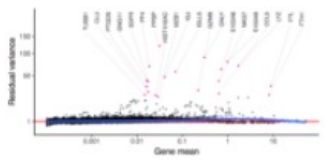
Mixscape



Explore new methods to analyze pooled single-celled perturbation screens.

GO

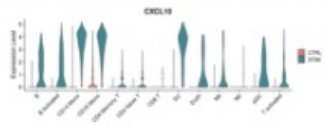
SCTransform



Examples of how to use the SCTransform wrapper in Seurat.

GO

SCTransform, v2 regularization



Examples of how to perform normalization, feature selection, integration, and differential expression with an updated version of sctransform.

GO

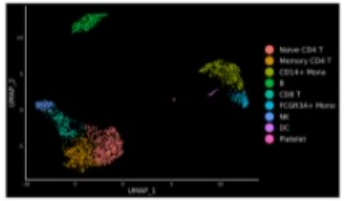
https://satijalab.org/seurat

Single Cell Analysis: Seurat is a Great Place to Start

Other

Here we provide a series of short vignettes to demonstrate a number of features that are commonly used in Seurat. We've focused the vignettes around questions that we frequently receive from users. Click on a vignette to get started.

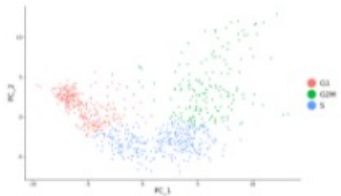
Visualization



An overview of the major visualization functionality within Seurat.

GO

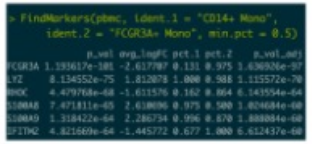
Cell Cycle Regression



Mitigate the effects of cell cycle heterogeneity by computing cell cycle phase scores based on marker genes.

GO

Differential Expression Testing



Perform differential expression (DE) testing in Seurat using a number of frameworks.

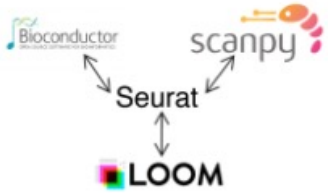
GO

Demultiplex Cell Hashing data

Learn how to work with data produced with Cell Hashing.

GO

Interoperability with Other Analysis Tools



Convert data between formats for different analysis tools.

GO

Parallelization

Speed up compute-intensive functions with parallelization.

GO

https://satijalab.org/seurat

Single Cell Analysis: Seurat is a Great Place to Start

SeuratWrappers

In order to facilitate the use of community tools with Seurat, we provide the Seurat Wrappers package, which contains code to run other analysis tools on Seurat objects. For the initial release, we provide wrappers for a few packages in the table below but would encourage other package developers interested in interfacing with Seurat to check out our contributor guide here.

Package	Vignette	Reference	Source
alevin	Import alevin counts into Seurat	Srivastava et. al., Genome Biology 2019	https://github.com/k3yavi/alevin-Rtools
ALRA	Zero-preserving imputation with ALRA	Linderman et al, bioRxiv 2018	https://github.com/KlugerLab/ALRA
CoGAPS	Running CoGAPS on Seurat Objects	Stein-O'Brien et al, Cell Systems 2019	https://www.bioconductor.org/packages/release/bioc/html/CoGAPS.htm
Conos	Integration of datasets using Conos	Barkas et al, Nature Methods 2019	https://github.com/hms-dbmi/conos
fastMNN	Running fastMNN on Seurat Objects	Haghverdi et al, Nature Biotechnology 2018	https://bioconductor.org/packages/release/bioc/html/scran.html
glmpca	Running GLM-PCA on a Seurat Object	Townes et al, Genome Biology 2019	https://github.com/willtownes/glmpca
Harmony	Integration of datasets using Harmony	Korsunsky et al, Nature Methods 2019	https://github.com/immunogenomics/harmony
LIGER	Integrating Seurat objects using LIGER	Welch et al, Cell 2019	https://github.com/MacoskoLab/liger
Monocle3	Calculating Trajectories with Monocle 3 and Seurat	Cao et al, Nature 2019	https://cole-trapnell-lab.github.io/monocle3
Nebulosa	Visualization of gene expression with Nebulosa	Jose Alquicira-Hernandez and Joseph E. Powell, Under Review	https://github.com/powellgenomicslab/Nebulosa
schex	Using schex with Seurat	Freytag, R package 2019	https://github.com/SaskiaFreytag/schex
scVelo	Estimating RNA Velocity using Seurat and scVelo	Bergen et al, bioRxiv 2019	https://scvelo.readthedocs.io/
Velocity	Estimating RNA Velocity using Seurat	La Manno et al, Nature 2018	https://velocyto.org
CIPR	Using CIPR with human PBMC data	Ekiz et. al., BMC Bioinformatics 2020	https://github.com/atakanekiz/CIPR-Package
miQC	Running miQC on Seurat objects	Hippen et. al., bioRxiv 2021	https://github.com/greenelab/miQC
tricycle	Running estimate_cycle_position from tricycle on Seurat Objects	Zheng et. al., bioRxiv 2021	https://www.bioconductor.org/packages/release/bioc/html/tricycle.html

https://satijalab.org/seurat

Single Cell Multi-Omics: The Data Challenges

=

	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

Transcriptome

+

	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

Methylome

+

	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

Proteome

+

	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

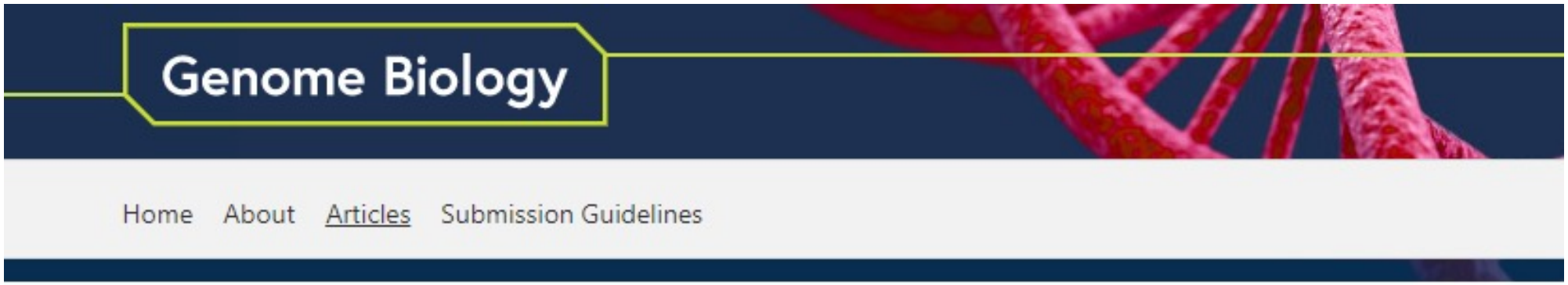
Spatial expression

+

	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

Chromatin (ATAC)

Challenges in Single Cell Data Analysis



Review | Open Access | Published: 07 February 2020

Eleven grand challenges in single-cell data science

David Lähnemann, Johannes Köster, [...] Alexander Schönhuth

Genome Biology **21**, Article number: 31 (2020) | Cite this article

Challenge I: Handling Sparsity in Single-Cell RNA Sequencing

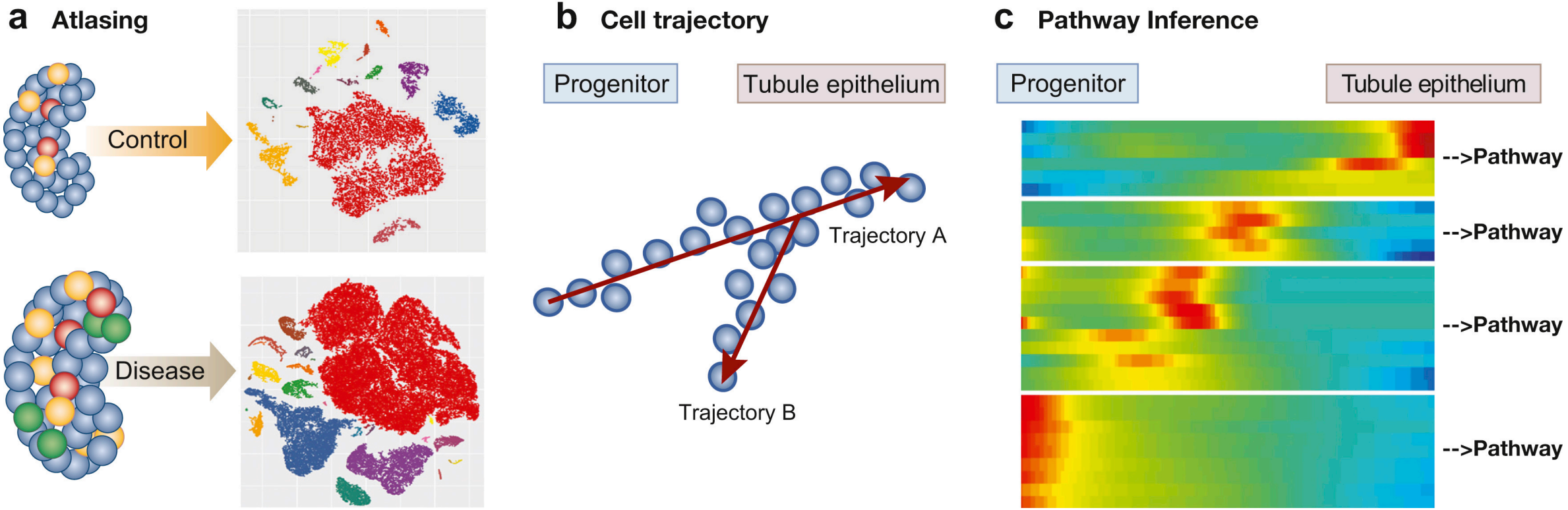
- scRNA-seq measurements typically suffer from large fractions of observed zeros, where a given gene in a given cell has no unique molecular identifiers or reads mapping to it.
- Sparsity pervades all aspects of scRNA-seq data analysis.
- The term “dropout” is often used to denote observed zero values in scRNA-seq data. But this term usually conflates two distinct types of zero values: those attributable to methodological noise, where a gene is expressed but not detected by the sequencing technology, and those attributable to biologically-true absence of expression.
- In general, two broad approaches can be applied to tackle this problem of sparsity: (i) use statistical models that inherently model the sparsity, sampling variation, and noise modes of scRNA-seq data with an appropriate data generative model (i.e., quantifying uncertainty or (ii) attempt to “impute” values for observed zeros (ideally the technical zeros; sometimes also non-zero values) that better approximate the true gene expression levels.

Challenge II: Defining Flexible Statistical Frameworks for Discovering Complex Differential Patterns in Gene Expression

- Beyond simple changes in average gene expression between cell types (or across bulk-collected libraries), scRNA-seq enables a high granularity of changes in expression to be unraveled.
- Most methods have focused on comparing average expression between groups
- The vast majority of differential expression detection methods assume that the groups of cells to be compared are known in advance
- While some methods exist to identify more general patterns of gene expression changes (e.g., variability, distributions), these methods could be further improved by integrating with existing approaches that account for confounding effects such as cell cycle and complex batch effects.

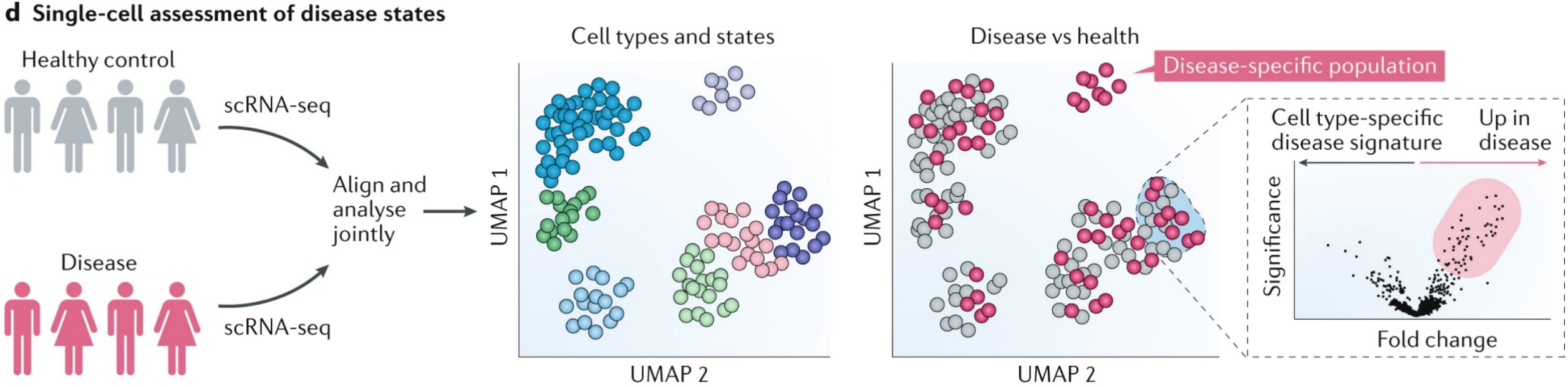
Genome Biology volume 21, Article number: 31 (2020)

Single Cell Application: Time Series in Disease and Tumors + Cell States

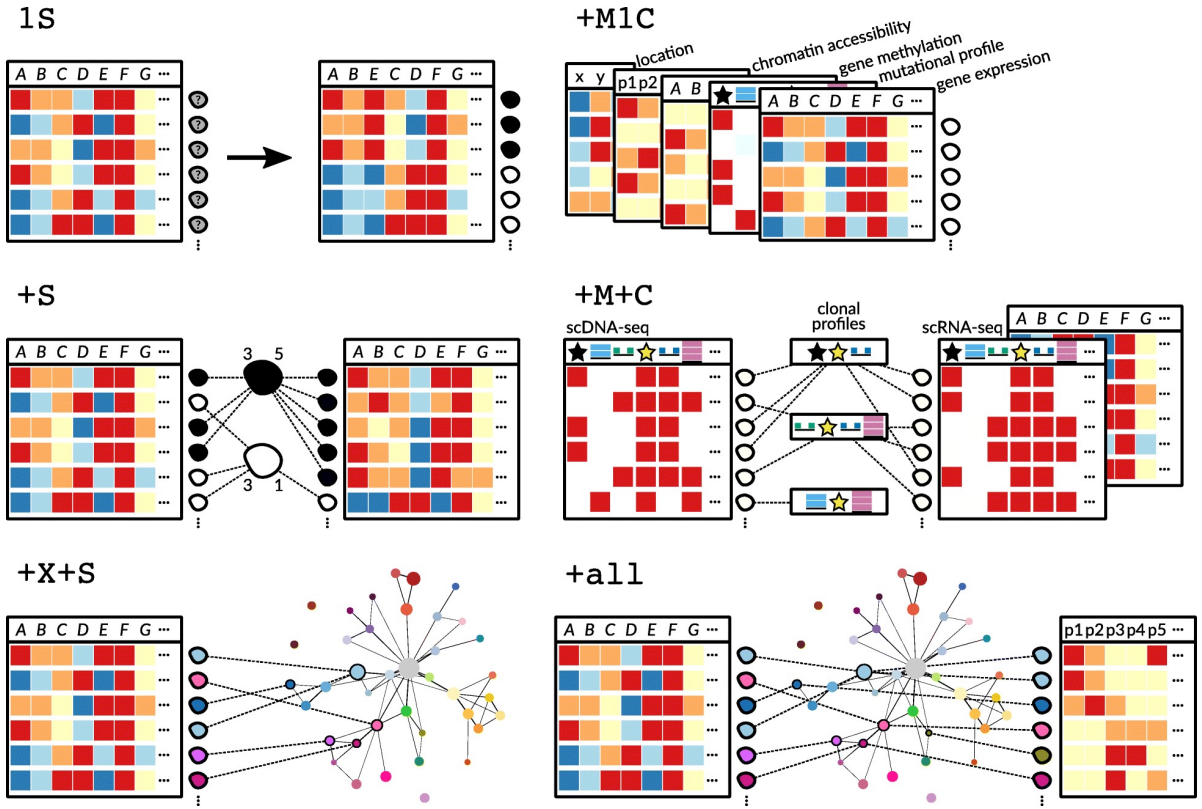


Kidney International Volume 96, Issue 4, October 2019, Pages 862-870
https://doi.org/10.1016/j.kint.2019.03.035

Disease Associated Cell Types



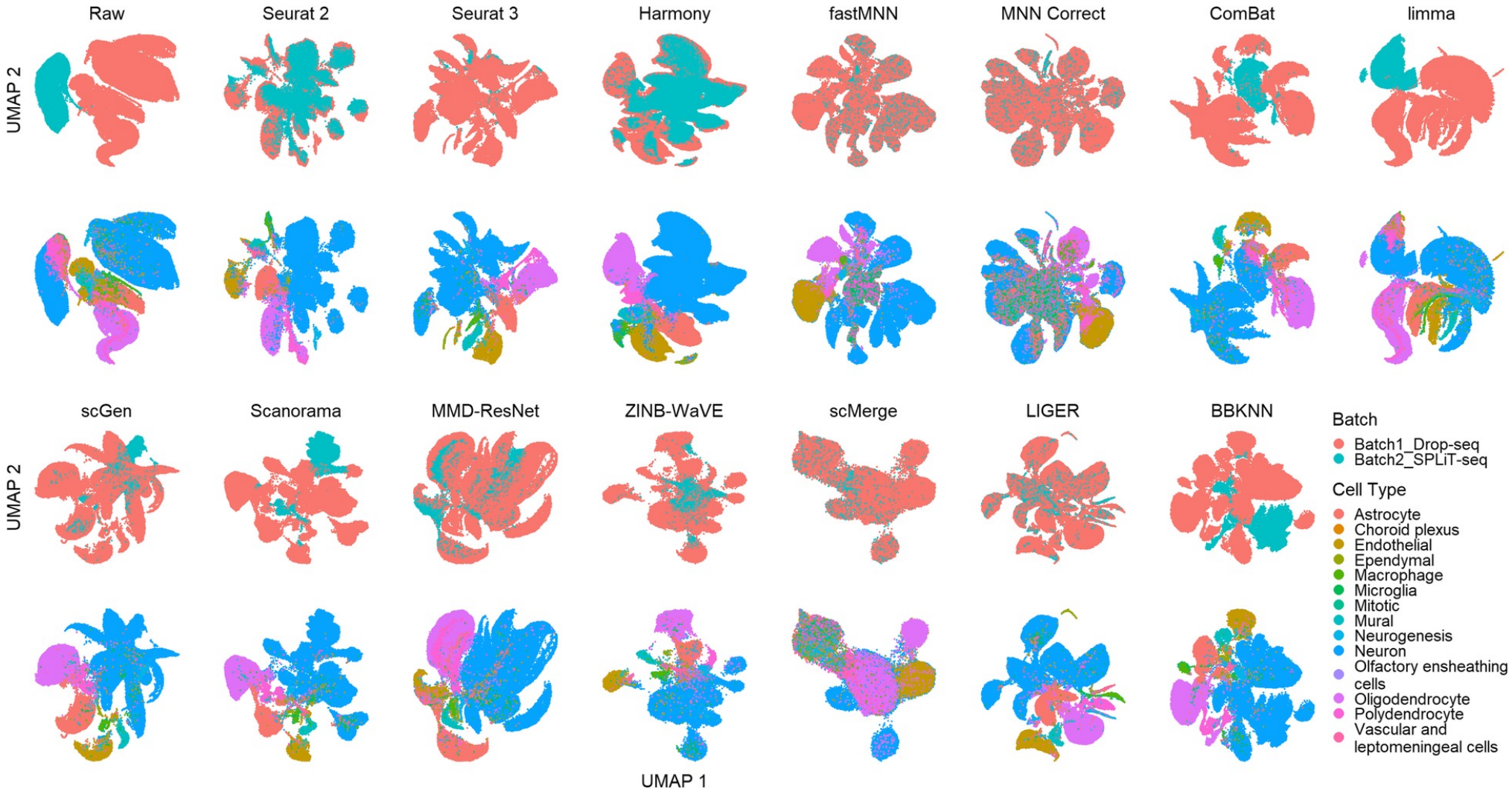
Stewart et al. Nature reviews nephrology (2019)



1.Abbreviations: "↓" same challenge also applies to all approaches below, *AM* analysis method, *exp(s)* experiment(s), *HCA* human cell atlas, *MT* measurement type, *smps* samples, *TCGA* The Cancer Genome Atlas

	Integration	Example MT combination	Example AMs	Promises	Challenges
1S	None	scDNA-seq	Clustering/unsupervised	Discover new subclones, cell types, or cell states	Technical noise ↓; data sparsity ↓
+S	Within 1 MT, within 1 exp, across >1 smps	scRNA-seq	Differential analyses, time series, spatial sampling	Identify effects across sample groups, time, and space	Batch effects ↓; validate cell type assignments ↓
+X+S	Within 1 MT, across >1 exp, across >1 smps	merFISH	Map cells to stable reference (cell atlas)	Accelerate analyses, increase sample size, generalize observations	Standards across experimental centers
+M1C	Across >1 MTs, within 1 exp, within 1 cell	scM&T-seq (scRNA-seq + methylome)	MOFA, DIABLO, MINT	Holistic view of cell state; quantify dependency of MTs	Scaling cell throughput; MT combinations limited; dependency of MTs ↓
+M+C	Across >1 MTs, within 1 exp, across >1 cells, within 1 cell pop	scDNA-seq + scRNA-seq	Cardelino, Clonealign, MATCHER	Use existing datasets (faster than +M1C); flexible experimental design	Validate cell/data matching; test assumptions for integrating data
+all	Across >1 MTs, across >1 exps, across >1 smps, within cells	Hypothetical (any combination)	Hypothetical (map cells to multi-omic HCA, single-cell TCGA)	Holistic view of biological systems	All from approaches +X+S, +M1C, and +M+C

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



Genome Biology volume 21, Article number: 12 (2020)

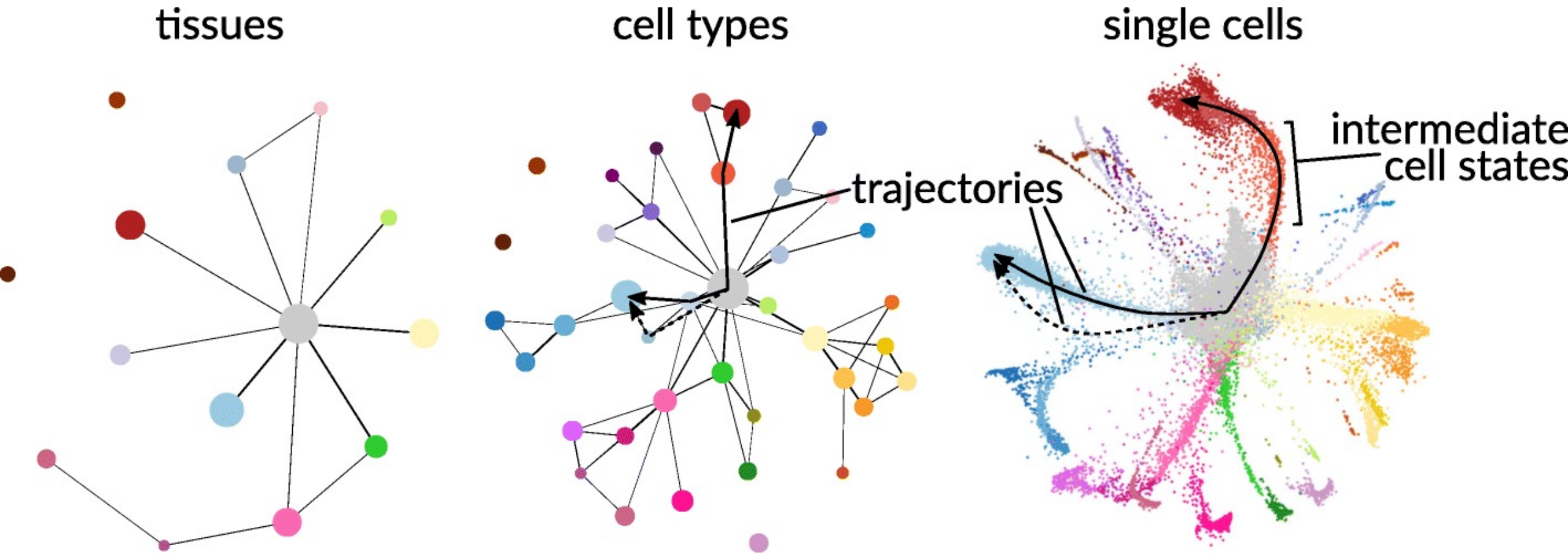
Challenge III: Mapping Single Cells to a Reference Atlas

- Classifying cells into cell types or states is essential for many secondary analyses. As an example, consider studying and classifying how expression within a cell type varies across different biological conditions.
- A computationally and statistically sound method for mapping cells onto atlases for a range of conceivable research questions will need to (i) enable operation at various levels of resolution of interest, and also cover continuous, transient cell states (ii) quantify the uncertainty of a particular mapping of cells of unknown type/state (iii) scale to ever more cells and broader coverage of types and states and (iv) eventually integrate information generated not only through scRNA-seq experiments, but also through other types of measurements, for example, scDNA-seq or protein expression data

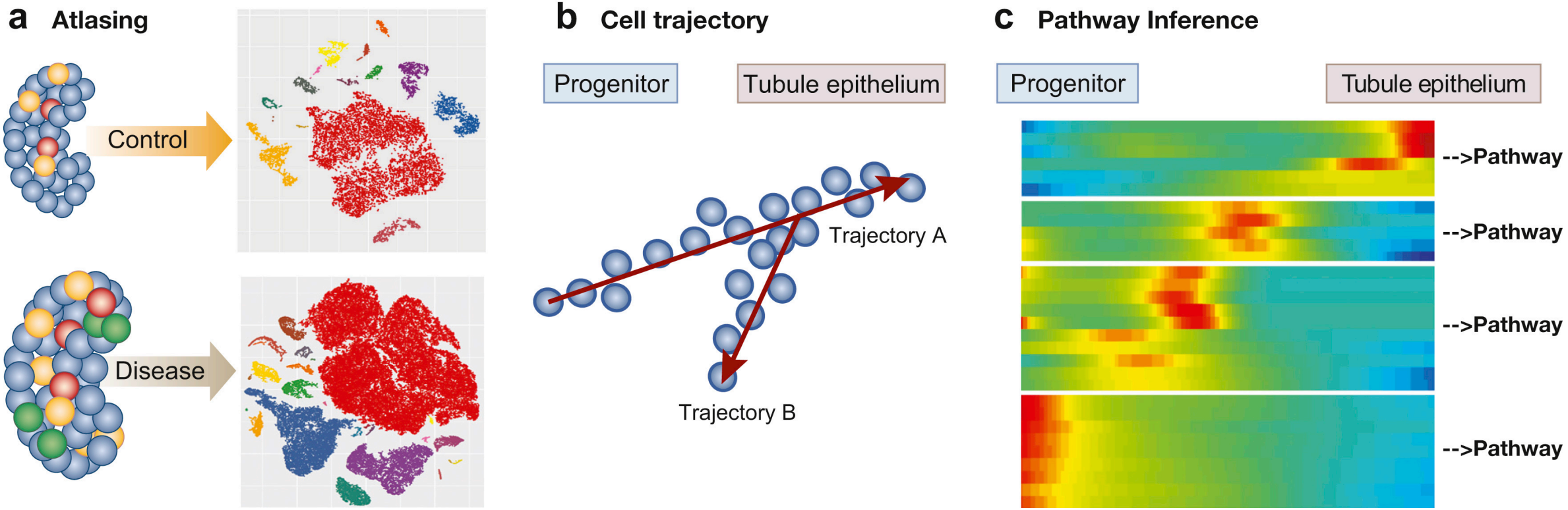
Challenge IV: Generalizing Trajectory Inference

- Several biological processes, such as differentiation, immune response, or cancer expansion, can be described and represented as continuous dynamic changes in cell type/state space using tree, graphical, or probabilistic models. A potential path that a cell can undergo in this continuous space is often referred to as a trajectory
- Trajectory inference is in principle not limited to transcriptomics. Nevertheless, modeling of other measurements, such as proteomic, metabolomic, and epigenomic, or even integrating multiple types of data is still at its infancy.
- We believe the study of complex trajectories integrating different data types, especially epigenetics and proteomics information in addition to transcriptomics data, will lead to a more systematic understanding of the processes determining cell fate.
- Trajectory methods start from a count matrix where genes are rows and cells are columns. First, a feature selection or dimensionality reduction step is used to explore a subspace where distances between cells are more reliable. Next, clustering and minimum spanning trees, principal curve or graph fitting, or random walks and diffusion operations on graphs are used to infer pseudo time and/or branching trajectories.

Different levels of resolution are of interest



Single Cell Application: Time Series in Disease and Tumors + Cell States



Kidney International Volume 96, Issue 4, October 2019, Pages 862-870
https://doi.org/10.1016/j.kint.2019.03.035

Challenge V: Finding patterns in spatially resolved measurements

- Single-cell spatial transcriptomics or proteomics technologies can obtain transcript abundance measurements while retaining spatial coordinates of cells or even transcripts within a tissue. With such data, the question arises of how spatial information can best be leveraged to find patterns, infer cell types or functions, and classify cells in a given tissue.

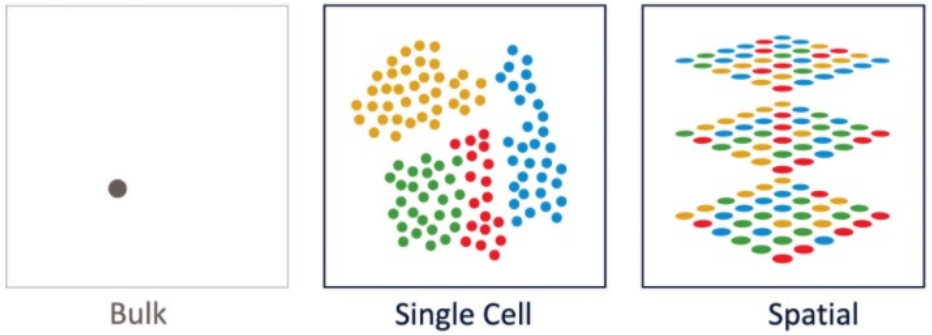
- The central problem is to consider gene or transcript expression and spatial coordinates of cells, and derive an assignment of cells to classes, functional groups, or cell types. Depending on the studied biological question, it can be useful to constrain assignments with expectations on the homogeneity of the tissue.

Single Cell Sequencing: Adding Spatial Component

=

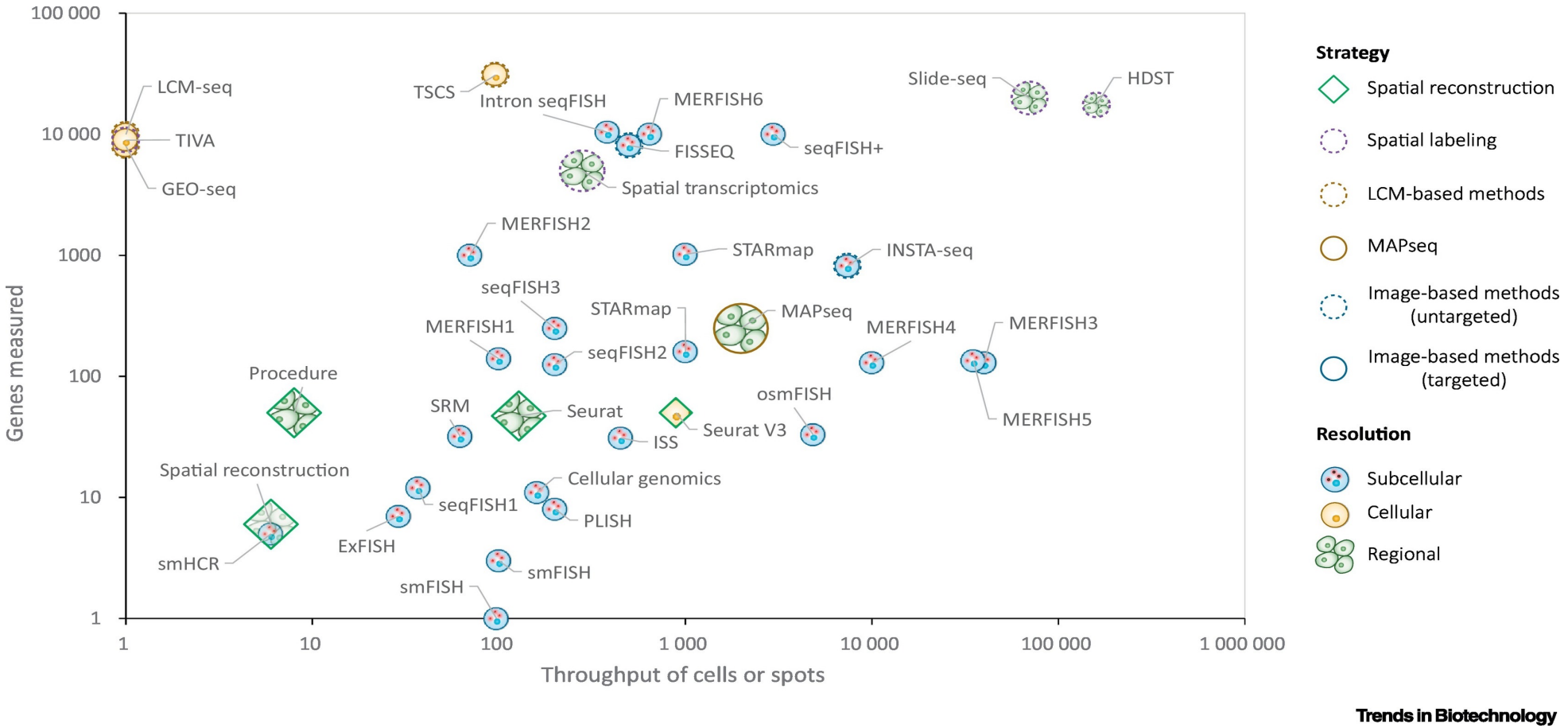
	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

+



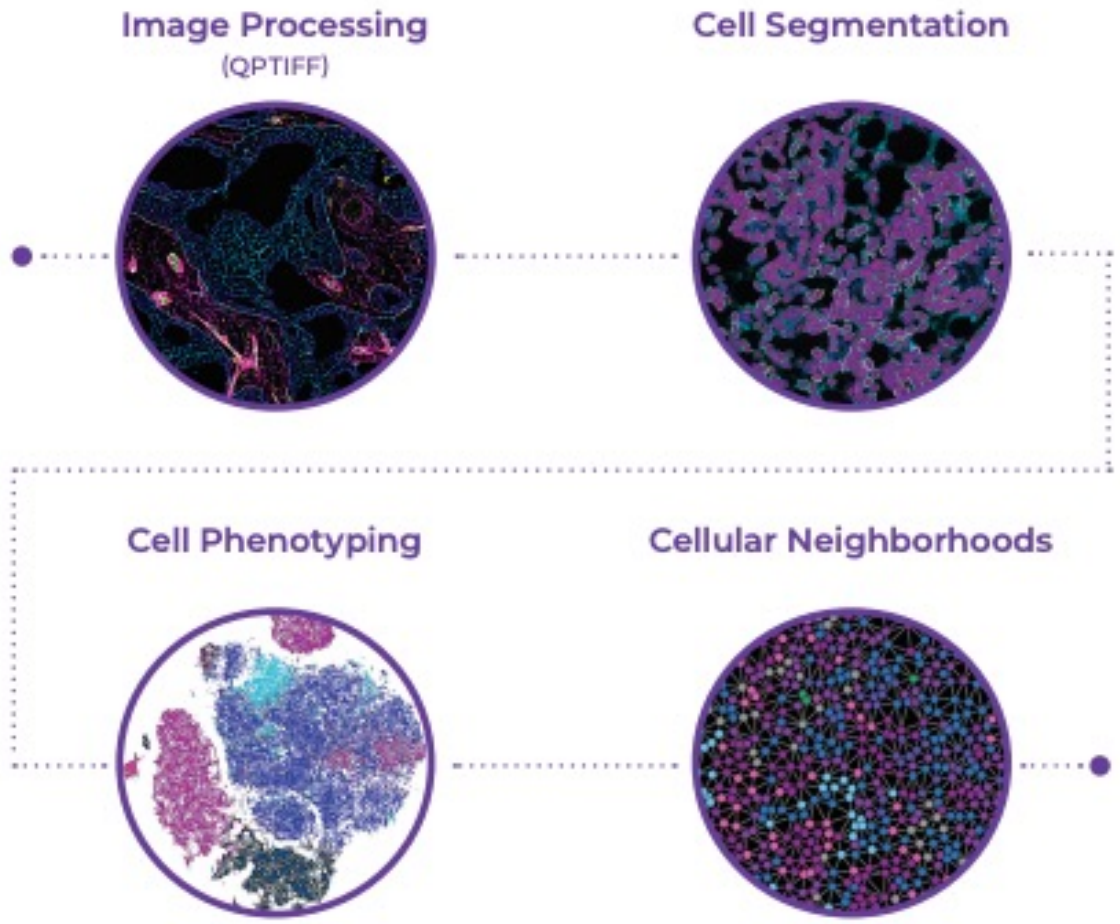
A dimensional comparison of bulk, single-cell, and spatial analyses. [10x Genomics]

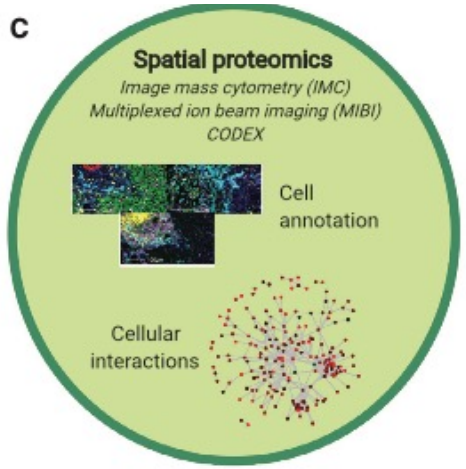
Emerging Spatial-seq Technologies

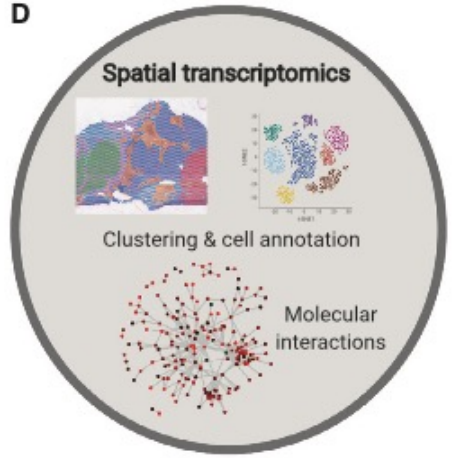


Trends in Biotechnology 2021 3943-58DOI: (10.1016/j.tibtech.2020.05.006)

Main Steps in Image Analysis

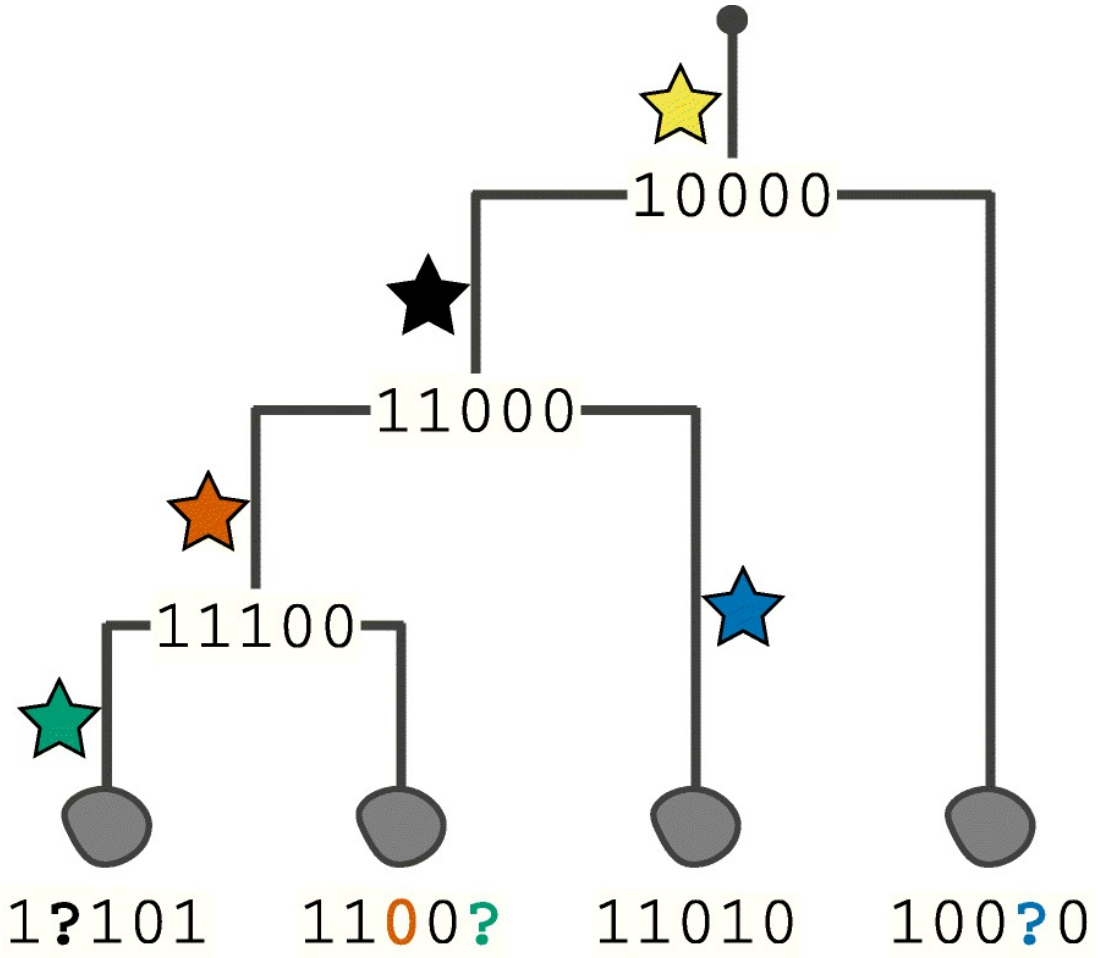






https://www.akoyabio.com/phenocycler/

Challenge VI: Dealing with errors and missing data in the identification of variation from single-cell DNA sequencing data

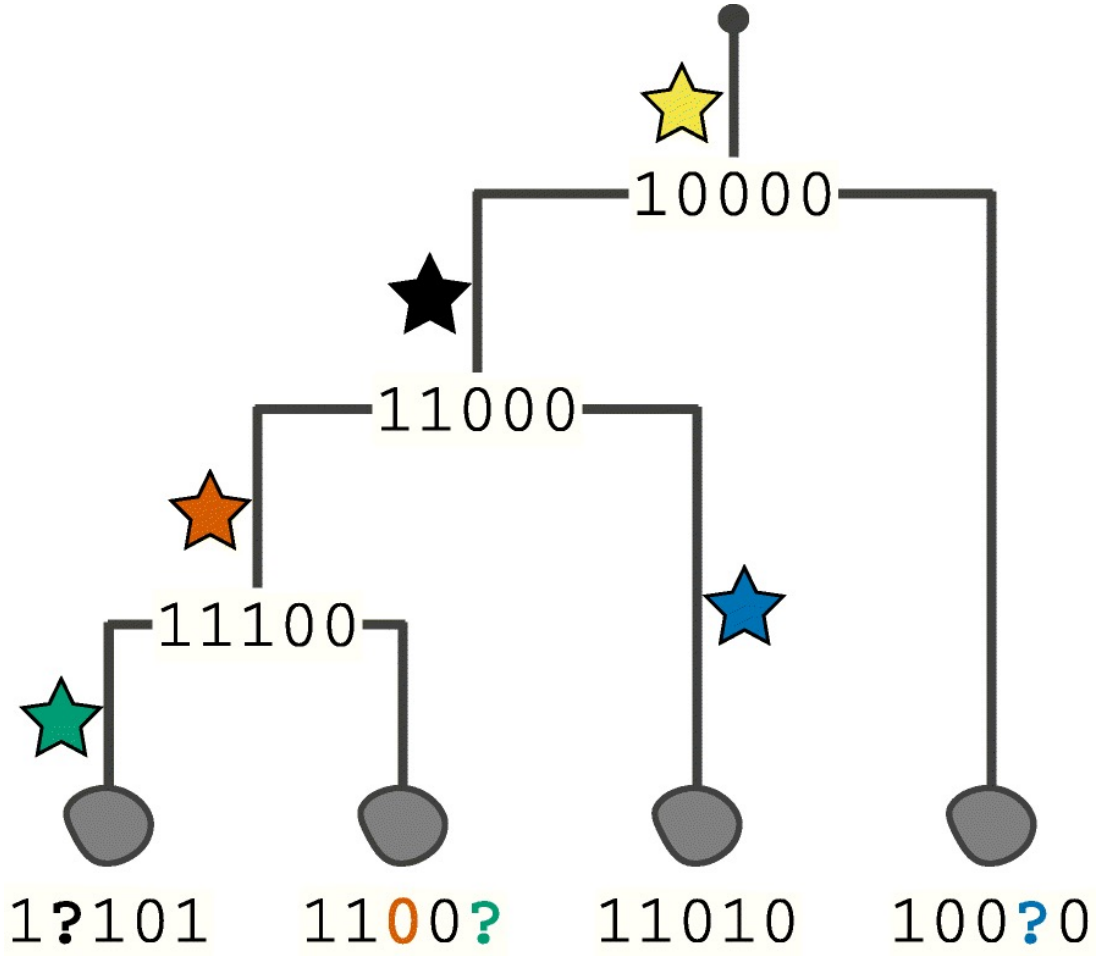


- The aim of scDNA-seq usually is to track somatic evolution at the cellular level, that is, at the finest resolution possible relative to the laws of reproduction. Examples are identifying heterogeneity and tracking evolution in cancer, as the likely most predominant use case, but also monitoring the interaction of somatic mutation with developmental and differentiation processes.

- To track genetic drifts, selective pressures, or other phenomena inherent to the development of cell clones or types—but also to stratify cancer patients for the presence of resistant subclones—it is instrumental to genotype and also phase genetic variants in single cells with sufficiently high confidence.

Genome Biology volume 21, Article number: 31 (2020)

Challenge VI: Dealing with errors and missing data in the identification of variation from single-cell DNA sequencing data



- Potential improvements in this area include (i) more explicit accounting for possible scDNA-seq error types, (ii) integrating with different data types with error profiles different from scDNA-seq (e.g., bulk sequencing or RNA sequencing), or (iii) integrating further knowledge of the process of somatic evolution, such as the constraints of phylogenetic relationships among cells, into variant calling models.

Genome Biology volume 21, Article number: 31 (2020)

Challenge VII: Scaling phylogenetic models to many cells and many sites

- Phylogenetic models of tumor evolution would still face the challenge of computational tractability, which is mainly induced by (i) the increasing numbers of cells that are sequenced in cancer studies and (ii) the increasing numbers of sites that can be queried per genome.

Challenge VIII: Integrating multiple types of variation into phylogenetic models

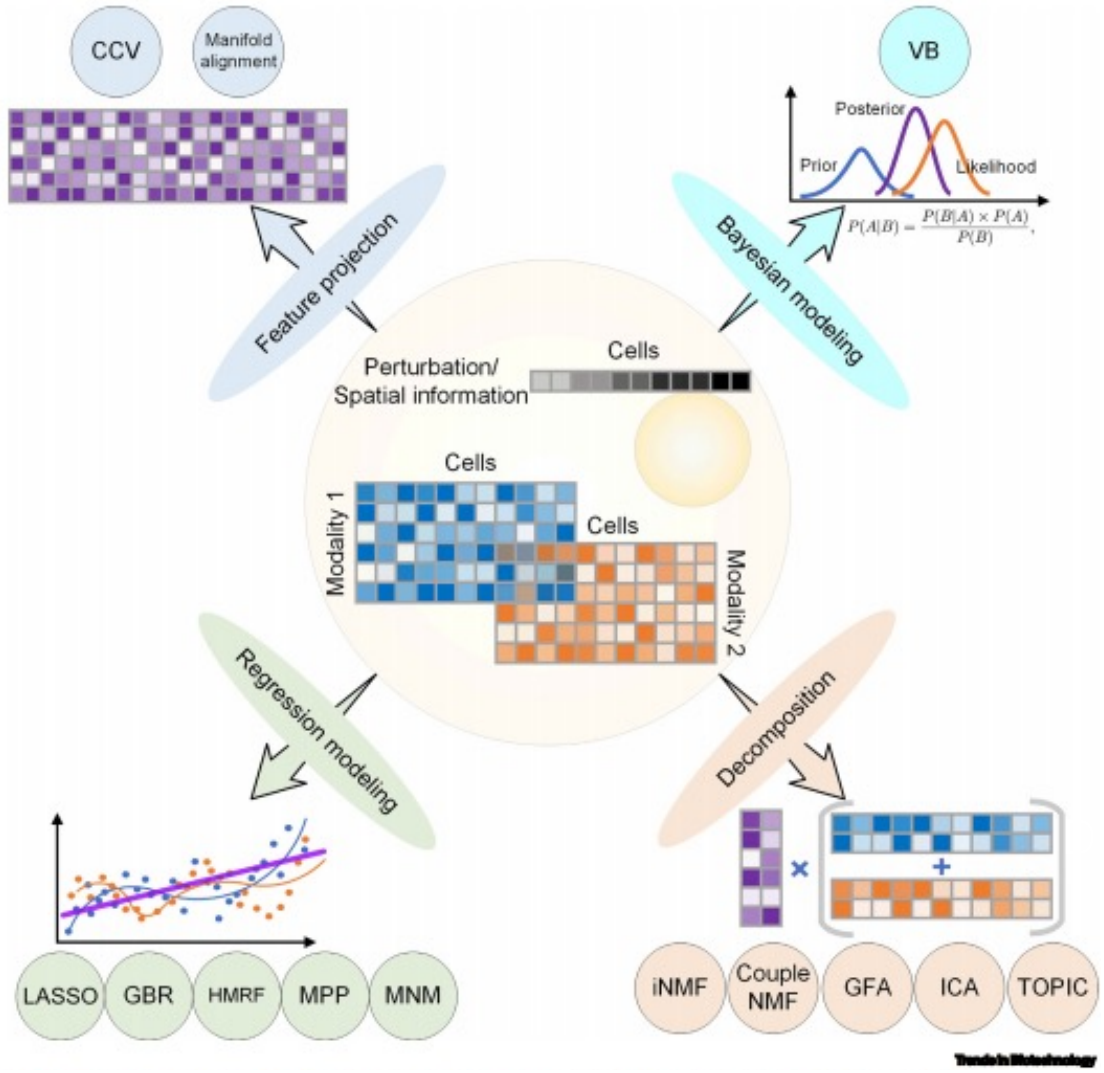
- Important it becomes to model all types of available signal in mathematical models of tumor evolution: from SNVs, over smaller insertions and deletions, to large structural variation and CNVs.

Challenge IX: Inferring population genetic parameters of tumor heterogeneity by model integration

Challenge X: Integration of single-cell data across samples, experiments, and types of measurement

- Biological processes are complex and dynamic, varying across cells and organisms. To comprehensively analyze such processes, different types of measurements from multiple experiments need to be obtained and integrated. Depending on the actual research question, such experiments can be different time points, tissues, or organisms. For their integration, we need flexible but rigorous statistical and computational frameworks.

Integration



Trends in Biotechnology 2020 DOI: (https://doi.org/10.1016/j.tibtech.2020.02.013)

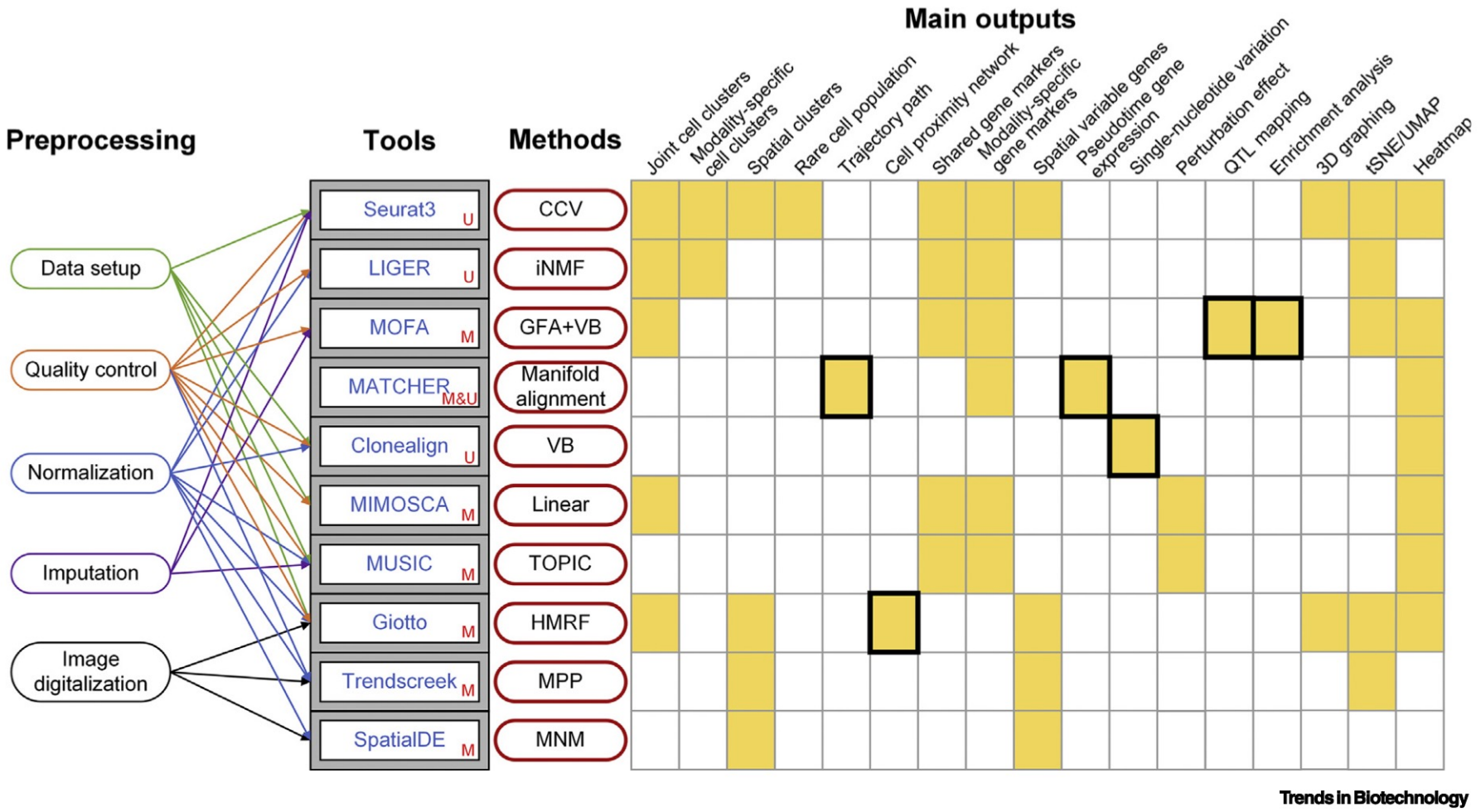
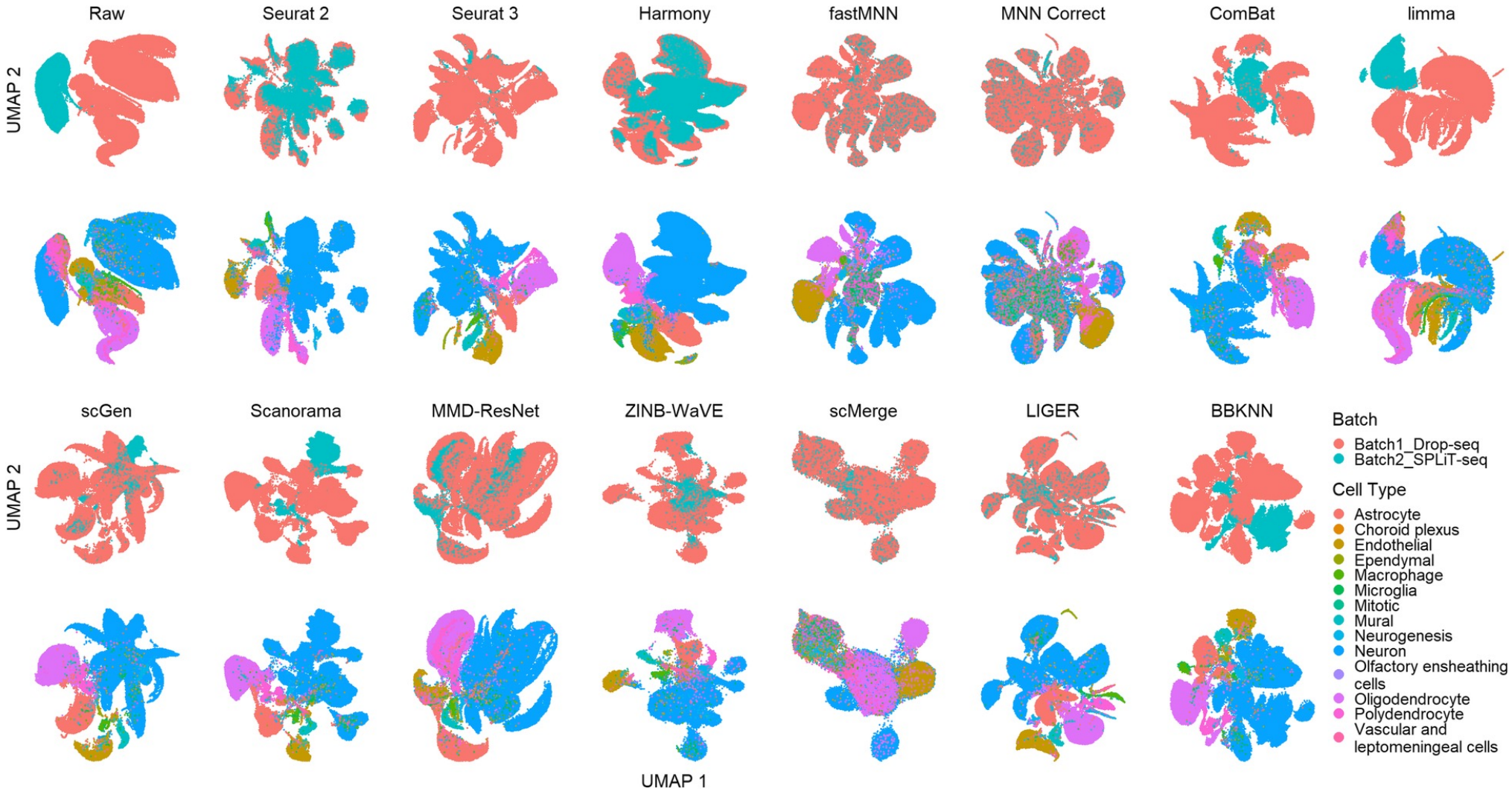


Figure 3. U (unmatched), M (matched), and M&U (both matched and unmatched) represent the data type that the paper describing the original tool claimed to support. The main outputs are summarized based on the original papers and tool tutorials from our investigations. Black frames indicate unique outputs. Abbreviations: CCV, canonical correlation vectorization; GFA, group factor analysis; HMRF, hidden Markov random field; ICA, independent component analysis; iNMF, integrative nonnegative matrix factorization; MNM, multivariate normal modeling; MPP, marked point process; tSNE, t-distributed stochastic neighbor embedding; VB, variational Bayes; UMAP, uniform manifold approximation and projection.

A Benchmark of Batch-effect Correction Methods for Single-cell RNA Sequencing Data



Genome Biology volume 21, Article number: 12 (2020)

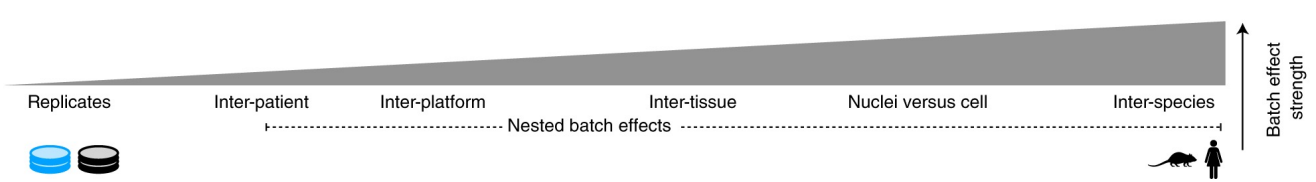
Guidelines to Choose an Integration Method

a

	Considerations	scANVI	Scanorama embed	scVI	FastMNN embed	scGen	Harmony	FastMNN gene	Seurat v3 RPCA	BBKNN	Scanorama gene	ComBat	MNN	Seurat v3 CCA	trVAE	Conos	DESC	LIGER	SAUCIE embed	SAUCIE gene
Input	Programming language	Python	Python	Python	R	Python	R	R	R	Python	Python	Python/R	Python/R	R	Python	R	Python	R	Python	Python
Input	Method runs without additional information	✗				✗														
Scib results	Consistent top performer	✓	✓	✓		✓														
Scib results	Top method on small/ simple tasks		✓		✓	✓	✓													
Scib results	Top method on large/ complex tasks	✓	✓	✓		✓														
Scib results	Top method on ATAC data	–		–			✓											✓		
Task details	Integrates strong batch effects	✓	–	–		✓			–	–				–						
Task details	Top method for recovery cell states or modules	✓	✓								✓	✓	✓							
Task details	Confounding of bio and batch variance	✓	–			✓														
Task details	Top method for trajectories	–	✓	–	✓	✓														
Task details	Method deals with varying compositions											✗								
Speed	Fast method for quick results									✓		✓								
Speed	Scales well to large datasets on CPU	✓	–	✓						✓	–								✓	✓
Speed	Method has GPU support	✓		✓		✓									✓		✓		✓	✓
Speed	Scales well to feature spaces beyond genes														✓	✓				
Output	Method shows corrected expression					✓		✓	✓		✓	✓	✓	✓						✓
Output	Method gives relative cell embeddings									✗						✗				

✓ Fulfills the criterion
– Partial fulfillment of criterion
✗ Does not fulfill criterion
Python
R

b



Nature Methods volume 19, pages41–50 (2022)

Challenge XI: Validating and Benchmarking Analysis Tools for Single-cell Measurements

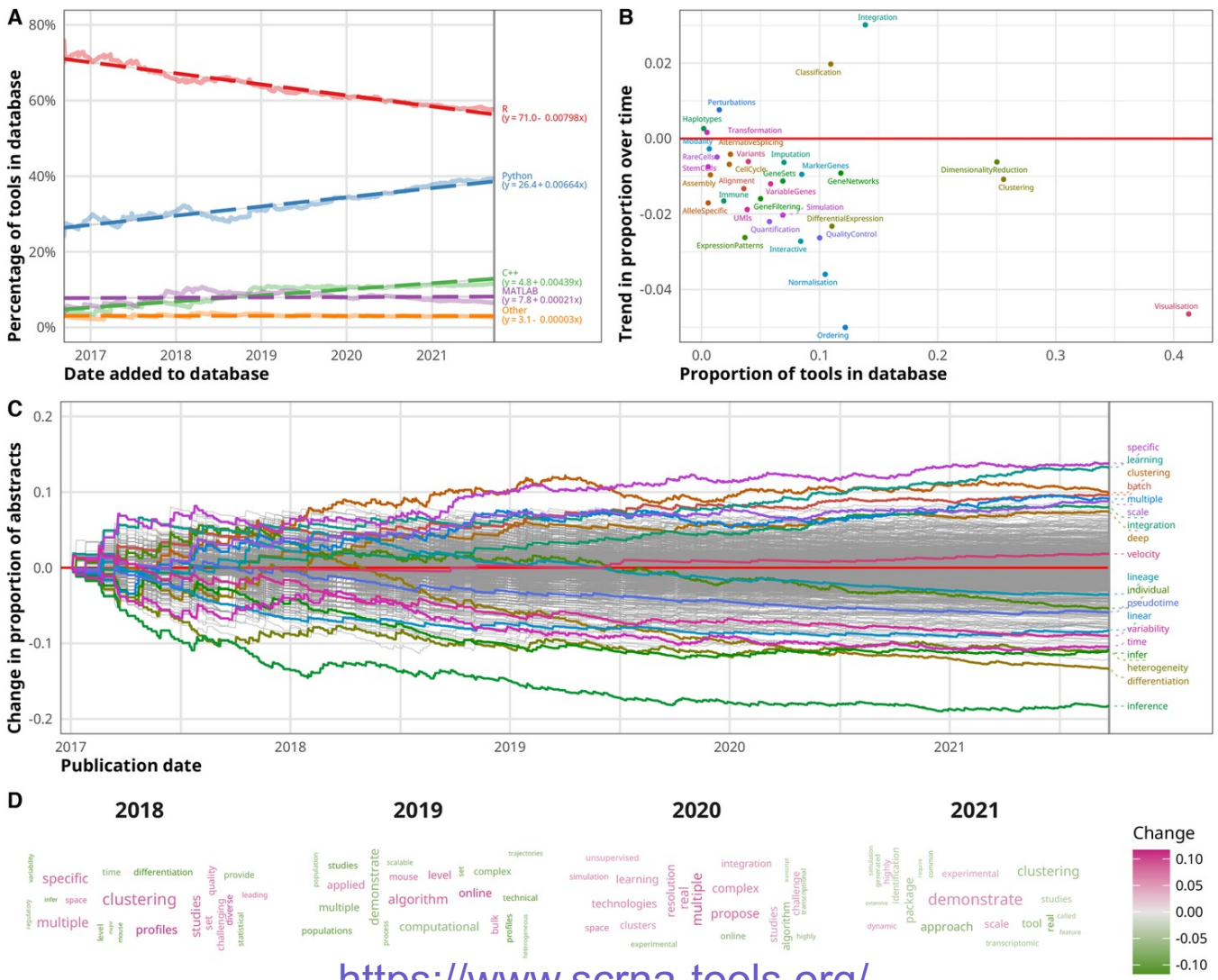
- With the advances in sc-seq and other single-cell technologies, more and more analysis tools become available for researchers, and even more are being developed and will be published in the near future.
- Thus, the need for datasets and methods that support systematic benchmarking and evaluation of these tools is becoming increasingly pressing.
- To be useful and reliable, algorithms and pipelines should be able to pass the following quality control tests: (i) They should produce the expected results (e.g., reconstruct phylogenies, estimate differential expressions, or cluster the data) of high quality and outperform existing methods, if such methods exist. (ii) They should be robust to high levels of sequencing noise and technological biases, including PCR bias, allele dropout, and chimeric signals. In addition, benchmarking should be conducted in a systematic way, following established recommendations.

Single Cell RNA Tools

https://www.scrna-tools.org/

Genome Biology volume 22, Article number: 301 (2021)

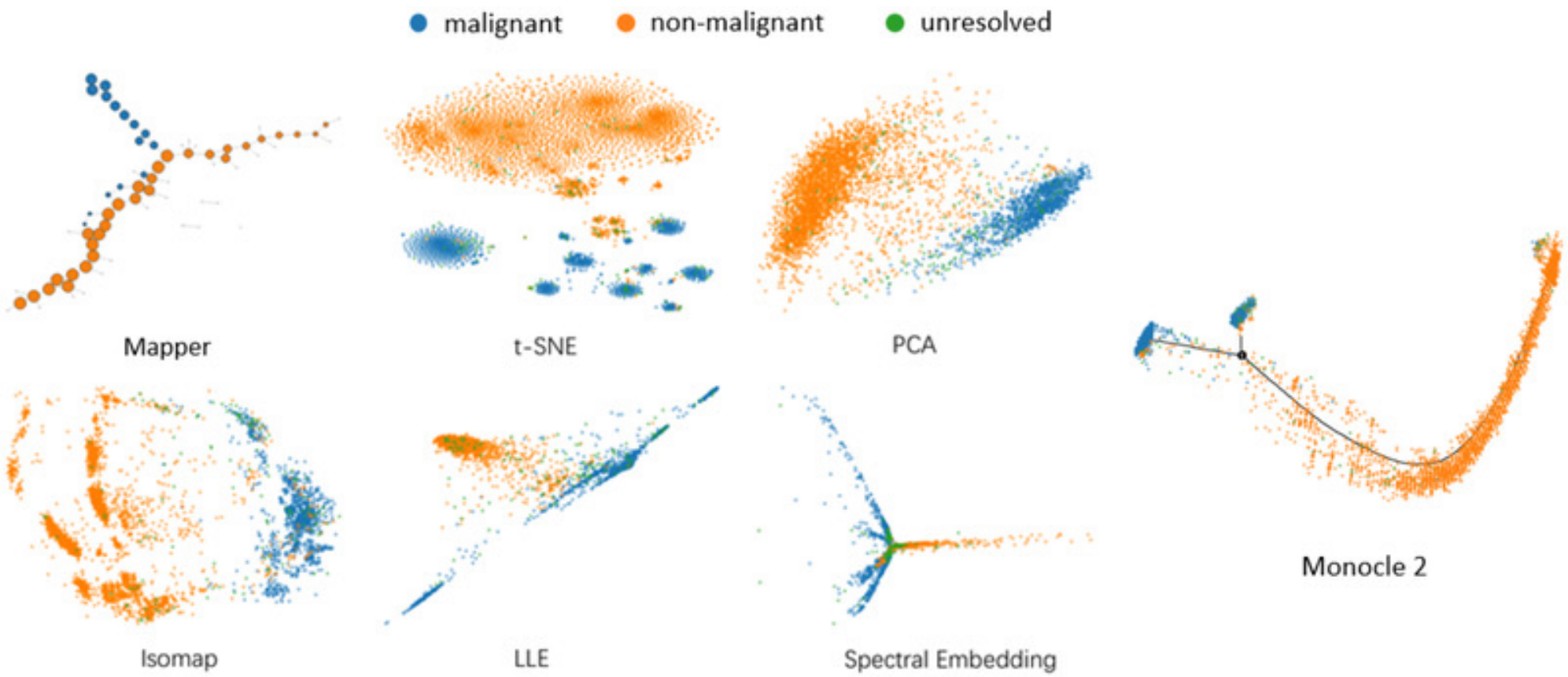
Single Cell RNA Tools



https://www.scrna-tools.org/

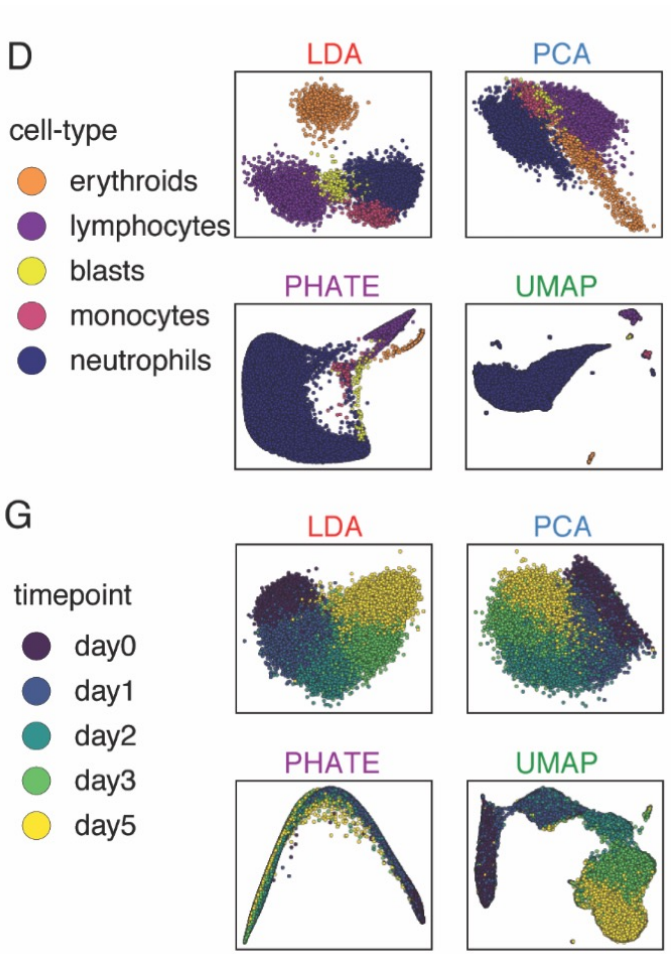
Genome Biology volume 22, Article number: 301 (2021)

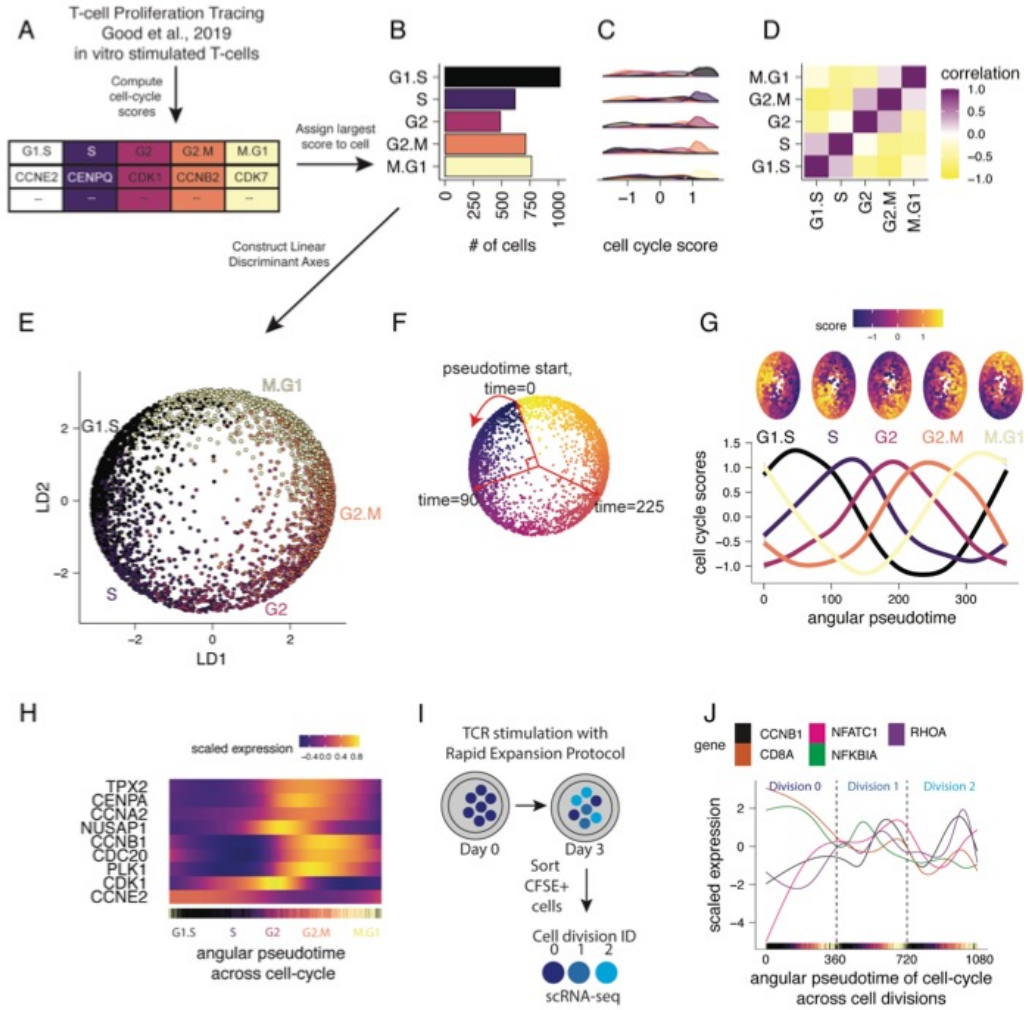
Challenge XII: Visualization



GePac Symp Biocomput. 2019; 24: 350–361.

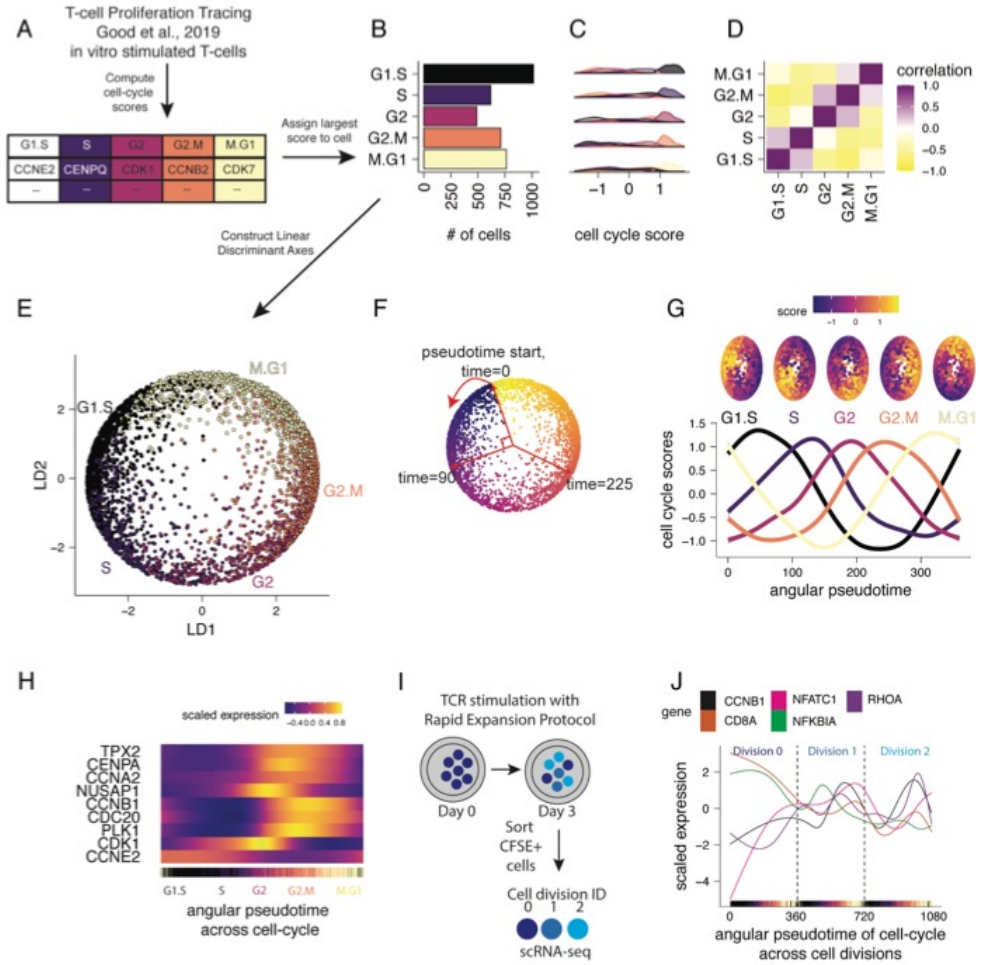
LDA is Computationally Efficient, Scalable, and Adequately Separates Class Labels





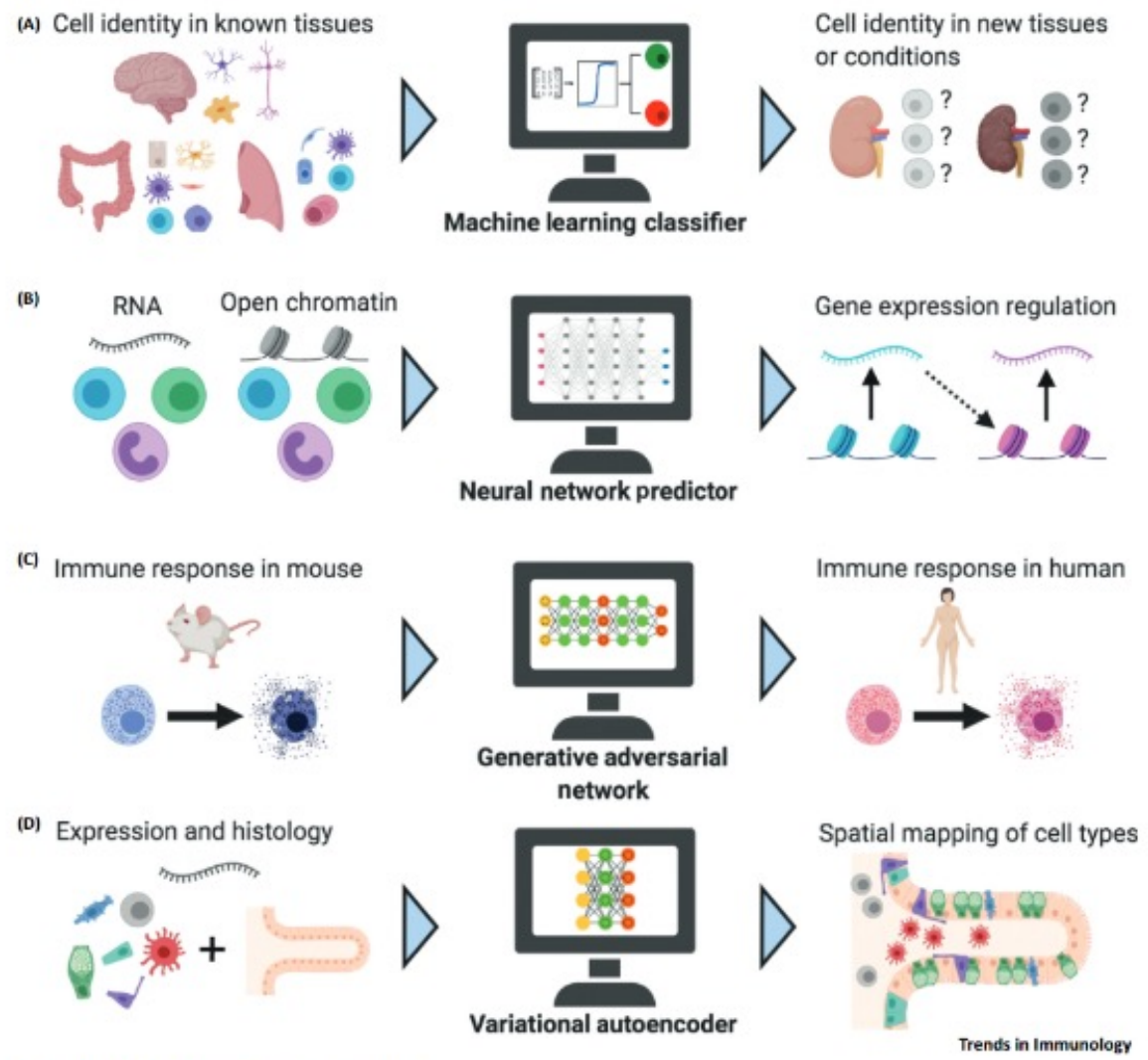
Amouzgar et. al., bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475279

LDA can Reconstruct Cyclical Trajectories using Single-cell RNAseq Data



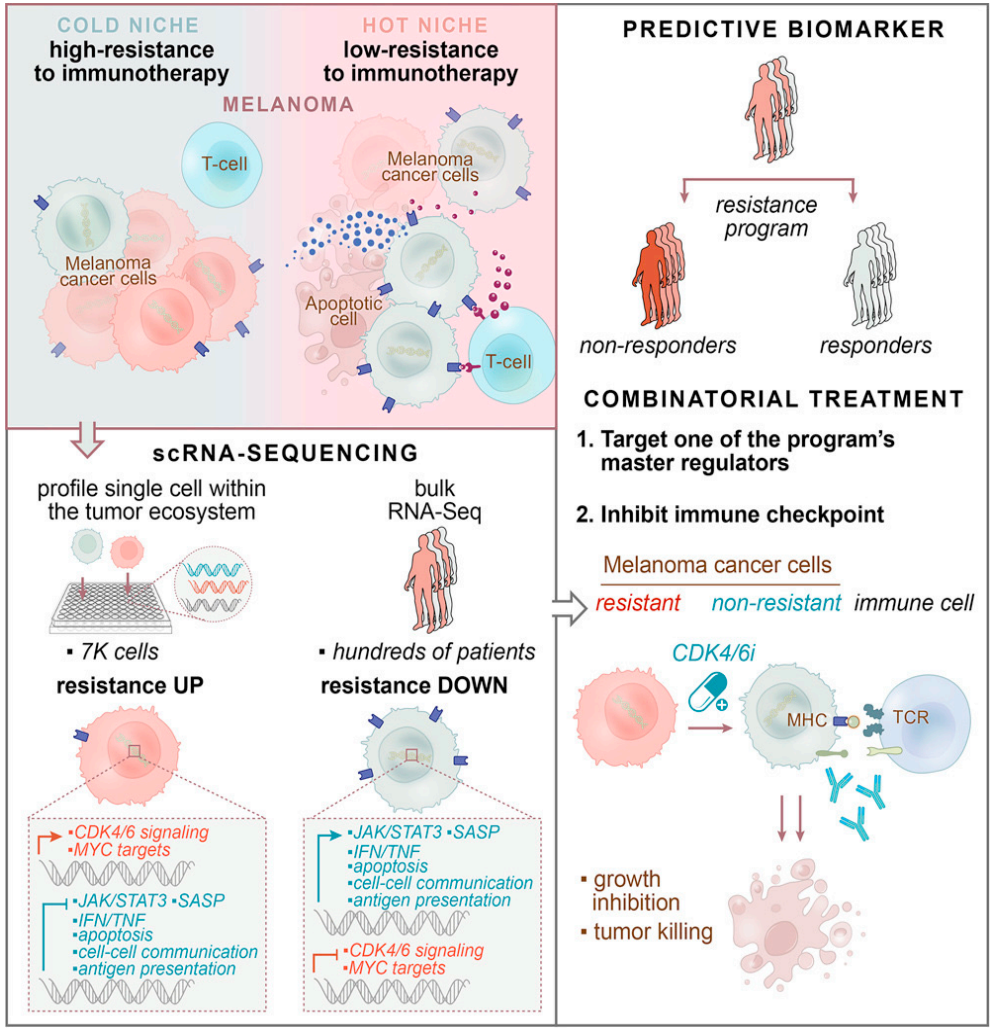
Amouzgar et. al., bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475279

AI Meets Single Cell Multi-Omics



Trends in Immunology 2019 401011-1021DOI: (10.1016/j.it.2019.09.004)

A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade



Single-cell RNA-seq identifies an immune resistance program in malignant cells

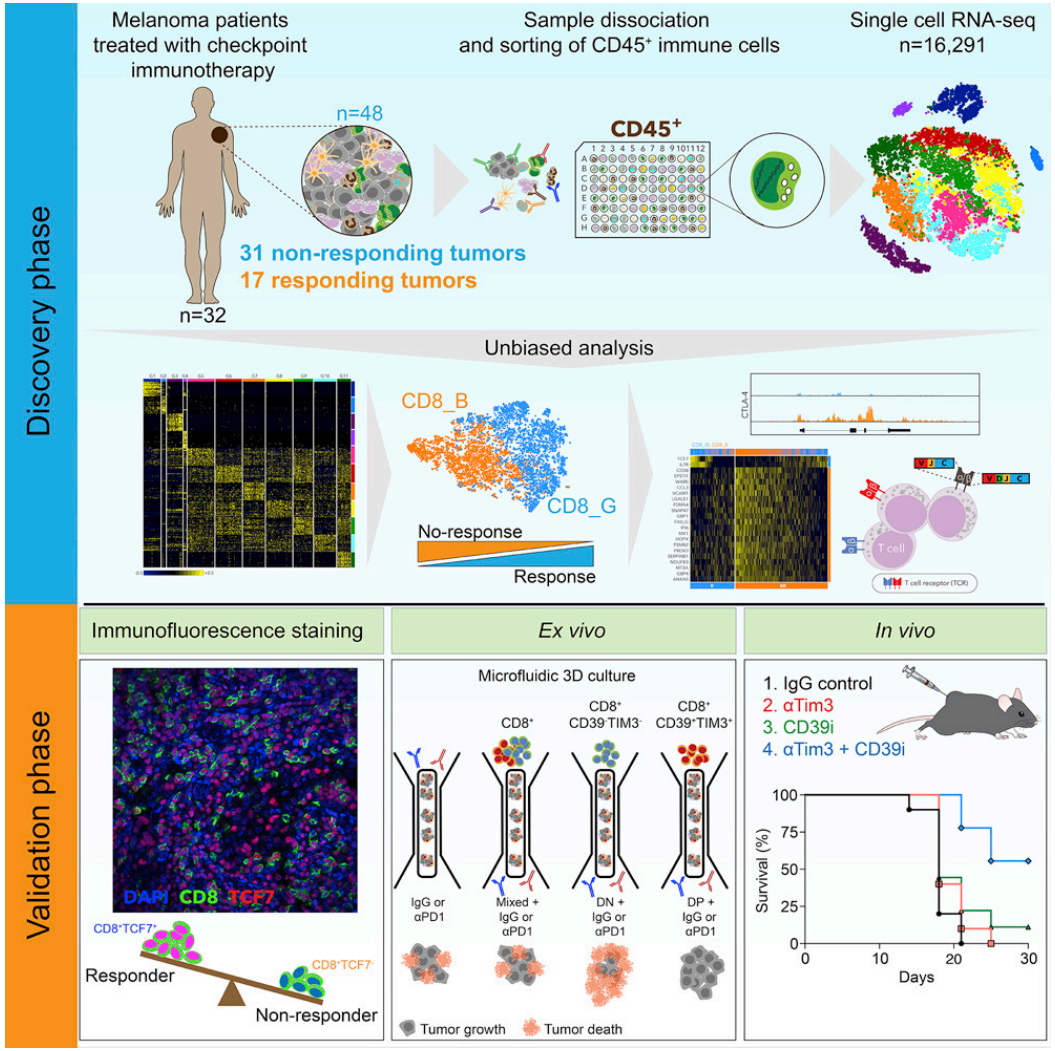
Multiple immune resistance mechanisms are co-regulated in the program

The program predicts clinical responses to immunotherapy in melanoma patients

CDK4/6 inhibitors repress the program and may sensitize melanoma to immunotherapy

Cell 2018 175984-997.e24DOI: (10.1016/j.cell.2018.09.006)

Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma



Single-cell RNA-seq reveals distinct CD45+ cells associated with clinical outcome

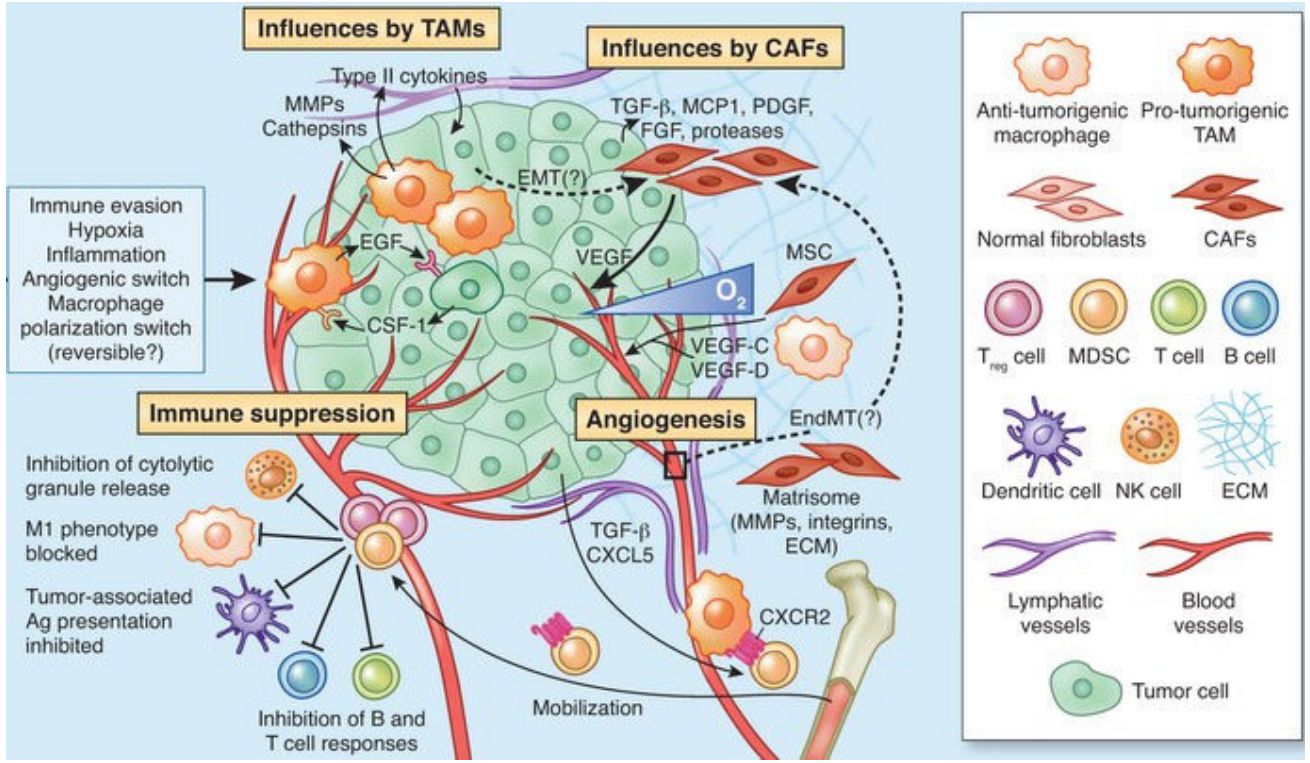
The balance between two CD8+ T cell states is linked with tumor regression

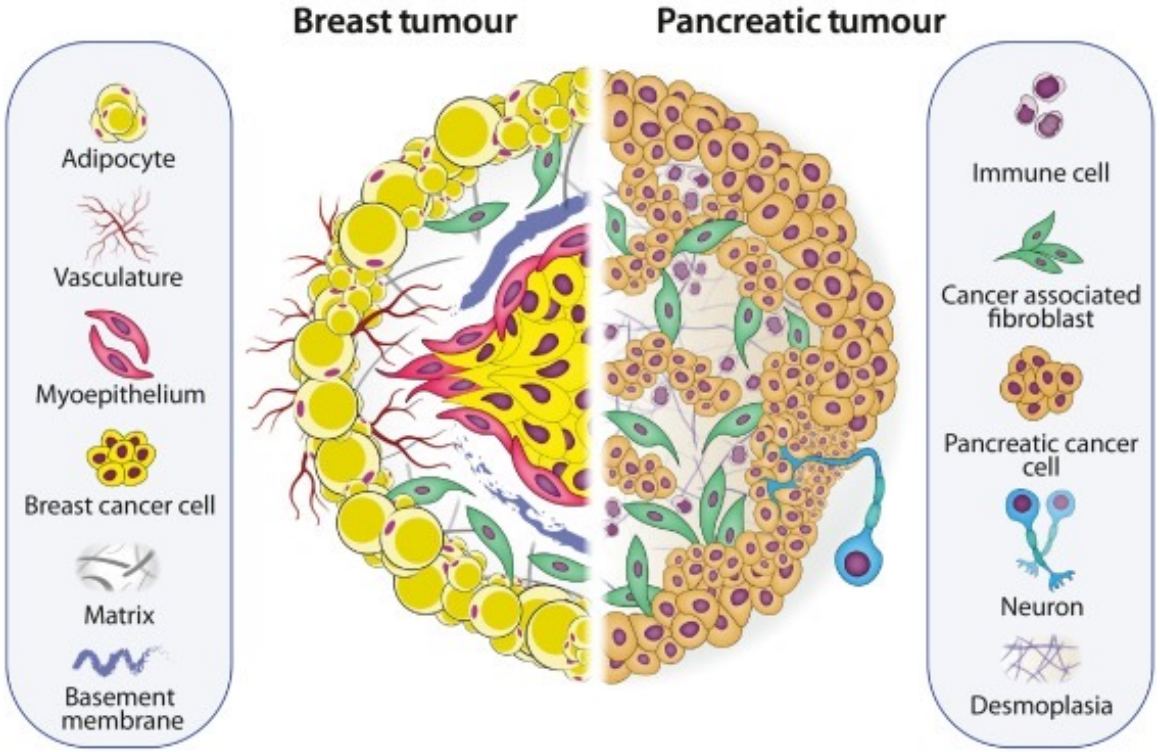
TCF7+CD8+ T cell frequency in tumor tissue predicts response and better survival

Dual blockade of CD39 with different checkpoint proteins enhances immunity

Cell 2018 175984-997.e24DOI: (10.1016/j.cell.2018.09.006)

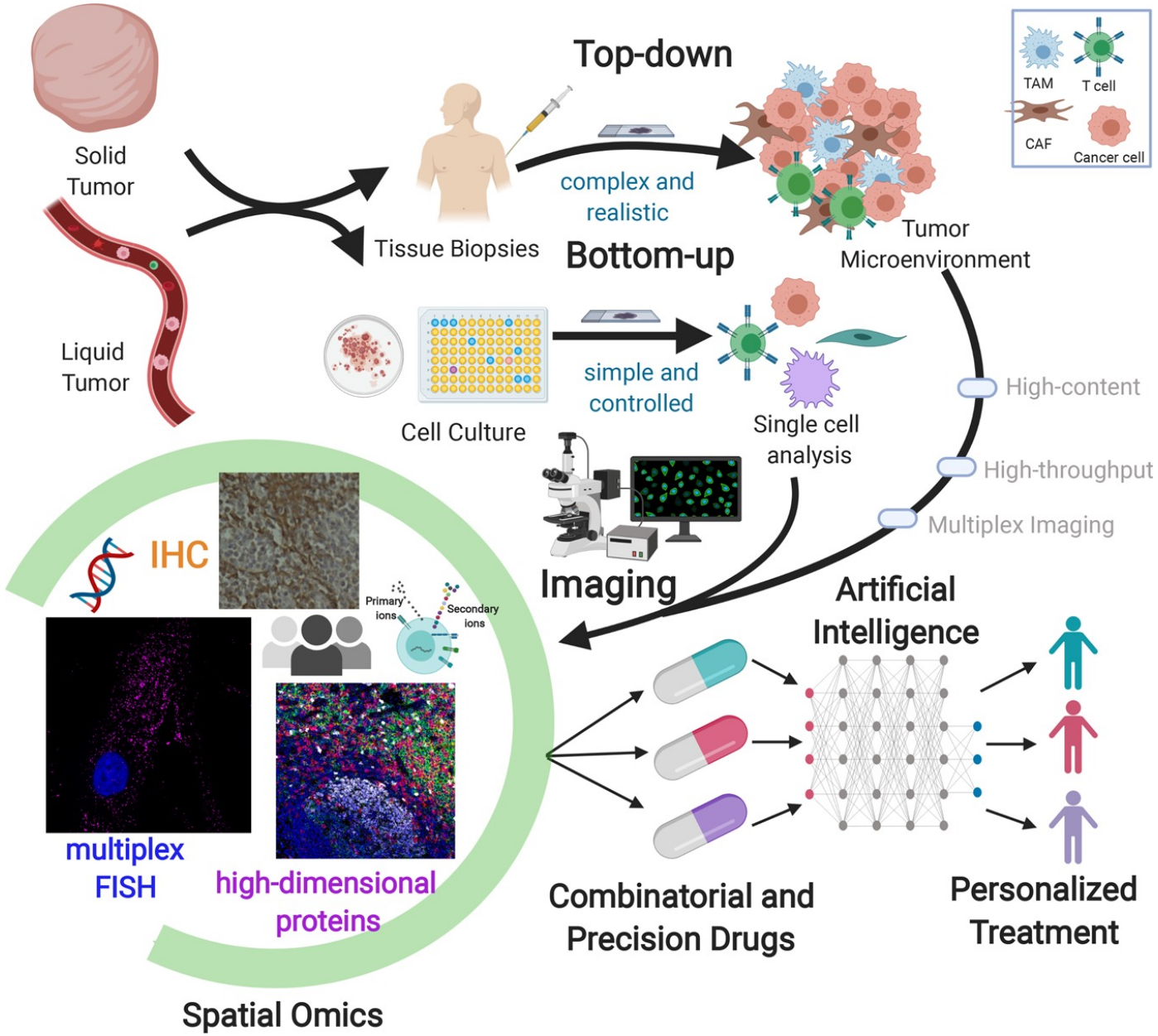
Spatial Aspects of TIME (Tumor Immune Micro Environment)





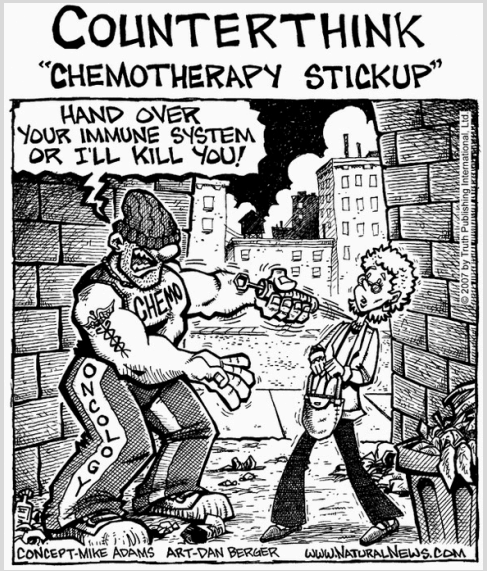
Chapman, Lacey. (2016). The Influence of GLUT-1 and ChREBP Tumour Biomarkers on Chemosensitivity to the mTOR Inhibitor Temsirolimus.

https://www.researchgate.net/publication/293176142_The_Influence_of_GLUT-1_and_ChREBP_Tumour_Biomarkers_on_Chemosensitivity_to_the_mTOR_Inhibitor_Temsirolimus


Top-down
Solid Tumor
Liquid Tumor
Tissue Biopsies
complex and realistic
Bottom-up
TAM
T cell
CAF
Cancer cell
Tumor Microenvironment
simple and controlled
Cell Culture
Single cell analysis
High-content
High-throughput
Multiplex Imaging
Imaging
IHC
Primary ions
Secondary ions
multiplex FISH
high-dimensional proteins
Spatial Omics
Artificial Intelligence
Combinatorial and Precision Drugs
Personalized Treatment

Changing the Paradigm

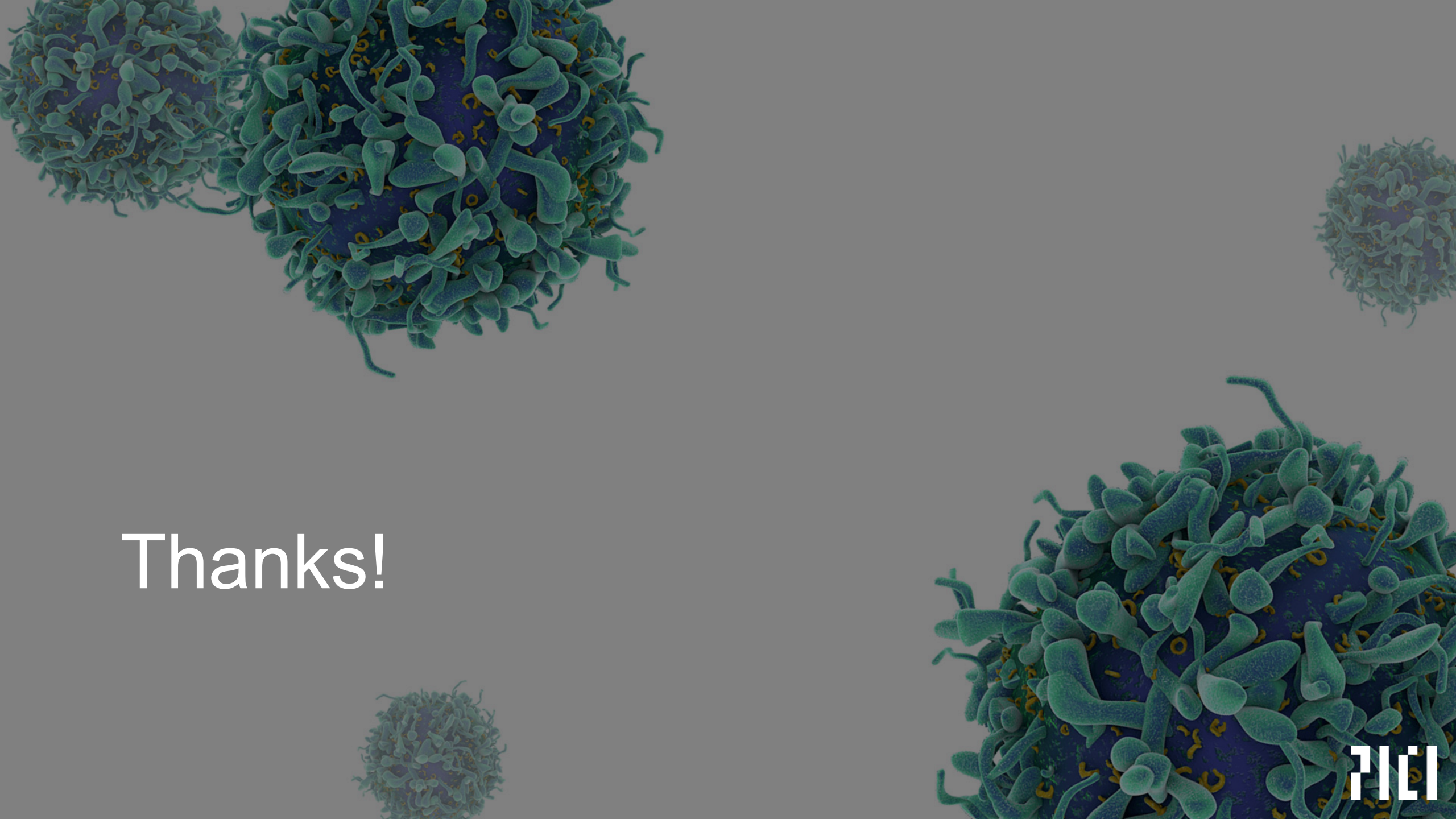
Chemo / Radiation / Surgery



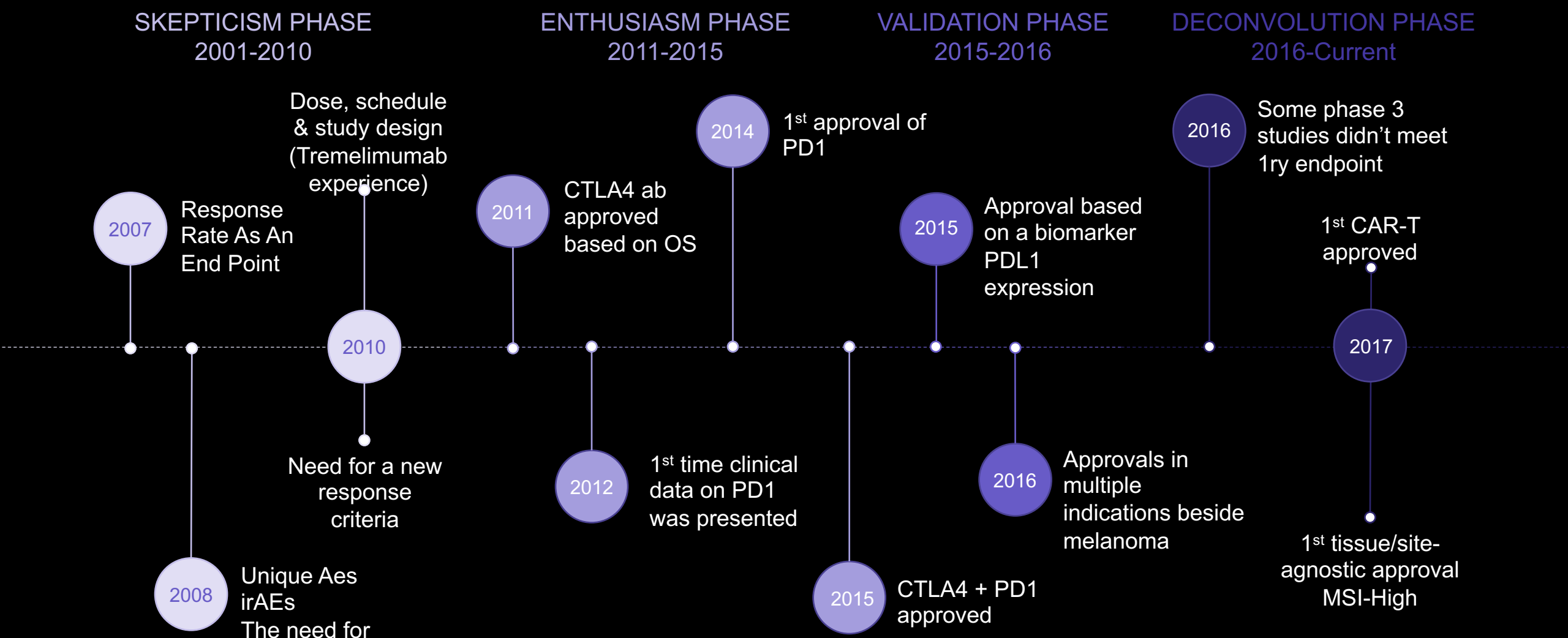
- Cut it out (if possible)
- Poison the tumor
- Wait for escape
- Poison again

Immunotherapy

- Re-educate the immune response to treat tumors as **non-self**
- Unleash the immune system brakes and turn on the gas
- Specificity, memory, durability and infectious anti-tumor activity


Thanks!

History and Evolution of Immunotherapy



AMADEUS:

UNDERSTAND HOT VS. COLD TUMORS

OPPORTUNITY

How do we figure out when cancer is most vulnerable to immunotherapy? We're taking a close look at cold vs. hot tumors, and searching for biomarkers to help answer the question.

KEY PARTNERS

Bristol-Myers Squibb

CANCER RESEARCH INSTITUTE®

LEAD INVESTIGATOR

Padmanee Sharma, MD, PhD
MD Anderson Cancer Center

PROGRESS

- Began enrolling patients in September 2018; almost 60 patients currently enrolled
- Study open at:
  - MD Anderson
  - Dana-Farber
  - Memorial Sloan Kettering
  - Stanford Medicine
  - UCLA
  - UCSF

PORTER:

TACKLE PROSTATE CANCER WITH NEW COMBINATIONS

OPPORTUNITY

Prostate cancer is the second leading cause of cancer death among men in the U.S. We will use a “platform” design to efficiently test several immunotherapy treatment combinations to best treat this deadly cancer.

KEY PARTNERS

Bristol-Myers Squibb

Celldex therapeutics

CANCER RESEARCH INSTITUTE

inovio PHARMACEUTICALS

Oncovir, Inc. Clinical Trials

LEAD INVESTIGATORS

Kristopher Wentzel, MD | Angeles Clinic
Matthew Galsky, MD | Mt. Sinai
Lawrence Fong, MD | UCSF
Julie Graff, MD | OHSU

PROGRESS

- Began enrolling patients in June 2019
- Study open at:
  - Angeles Clinic
  - Icahn School of Medicine, Mt. Sinai
  - UCSF
  - Oregon Health & Science University

Our Partners in PICI Bruce Program

MEMBER INSTITUTIONS + RESEARCHERS

UCSF

City of Hope®

Stanford MEDICINE

UCLA

PARTNERS + COLLABORATORS

CANCER RESEARCH INSTITUTE®

ACGT® ALLIANCE FOR CANCER GENE THERAPY