Advanced topics in single-cell transcriptomics

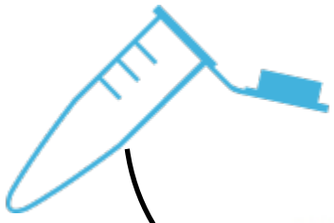
Avi Srivastava

New York Genome Center

May 27, 2020

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

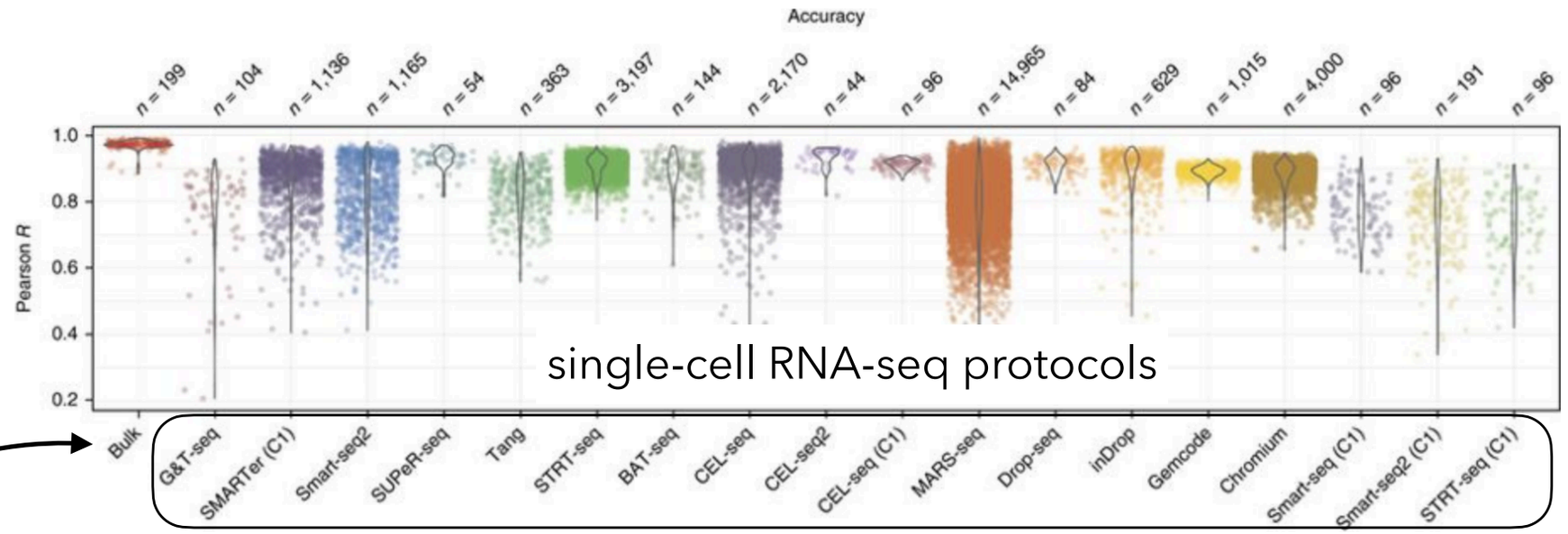
Note: All the slides with SBU Red heading, courtesy of Rob Patro

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)

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.

The computational challenge(s)

Scalability

Muir *et al. Genome Biology* (2016) 17:53
DOI 10.1186/s13059-016-0917-0

Genome Biology

OPINION Open Access

The real cost of sequencing: scaling computation to keep pace with data generation

Paul Muir[1,2,3], Shantao Li[4], Shaoke Lou[4,5], Daifeng Wang[4,5], Daniel J Spakowicz[4,5], Leonidas Salichos[4,5], Jing Zhang[4,5], George M. Weinstock[6], Farren Isaacs[1,2], Joel Rozowsky[4,5] and Mark Gerstein[4,5,7*]

"This new regime, in which costs scale with the amount of computational processing time, places a premium on driving down the average cost by developing efficient algorithms for data processing."

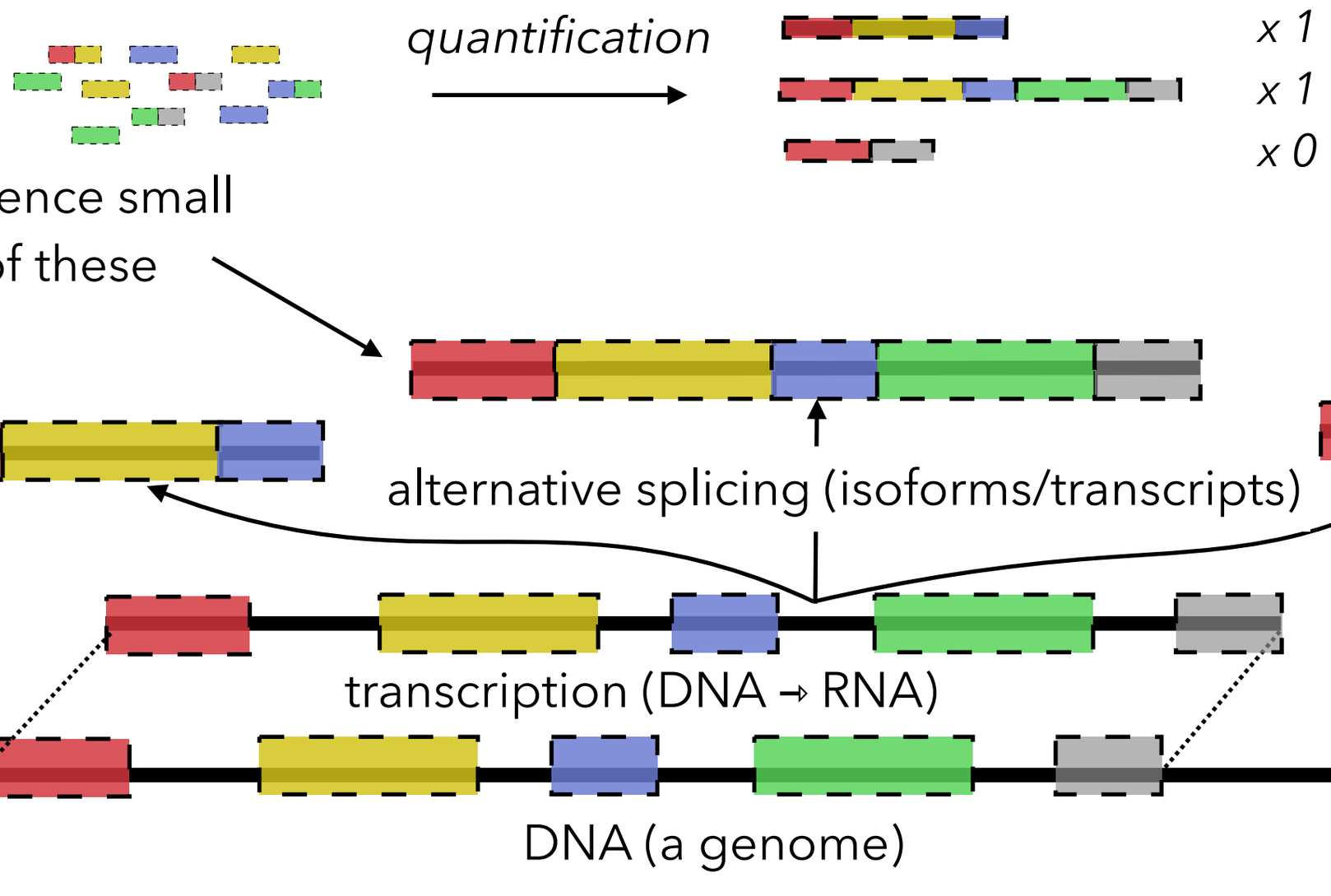
Also, it's not just "new" data that is the problem:

In addition to new data, re-analysis of existing experiments often desired: In light of new annotations, discoveries, and methodological advancements.

Expression Study basics in one slide

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

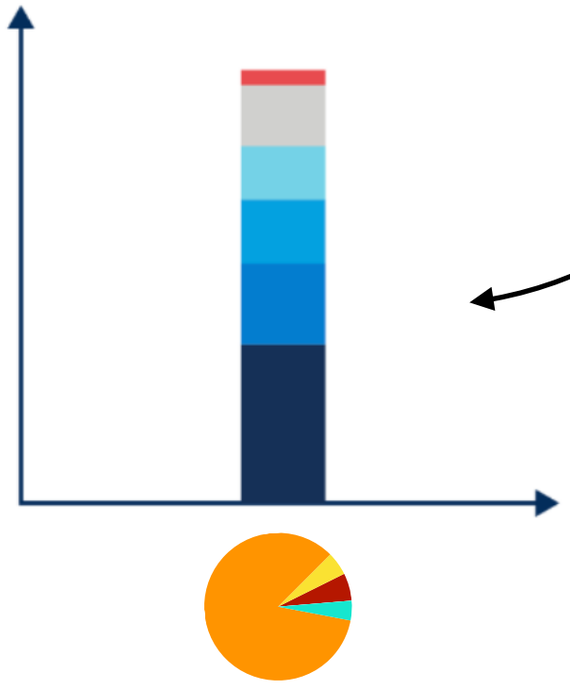
- What genes / transcripts are turned on?
- At what levels are they expressed?
- How do they respond to environment / stimuli?



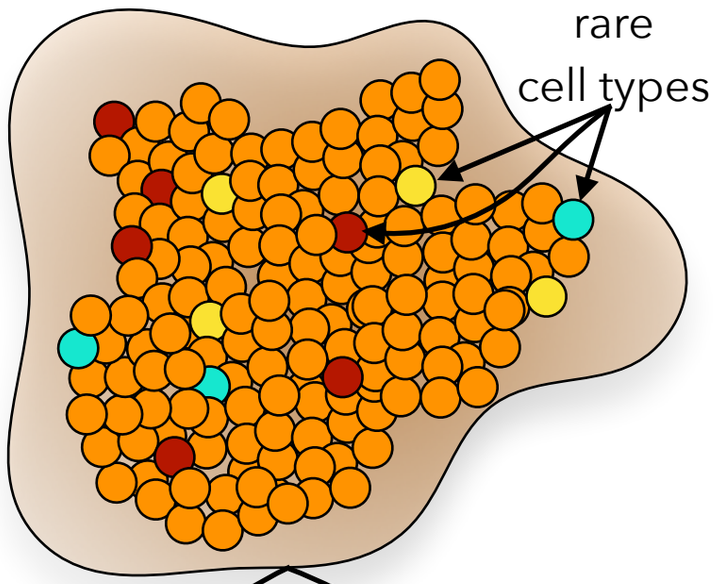
RNA-seq ⇒ single-cell RNA-seq

Bulk RNA-seq

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



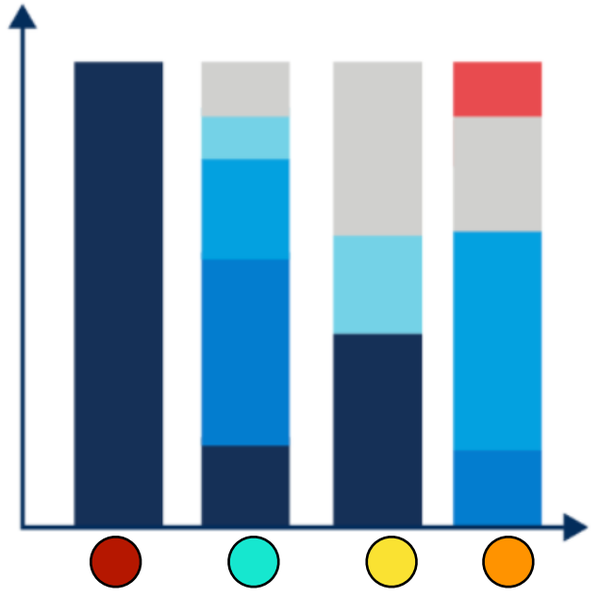
Tissue



These present **very** different challenges

single-cell RNA-seq

- Typically tens of thousands of cells
- Low-coverage (reads / cell)
- Measure gene abundance at the single-cell level, specially useful for rare cell types.



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)

Single Cell Protocols

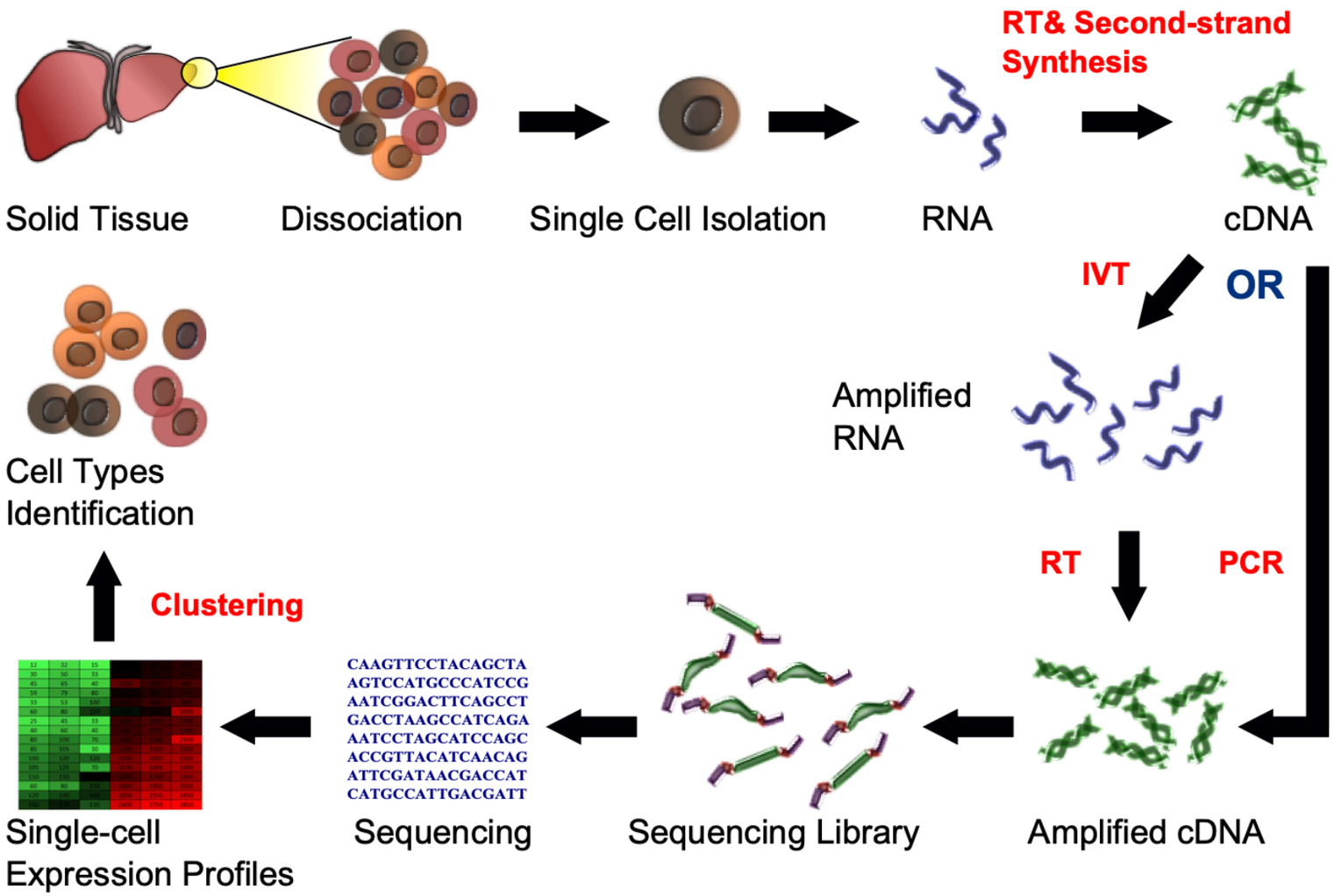
Single Cell Isolation:

* **Micro-well:** Requires human pipetting but are surface-markable however has very low-throughput.

* **Micro-fluid (Fluidigm's C1):** It's 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.

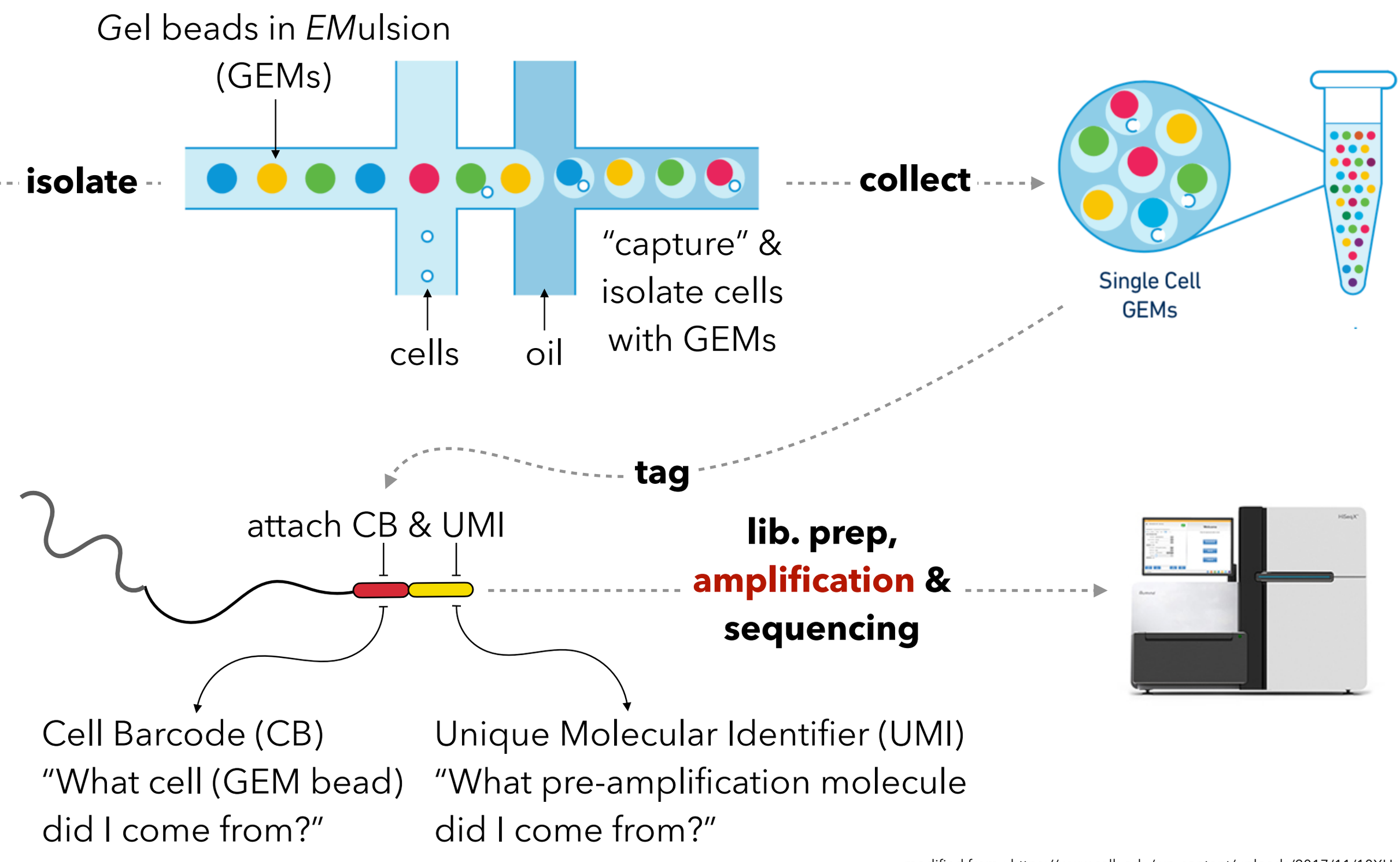
Single Cell RNA-sequencing workflow:



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.



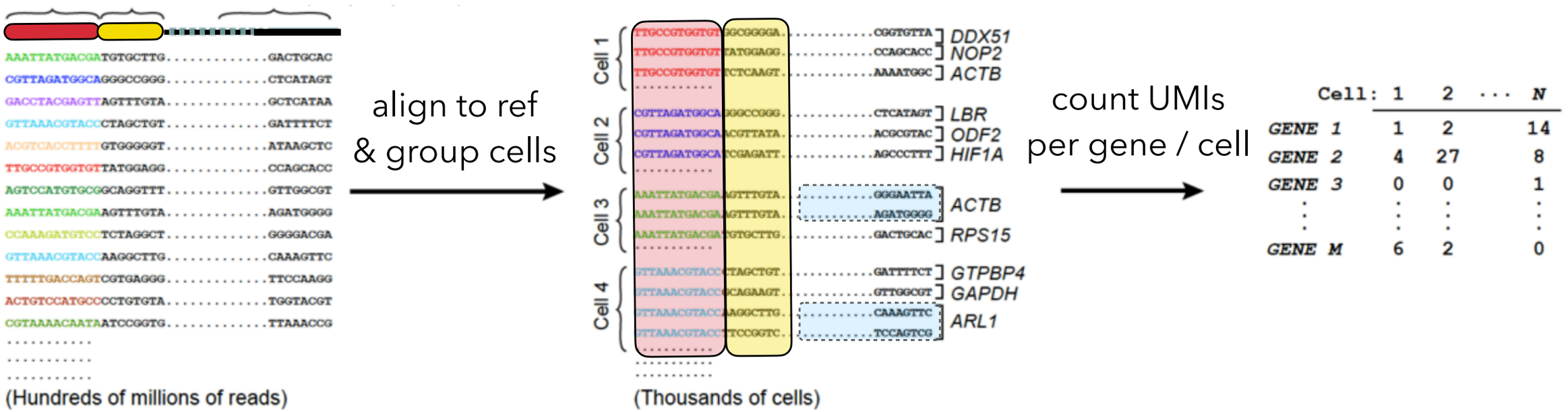
modified from : https://www.salk.edu/wp-content/uploads/2017/11/10XUserGuide.pdf

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!

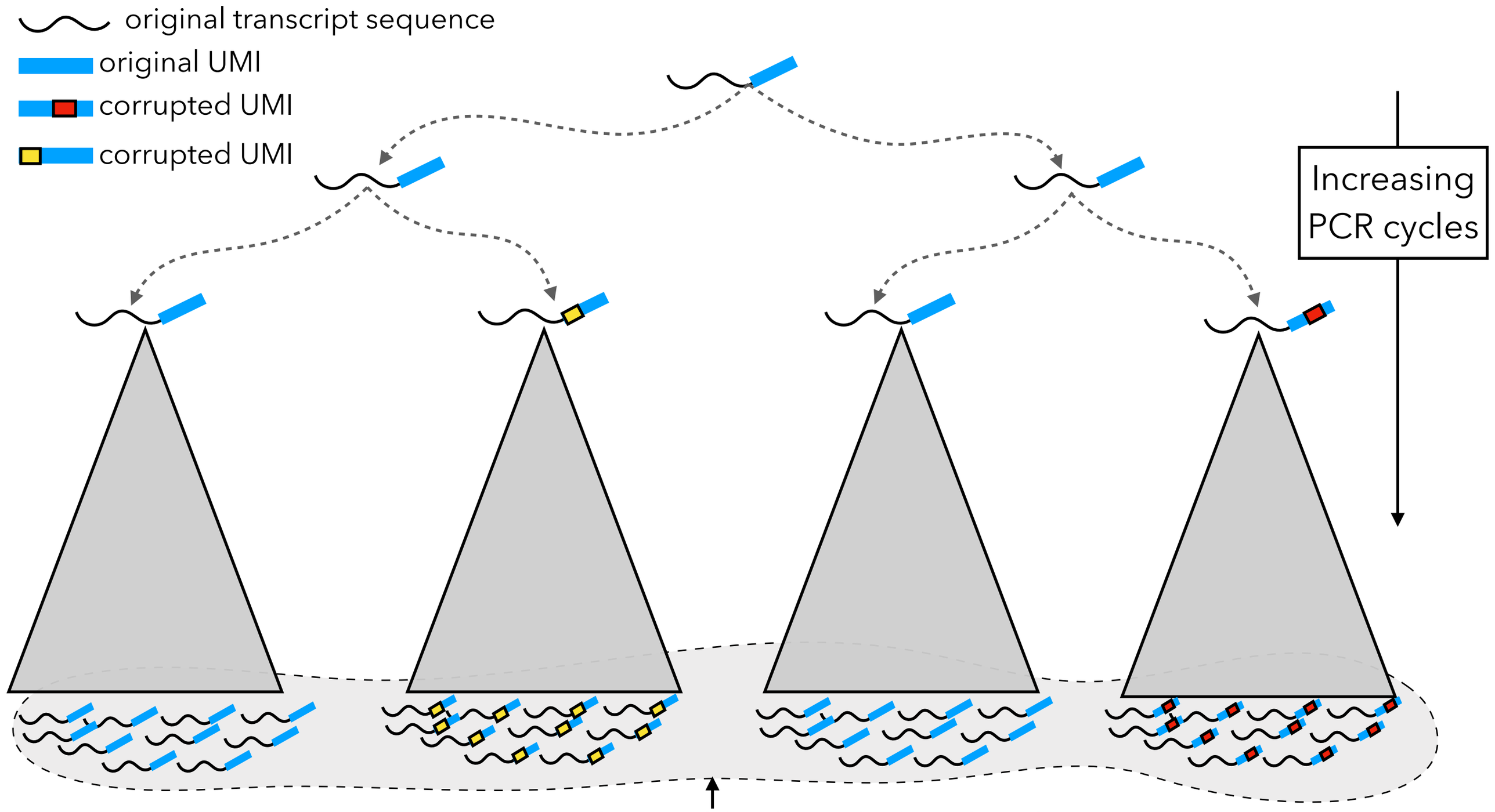
… how many lines is that in Pandas?



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

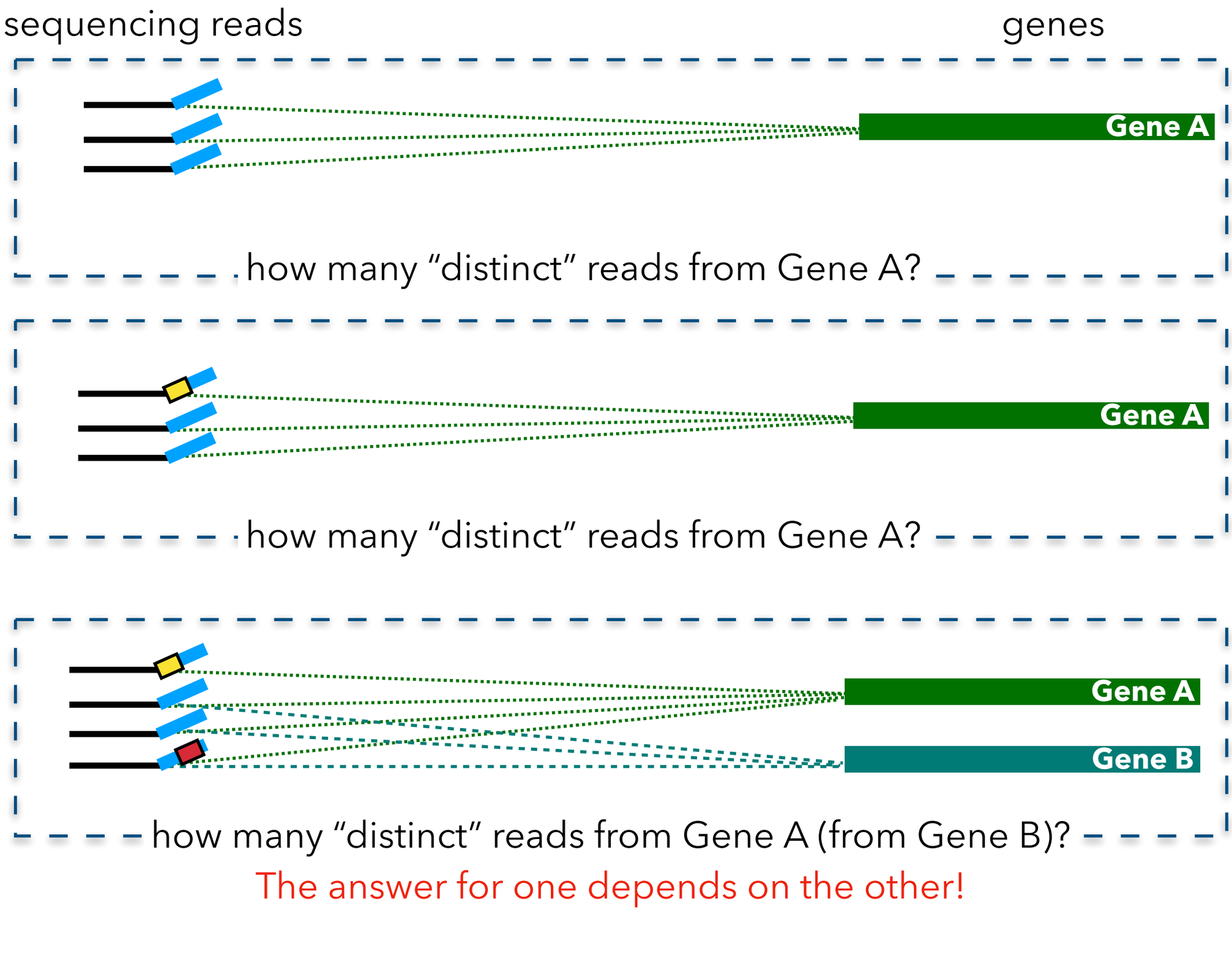


Sample reads from this exponentially amplified pool of molecules

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

Multi-mapping 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:

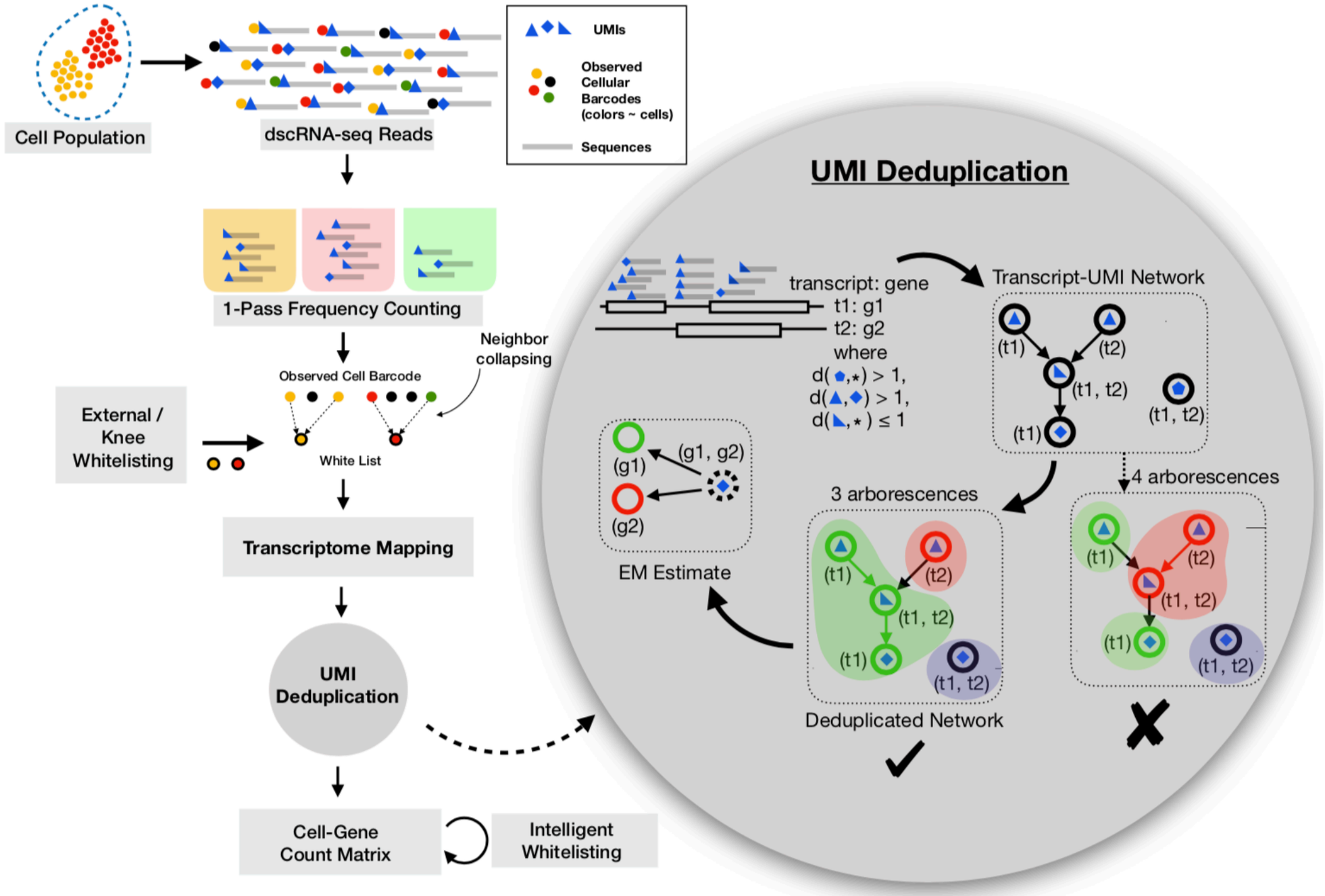
- 1 UMI ≠ 1 pre-PCR molecule (often); because of small material per-cell, typical protocols involve many rounds of PCR; UMI tags are subject to PCR and sequencing error.

- 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

alevin: dscRNA-seq quantification

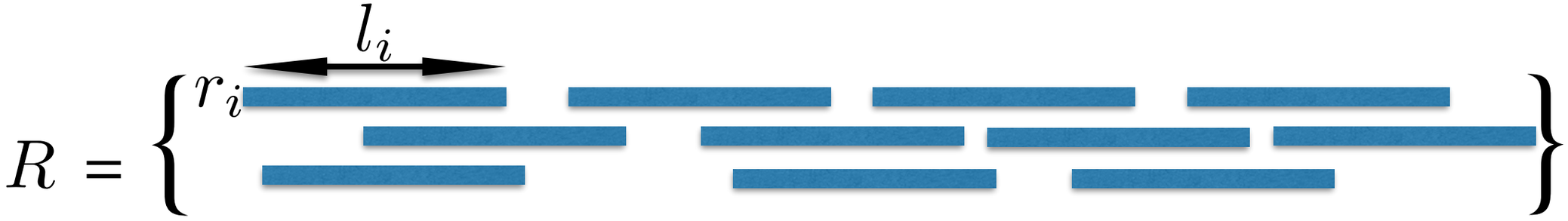


Srivastava, Avi, et al. "Alevin efficiently estimates accurate gene abundances from dscRNA-seq data." *Genome biology*20.1 (2019): 65.

The read-alignment problem

Given: A collection of sequencing reads, and some target sequence (e.g. a genome)

Find: For each read, all locations where the read is within edit distance ϵ of the reference, and the edits that achieve this distance.



$d : \Sigma^{|u|} \times \Sigma^{|v|} \rightarrow \mathbb{Z}$ and ϵ (maximum edit distance)

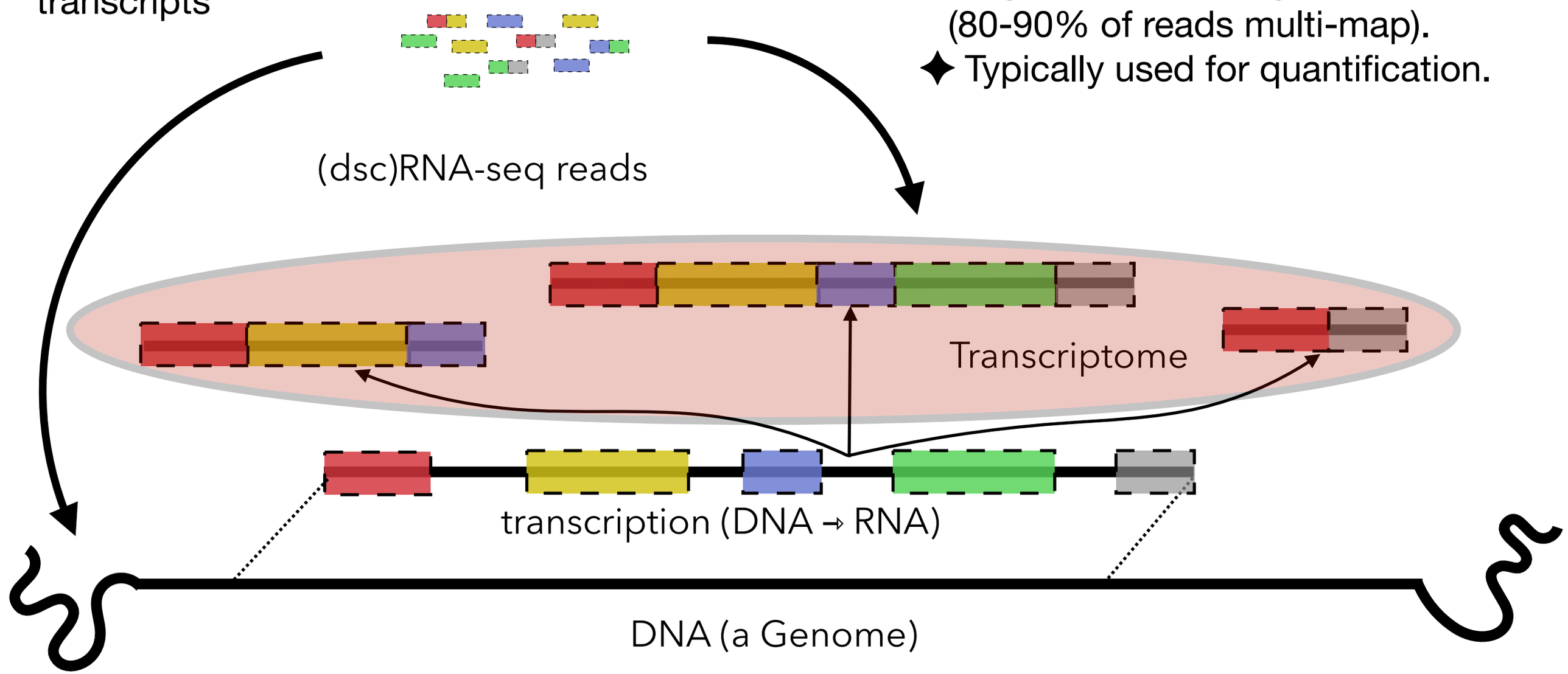
Read-Alignment Strategies

Genome Mapping

- Tools like: STAR, HISAT.
- Maps to full Genome (~3G for Human).
- Less per read multi-mapping.
- Typical use case includes finding new transcripts

Transcriptome Mapping

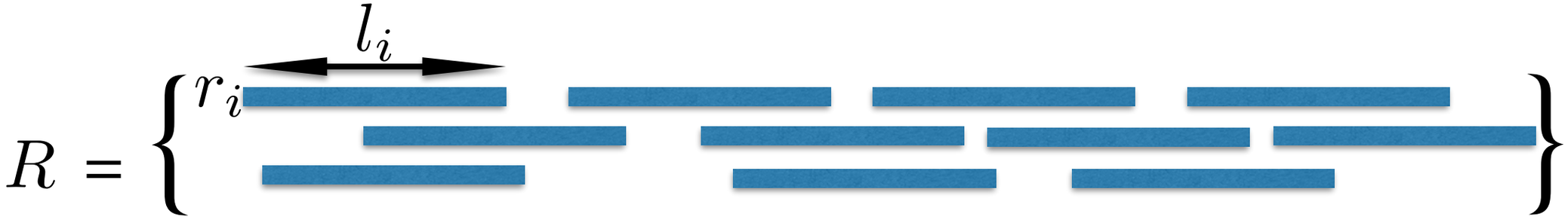
- Tools like: Bowtie, RapMap, Selective-alignment.
- Maps to Transcriptome (~300M for Human).
- High multi-mapping rate (80-90% of reads multi-map).
- Typically used for quantification.



The read-MAPPING problem

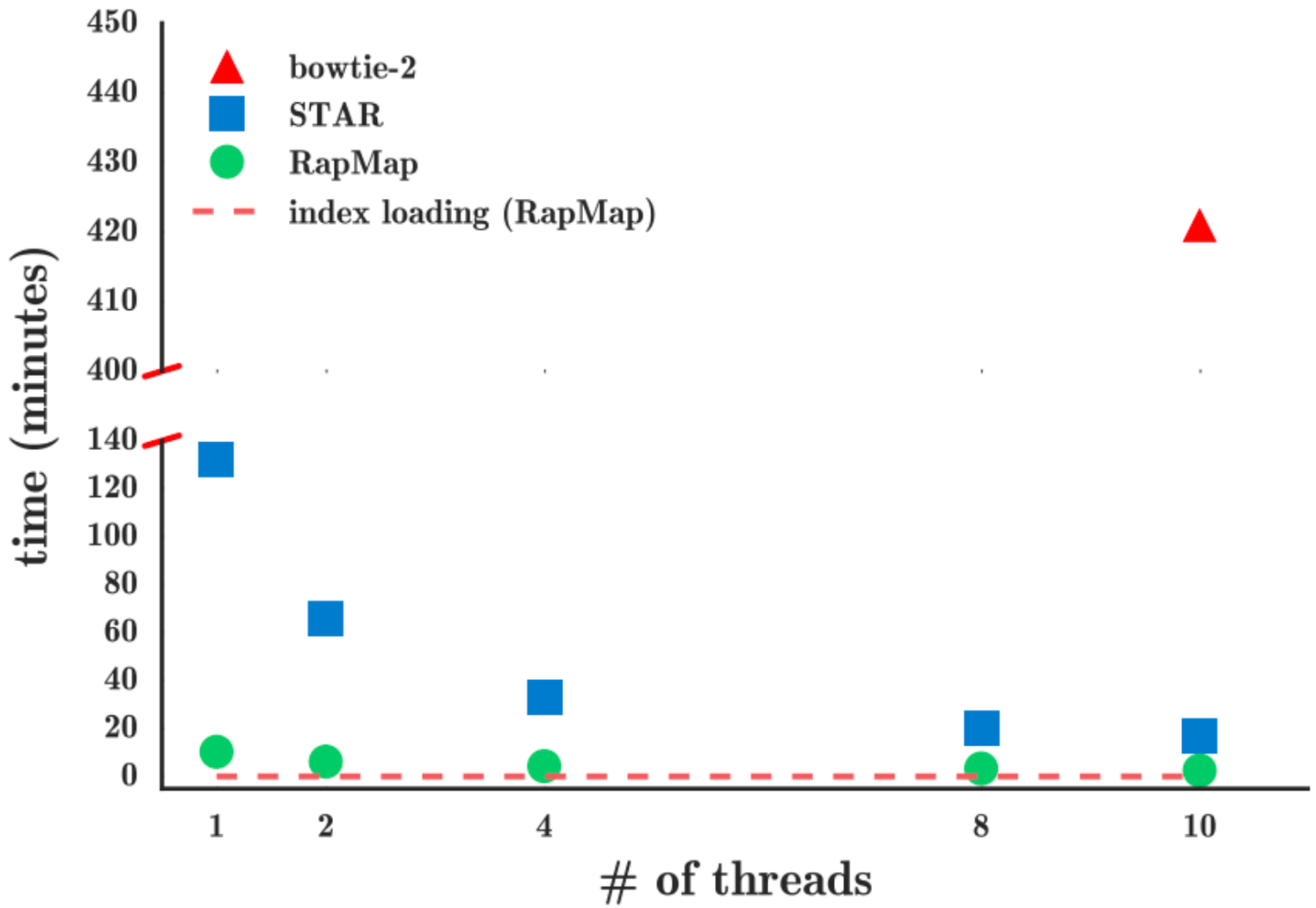
Given: A collection of sequencing reads, and some target sequence (e.g. a genome)

Find: For each read, all locations where the read is within edit distance ϵ of the reference, ~~and the edits that achieve this distance.~~



$d : \Sigma^{|u|} \times \Sigma^{|v|} \rightarrow \mathbb{Z}$ and ϵ (maximum edit distance)

Why relax the problem ?



- Analysis on 75Million (76bp) PE reads to human transcriptome; RapMap does in just **minutes**.

Phylogeny of read-alignment

Aligning (Mapping) NGS Reads

DNA-sequencing

Genome (Spliced)

- Aligns RNA-seq reads to genome
- Challenge of **Spliced Alignment**
- Example: topHat, STAR, HISAT(1/2)

RNA-sequencing

Transcriptome

- Aligns RNA-seq reads to transcriptome
- Challenge of **high multi-mapping rate**
- Example: Bowtie(1/2), BWA(SW/MEM)

Aligner

- Base-to-Base Alignment (CIGAR string)
- Selective-Alignment

Mapper

- **NO** CIGAR string
- RapMap

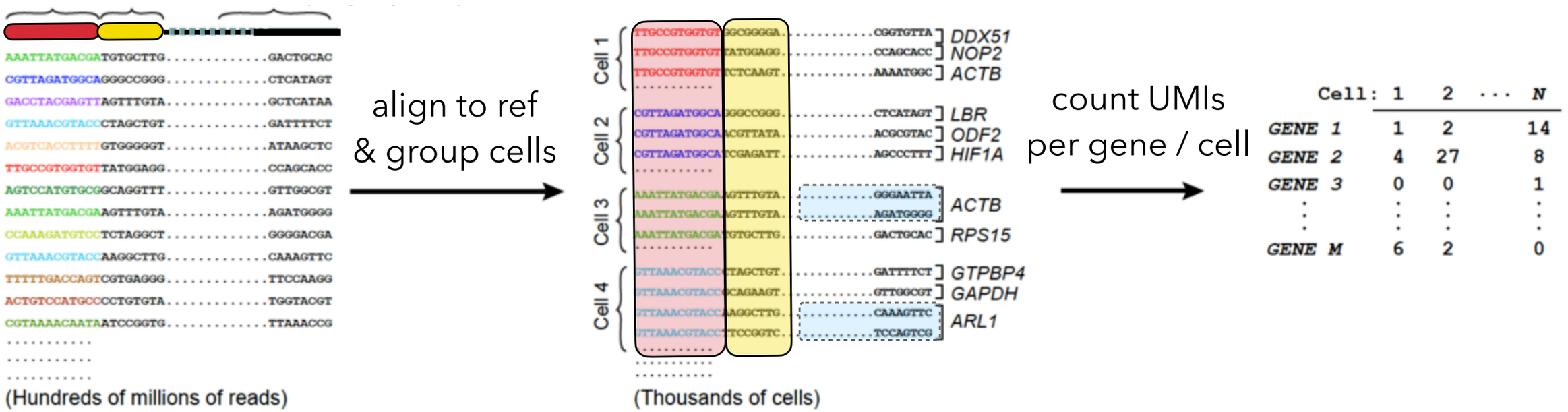
Srivastava, Avi, et al. "Alignment and mapping methodology influence transcript abundance estimation." *BioRxiv* (2019): 657874.

Recap

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

dscRNA-seq quantification methods

CellRanger, (2017) ** – STAR aligner

Zheng, Grace XY, et al. "Massively parallel digital transcriptional profiling of single cells." *Nature communications* 8.1 (2017): 1-12.

UMI-tools, (2017) – STAR aligner

Smith, Tom, Andreas Heger, and Ian Sudbery. "UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy." *Genome research* 27.3 (2017): 491-499.

Alevin, (2018) – Salmon's quasi-mapping / selective -alignment

Srivastava, Avi, et al. "Alevin efficiently estimates accurate gene abundances from dscRNA-seq data." *Genome biology* 20.1 (2019): 65.

STAR-solo, (2019) * – STAR aligner

Dobin, Alexander, et al. "STAR: ultrafast universal RNA-seq aligner." *Bioinformatics* 29.1 (2013): 15-21. [No preprint for single-cell version]

Hera-T, (2019) * – Hera's aligner

Tran, Thang, et al. "Hera-T: an efficient and accurate approach for quantifying gene abundances from 10X-Chromium data with high rates of non-exonic reads." *bioRxiv* (2019): 530501.

bustools, (2019) * – Kallisto's pseudo-alignment

Melsted, Páll, et al. "Modular and efficient pre-processing of single-cell RNA-seq." *BioRxiv* (2019): 673285.

This is not an exhaustive list, methods were selected based on read-alignment strategy

*** Not Peer-reviewed**
**** No methods paper**

One shortcoming with current dscRNA-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." *Genome biology*20.1 (2019): 65.

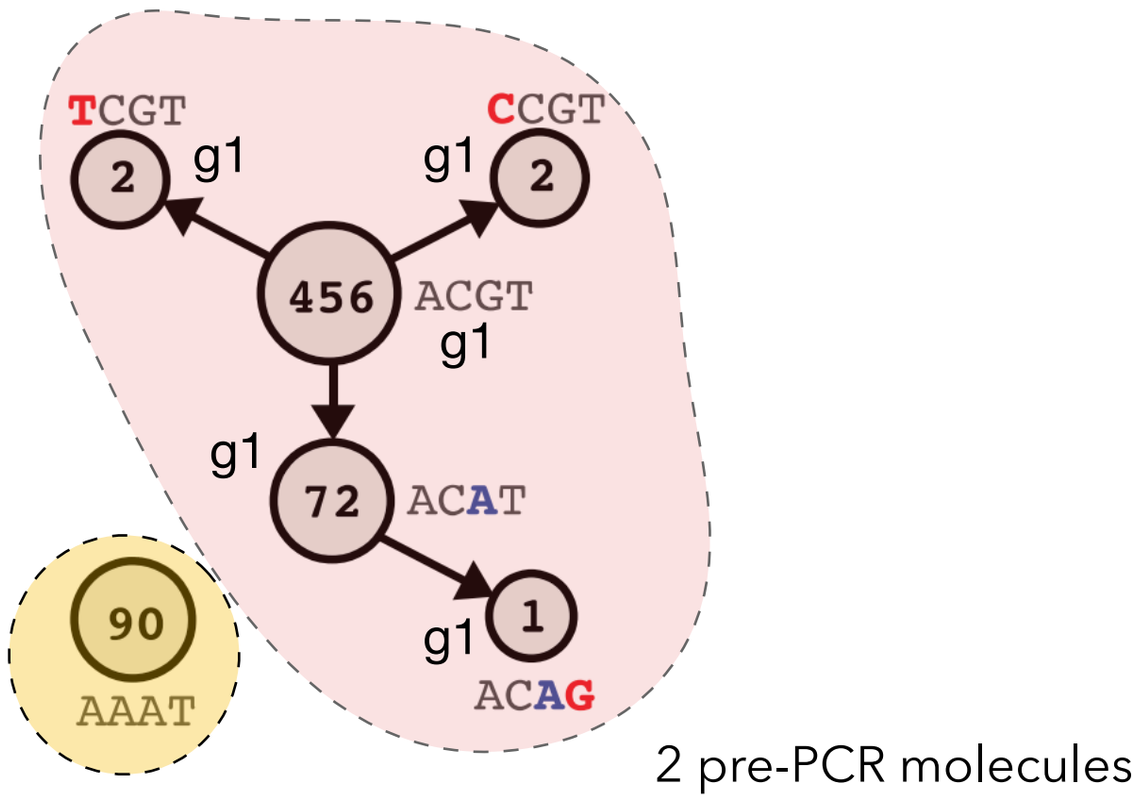
UMI-Deduplication

Given: For each cell, a set of UMI's with their frequencies based on gene mapping.

Find: Number of pre-PCR molecules by deduplicating the set of UMIs.

UMI-tools directional approach

- Start with highest frequency UMI & recursively add an edge to all the UMIs within 1-edit distance.
- The direction of the edge is from high frequency node to low frequency node.
- Collapse the connected component into 1 pre-PCR molecule.



2 pre-PCR molecules

Smith, Tom, Andreas Heger, and Ian Sudbery. "UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy." *Genome research* 27.3 (2017): 491-499.

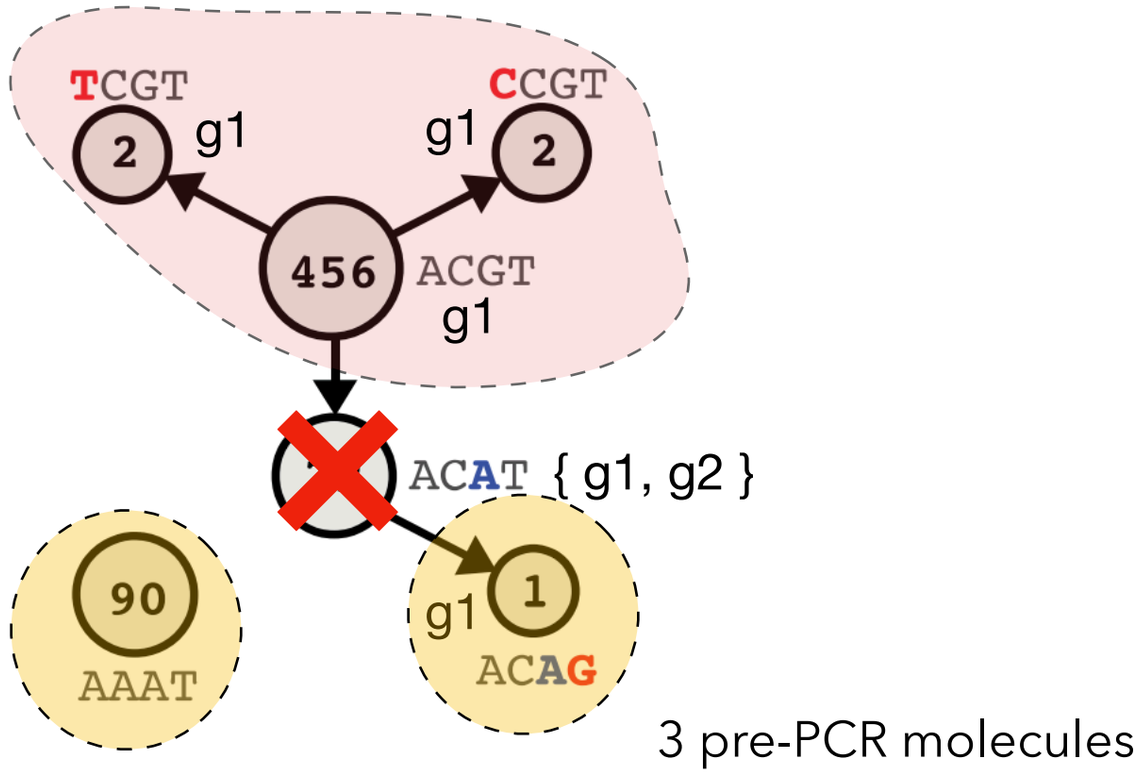
UMI-Deduplication

Given: For each cell, a set of UMI's with their frequencies based on gene mapping.

Find: Number of pre-PCR molecules by deduplicating the set of UMIs.

UMI-tools directional approach

- Start with highest frequency UMI & recursively add an edge to all the UMIs within 1-edit distance.
- The direction of the edge is from high frequency node to low frequency node.
- Collapse the connected component into 1 pre-PCR molecule.
- Dropping multi-mapping reads breaks the network.



3 pre-PCR molecules

Smith, Tom, Andreas Heger, and Ian Sudbery. "UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy." *Genome research* 27.3 (2017): 491-499.

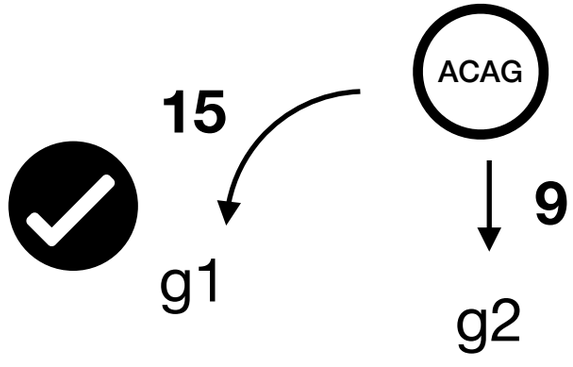
UMI-Deduplication

Given: For each cell, a set of UMI's with their frequencies based on gene mapping.

Find: Number of pre-PCR molecules by deduplicating the set of UMIs.

CellRanger (STAR-solo) approach

- CellRanger follows bottom-up instead of top-down approach i.e. unlike UMI-tools way of assigning UMI to gene, cellranger assign genes to UMIs.
- In the first round, within each cell, cellranger assigns only the "gene-unique" mapped read to a UMI, generating a frequency distribution of genes for each UMI.
- In the second round of processing, each UMI is assigned to the highest frequency gene from its distribution.



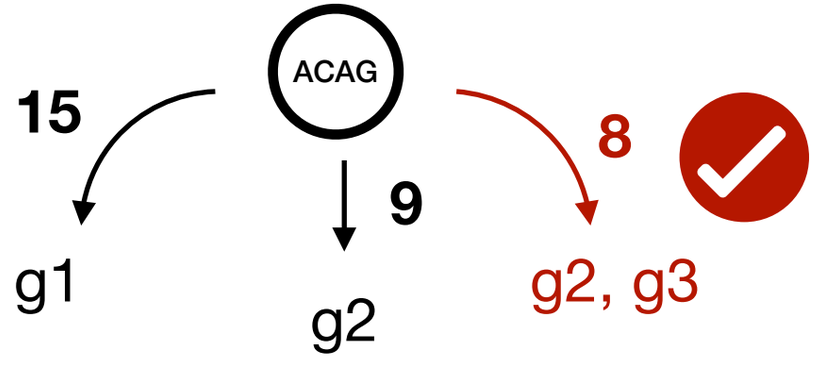
UMI-Deduplication

Given: For each cell, a set of UMI's with their frequencies based on gene mapping.

Find: Number of pre-PCR molecules by deduplicating the set of UMIs.

CellRanger (STAR-solo) approach

- CellRanger follows bottom-up instead of top-down approach i.e. unlike UMI-tools way of assigning UMI to gene, cellranger assign genes to UMIs.
- In the first round, within each cell, cellranger assigns only the "gene-unique" mapped read to a UMI, generating a frequency distribution of genes for each UMI.
- In the second round of processing, each UMI is assigned to the highest frequency gene from its distribution.



Parsimonious UMI Graph (PUG) resolution

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

Each **vertex** (v ∈ V) is a tuple:

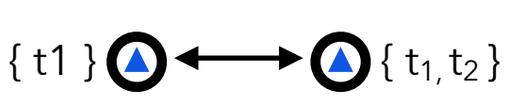
- **eq_v** : the set of transcripts to which the read maps
- **s_v** : the UMI sequence tagging the read

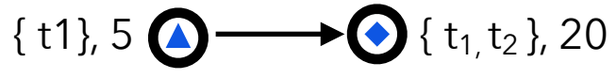
Each v has count, *c*(v) – number of "equivalent" reads.

"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* v → u)





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*.

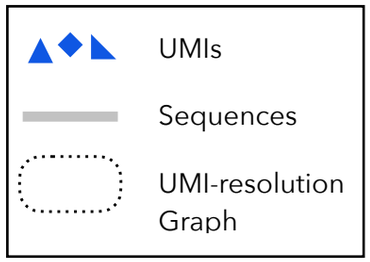
We seek <u>parsimony</u>

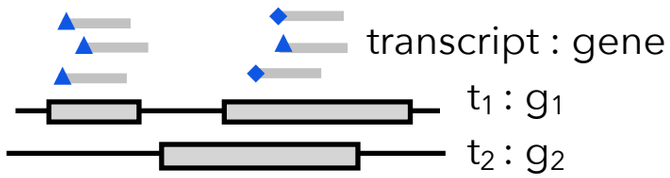
Each component can be ascribed to the reads sequenced from a *single initial molecule* (before amplification).

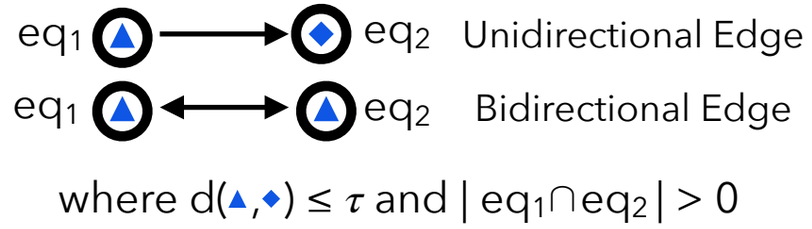
Each *monochromatic arborescence* is a set of vertices that can all be described as "reads" coming from the same, original transcript.

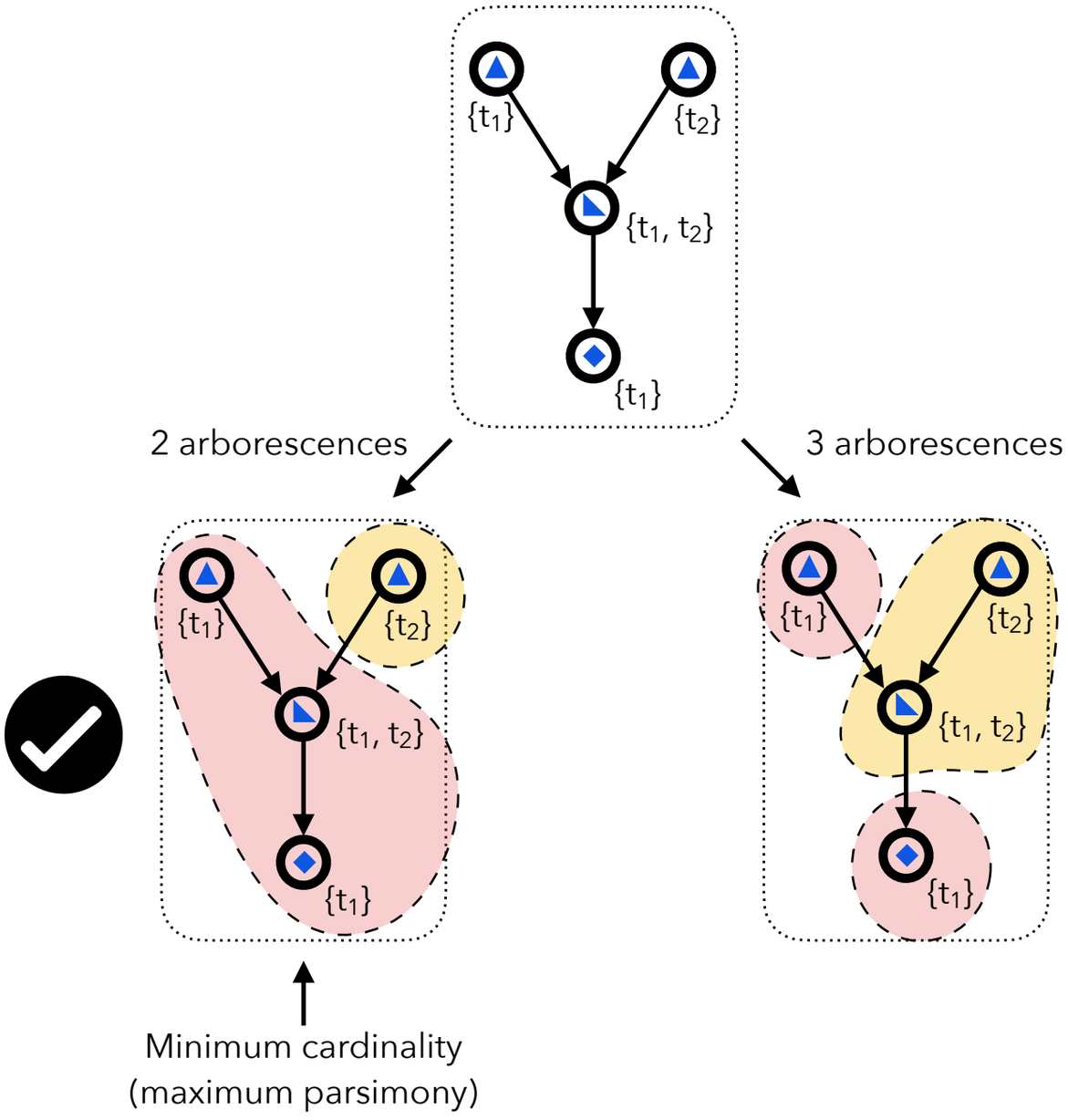
The decision problem is *NP-complete* (reduction from DOMINATING SET); but experimental instances are usually *simple* and we adopt a *greedy heuristic*. (solutions are known to be optimal in > 80% of cases)

UMI Resolution

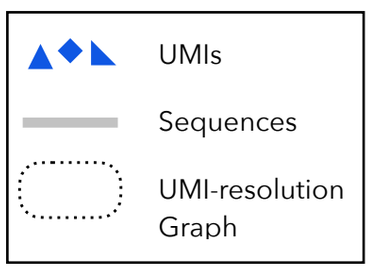


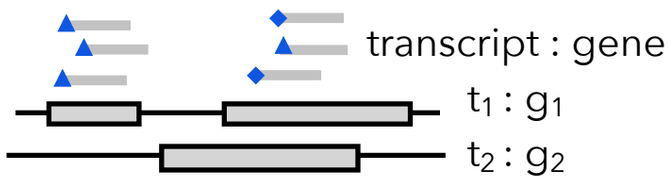


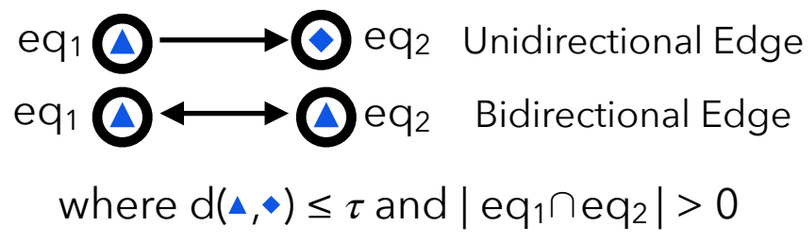


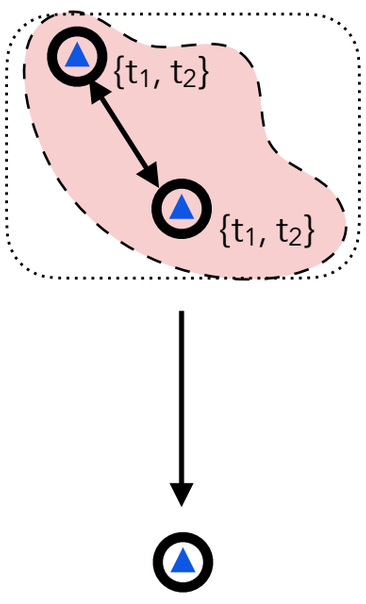


UMI Resolution





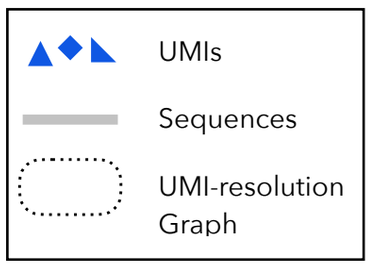


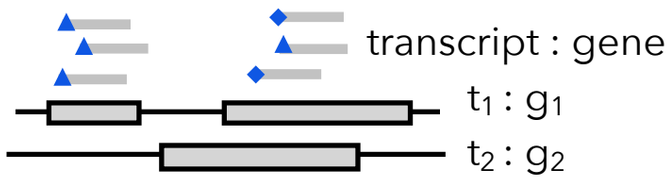


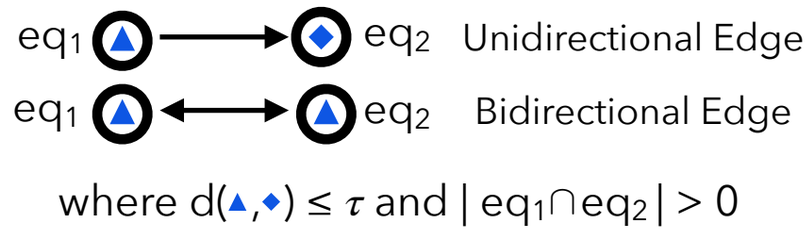
Equally-parsimonious under t1 **or** t2

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

EM & Tiers Characterization



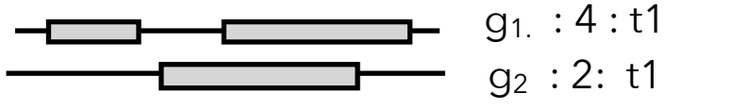




tier 1

{g1} x 4

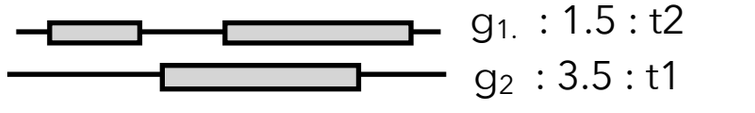
{g2} x 2



tier 2

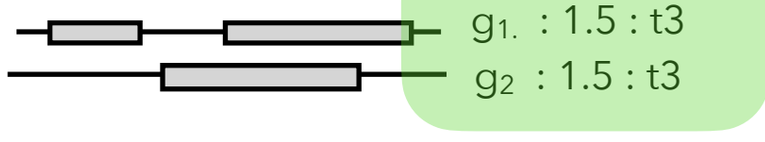
{g1, g2} x 3

{g2} x 2

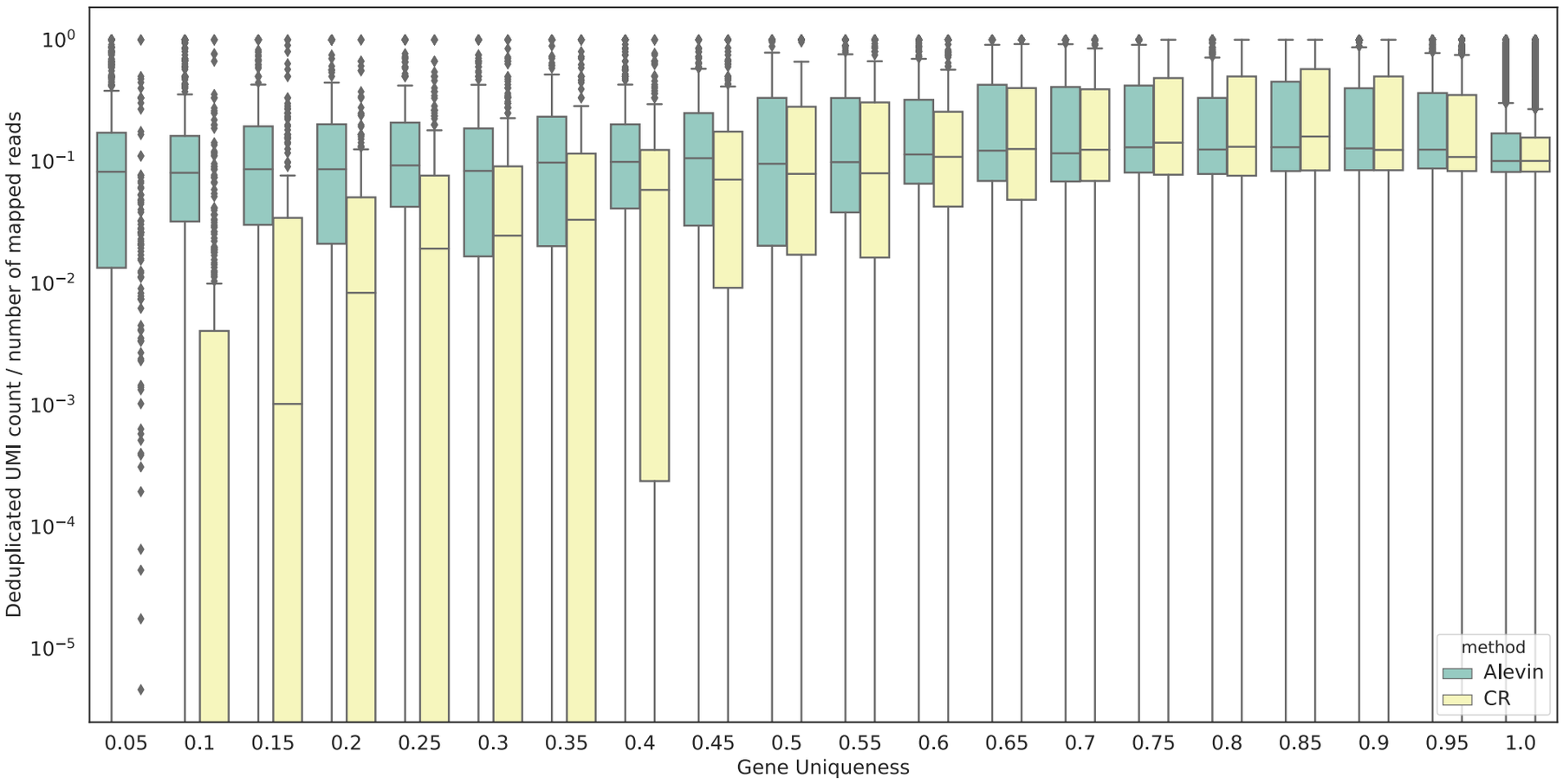


tier 3

{g1, g2} x 3



Discarding gene-ambiguous reads does not affect all genes equally

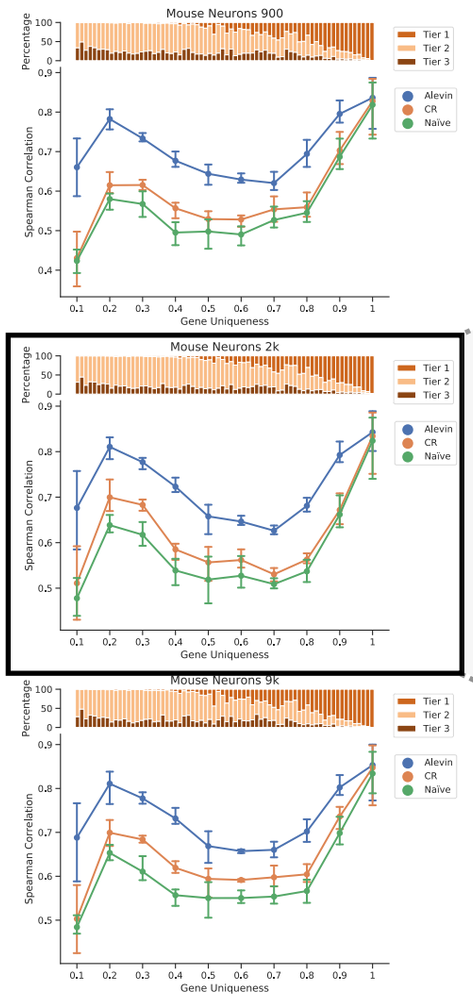


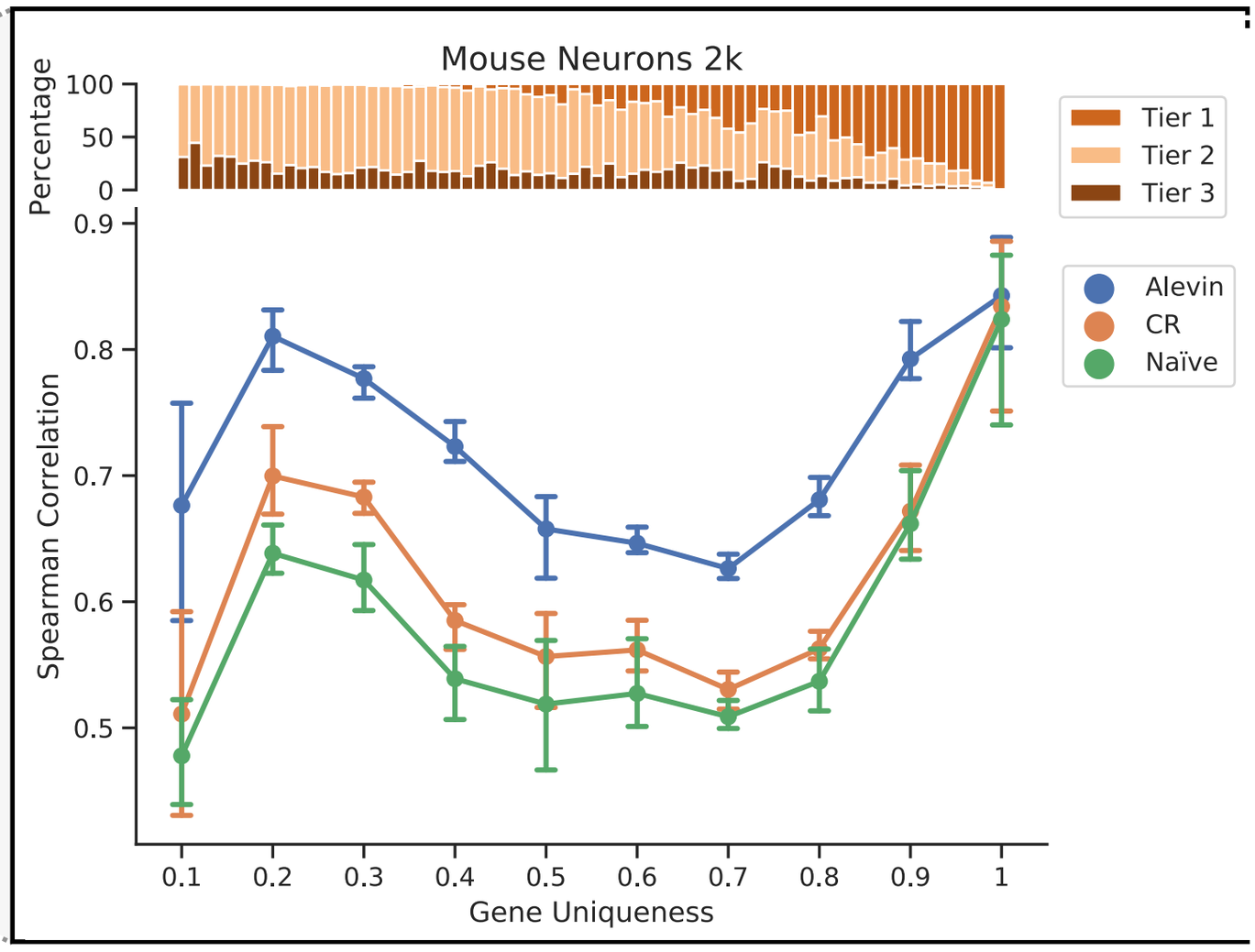
Bin number	Human	Mouse
1	2538	3984
2	1094	1398
3	813	996
4	852	944
5	908	1137
6	1044	1178
7	1339	1392
8	1943	1615
9	3360	2226
10	44397	38509

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+





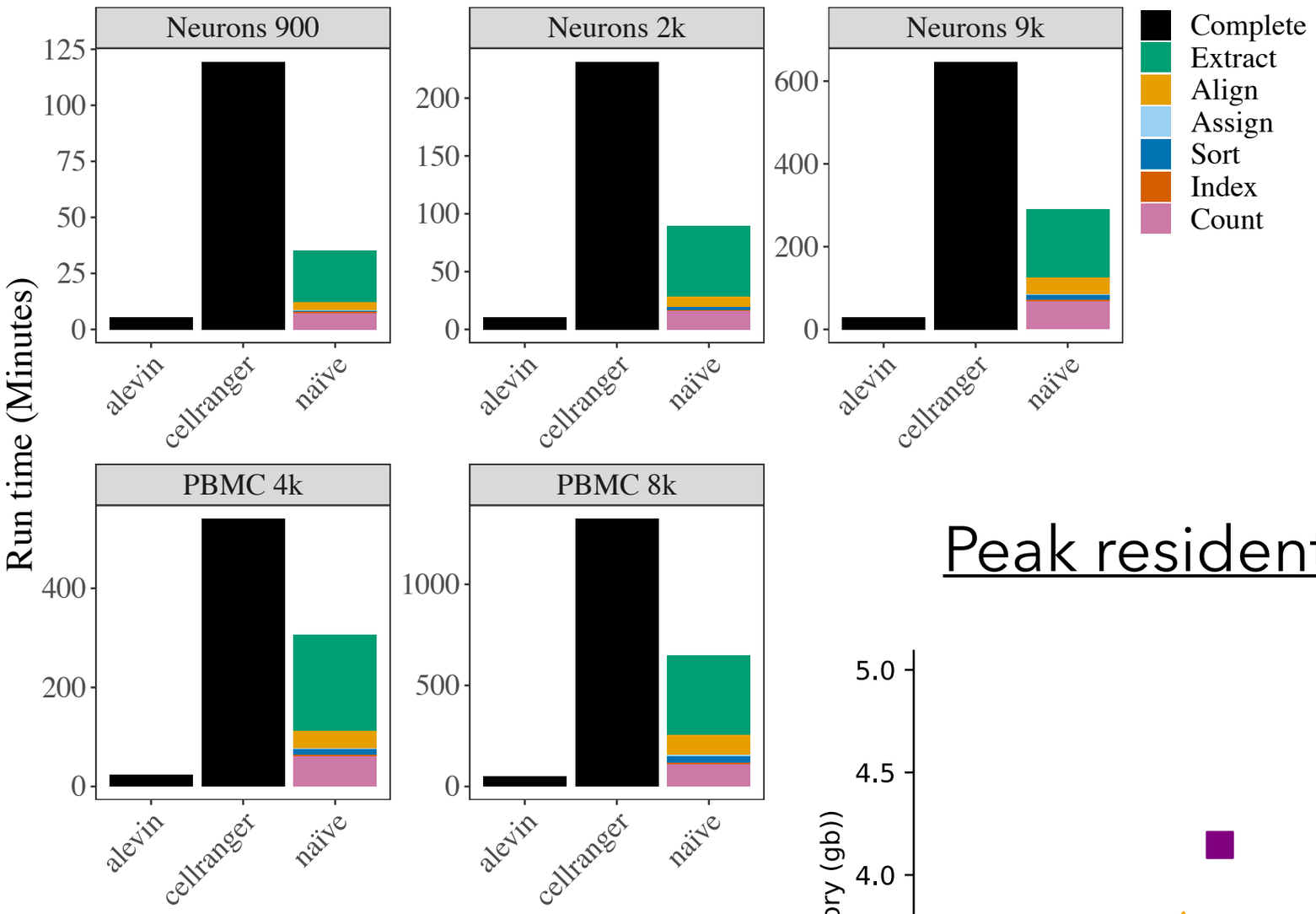
For each scRNA-seq dataset, we obtained multiple bulk samples from the same tissue. We compare gene-level quants in bulk (quantified with RSEM*) to average expressions across single cells.

*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.

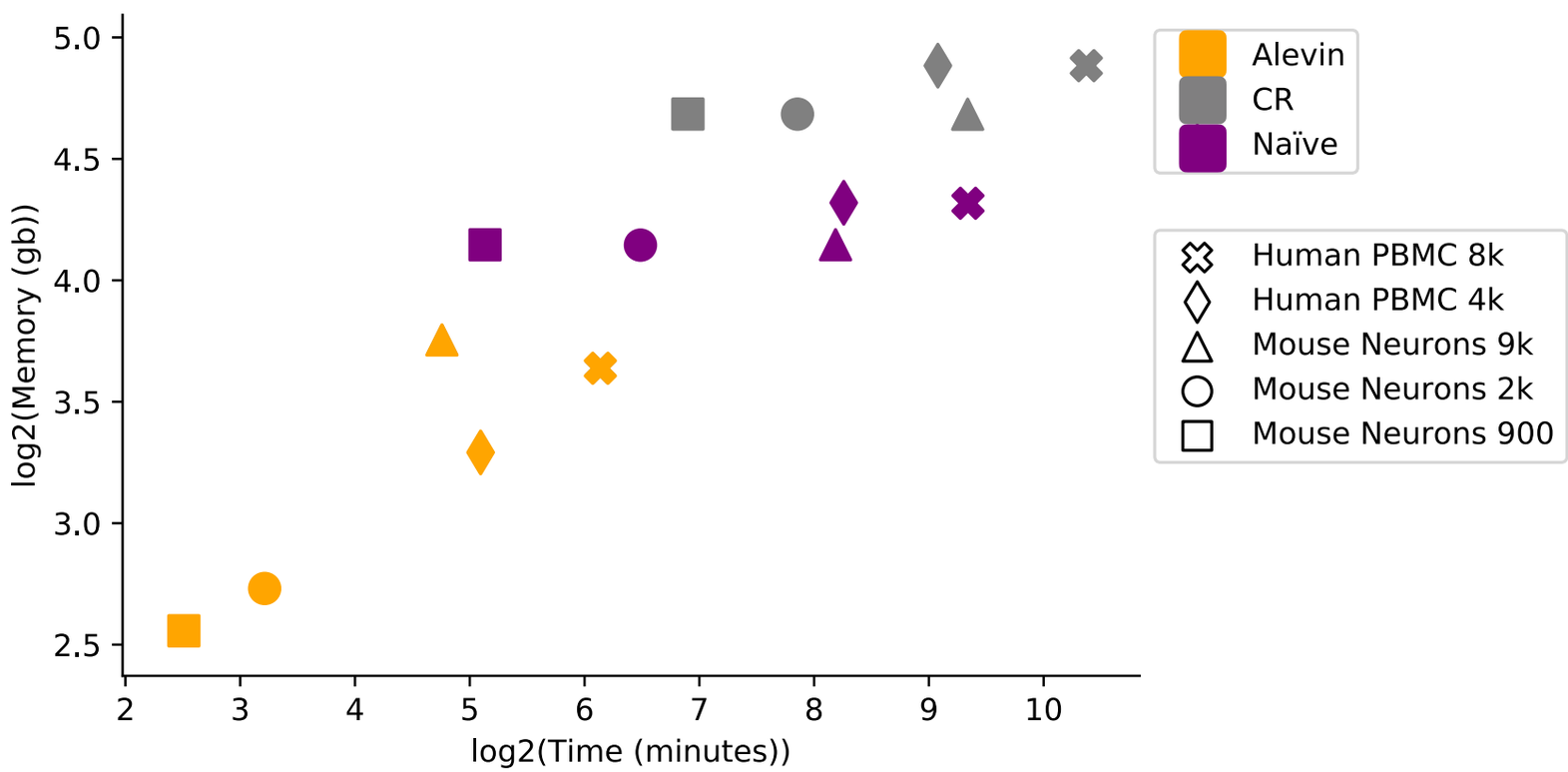
+Zheng, Grace XY, et al. "Massively parallel digital transcriptional profiling of single cells." Nature communications 8 (2017): 14049.

alevin is fast & efficient

Wall clock time (16 threads)



Peak resident memory (16 threads)



alevin-tutorial

https://combine-lab.github.io/alevin-tutorial/

Alevin w/ Feature Barcodes

Feature Barcoding based Single-Cell Quantification with alevin …

20 Apr 2020 • on alevin

Spatial Alevin

Spatial Single-Cell Quantification with alevin …

6 Apr 2020 • on alevin

How to use alevin with iSEE

Preparing alevin output for exploration with iSEE …

25 Mar 2020 • on alevin

Alevin Velocity

RNA Velocity with alevin …

22 Mar 2020 • on alevin

Selective Alignment

Fast is Good but Fast and accurate is better ! …

30 Oct 2019 • on selective-alignment