Single-cell transcriptome analysis: opportunities and computational challenges

Rob Patro

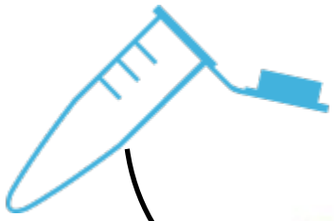
CBCB
Center for Bioinformatics & Computational Biology

UMIACS
University of Maryland Institute for Advanced Computer Studies

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

The Molecular Microscope of the (early) 21st Century

Sample collection & prep



Sequencing

Sequencing Output Per Flow Cell

Flow Cell Type	S4
2 × 100 bp	1600–2000 Gb†
2 × 150 bp	2400–3000 Gb

4.8-6 Tbases / run

x2

"Analysis"

```
@SRR1215997.1 HWI-ST1311:58:C132FACXX:3:1101:3092:2249/1
GAAGATGAGTGCATTGGAGGCTCTGGTGATGGAATGATAAAAGTGAAGAATCAGCTCCGCTTCCTTGCAGAACTGGCCTATGATCTGGATGTGGATGATG
+
7<@DD;DAHDFB?G4A?BAFHFGFEH<C<E<?CFEFBGCH<DDBE<DGCF>?F<D*?FFFHIGH@GGIGII9)=;=?DED>@C>C3@CC>@@C:3:3:5@
@SRR1215997.2 HWI-ST1311:58:C132FACXX:3:1101:3435:2101/1
CTTNTGACGCACTCCTCTAATTTCGCCATATCTGTCTCATCATCCCAAGGTTTCACATCTAGTAAGATGGAAGACTTGGCAACAAGTGCAGGTTTTTTGG
+
?@@#4ADACFHFHBGHHIEHIIIHHGIIJIIGIGGBGIGE4?BBBFGGII./8C8)/@CHGIHHGEHIDEHGCHEDDEFEECCDDCCACCCD<>A??BB<
@SRR1215997.3 HWI-ST1311:58:C132FACXX:3:1101:3410:2170/1
TGCTCAACGGCTCTCAGCTGGTGCTCCATGGCCTGGAACTGGGCCACAGTGGCCTCTACGCCTGCTTCCACCGTGACTCCTGGCACCTGCGCCACCAAGT
```

Unparalleled resolution & throughput, but …

Sequencing is the medium for *many different types of assays*

Different measurements often require new methods

Produces so much data that *analysis becomes the bottleneck*

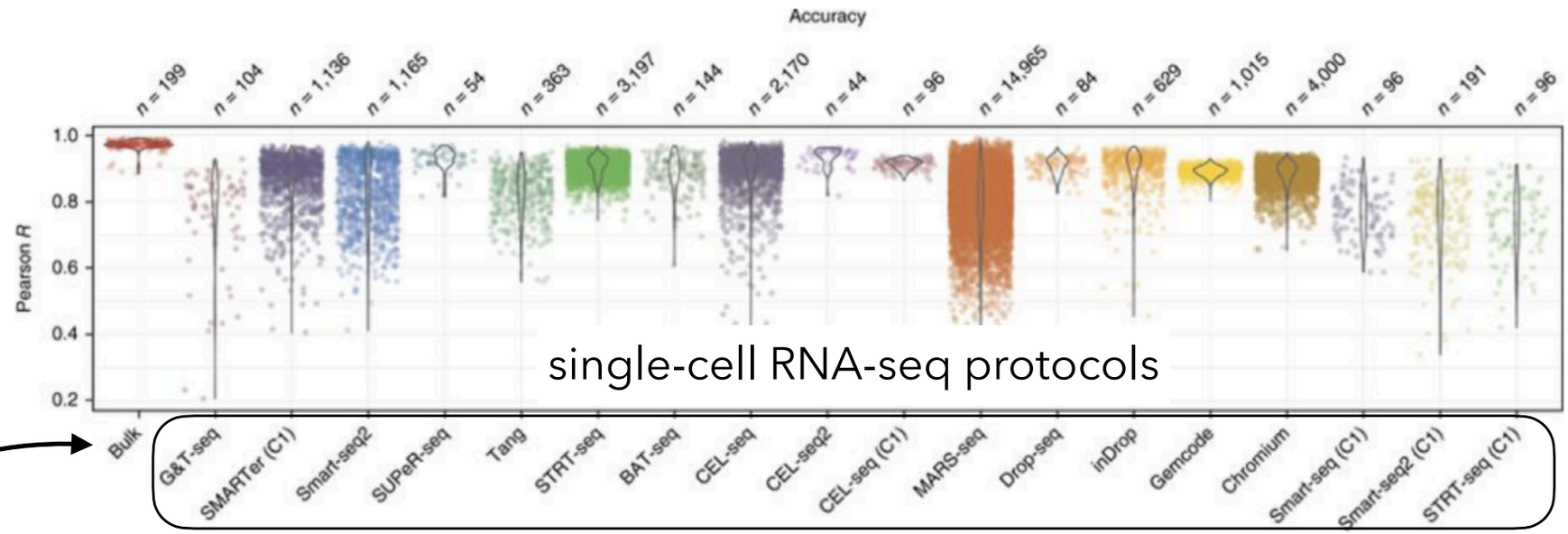
Magnitude of data requires fundamentally new approaches

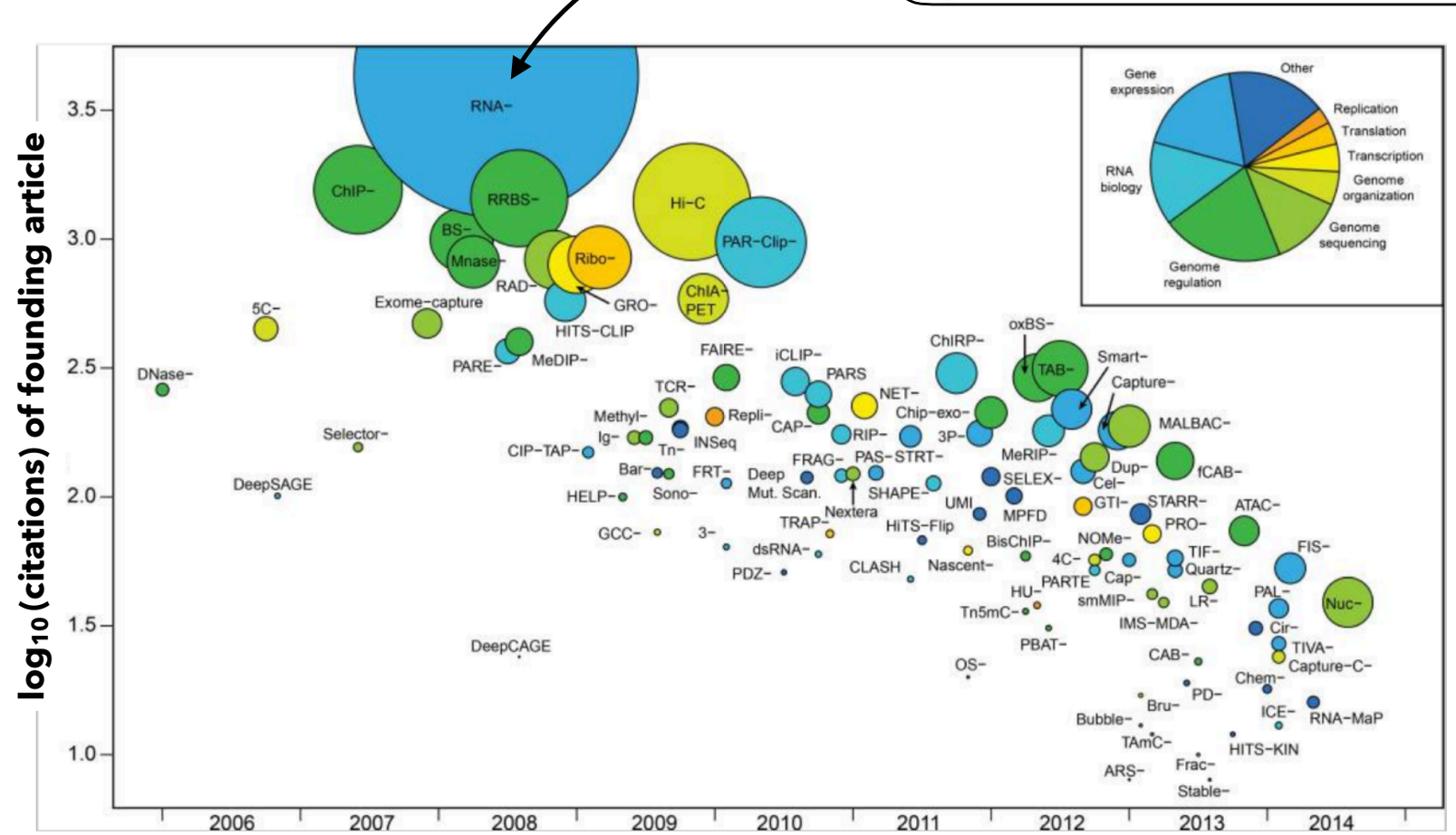
https://emea.illumina.com/systems/sequencing-platforms/novaseq/specifications.html

Challenges of studying RNA

Understanding *new* types of data

Even just a single "seq" type (e.g. RNA-seq) has many different variations that cannot all be processed in the same way





Proliferation of new technologies

Measuring *different things*

Measuring in *different ways*

High-Throughput sequencing is the common factor (the medium)

Note: All the slides with SBU Red heading, Courtesy of Rob Patro or Hirak Sarkar

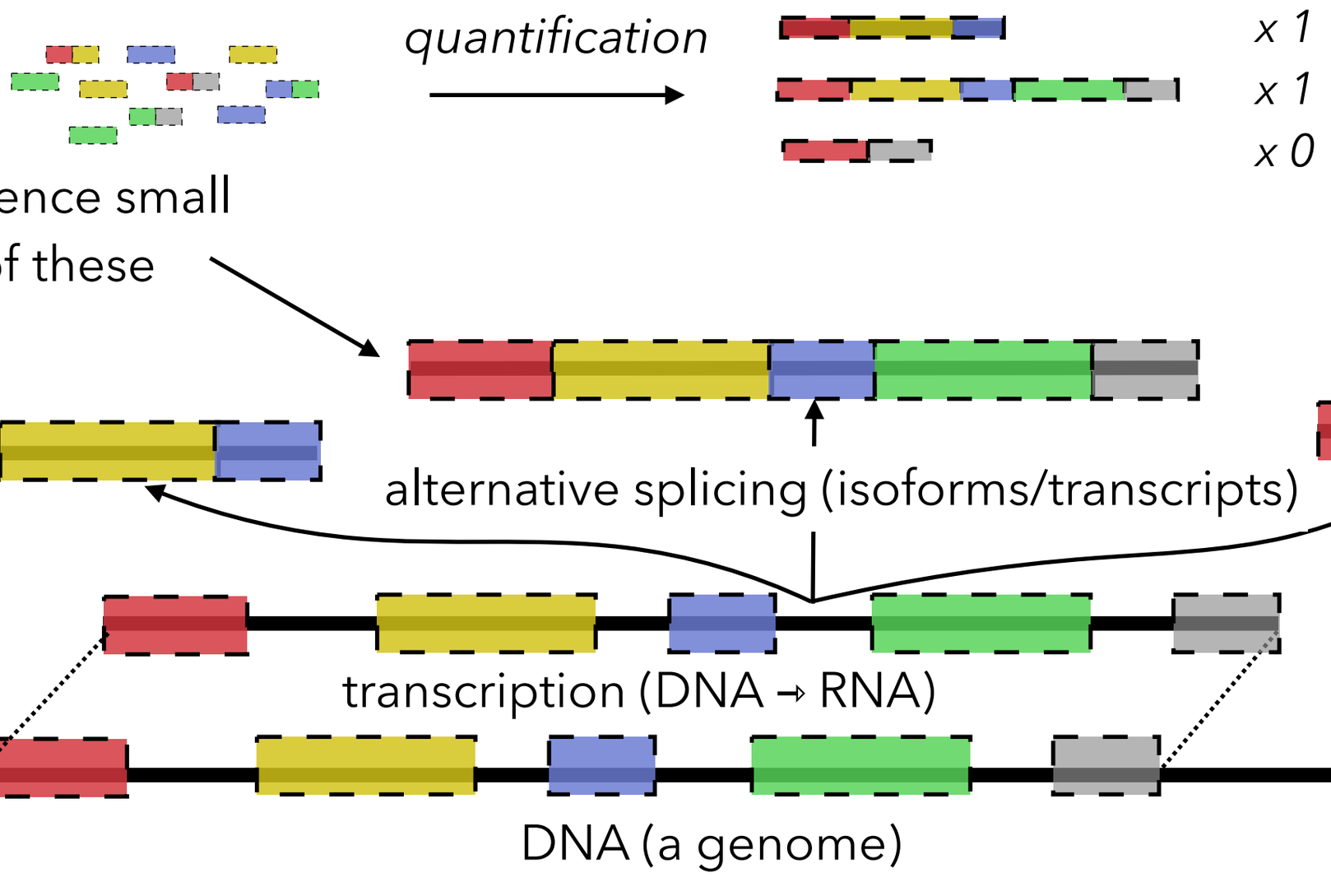
Svensson, Valentine, et al. "Power analysis of single-cell RNA-sequencing experiments." *Nature methods* 14.4 (2017): 381.

Reuter, Jason A., Damek V. Spacek, and Michael P. Snyder. "High-throughput sequencing technologies." *Molecular cell* 58.4 (2015): 586-597.

Expression Study basics in one slide

Lets us (among other things) quantify expression in a tissue:

- What genes / transcripts are turned on?
- At what levels are they expressed?
- How do they respond to environment / stimuli?

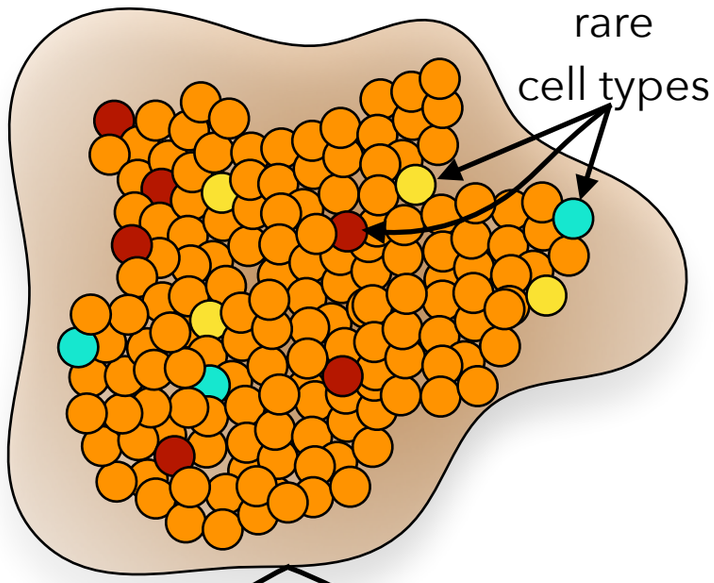


RNA-seq ⇒ single-cell RNA-seq: a brave new world

Bulk RNA-seq

- Typically millions or 10s of millions of cells
- High-fidelity & high-sensitivity
- Measure transcript abundance at the population-level

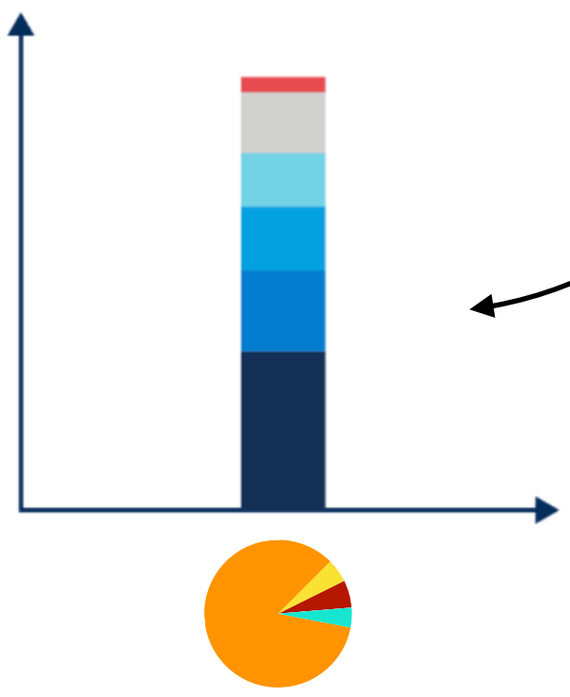
Tissue

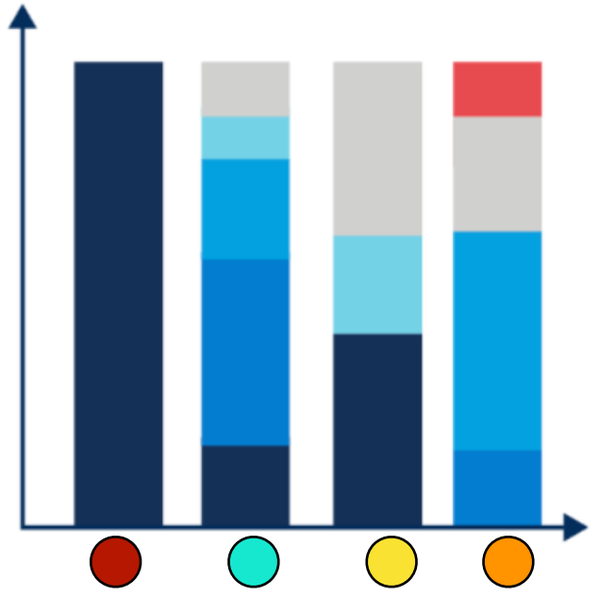


These present **very** different challenges

single-cell RNA-seq

- Typically tens of thousands of cells
- Low-coverage (reads / cell)
- Measure transcript abundance at the single-cell level





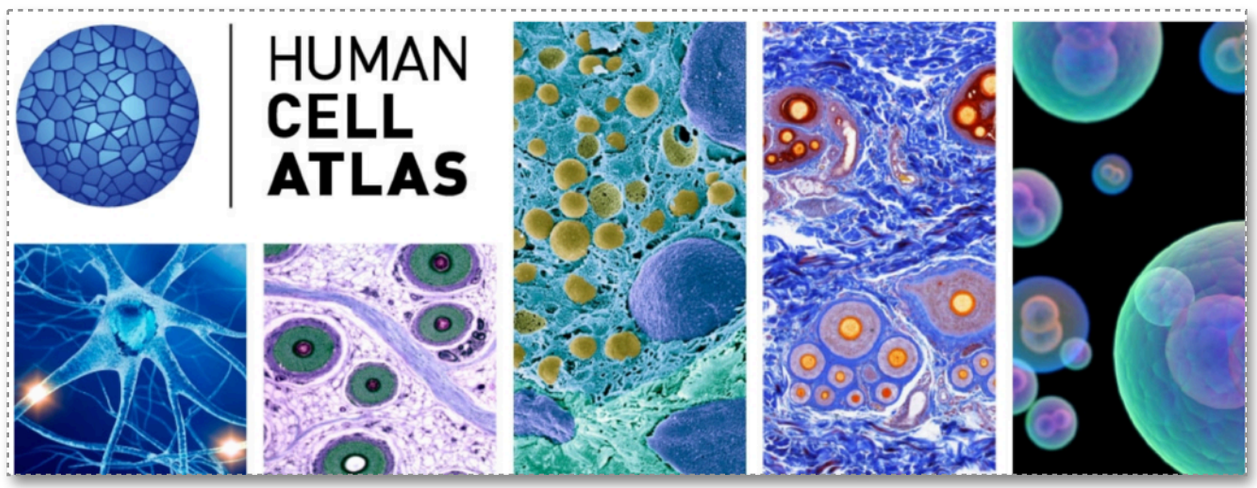
involves cell-type identification (supervised or unsupervised)

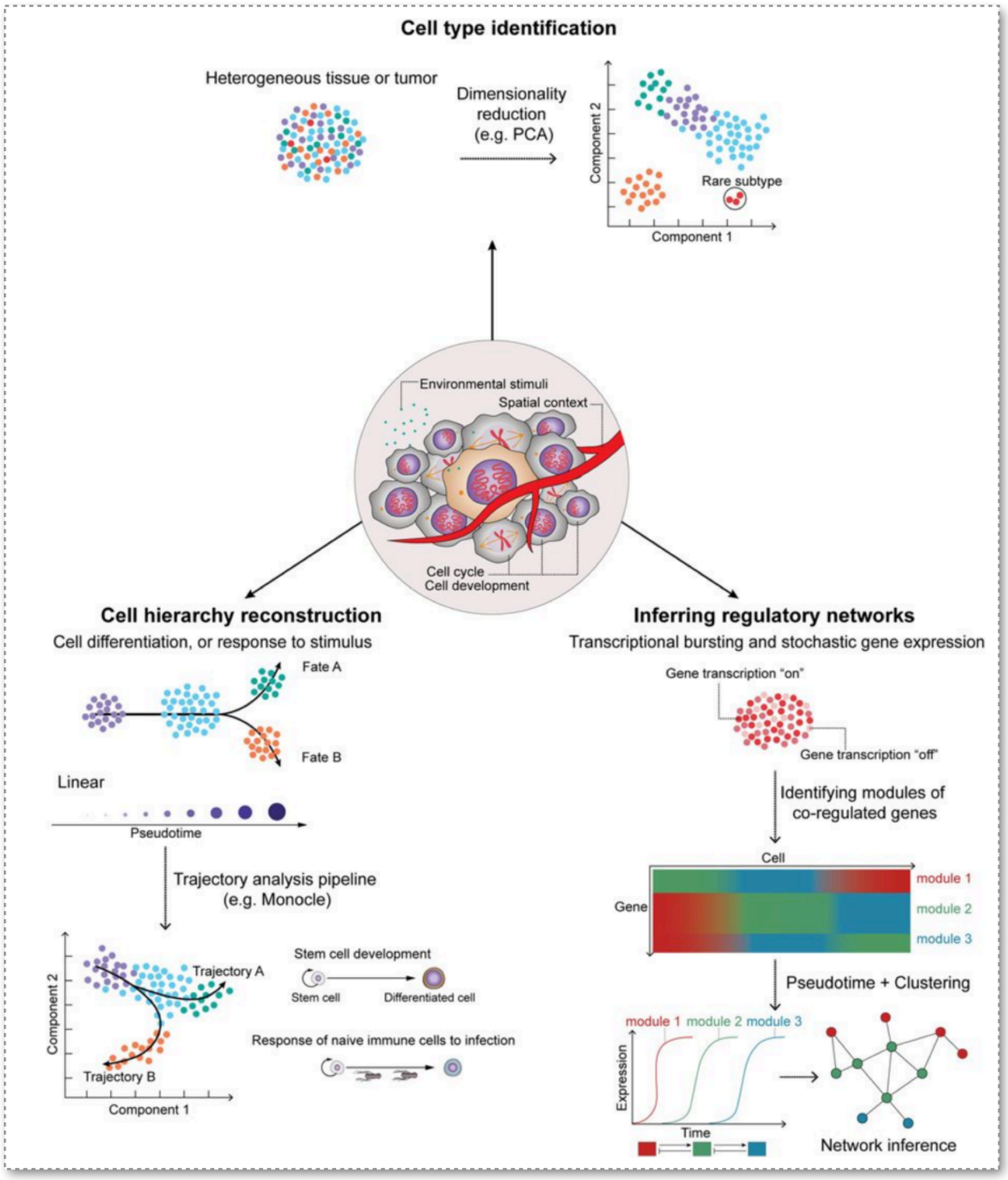
https://www.slideshare.net/QIAGENscience/why-scientist-analyze-single-cells (smoothie)
https://community.10xgenomics.com/t5/10x-Blog/Single-Cell-RNA-Seq-An-Introductory-Overview-and-Tools-for/ba-p/547 (bar chart images)

Why is single-cell exciting?

Explore biology at *unprecedented resolution*.
Some *transformative* applications:

- Study *treatment-resistant* cells in disease (cancer)
- Understand tissue / organism *development* (cell fate)
- Understand immune response at the cellular level
- Understand dynamic cellular processes
- Learn how expression is regulated (regulatory net.)
- Characterize new cell types & cell states (HCA)





Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, *50*(8), 96.
https://www.broadinstitute.org/research-highlights-human-cell-atlas

Example of the types of large-scale questions this tech. Allows

ARTICLE

https://doi.org/10.1038/s41586-019-0969-x

The single-cell transcriptional landscape of mammalian organogenesis

Junyue Cao[1,2,10], Malte Spielmann[1,10], Xiaojie Qiu[1,2], Xingfan Huang[1,3], Daniel M. Ibrahim[4,5], Andrew J. Hill[1], Fan Zhang[6], Stefan Mundlos[4,5], Lena Christiansen[6], Frank J. Steemers[6], Cole Trapnell[1,7,8]* & Jay Shendure[1,7,8,9]*

A single-cell view of mouse organogenesis

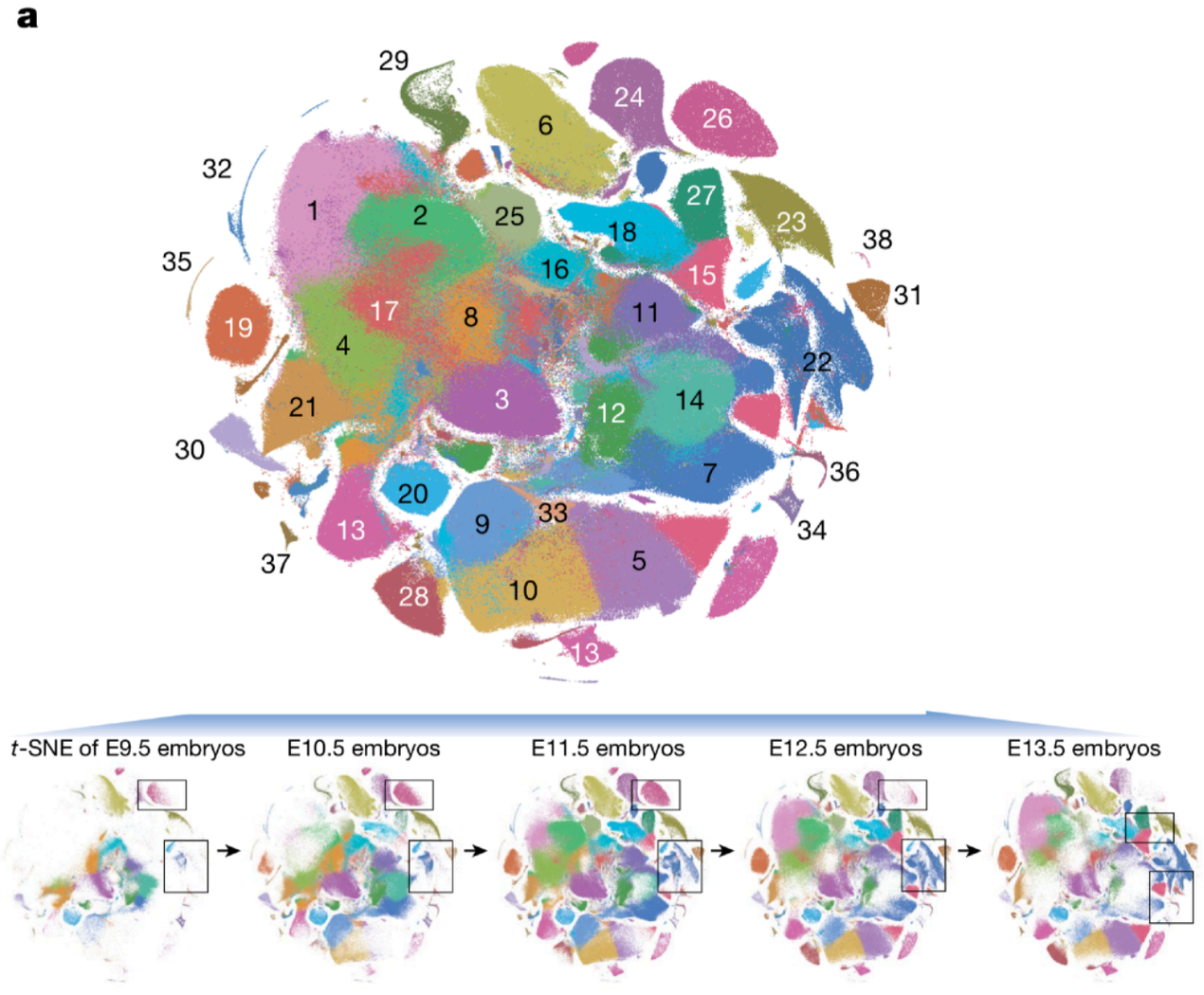
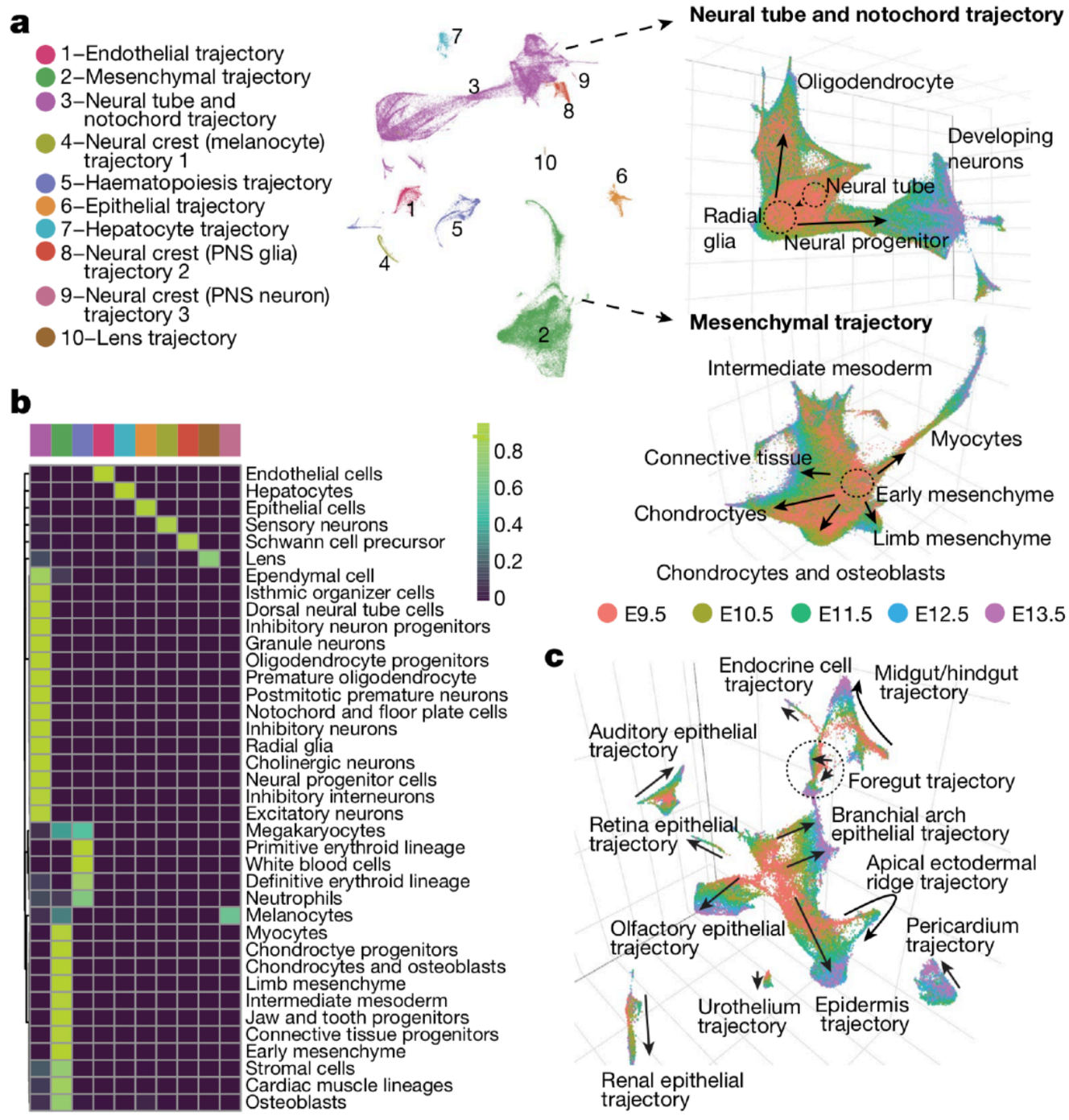


Fig. 2 | Identifying the major cell types of mouse organogenesis. **a**, *t*-SNE visualization of 2,026,641 mouse embryo cells (after removing a putative doublet cluster), coloured by cluster identity (ID) from Louvain clustering (in **b**), and annotated on the basis of marker genes. The same *t*-SNE is plotted below, showing only cells from each stage (cell numbers from left to right: $n = 151{,}000$ for E9.5; 370,279 for E10.5; 602,784

Elucidating development trajectories

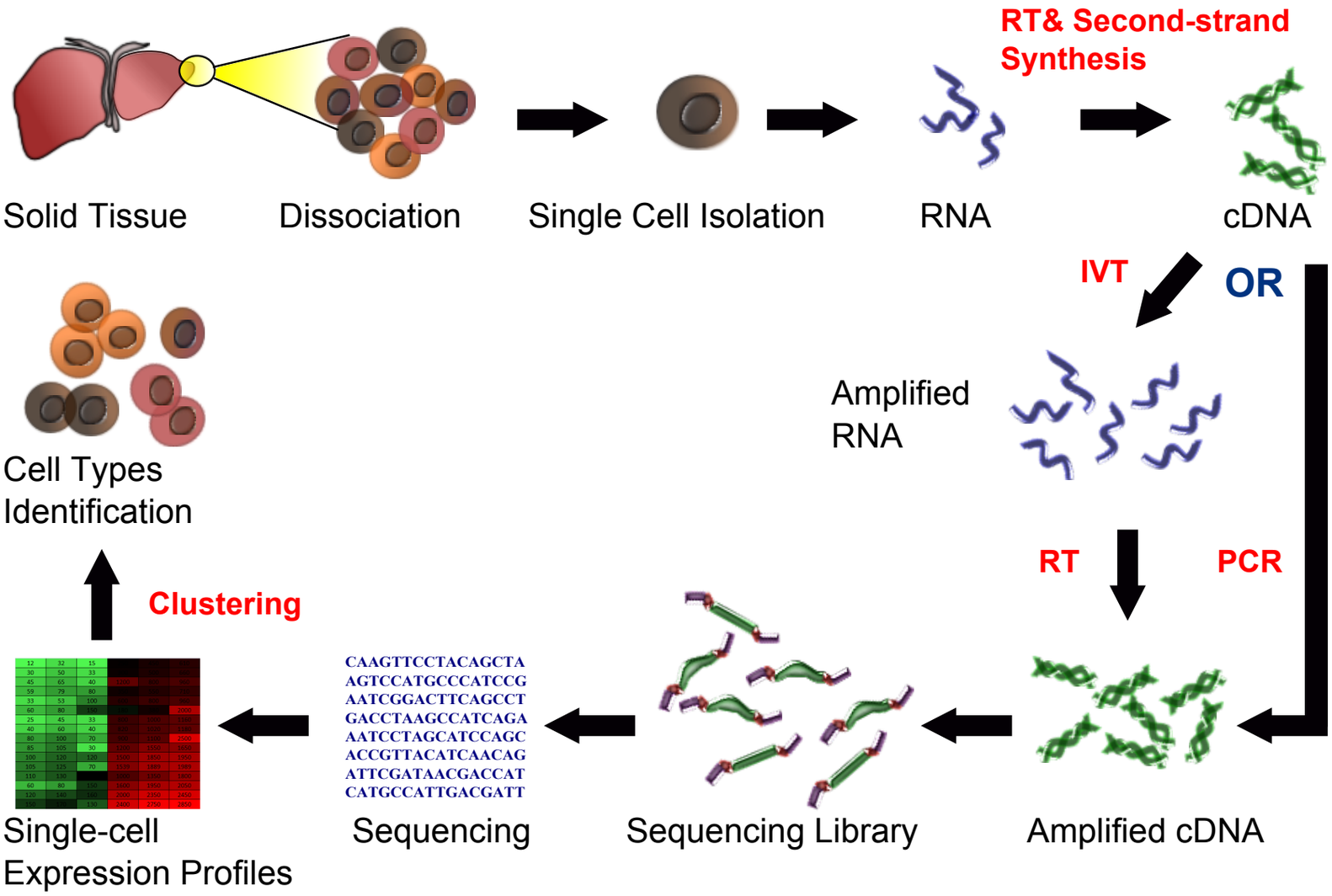


Single Cell Protocol

Single Cell Isolation:

- **Micro-well:** Requires human pipetting & are surface-markable, however has very low-throughput.
- **Micro-fluid (Fluidigm's C1):** Based on an Integrated system which has higher capture-rate than micro-well, however throughput is still low.
- **droplets**: This method has very high-throughput, add cellular-barcoding but has the caveat of higher sequencing cost because of low coverage of transcripts.

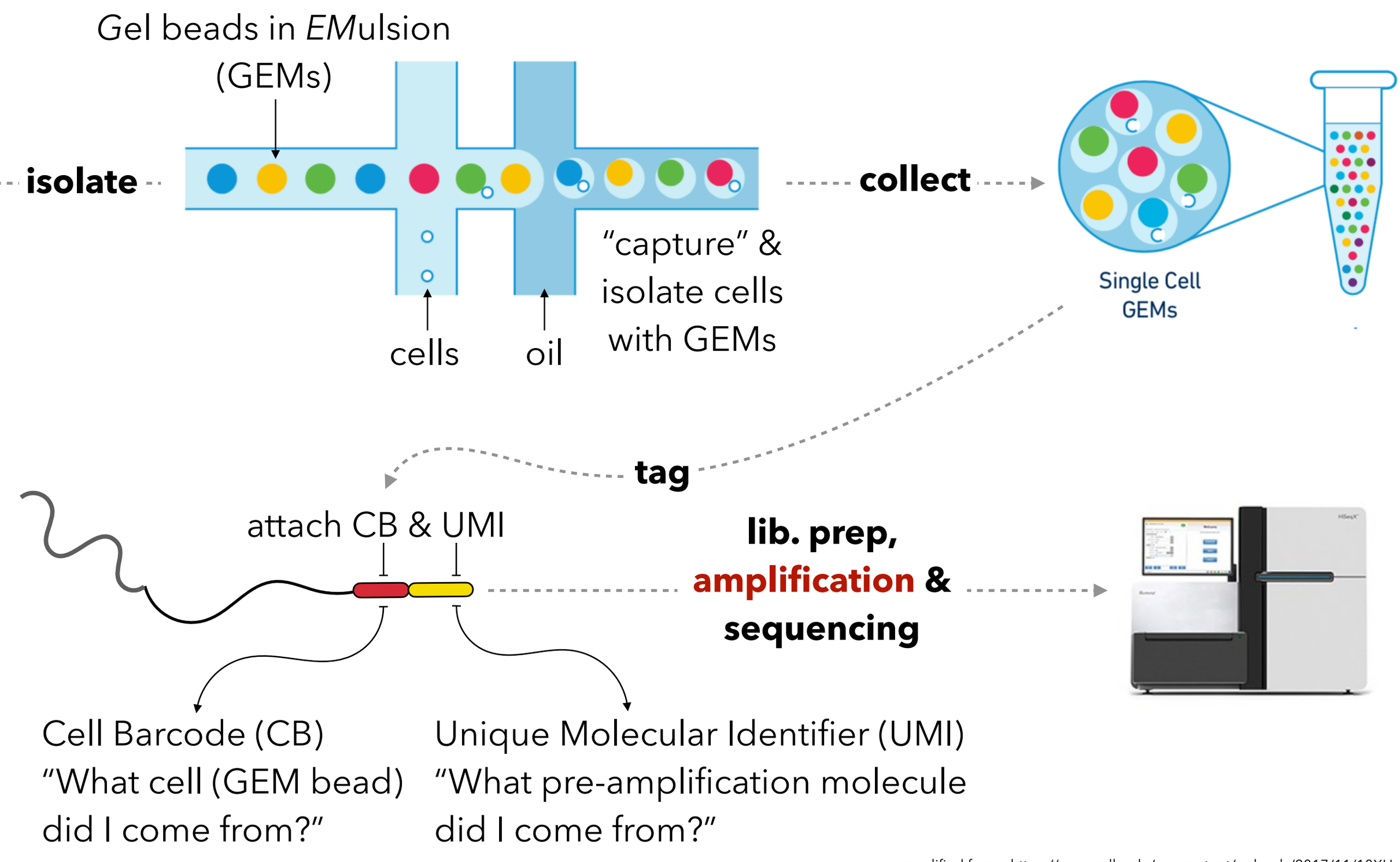
Single Cell RNA-sequencing workflow:



https://en.wikipedia.org/wiki/File:RNA-Seq_workflow-5.pdf

How does it work?

Many different protocols – here is a gist of droplet-based (microfluidic) techniques.



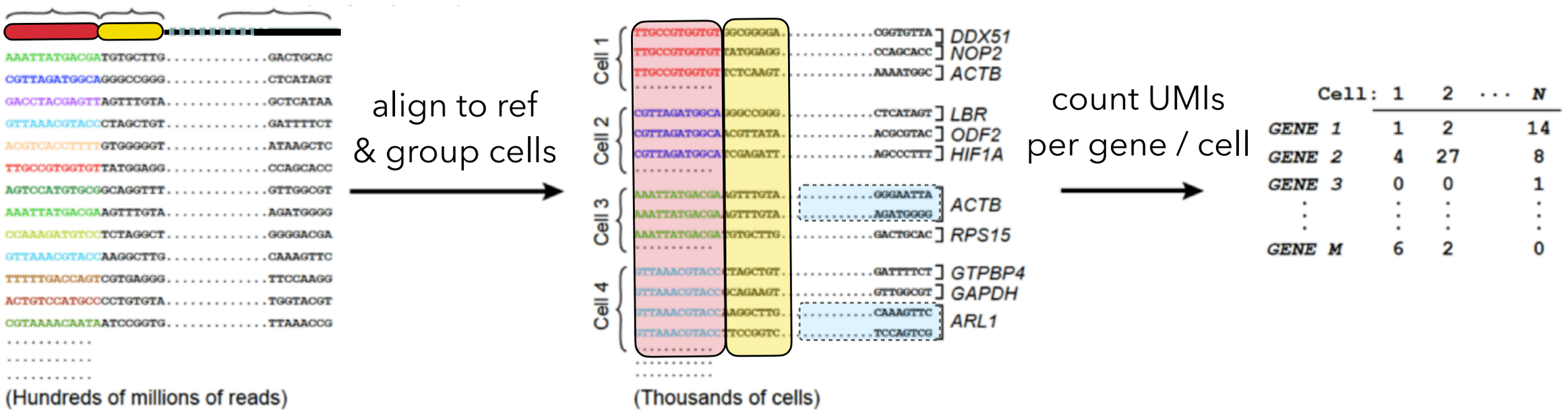
modified from : https://www.salk.edu/wp-content/uploads/2017/11/10XUserGuide.pdf

How does it work?

In theory using the CB & UMI to estimate gene expression is easy:

1. Group sequenced reads by their CB (read → cell)
2. Collapse all reads with the same UMI (read in cell → molecule)
3. Count!

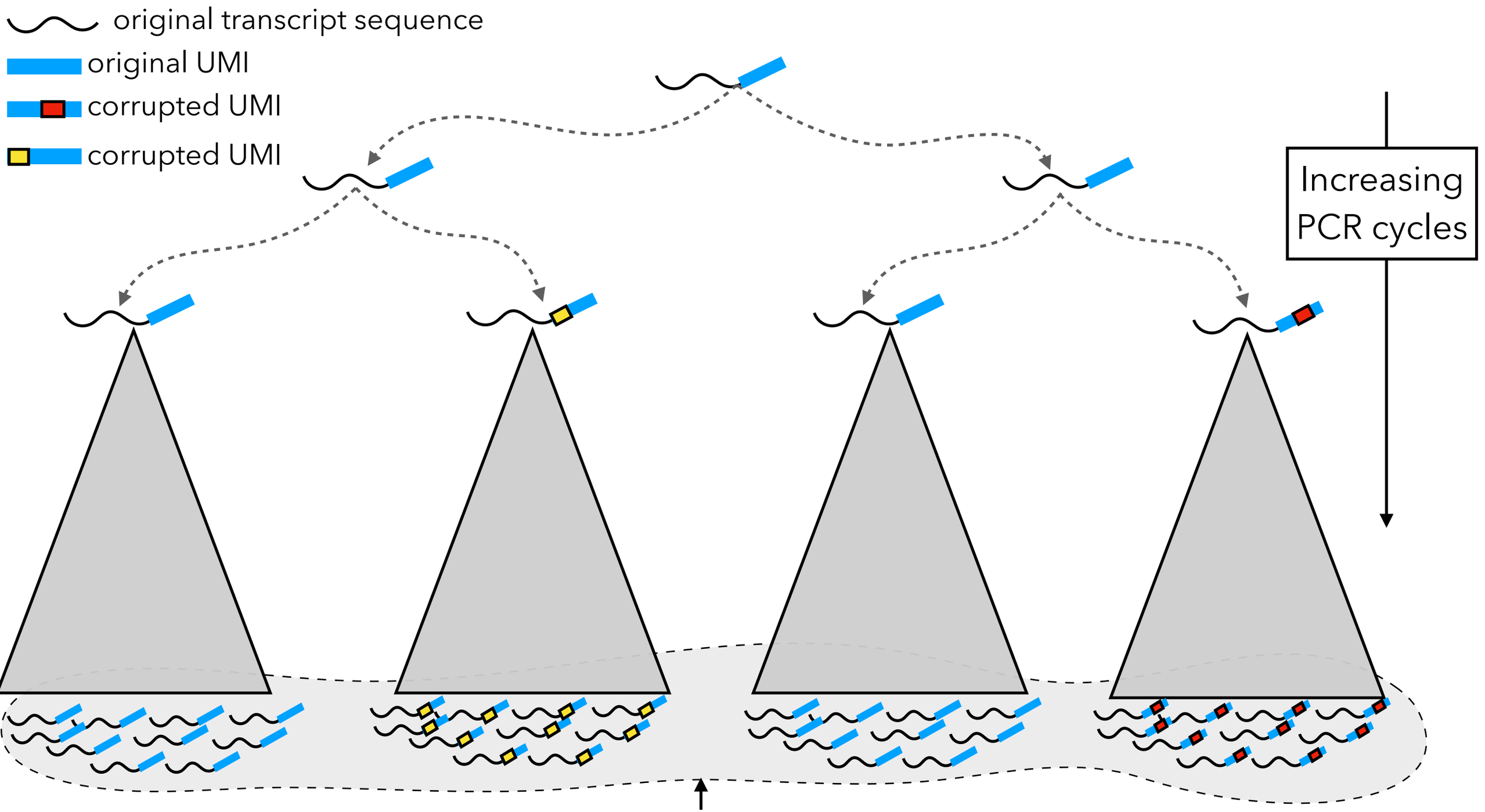
… how many lines is that in Pandas?



In practice the process is ***much*** more complicated

http://mccarrolllab.org/wp-content/uploads/2015/05/figure-2D.png

The cost of small input material

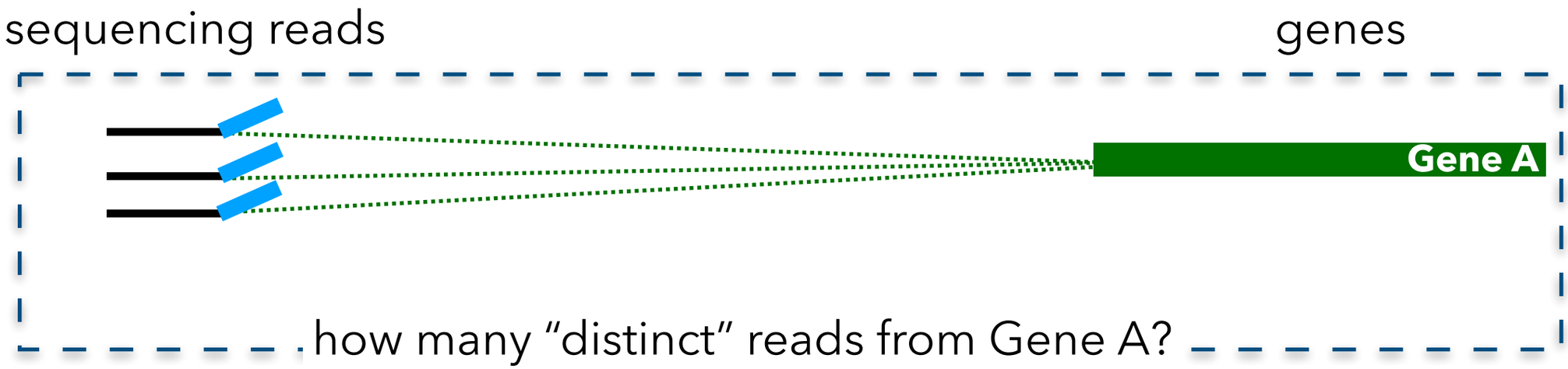


Sample reads from this exponentially amplified pool of molecules

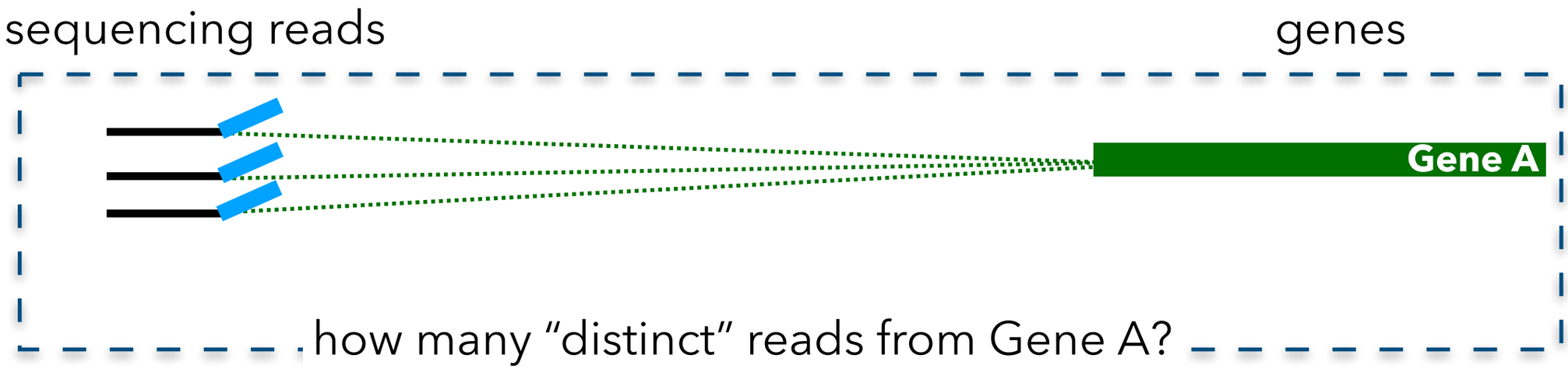
PCR “duplicates” of the same molecule can now have distinct UMIs – but probably *similar*

Must “collapse” similar UMIs, or we will vastly mis-estimate # of original molecules

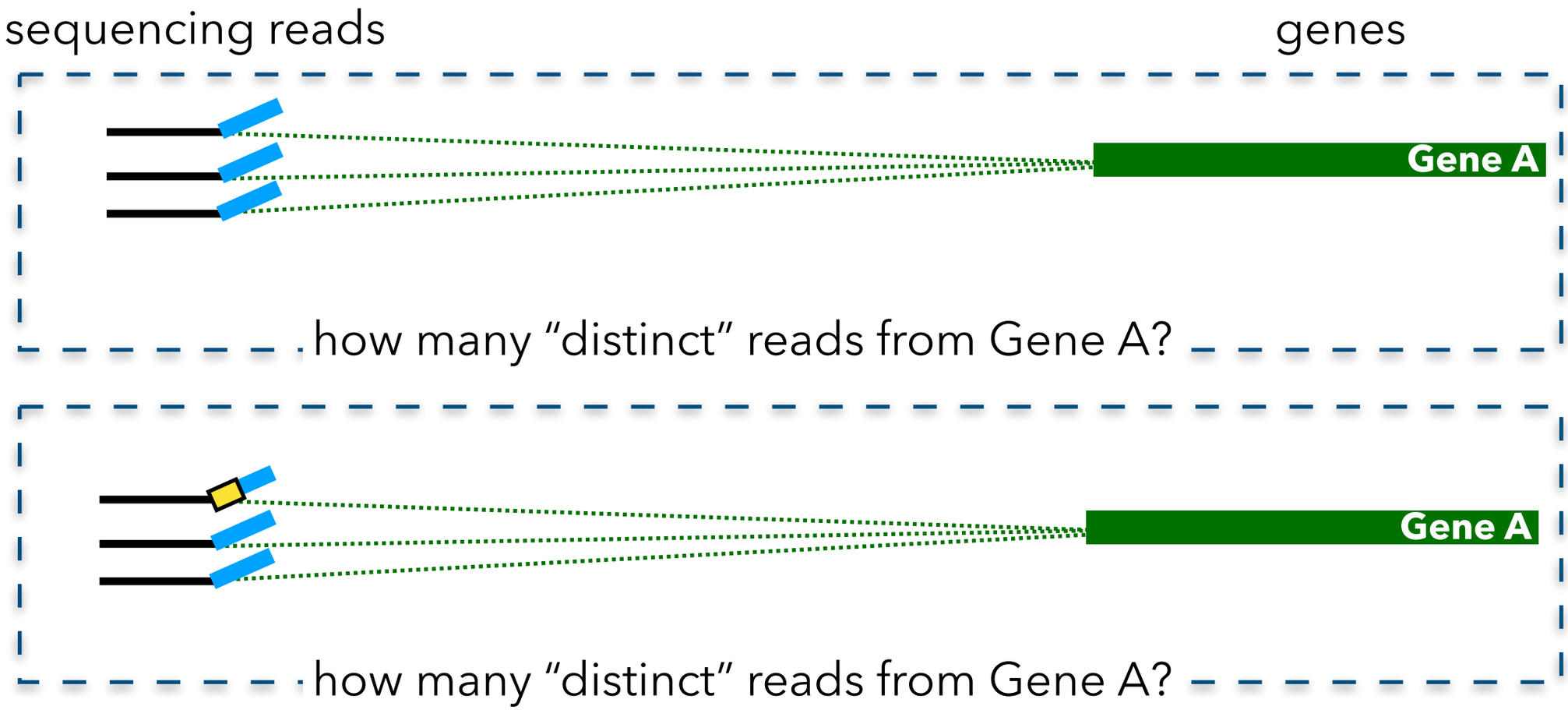
Multimapping is the real culprit



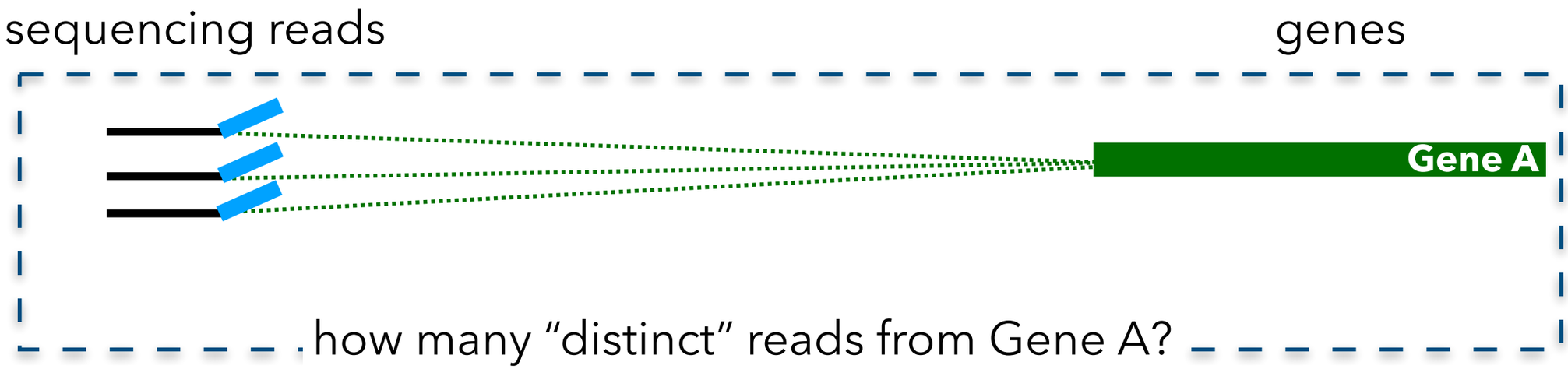
Multimapping is the real culprit



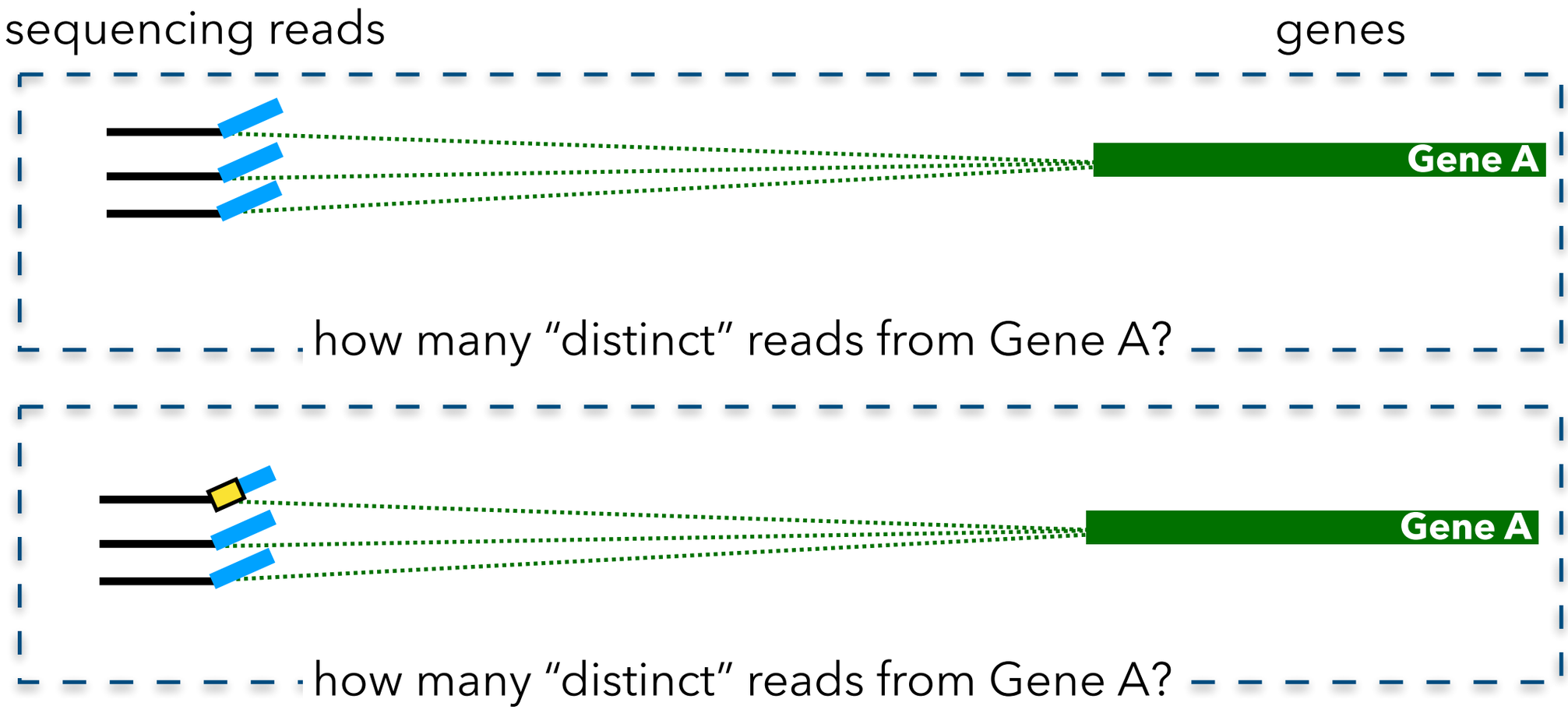
Multimapping is the real culprit



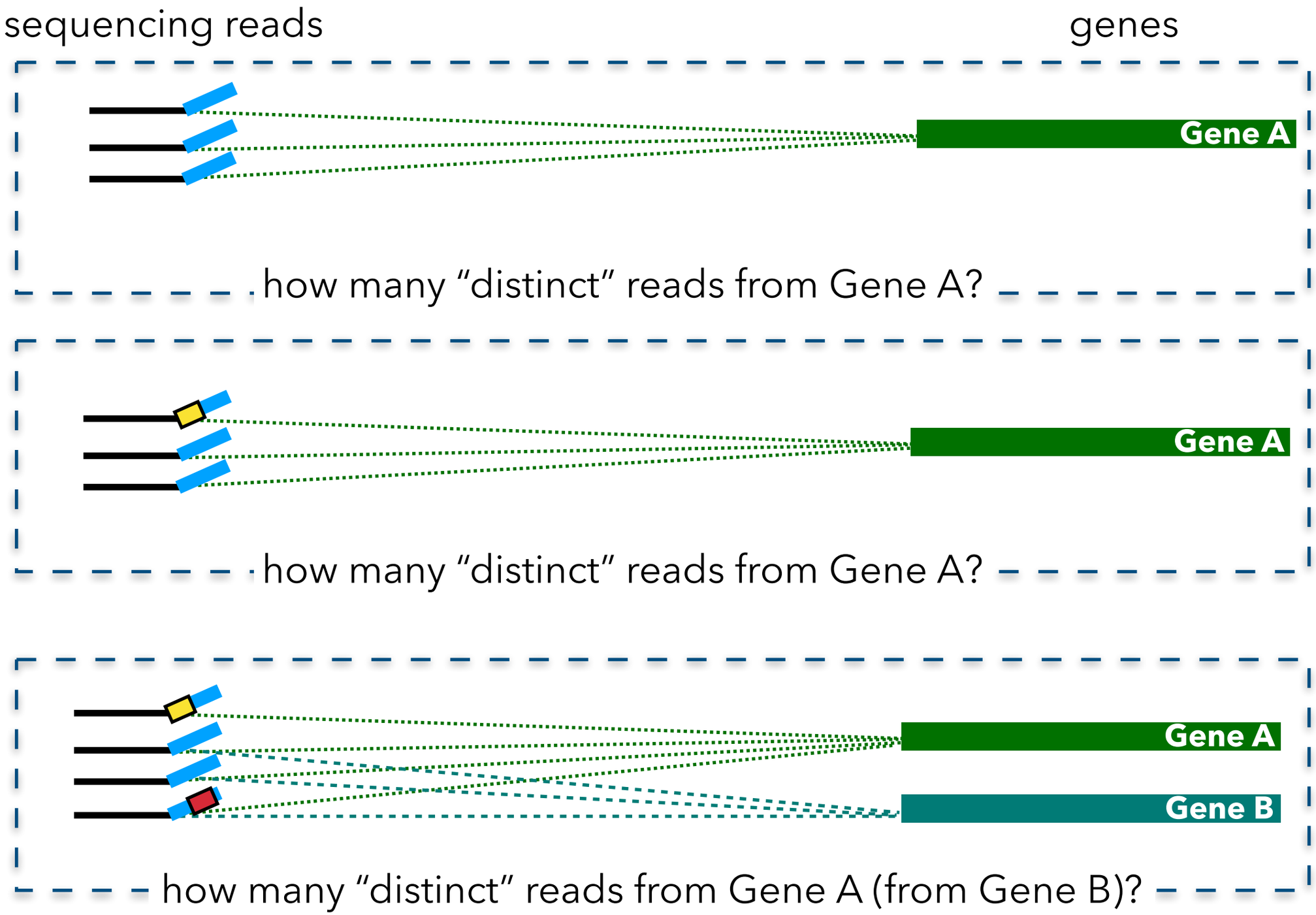
Multimapping is the real culprit



Multimapping is the real culprit



Multimapping is the real culprit



The answer for one depends on the other!

single-cell quantification is challenging

There are *many* reasons single-cell quantification is challenging!

most of these stem from the difficulty of cell isolation & small quantities of genetic material:

- 1 UMI ≠ 1 pre-PCR molecule (often); because of small material per-cell, typical protocols involve many rounds of PCR; UMI tags are subject to PCR and sequencing error.

- 1 Cell Barcode ≠ 1 Cell (sometimes); capture-related issues (e.g. doublets & empty droplets)

- "dropout" & bias due to sampling of relatively small number of reads from highly-amplified pool of original molecules

- UMI "collisions" are possible due to limited UMI pool

One shortcoming with current tagged-end scRNA-seq processing techniques

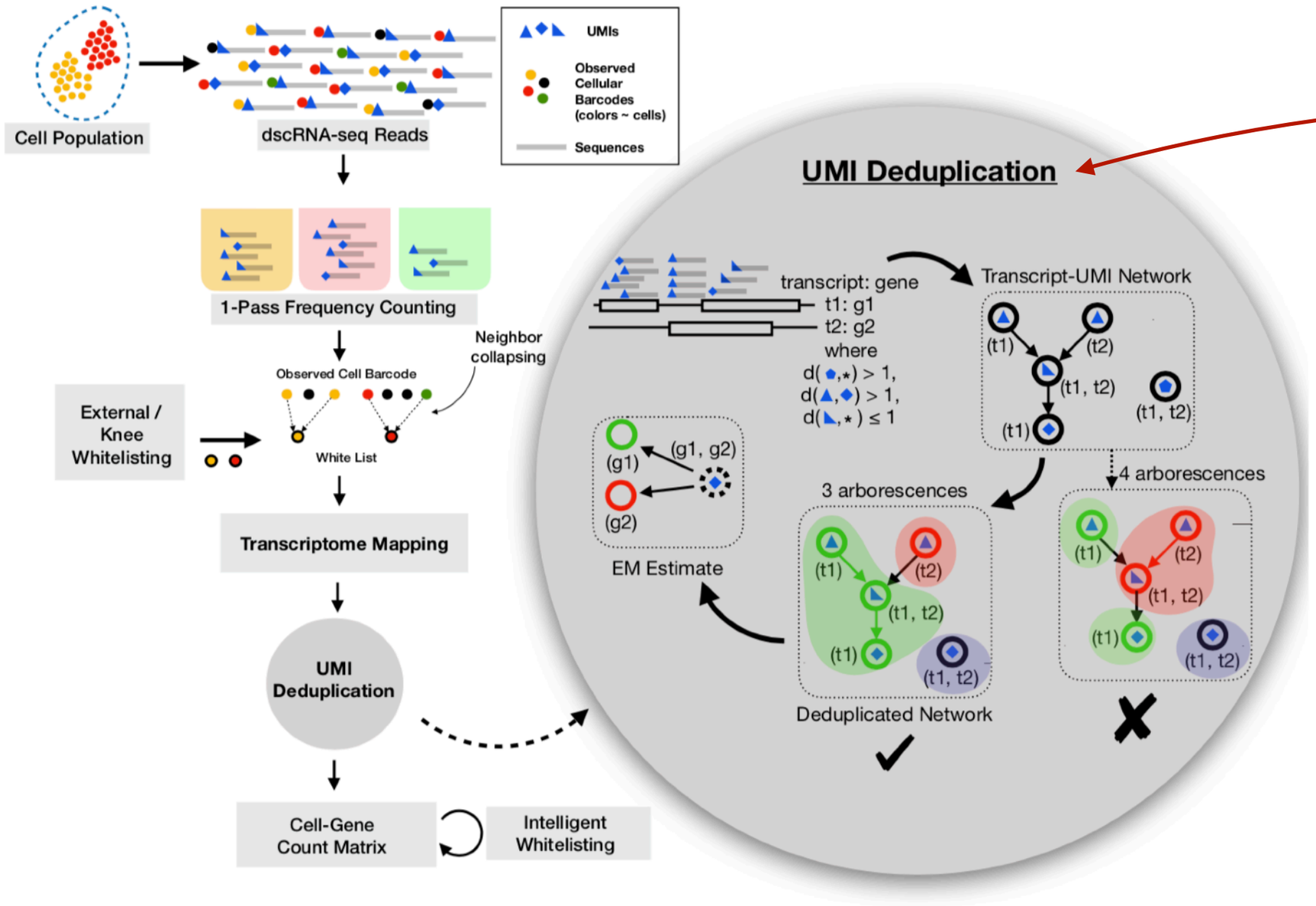
Unfortunately, current approaches have no principled way to deal with reads that map between multiple genes … simply **discard them**.

This may seem like a minor issue, but, in a typical dataset, this is 13-23% of the reads!

Sample	Percentage
Human PBMC 4k	13.8
Human PBMC 8k	13.8
Mouse Neurons 900	21.6
Mouse Neurons 2k	22.5
Mouse Neurons 9k	17.1

Srivastava, Avi, et al. "Alevin efficiently estimates accurate gene abundances from dscRNA-seq data." *Biorxiv* 1: 12.

alevin: dscRNA-seq done right



Here: I will only talk about how we model & solve this problem

Implemented as part of our salmon tool

Alevin efficiently estimates accurate gene abundances from dscRNA-seq data

Avi Srivastava *[1], Laraib Malik †[1], Tom Smith ‡[2], Ian Sudbery §[3], and Rob Patro ¶‖[1]

[1] *Department of Computer Science, Stony Brook University, Stony Brook, USA*
[2] *Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, CB2 1GA Cambridge, UK*
[3] *Sheffield Institute for Nucleic Acids, Department of Molecular Biology and Biotechnology, The University of Sheffield, Sheffield, S10 2TN, UK*

COMBINE-lab / salmon

View Repository | Unwatch 39 | Unstar 249 | Fork 80

Code | Issues 106 | Pull requests 1 | Projects 0 | Wiki | Insights | Settings

Highly-accurate & wicked fast transcript-level quantification from RNA-seq reads using lightweight alignments

Edit

https://combine-lab.github.io/salmon

quasi-mapping | bioinformatics | rna-seq | rnaseq | salmon | quantification | sailfish | c-plus-plus | gene-expression | scrna-seq | single-cell | 10x

Parsimonious UMI Graph (PUG) resolution

Represent the UMIs / transcripts relationship as a graph G = (V, E) .

Each **vertex** (v ∈ V) is a tuple:

- **eq_v** : the set of transcripts to which the read maps
- **s_v** : the UMI sequence tagging the read

Each v has count, *c*(v) – number of "equivalent" reads.

"equivalent" means aligns to same transcripts and has same UMI.

There is an **edge** (e = {u,v} ∈ E) if, for some chosen edit distance τ :

- $s_u = s_v$ and $|eq_u \cap eq_v| > 0$ (*bidirected* edge)
- $d(s_u, s_v) < \tau$, $c(u) \sim c(v)$ and $|eq_u \cap eq_v| > 0$ (*bidirected* edge)
- $d(s_u, s_v) < \tau$, $c(u) > 2c(v)+1$ and $|eq_u \cap eq_v| > 0$ (*directed* u → v)

Smith, T. S., Heger, A., & Sudbery, I. (2017). UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research*

A parsimony-guided approach

We will attempt to *explain* the PUG using the *minimum number* of pre-PCR molecules. More formally:

Given: UMI-resolution graph G = (V, E)

Find: a *minimum cardinality* cover by *monochromatic arborescences*.

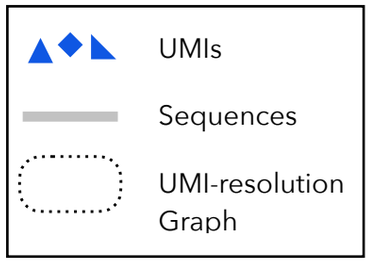
We seek <u>parsimony</u>

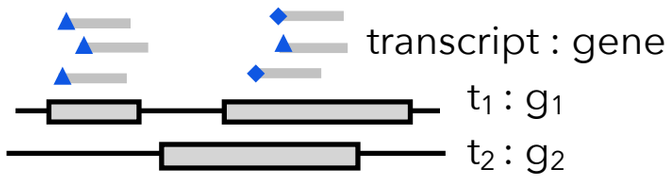
Each component can be ascribed to the reads sequenced from a *single initial molecule* (before amplification).

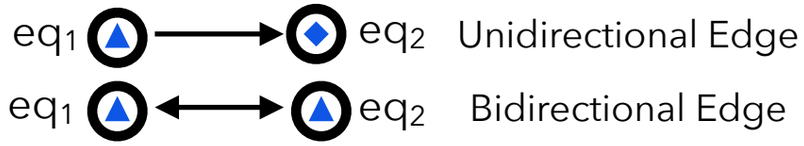
Each *monochromatic arborescence* is a set of vertices that can all be described as "reads" coming from the same, original transcript.

The decision problem is *NP-complete* (reduction from DOMINATING SET); but experimental instances are usually *simple* and we adopt a *greedy heuristic*. (solutions are known to be optimal in > 80% of cases)

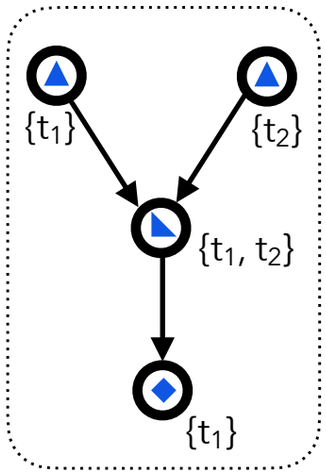
UMI Resolution



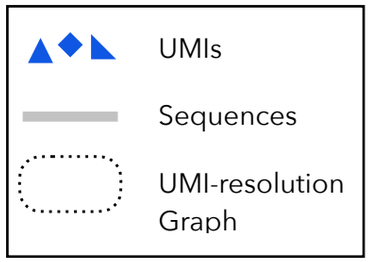


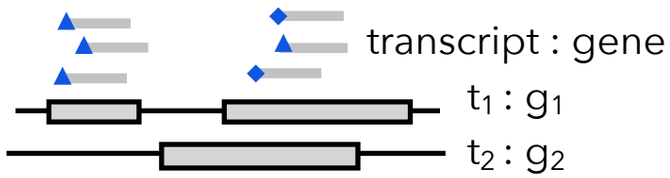


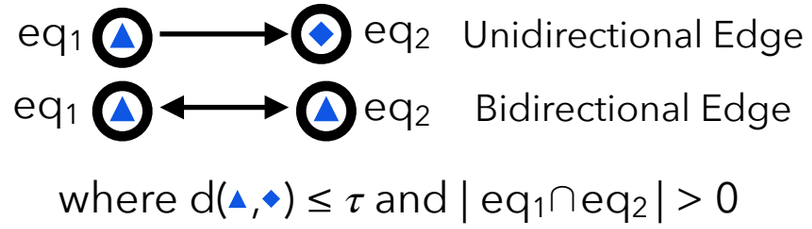
where d(▲,◆) ≤ τ and | eq$_1$∩eq$_2$ | > 0

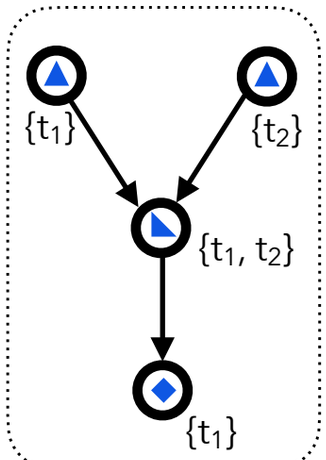


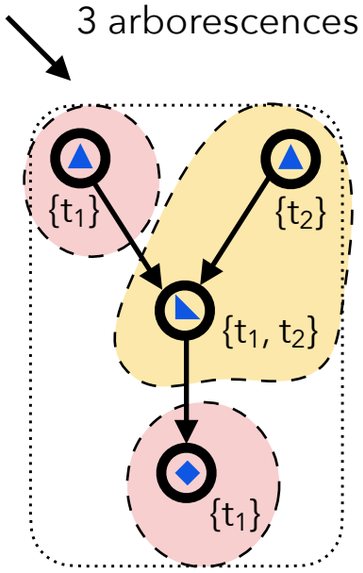
UMI Resolution



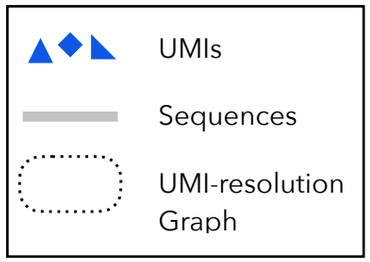


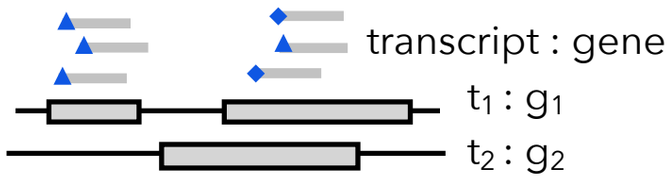


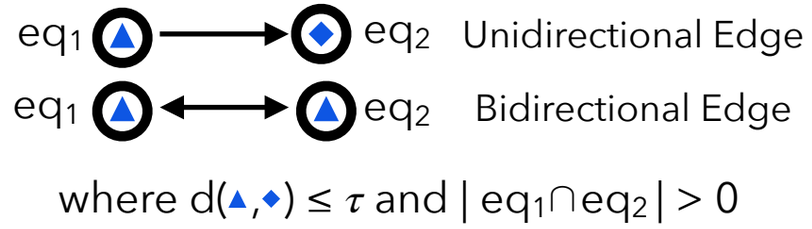


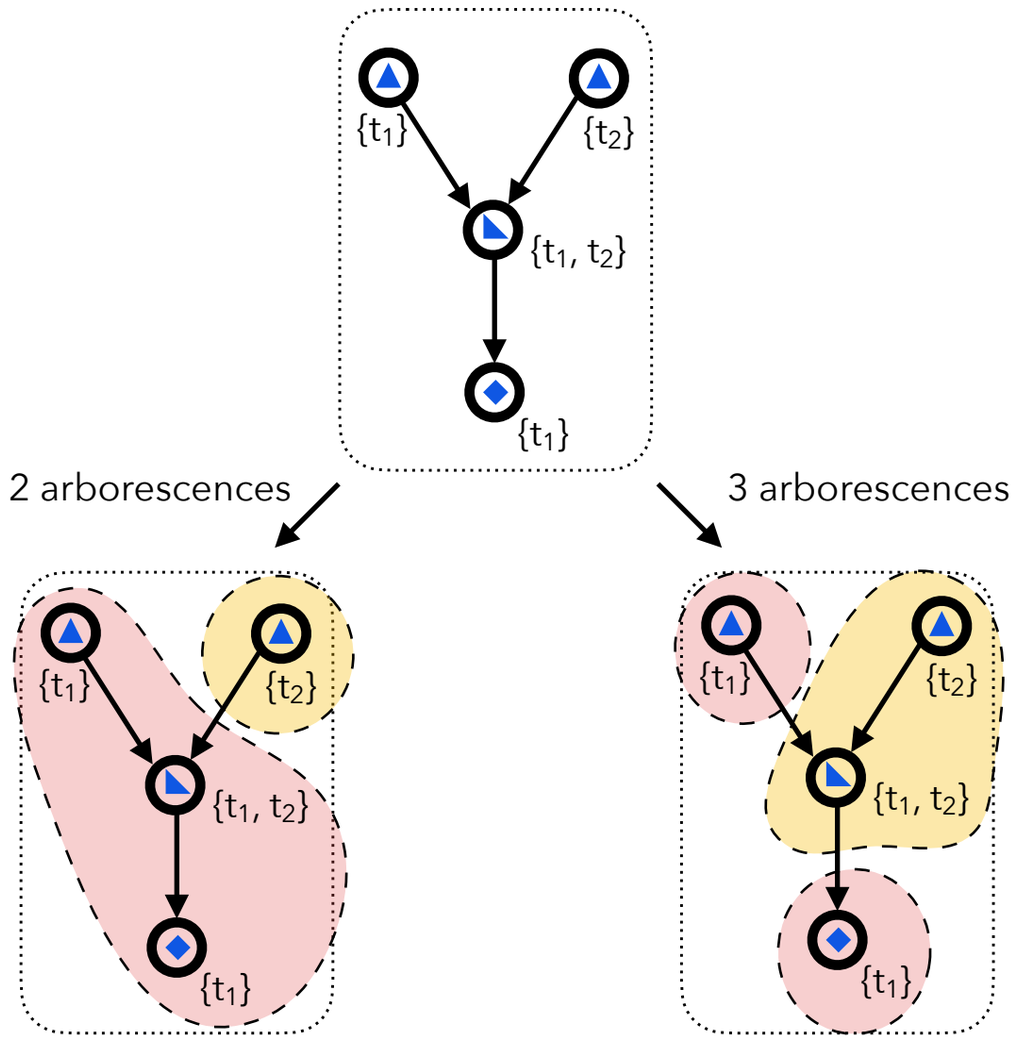


UMI Resolution

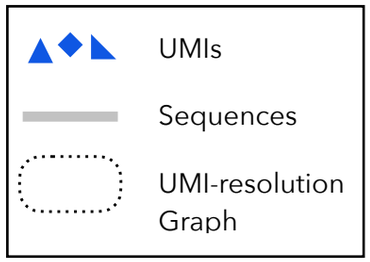


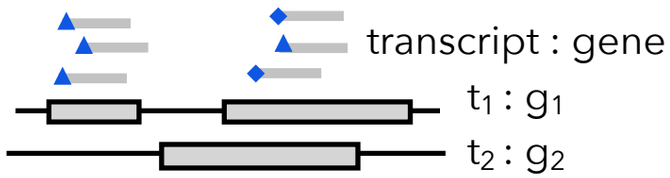


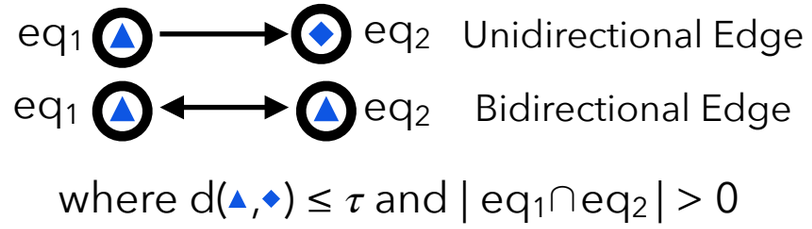


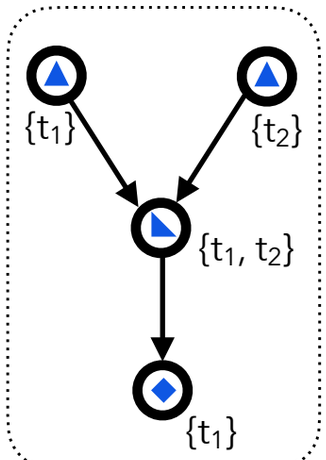


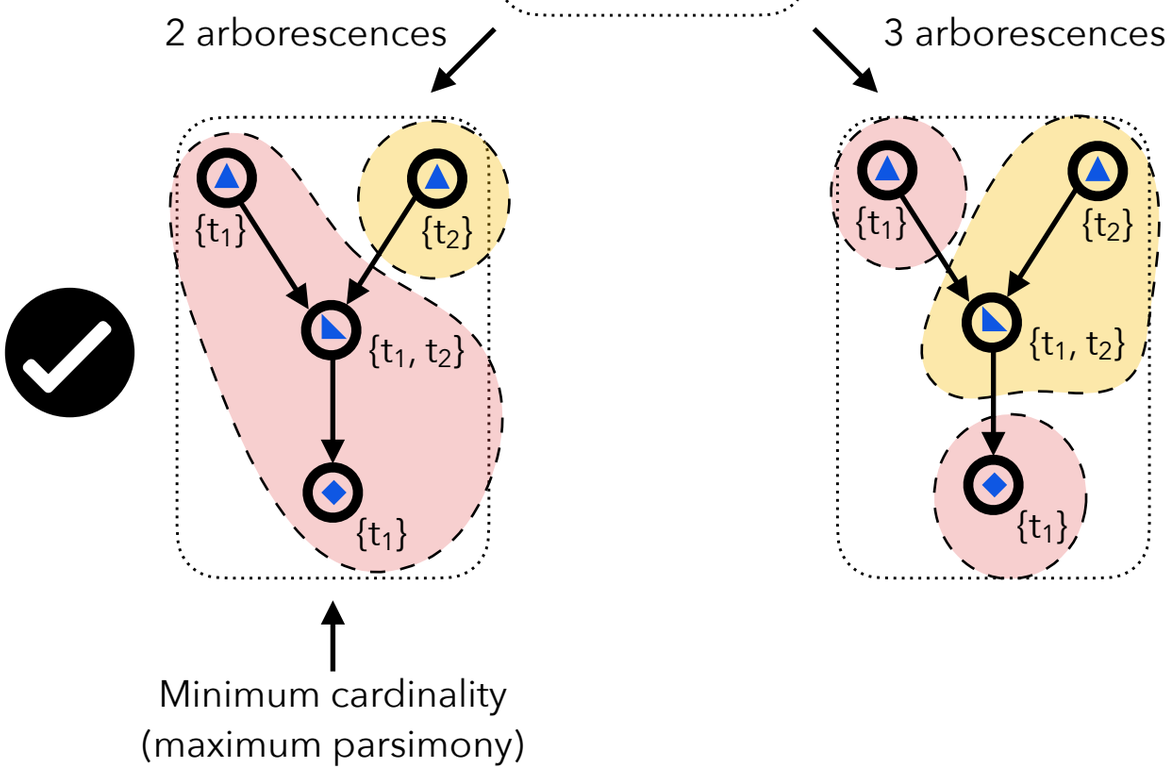
UMI Resolution



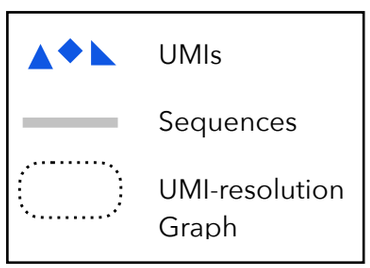


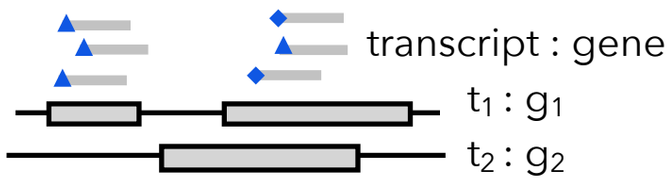


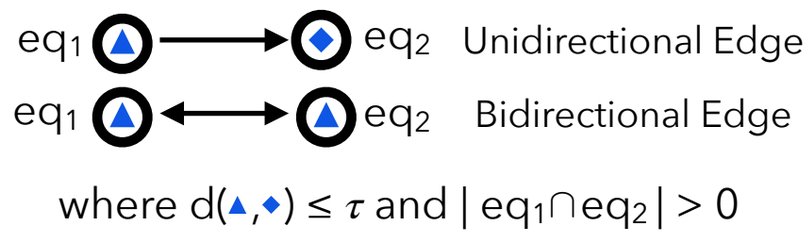


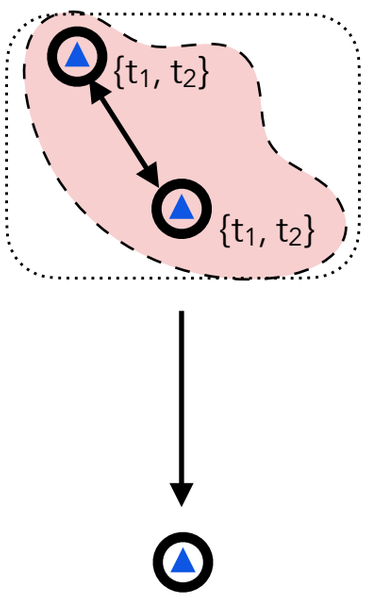


UMI Resolution





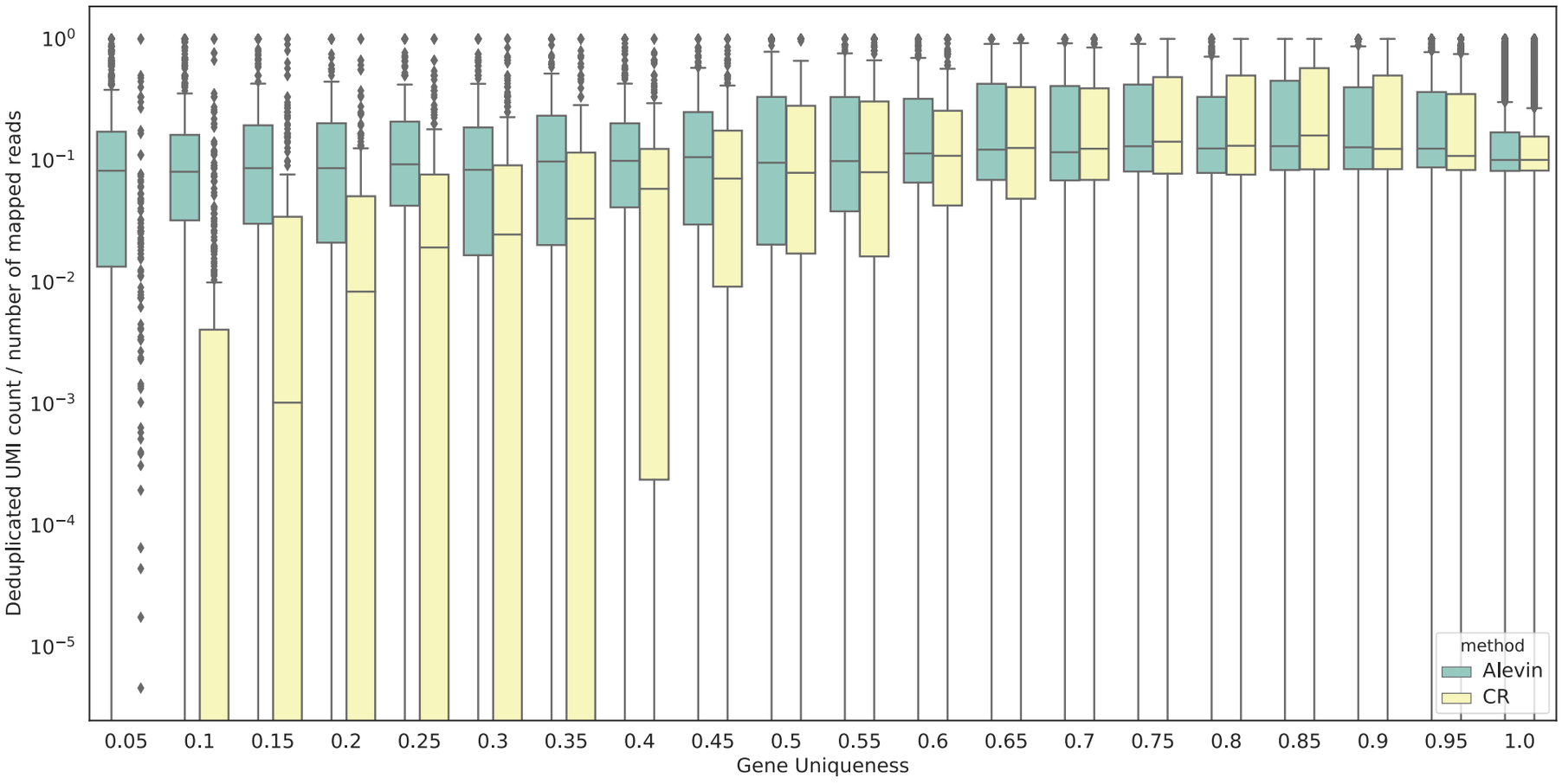




Equally-parsimonious under t1 **or** t2

- Parsimony cannot resolve the gene of origin here
- We treat UMI as gene-ambiguous & resolve via EM-algorithm

Discarding gene-ambiguous reads does not affect all genes equally

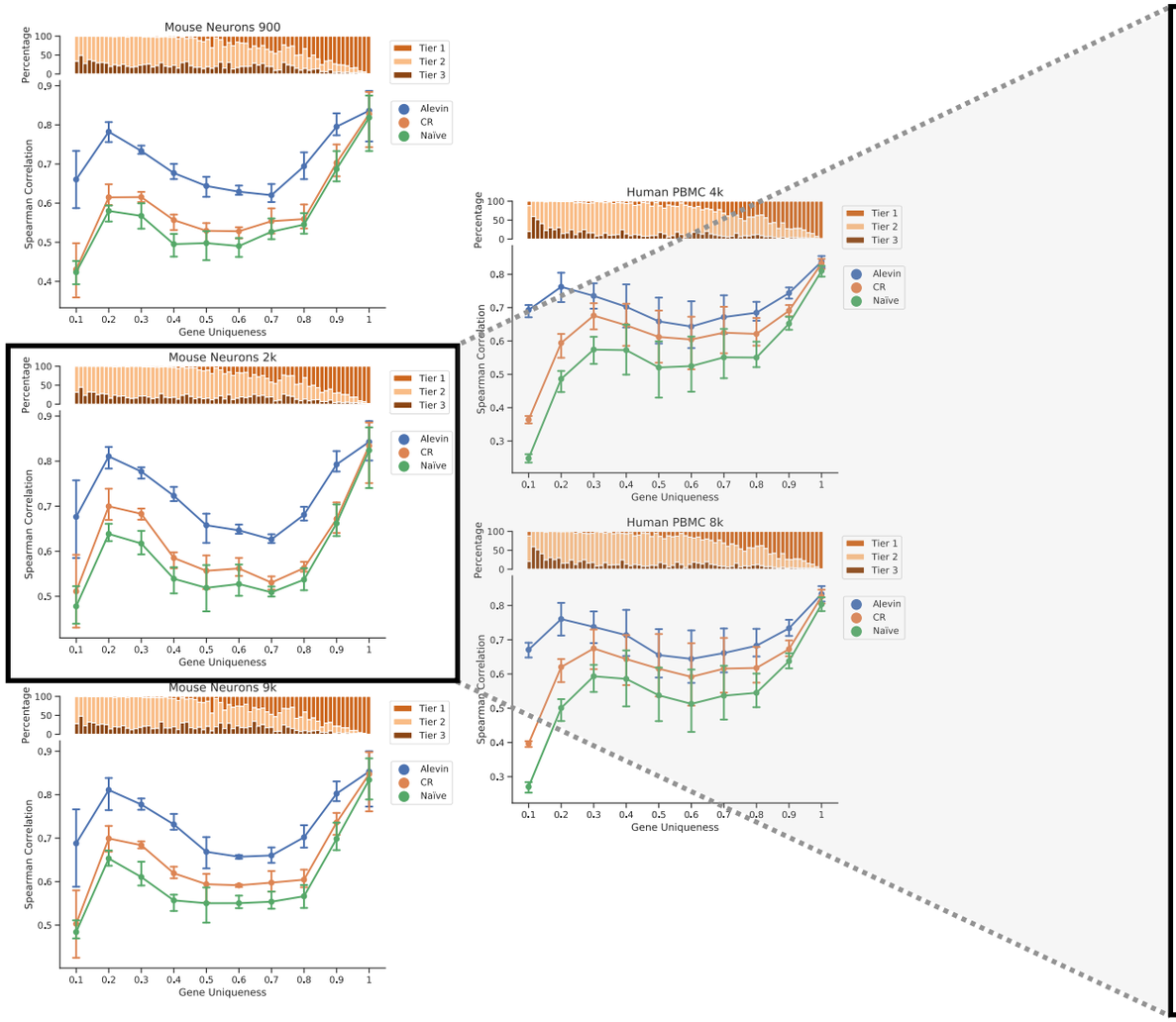


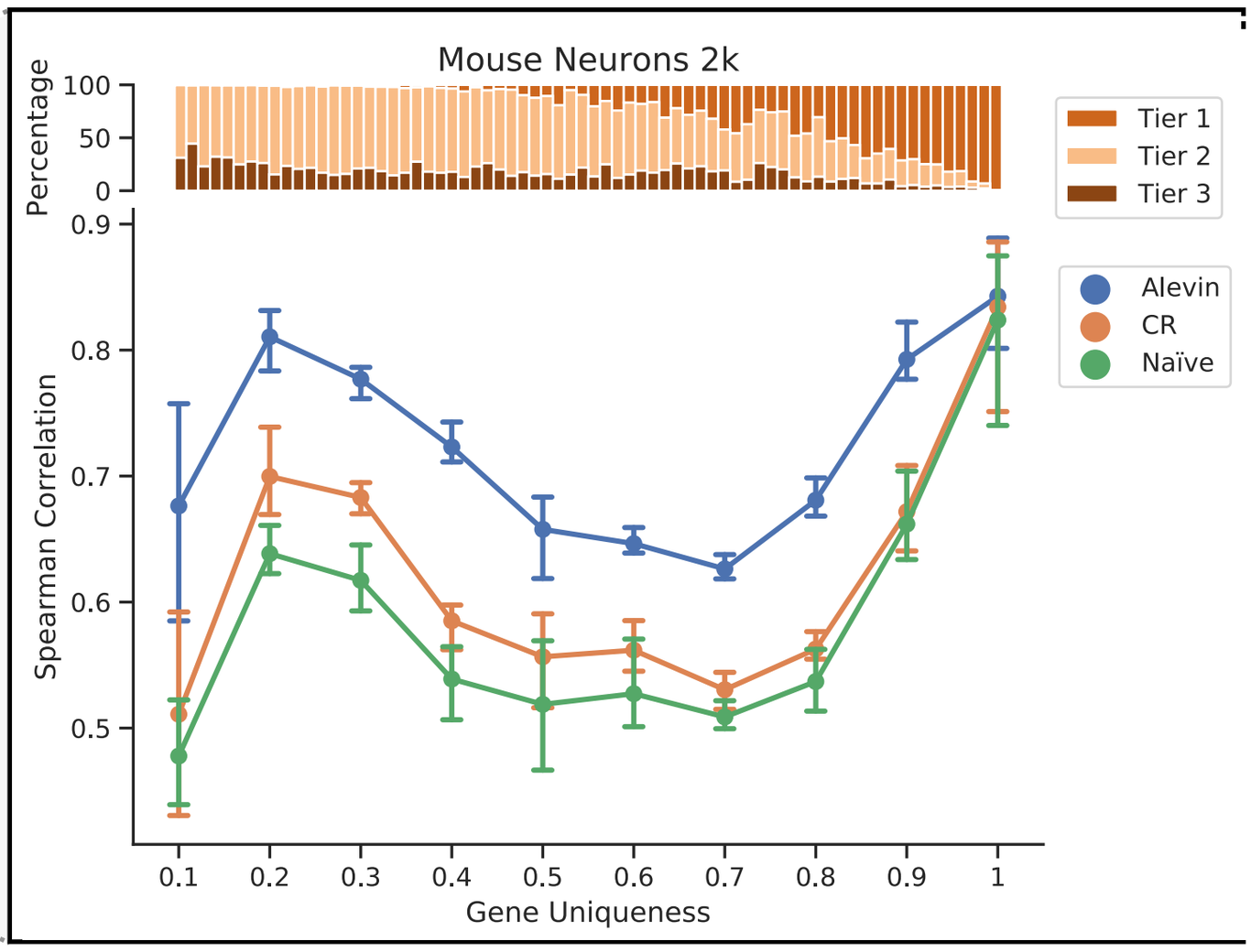
Bin number	Human	Mouse
1	2538	3984
2	1094	1398
3	813	996
4	852	944
5	908	1137
6	1044	1178
7	1339	1392
8	1943	1615
9	3360	2226
10	44397	38509

Stratify genes by sequence uniqueness and look at distribution of # of de-duplicated UMIs per input read (PBMC 4k). Large effect for genes with lots of sequence homology (including important paralog families). Thus, the discarding of gene multi-mapping reads leads to systematic bias.

Accounting for gene-ambigous improves correlation with bulk RNA-seq

Trend across 5 public 10x datasets+





For each scRNA-seq dataset, we obtained multiple bulk samples from the same tissue. We compare gene-level quants in bulk (quantified with RSEM*) to average expressions across single cells.

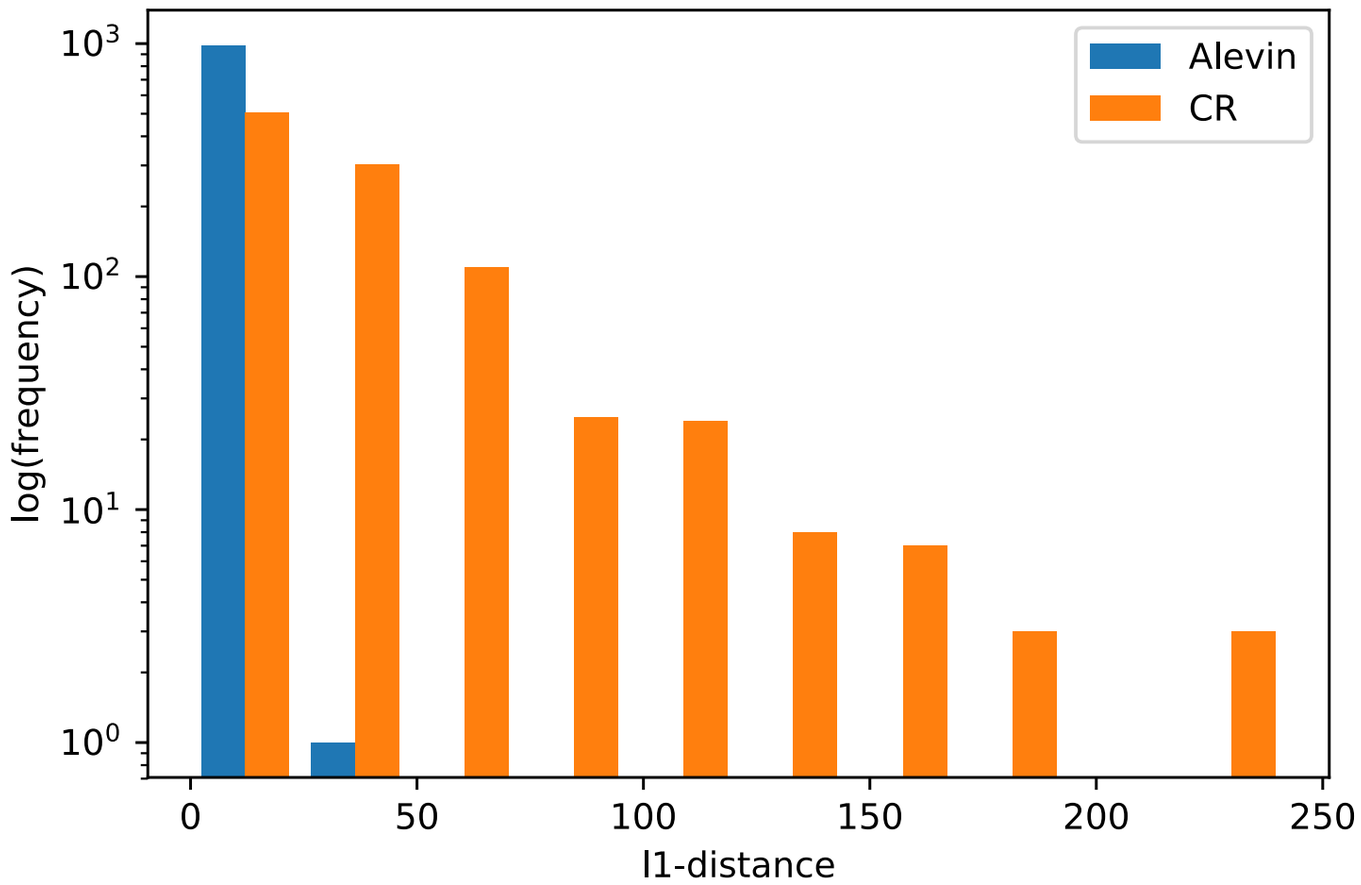
*Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, *12*(1), 323.

+Zheng, Grace XY, et al. "Massively parallel digital transcriptional profiling of single cells." Nature communications 8 (2017): 14049.

Robustness to *in silico* homology

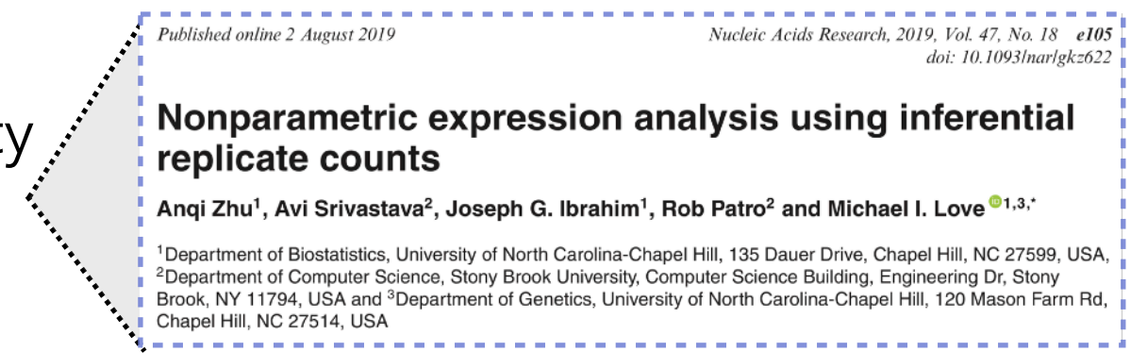
Compare quantifications of a mouse neuron sample:

aligned and quantified against **mouse reference** vs. **mouse + human reference**



> "A cow is … basically a human" (paraphrased)
> — Bruce Futcher (yeast biologist, extraordinaire)

Also: alevin can provide <u>bootstrap</u> uncertainty estimates of gene counts to be used in downstream analysis



Published online 2 August 2019 — *Nucleic Acids Research, 2019, Vol. 47, No. 18* **e105**
doi: 10.1093/nar/gkz622

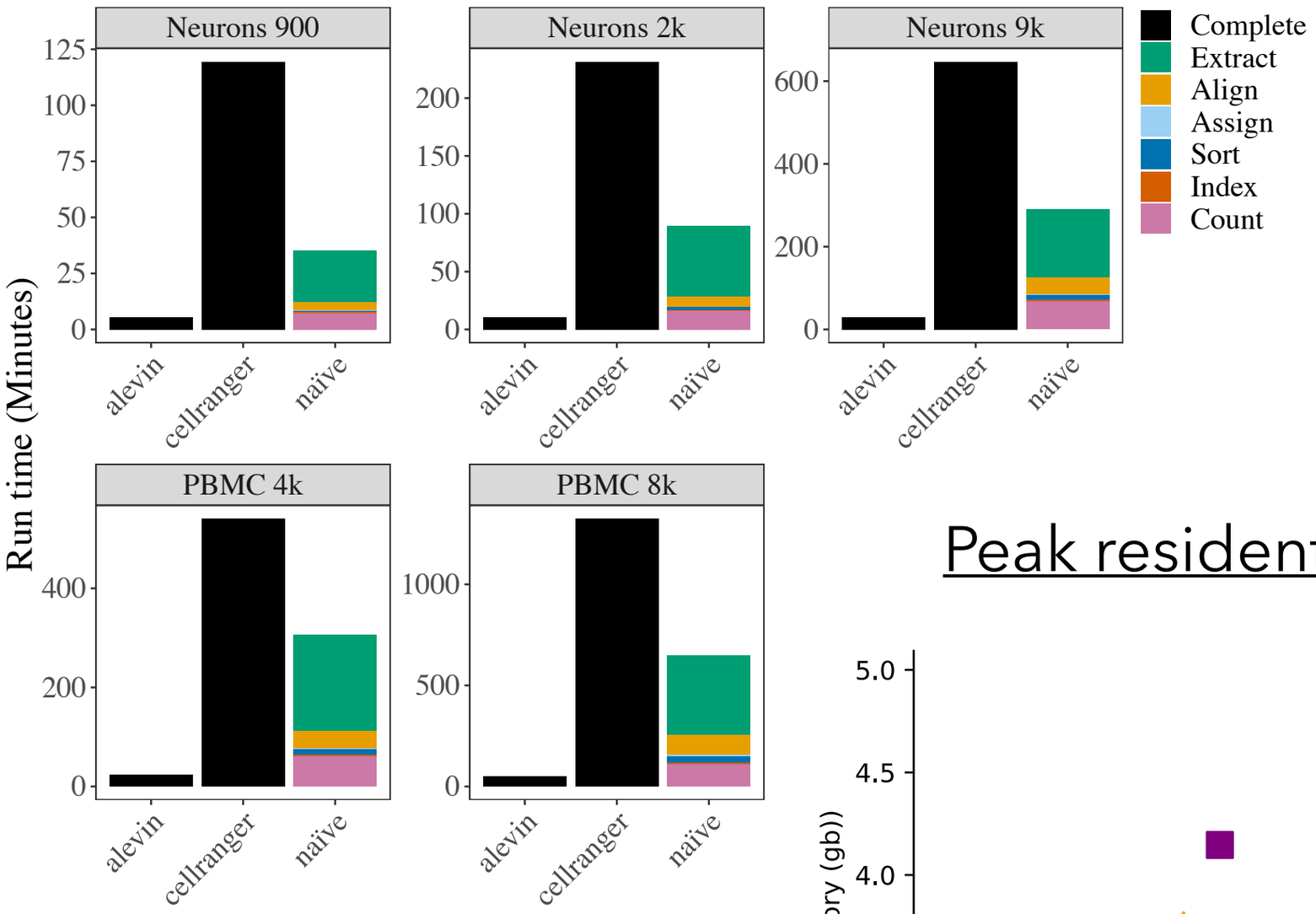
Nonparametric expression analysis using inferential replicate counts

Anqi Zhu[1], Avi Srivastava[2], Joseph G. Ibrahim[1], Rob Patro[2] and Michael I. Love[1,3,*]

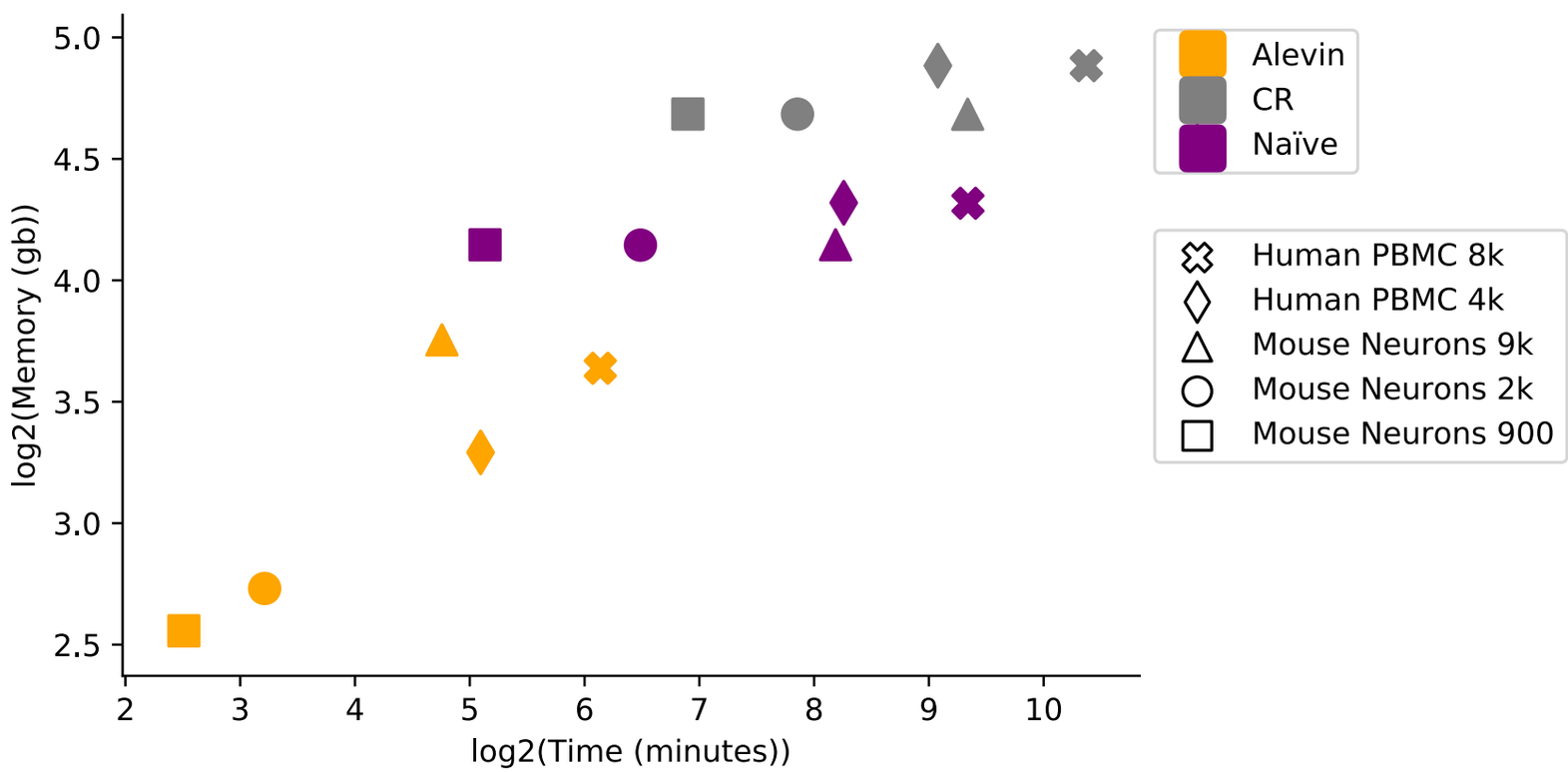
[1]Department of Biostatistics, University of North Carolina-Chapel Hill, 135 Dauer Drive, Chapel Hill, NC 27599, USA, [2]Department of Computer Science, Stony Brook University, Computer Science Building, Engineering Dr, Stony Brook, NY 11794, USA and [3]Department of Genetics, University of North Carolina-Chapel Hill, 120 Mason Farm Rd, Chapel Hill, NC 27514, USA

alevin is fast & efficient

Wall clock time (16 threads)



Peak resident memory (16 threads)



what's missing ?

Simulations

Bulk RNA-seq

RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome

Bo Li and Colin N Dewey

Modelling and simulating generic RNA-Seq experiments with the flux simulator

Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, Michael Sammeth Author Notes

***Polyester*: simulating RNA-seq datasets with differential transcript expression**

FREE

Alyssa C. Frazee, Andrew E. Jaffe, Ben Langmead, Jeffrey T. Leek Author Notes

… and more

Single Cell RNA-seq

Splatter: simulation of single-cell RNA sequencing data

Luke Zappia , Belinda Phipson and Alicia Oshlack

Simulating multiple faceted variability in single cell RNA sequencing

Xiuwei Zhang, Chenling Xu & Nir Yosef

*but they are **count simulators***

*** slide courtesy of Avi Srivastava**

what's missing ?

Simulations

Bulk RNA-seq

RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome

Bo Li and Colin N Dewey

Modelling and simulating generic RNA-Seq experiments with the flux simulator

Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, Michael Sammeth Author Notes

***Polyester*: simulating RNA-seq datasets with differential transcript expression**

Alyssa C. Frazee, Andrew E. Jaffe, Ben Langmead, Jeffrey T. Leek Author Notes

… and more

Single Cell RNA-seq

Splatter: simulation of single-cell RNA sequencing data

Luke Zappia , Belinda Phipson and Alicia Oshlack

Simulating multiple faceted variability in single cell RNA sequencing

Xiuwei Zhang, Chenling Xu & Nir Yosef

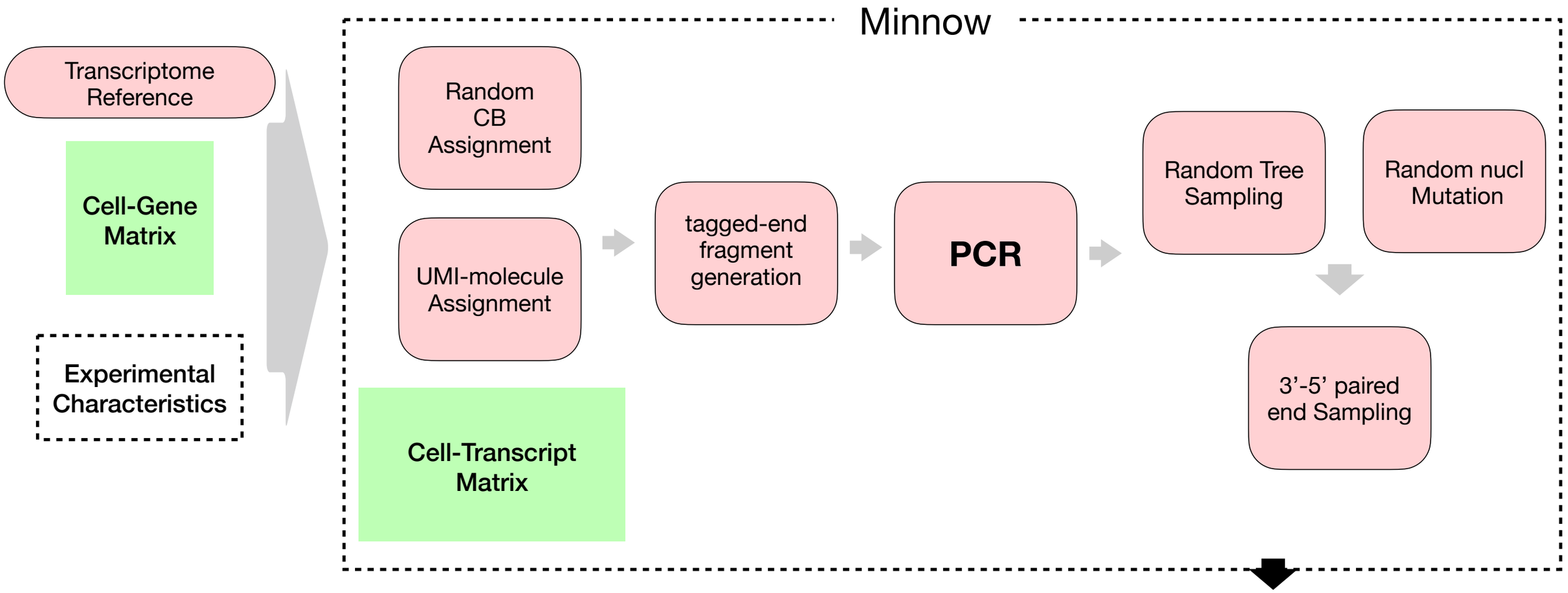
*but they are **count simulators***

***Minnow*: a principled framework for rapid simulation of dscRNA-seq data at the read level**

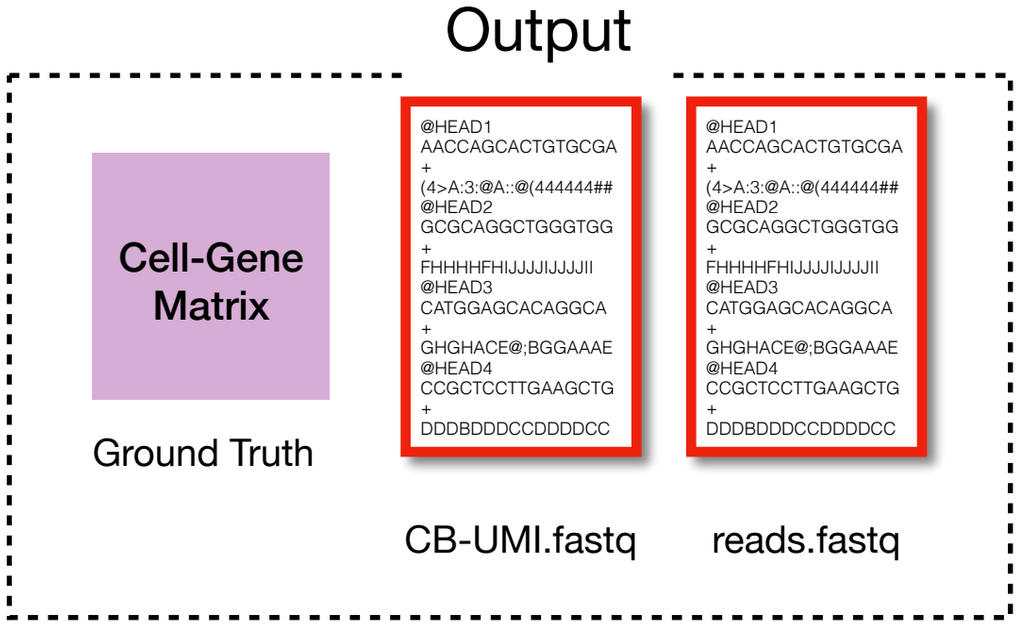
Hirak Sarkar, Avi Srivastava, Rob Patro

*** slide courtesy of Avi Srivastava**

Outline of Minnow



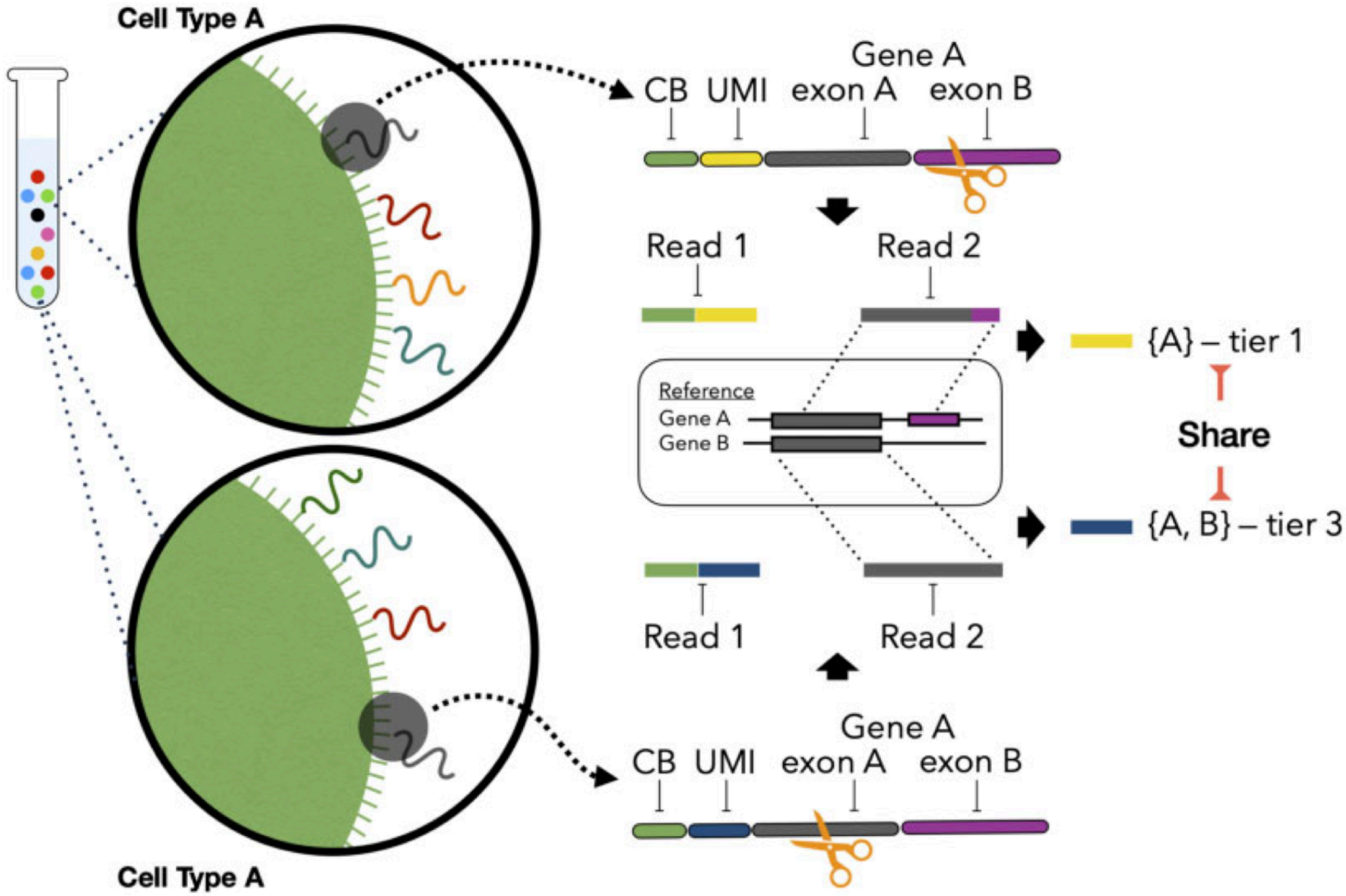
☞ Minnow generates *read sequences* along with the cell-gene count matrices which is crucial for comparing quantification methods.



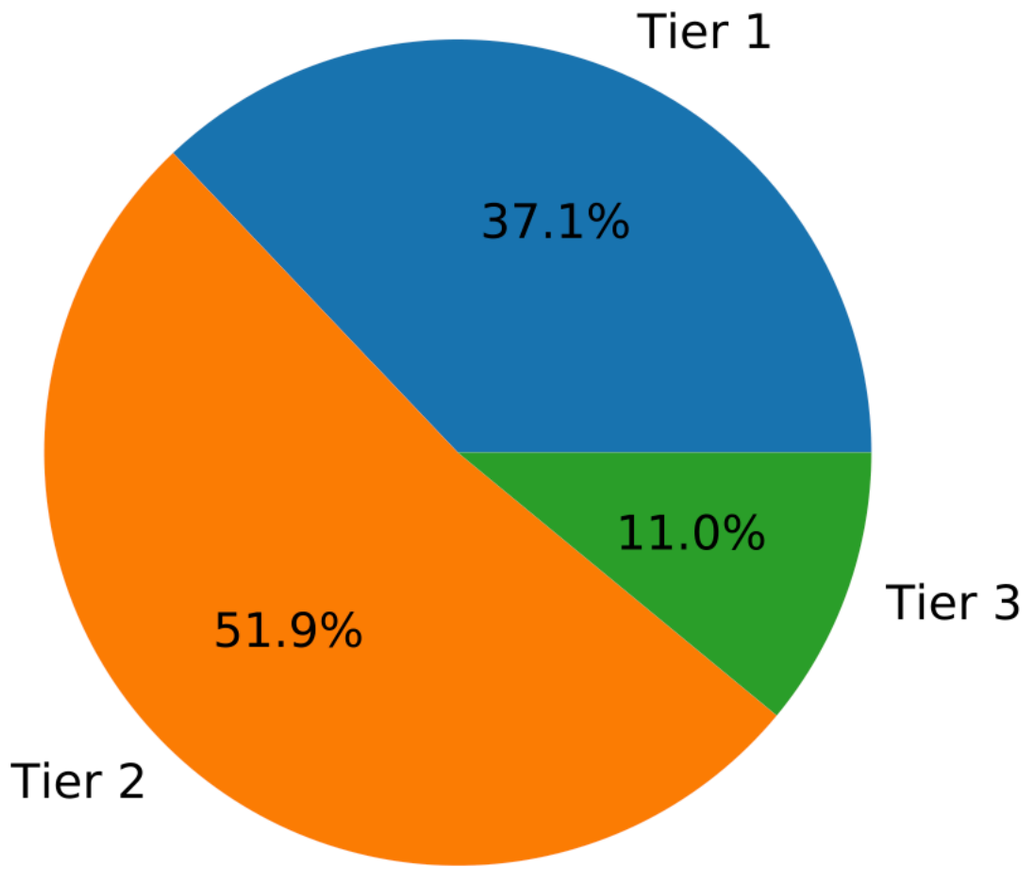
Sarkar, Hirak, Avi Srivastava, and Rob Patro. "Minnow: a principled framework for rapid simulation of dscRNA-seq data at the read level." *Bioinformatics* 35.14 (2019): i136-i144.

*** slide courtesy of Avi Srivastava**

Doing even better by sharing information



Full ambiguity is usually cell-specific



The percentage of cells assigning each tier to the genes, showing that the degree of confidence in the quantification estimates varies across cells even for a single gene. This plot is made using 7484 genes that have been assigned Tier 3 in at least one cell

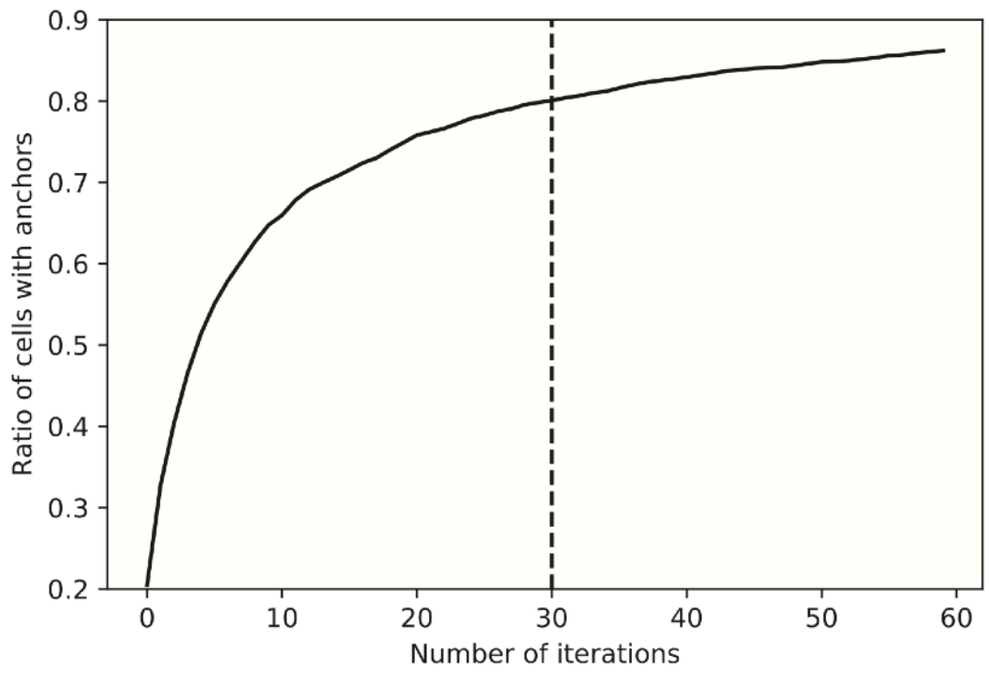
Using “similar” cells to share information

Split the dataset (cells) into halfs, randomly

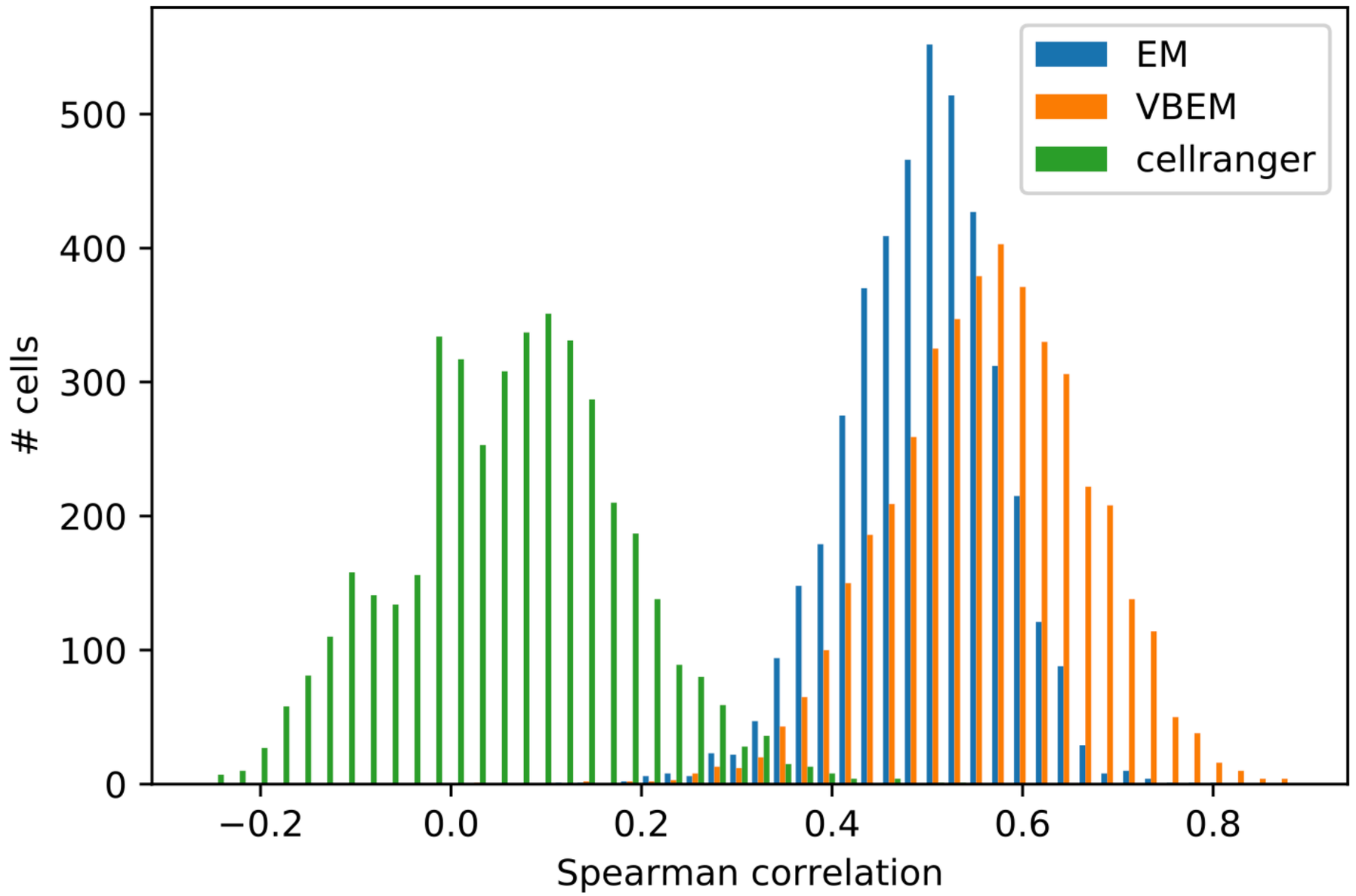
Use the anchoring algorithm of Seurat to anchor cells from one half into the other half

Repeat the above process many times.

The empirical prior for tier 3 genes in cell i becomes the average abundance of this gene in the list of anchors for cell i



This improves quantification accuracy for tier 3 genes



Most work has focused on “downstream” analysis

- Denoising : removal of “empty” droplets, detection of “doublets”
- Imputation : detection and inference of “dropouts”*
- Clustering and cell-type identification : Which cells are of the same “type”? What types of cells are present in my experiment?
- Pseudotime analysis: What expression changes are important / determinative in development? What are developmental branchpoints? What is the Waddington landscape?
- Dimensionality reduction for visualization : tSNE and UMAP are currently most popular. How do we visualize 10^3 — 10^6 cells?

*** “dropout” is a somewhat contentious term / idea — really just the implications of finite sampling from a limited population**

Most work has focused on "downstream" analysis

- Denoising : removal of "empty" droplets, detection of "doublets"

Cell Systems
Volume 8, Issue 4, 24 April 2019, Pages 281-291.e9

Article

Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data

Samuel L. Wolock [1], Romain Lopez [1, 2, 3], Allon M. Klein [1, 4]

Show more

https://doi.org/10.1016/j.cels.2018.11.005

Get rights and content

Lun et al. Genome Biology (2019) 20:63
https://doi.org/10.1186/s13059-019-1662-y

Genome Biology

METHOD Open Access

EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data

Aaron T. L. Lun[1*†], Samantha Riesenfeld[2†], Tallulah Andrews[3†], The Phuong Dao[4†], Tomas Gomes[3†], participants in the 1st Human Cell Atlas Jamboree and John C. Marioni[1,3,5*]

BRIEF REPORT | VOLUME 8, ISSUE 4, P329-337.E4, APRIL 24, 2019

DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors

Christopher S. McGinnis • Lyndsay M. Murrow • Zev J. Gartner [4] • Show footnotes

Published: April 03, 2019 • DOI: https://doi.org/10.1016/j.cels.2019.03.003 • Check for updates

Most work has focused on "downstream" analysis

- Imputation : detection and inference of "dropouts"*

Peng *et al. Genome Biology* (2019) 20:88
https://doi.org/10.1186/s13059-019-1681-8

Genome Biology

METHOD Open Access

SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data

Tao Peng[1], Qin Zhu[2], Penghang Yin[3] and Kai Tan[1,2,4,5,6,7*]

Research | Open Access | Published: 27 August 2020

A systematic evaluation of single-cell RNA-sequencing imputation methods

Wenpin Hou, Zhicheng Ji, Hongkai Ji & Stephanie C. Hicks

Genome Biology **21**, Article number: 218 (2020) | Cite this article
808 Accesses | **89** Altmetric | Metrics

nature COMMUNICATIONS

ARTICLE
DOI: 10.1038/s41467-018-03405-7 OPEN

An accurate and robust imputation method scImpute for single-cell RNA-seq data

Wei Vivian Li[1] & Jingyi Jessica Li[1,2]

nature methods BRIEF COMMUNICATION
https://doi.org/10.1038/s41592-018-0033-z

SAVER: gene expression recovery for single-cell RNA sequencing

Mo Huang[1], Jingshu Wang[1], Eduardo Torre[2,3], Hannah Dueck[4], Sydney Shaffer[3], Roberto Bonasio[5], John I. Murray[4], Arjun Raj[3,4], Mingyao Li[6] and Nancy R. Zhang[1*]

We found that the majority of scRNA-seq imputation methods outperformed no imputation in recovering gene expression observed in bulk RNA-seq. However, the majority of the methods did not improve performance in downstream analyses compared to no imputation, in particular for clustering and trajectory analysis, and thus should be used with caution.

Most work has focused on "downstream" analysis

- Clustering and cell-type identification : Which cells are of the same "type"? What types of cells are present in my experiment?

nature methods

Brief Communication | Published: 27 March 2017

SC3: consensus clustering of single-cell RNA-seq data

Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green & Martin Hemberg

Nature Methods **14**, 483–486 (2017) | Download Citation

Published online 21 June 2019

Nucleic Acids Research, 2019, Vol. 47, No. 16 **e95**
doi: 10.1093/nar/gkz543

CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing

Jurrian K. de Kanter†, Philip Lijnzaad†, Tito Candelli, Thanasis Margaritis and Frank C.P. Holstege*

Princess Máxima Center for Pediatric Oncology, Heidelberglaan 25, 3584 CS, Utrecht, The Netherlands

Received February 22, 2019; Revised June 05, 2019; Editorial Decision June 06, 2019; Accepted June 08, 2019

Clustering and classification methods for single-cell RNA-sequencing data

Ren Qi, Anjun Ma, Qin Ma, Quan Zou

Briefings in Bioinformatics, bbz062, https://doi.org/10.1093/bib/bbz062

Published: 04 July 2019 **Article history** ▾

Most work has focused on "downstream" analysis

- Pseudotime analysis: What expression changes are important / determinative in development? What are developmental branchpoints? What is the Waddington landscape?

nature methods

Brief Communication | Published: 21 August 2017

Reversed graph embedding resolves complex single-cell trajectories

Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A Pliner & Cole Trapnell

Nature Methods **14**, 979–982 (2017) | Download Citation
5211 Accesses | **230** Citations | **31** Altmetric | Metrics

nature COMMUNICATIONS

Article | Open Access | Published: 23 April 2019

Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM

Huidong Chen, Luca Albergante, Jonathan Y. Hsu, Caleb A. Lareau, Giosuè Lo Bosco, Jihong Guan, Shuigeng Zhou, Alexander N. Gorban, Daniel E. Bauer, Martin J. Aryee, David M. Langenau, Andrei Zinovyev, Jason D. Buenrostro, Guo-Cheng Yuan & Luca Pinello

Nature Communications **10**, Article number: 1903 (2019) | Download Citation
8071 Accesses | **2** Citations | **44** Altmetric | Metrics

nature COMMUNICATIONS

Article | Open Access | Published: 22 June 2018

Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data

Kieran R Campbell & Christopher Yau

Most work has focused on “downstream” analysis

- Dimensionality reduction for visualization : tSNE and UMAP are currently most popular. How do we visualize 10^3 — 10^6 cells?

nature biotechnology

Analysis | Published: 03 December 2018

Dimensionality reduction for visualizing single-cell data using UMAP

Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux & Evan W Newell

nature COMMUNICATIONS

Article | Open Access | Published: 21 May 2018

Interpretable dimensionality reduction of single cell transcriptome data with deep generative models

Jiarui Ding, Anne Condon & Sohrab P. Shah

Nature Communications **9**, Article number: 2002 (2018) | Download Citation

6813 Accesses | **25** Citations | **62** Altmetric | Metrics »

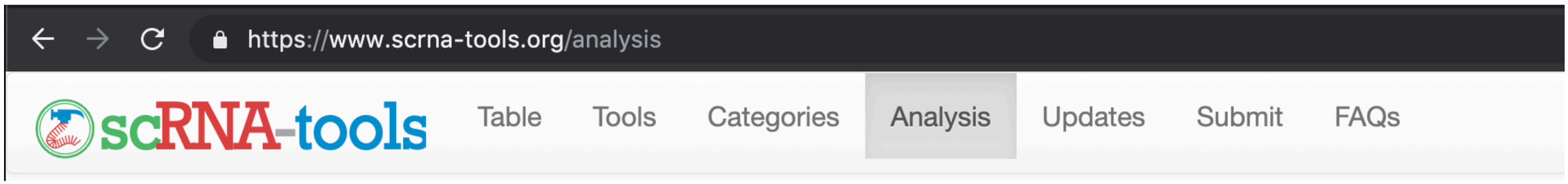
Journal of Machine Learning Research 9 (2008) 2579-2605 Submitted 5/08; Revised 9/08; Published 11/08

Visualizing Data using t-SNE

Laurens van der Maaten LVDMAATEN@GMAIL.COM
TiCC
Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

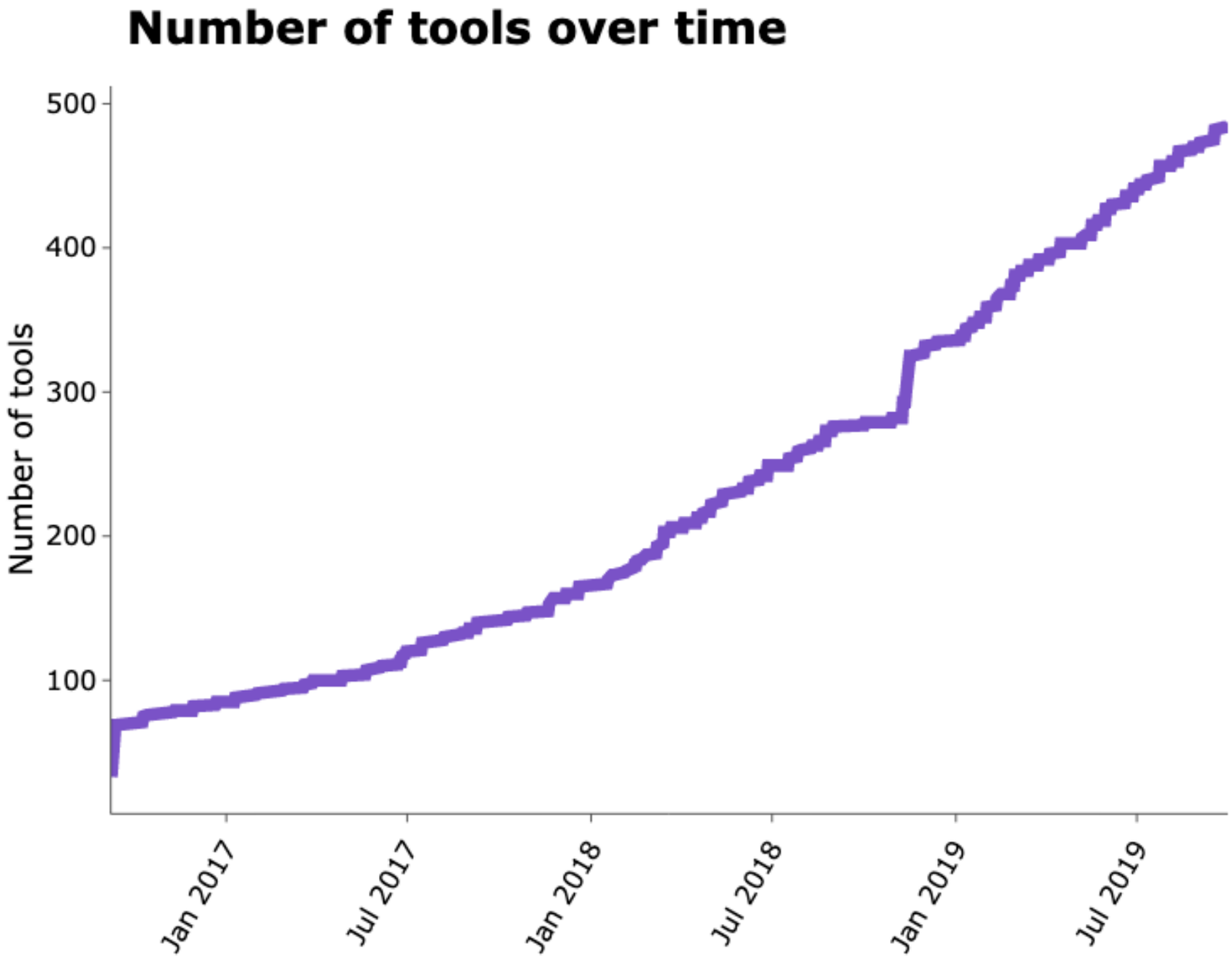
Geoffrey Hinton HINTON@CS.TORONTO.EDU
Department of Computer Science
University of Toronto
6 King's College Road, M5S 3G4 Toronto, ON, Canada

Simply keeping track of new methods is difficult!



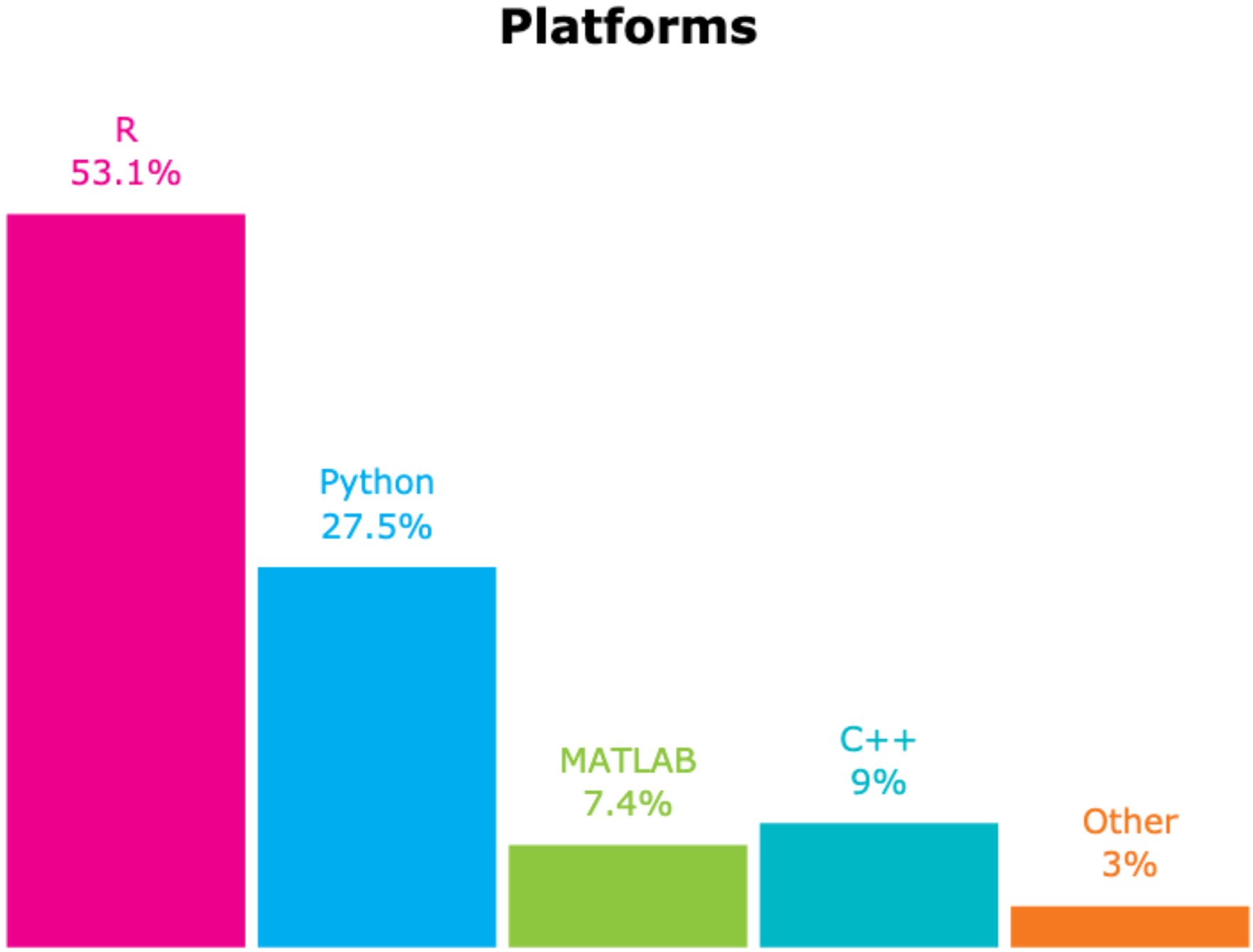
scRNA-tools: website run out of Alicia Oshlack's lab. Maintains a constantly-updated database of different tools for processing scRNA-seq data. Let's take a look!

Simply keeping track of new methods is difficult!



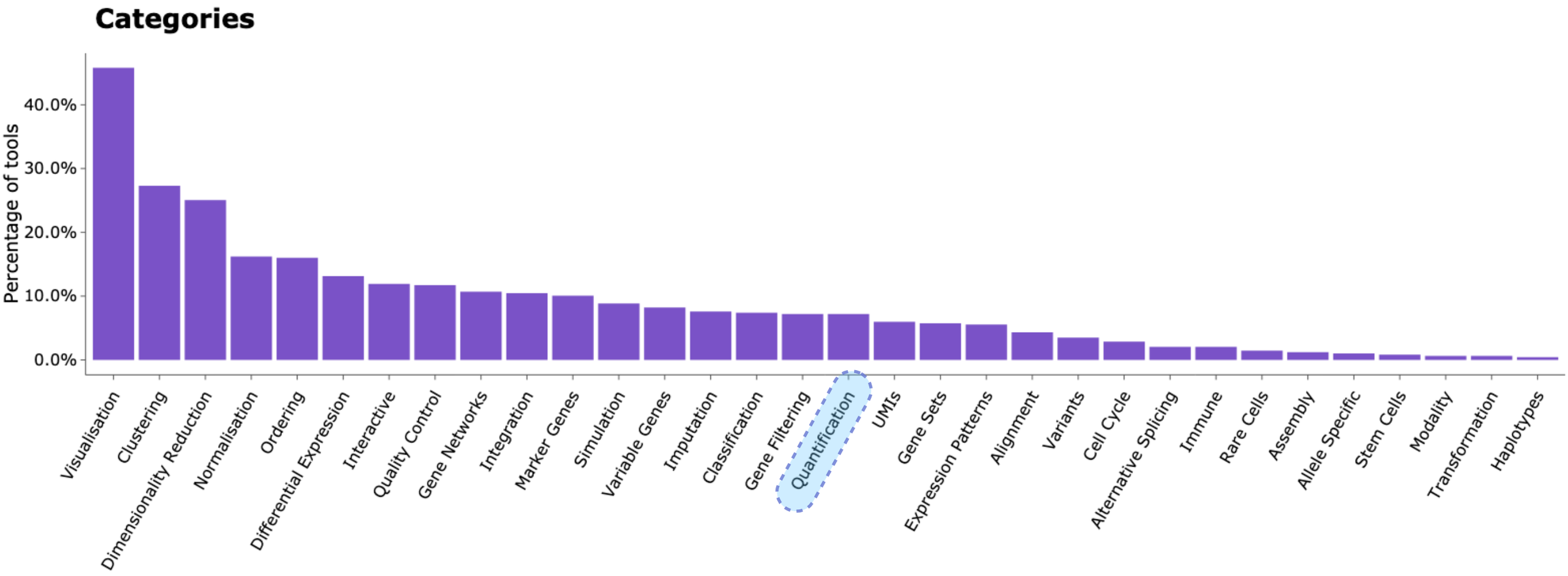
The number of tools is increasing rapidly

Simply keeping track of new methods is difficult!



Most such tools are written in “R” —
perform visualization & statistical analysis of count matrices

Simply keeping track of new methods is difficult!



As I mentioned, most tools deal with problems “downstream” of quantification

Recent work aims to tackle multiple such problems at once

Many variants on the same idea : try to learn a “latent” space of single-cell expression, from which many different questions can be answered.

nature methods

ARTICLES

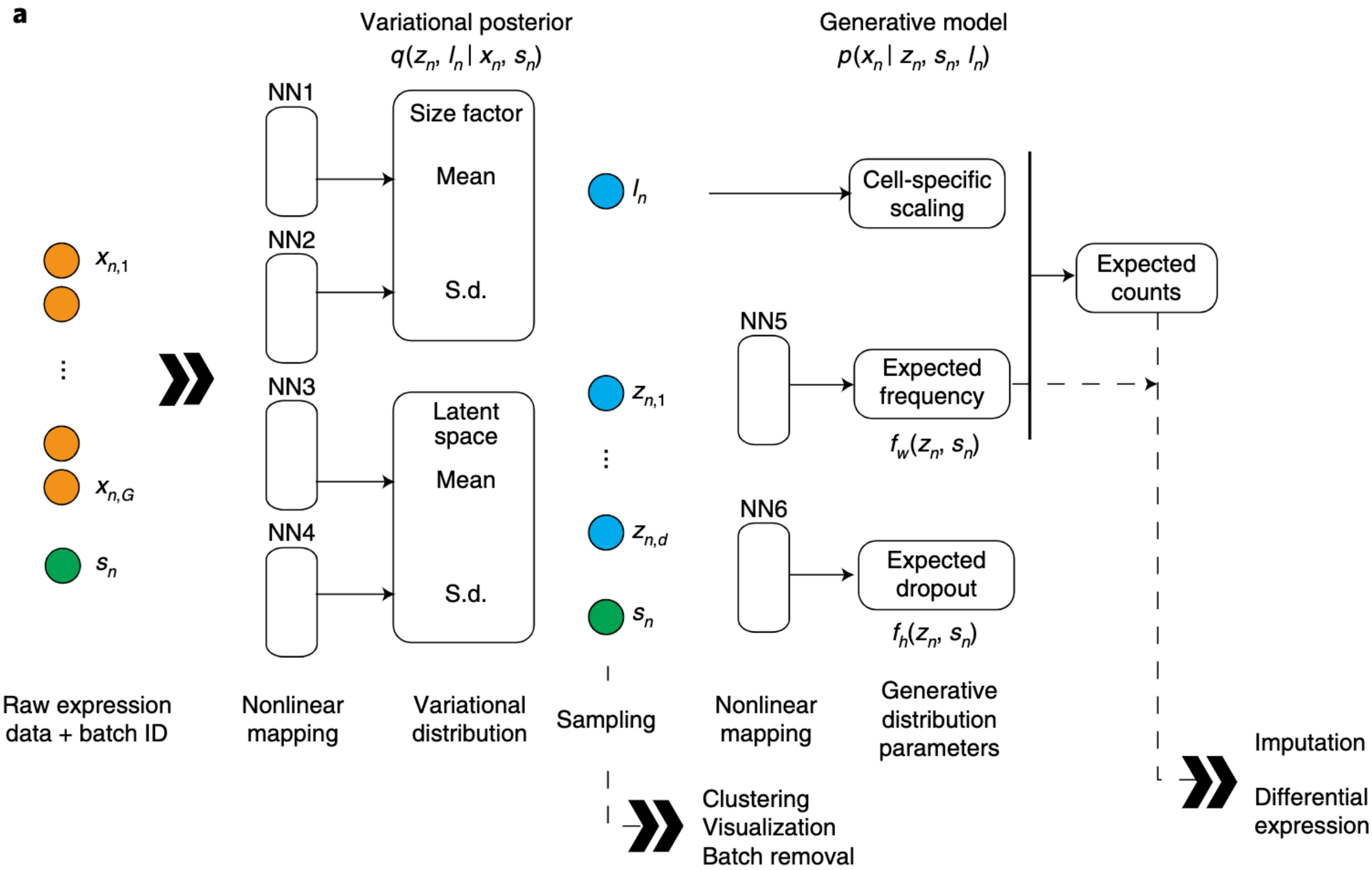
https://doi.org/10.1038/s41592-018-0229-2

Deep generative modeling for single-cell transcriptomics

Romain Lopez[1], Jeffrey Regier[1], Michael B. Cole[2], Michael I. Jordan[1,3] and Nir Yosef[1,4,5*]

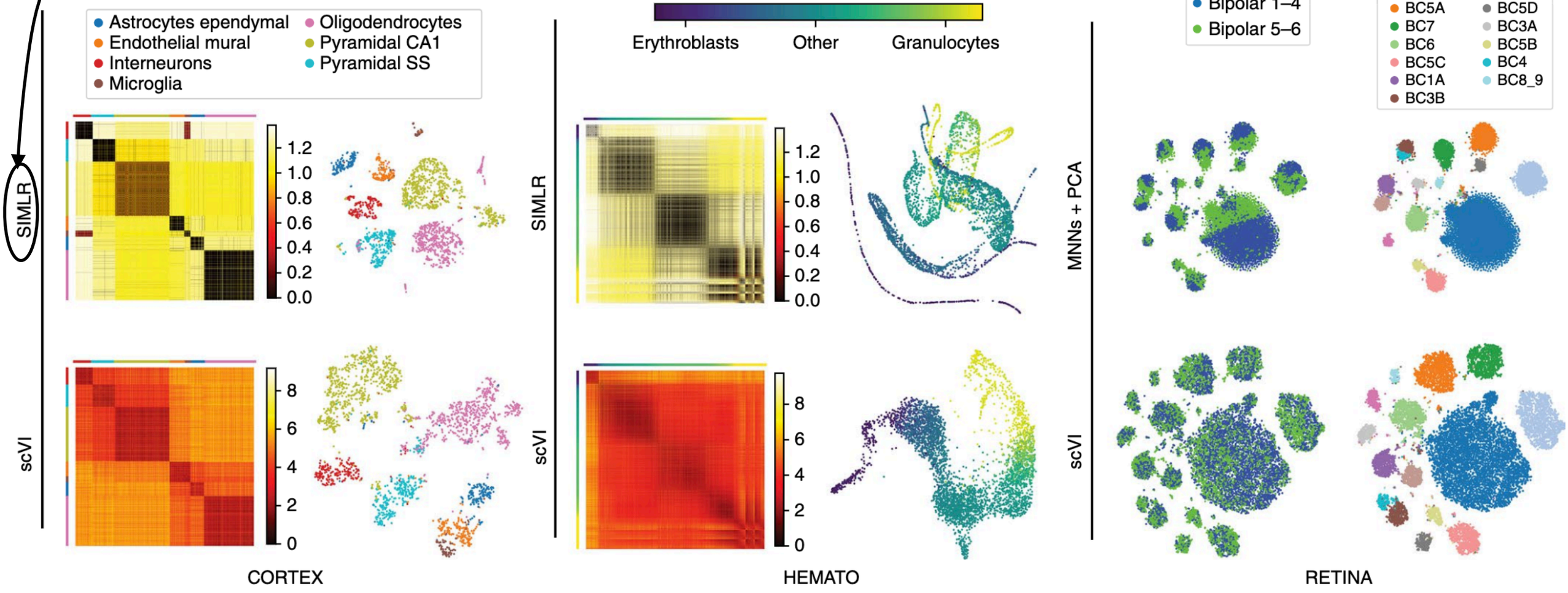
Single-cell transcriptome measurements can reveal unexplored biological diversity, but they suffer from technical noise and bias that must be modeled to account for the resulting uncertainty in downstream analyses. Here we introduce single-cell variational inference (scVI), a ready-to-use scalable framework for the probabilistic representation and analysis of gene expression in single cells (https://github.com/YosefLab/scVI). scVI uses stochastic optimization and deep neural networks to aggregate information across similar cells and genes and to approximate the distributions that underlie observed expression values, while accounting for batch effects and limited sensitivity. We used scVI for a range of fundamental analysis tasks including batch correction, visualization, clustering, and differential expression, and achieved high accuracy for each task.

scVI overview



scVI for clustering & vis

Method for joint clustering and dim-reduction

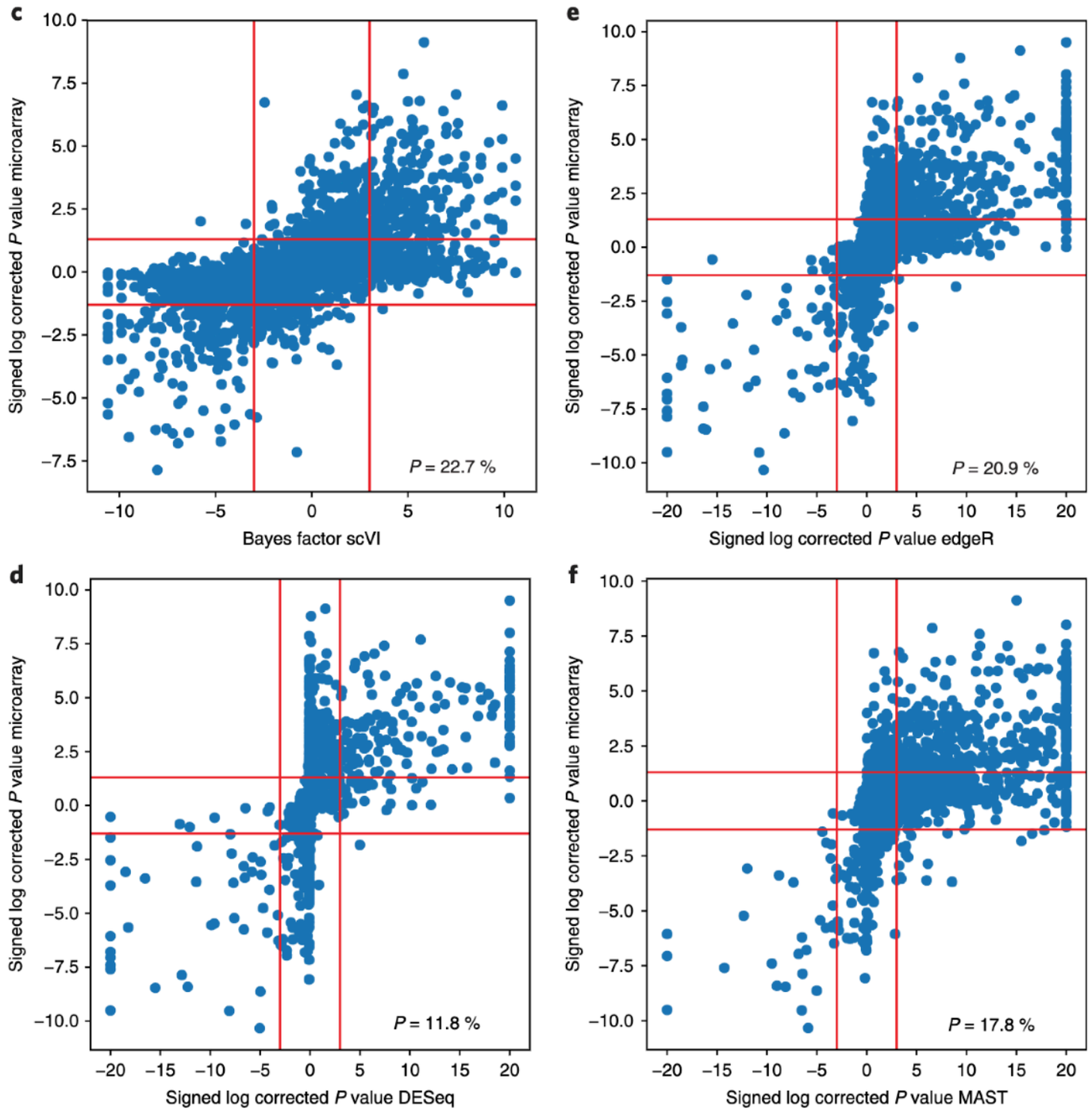


Discrete clusters

Developmental gradient

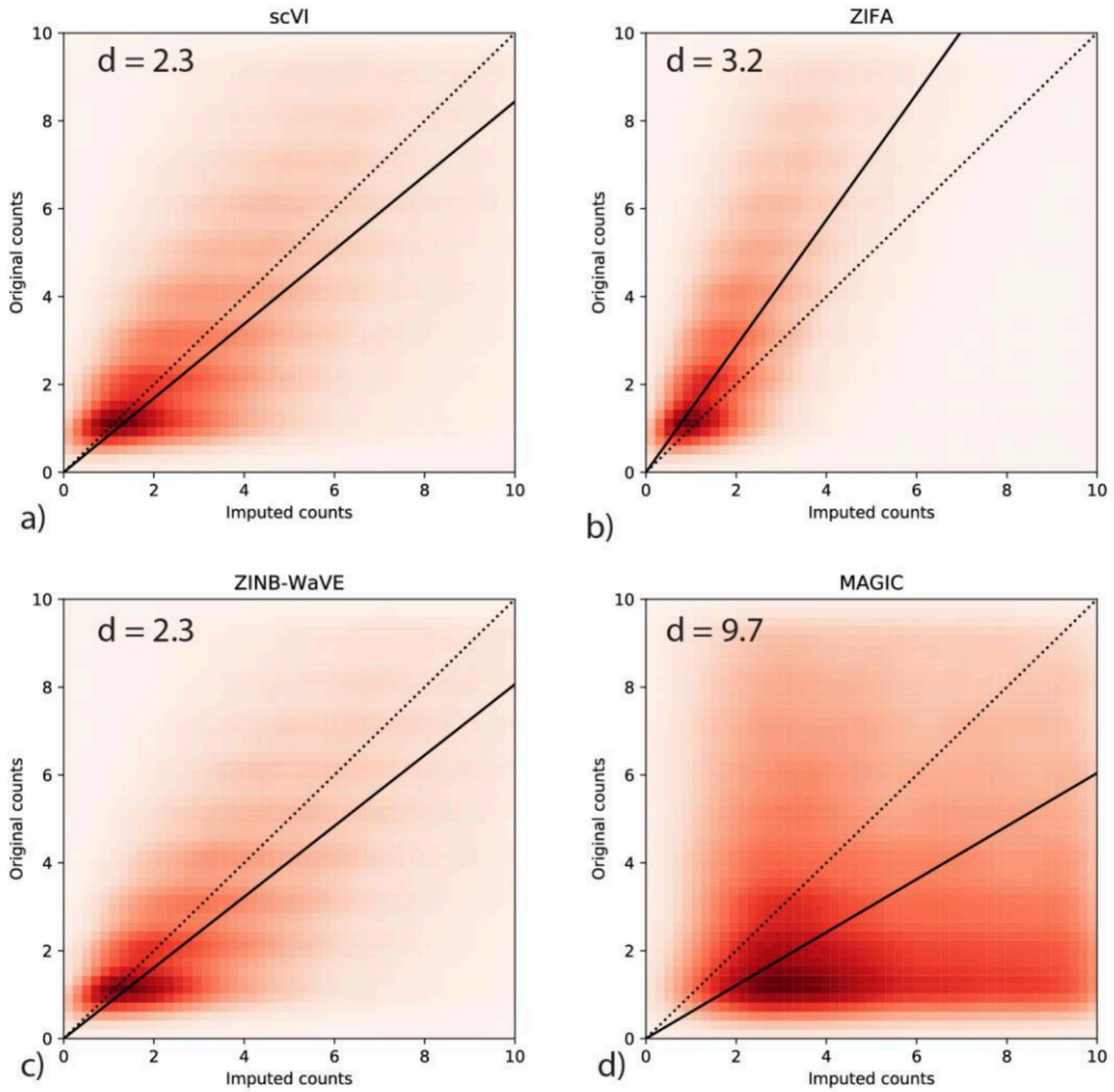
Clustering of retinal cells

scVI for differential expression



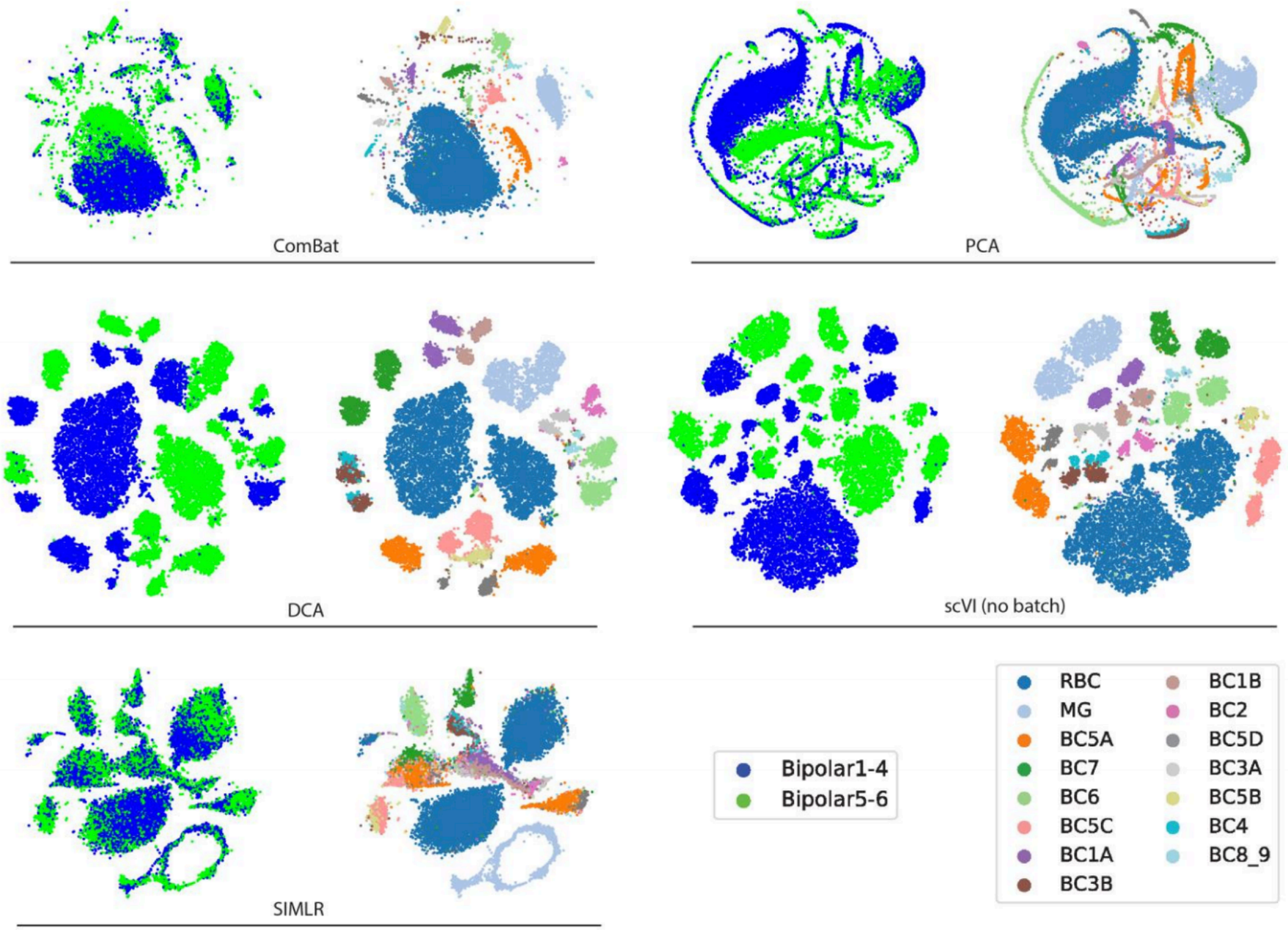
Significance levels of differential expression between B cells and dendritic cells. Points represent individual genes (n= 3,346). Bayes factors or BH-corrected P-values on scRNA-seq data are compared with bulk microarray-based BH-corrected P-values. Horizontal lines denote the significance threshold of 0.05 for corrected P-values. Vertical lines denote the significance threshold for the Bayes factor of scVI (c) or 0.05 for corrected P-values for DESeq2 (d), edgeR (e), and MAST (f).

scVI for imputation

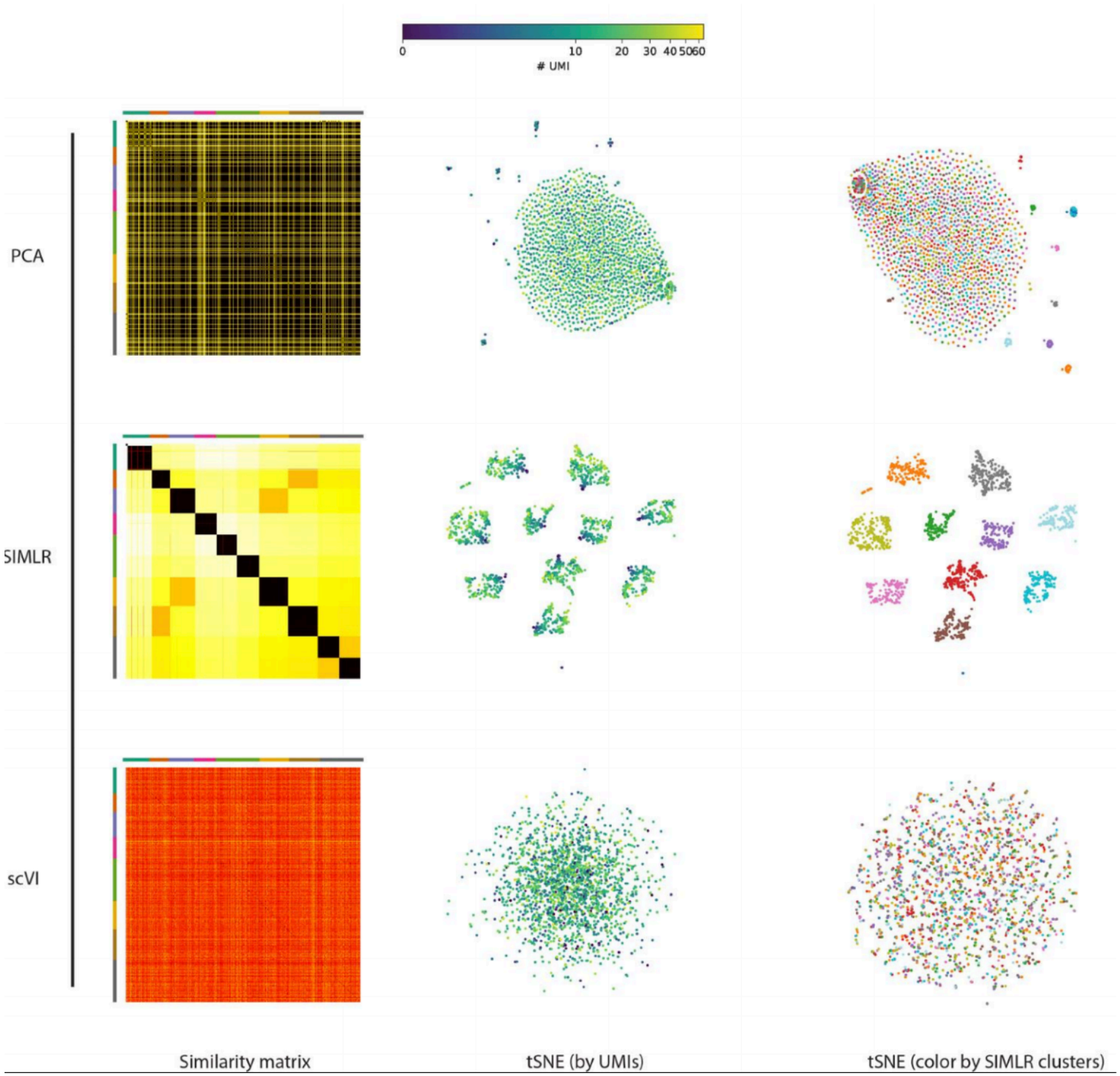


(a-d) The heat maps represent density plots of imputed values (by scVI, ZIFA, ZINB-WaVE, and MAGIC respectively) on a downsampled version versus the original (nonzero) values prior to downsampling. The reported score d is the median imputation error across all the hidden entries (lower is better; see Methods). Each density plot was computed using n = 55,932 independently perturbed entries from the original matrix.

scVI for batch effect removal



scVI latent representation doesn't introduce "false structure"



The quickly-growing repository of scRNA-seq data poses other challenges

Fast searches of large collections of single cell data using scfind

Jimmy Tsz Hang Lee[+], Nikolaos Patikas[+], Vladimir Yu Kiselev and Martin Hemberg[*]

Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, UK
[+]These authors contributed equally
[*]Corresponding author: mh26@sanger.ac.uk

Index single-cell datasets to quickly find all cells expressing certain genes, or patterns of genes (without know which patterns to search ahead of time).

Find frequently-recurring sub-patterns as indicators of cell-type.

Some of *our* ongoing work in this area

- Alternative models for resolving UMIs (*likelihood* vs. *parsimony*)
- Further improving quantification estimates by sharing information across cells (e.g. empirical Bayes & hierarchical models)
- Improved estimation of splicing rate within alevin model (including unspliced reads for e.g. RNA-velocity)
- Improvement of tricky alignment / mapping cases
- Better data structures and algorithms for indexing and searching large corpora of expression data

Just as single-cell technology leads to exciting new biology, it opens the door to a host of computational challenges and opportunities!