Introduction

Caltech Bioinformatics Symposium
February 14, 2019

Overview

- 9:00am – 10:00am **Lior Pachter** (Introduction)
- 10:00am – 10:15am Break
- 10:15am – 11:00am **Páll Melsted** (BUS format and single-cell RNA-seq processing)
- 11:00am – 11:30am **Fan Gao** (Benchmarks of kallisto and bustools)

- Afternoon (1pm—3pm) workshop in the Morgan Library with Eduardo Beltrame, Fan Gao, Dongyi (Lambda) Lu, Páll Melsted, Lior Pachter
- **Thanks to Fan Gao and Lisa Sledd for organization**

Bioinformatics Resource Center

Bioinformatics
Resource Center

HOME ABOUT SERVICES TEAM EVENTS CONTACT

bioinformatics.caltech.edu

Fan Gao - bioinformatics
Ingileif Hallgrimsdóttir - statistics

The growth of RNA-seq

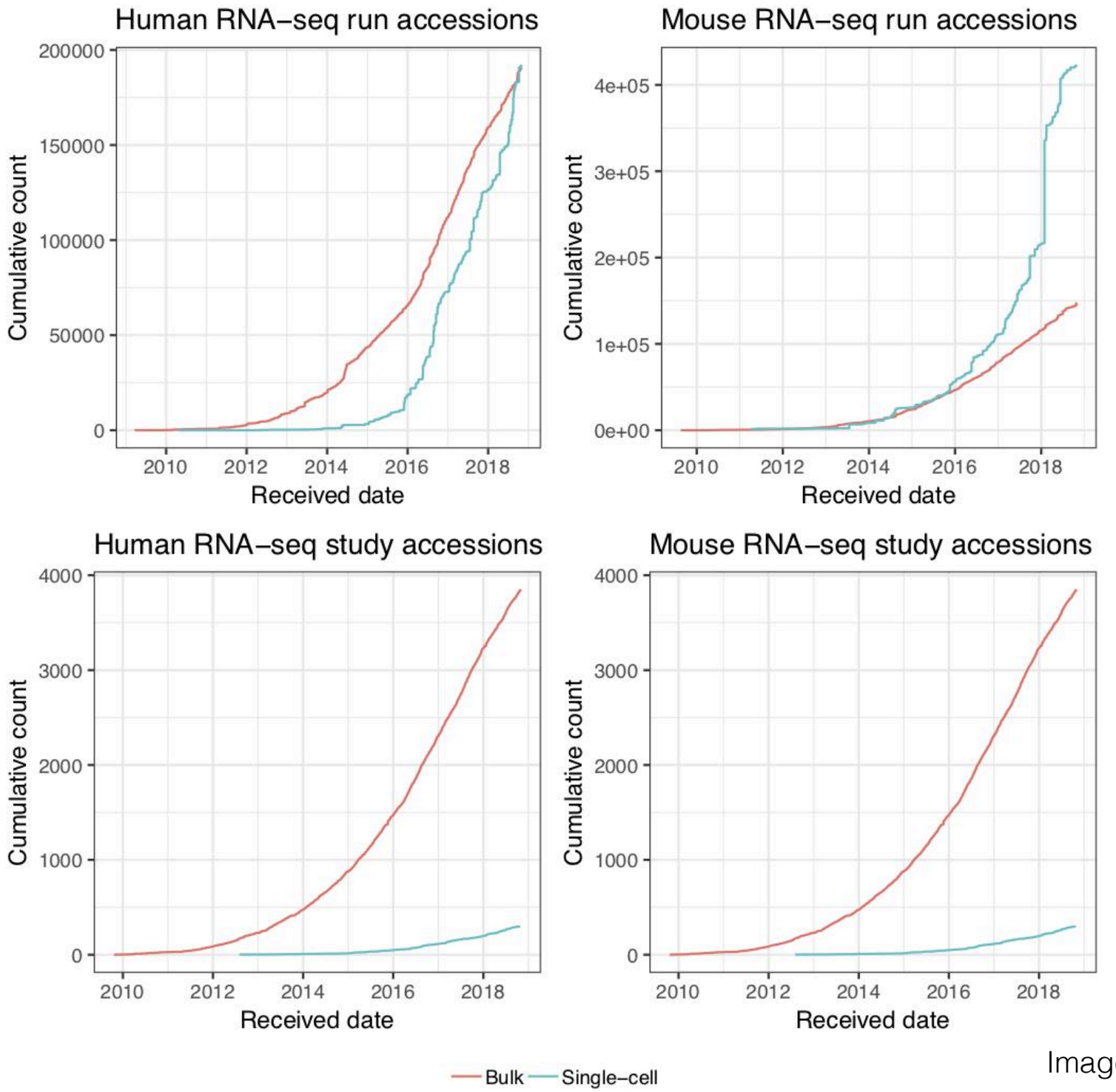
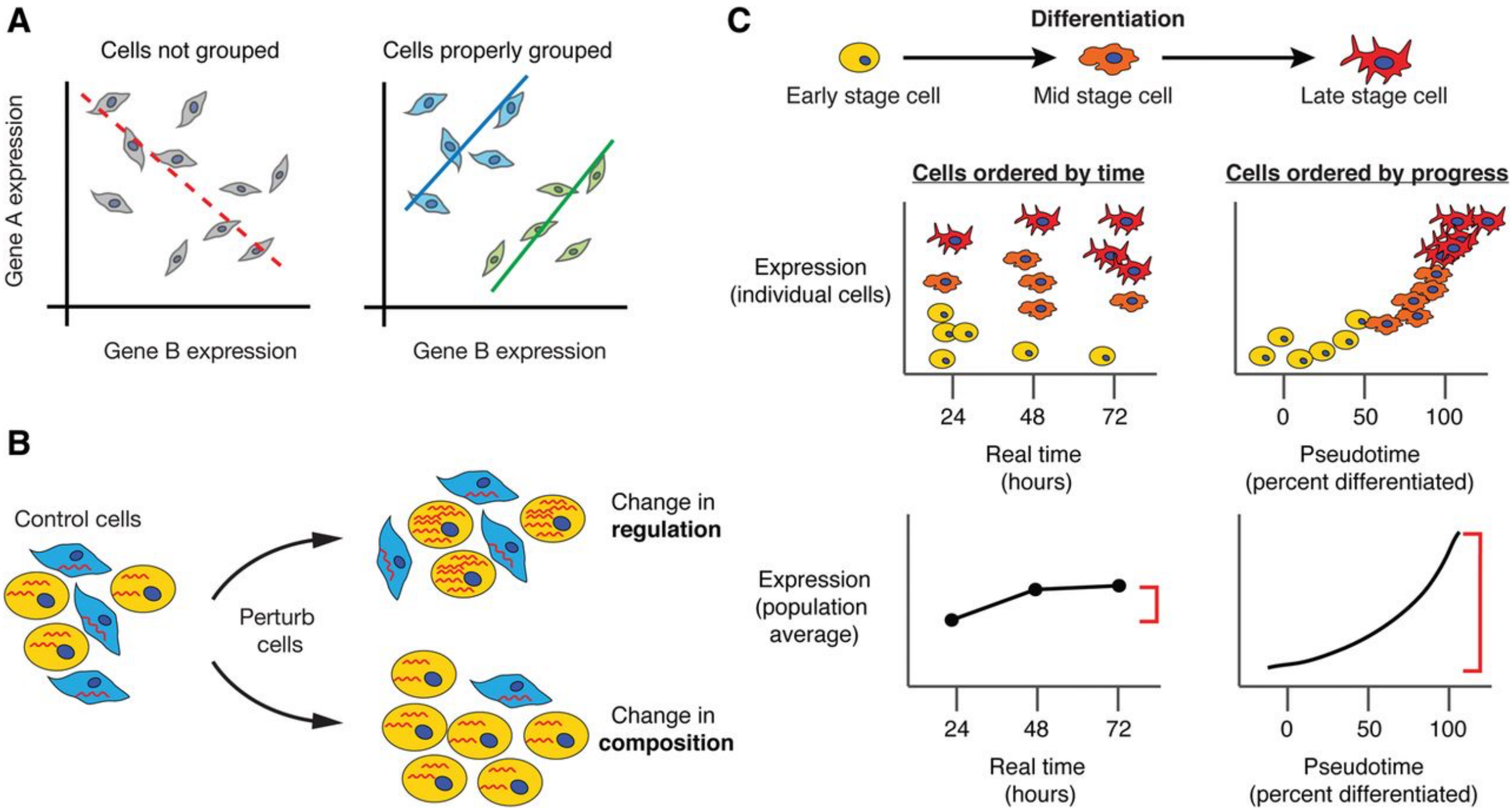


Image credit: Ben Langmead

Why?



Defining cell types and states with single cell genomics, Cole Trapnell 2015

Deluge of data

- Datasets are growing not only in number but in size and complexity
- With single-cell RNA-seq technology individual researchers can easily generate hundreds of millions of reads worth of RNA-seq data **per experiment**.
- The scale of projects involving single-cell RNA-seq today can rival that of projects that only a few years ago required a consortium of scientists for analysis.
- Traditional analysis of such data is computationally intensive, expensive, and complicated.

Democratizing analysis

- When analysis requires computational power beyond what's easily available to the average biologist, this adds a barrier between them and their data.
- The ability to analyze one's own data can reduce dependence on external support.
- The ability to explore one's data can lead to new questions and discoveries.

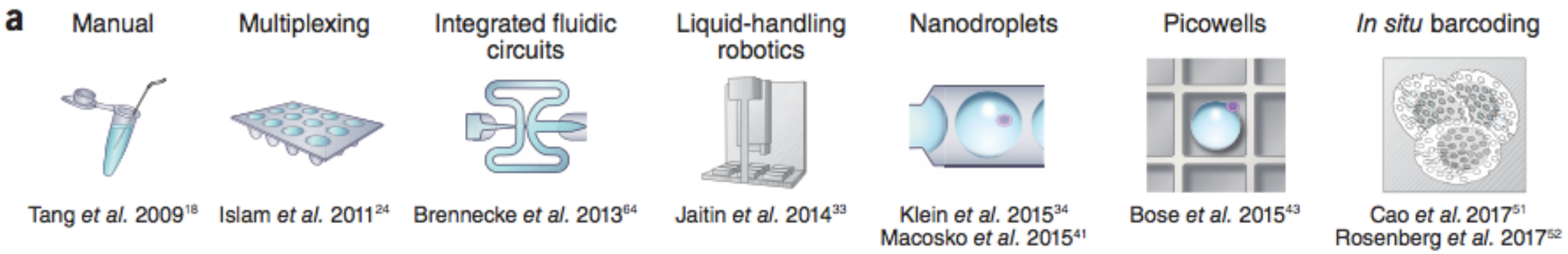
Usability

- We want data analysis to be not just *possible* but *usable*
- Analysis should make it easy to say “What if…”
- When an analysis takes huge amounts of computer time, it limits exploration
- Personal anecdote (Nicolas Bray): Once waited two weeks for an analysis of a large RNA-seq dataset to finish, only to have a new genome annotation be released the next day

An introduction to the technologies

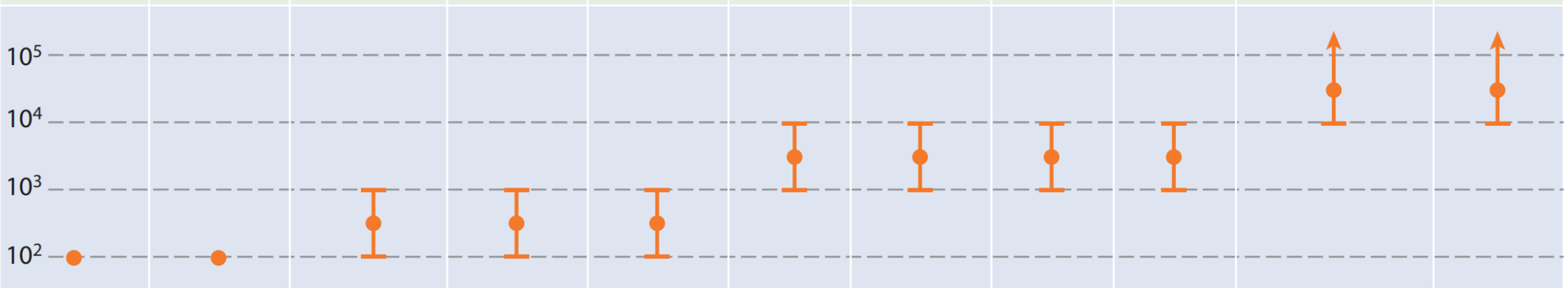
- ***Universal*** in terms of cell size, type and state.
- ***In situ*** measurements.
- No ***minimum input*** of number of cells to be assayed.
- Every cell is assayed, i.e.100% ***capture rate***.
- Every transcript in every cell is detected, i.e.100% ***sensitivity***.
- Every transcript is identified by its ***full-length sequence***.
- Transcripts are assigned correctly to cells, e.g. no ***doublets***.
- Additional ***multimodal*** *measurements*
- ***Cost*** effective per cell.
- ***Easy*** to use.

A decade of single-cell RNA-seq



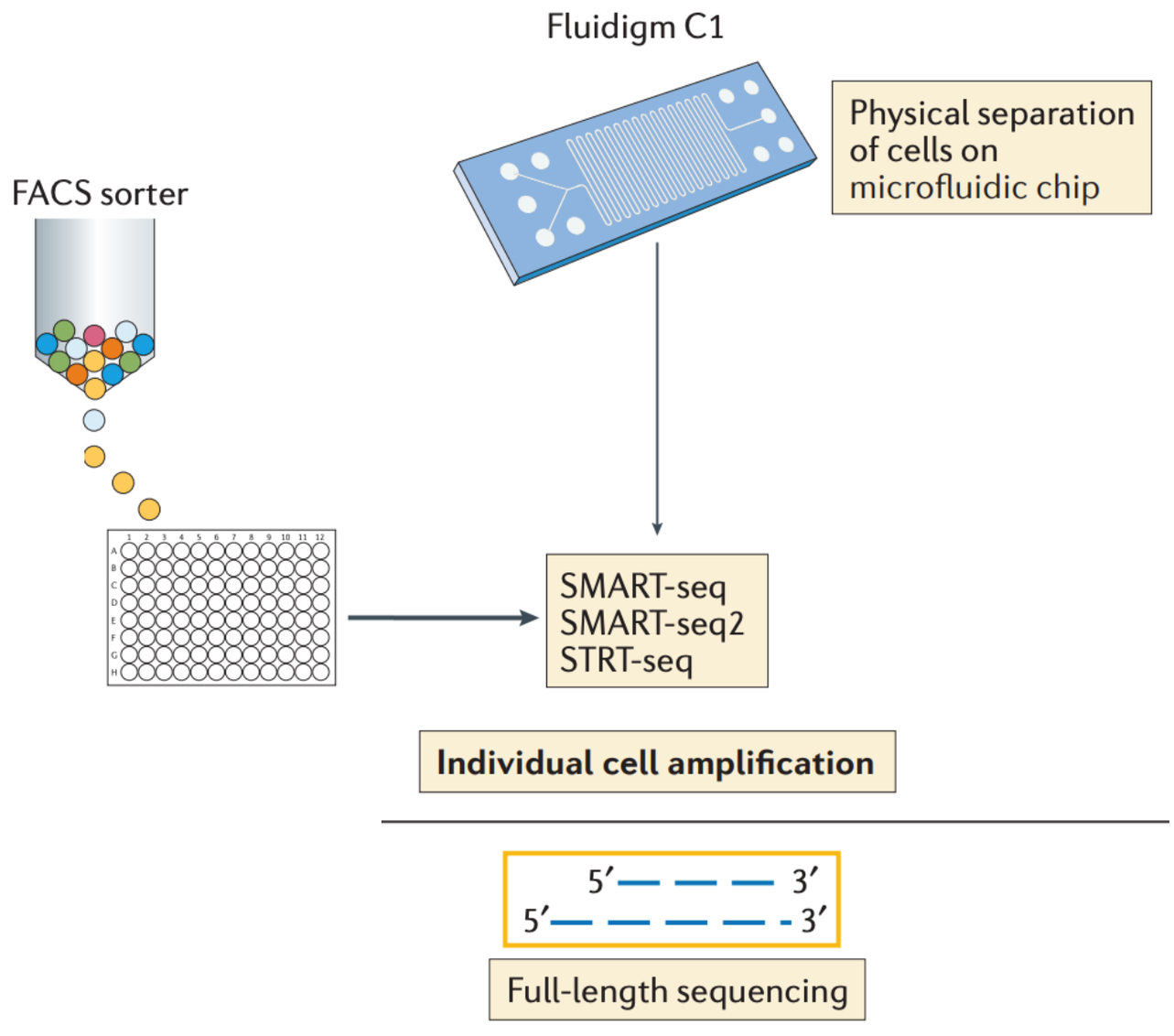
Svensson et al. 2018. DOI:10.1038/nprot.2017.149

Popular single-cell RNA-seq protocols

	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H and DNA pol I	TSO	PolyA tailing and primer ligation	RNase H and DNA pol I	TSO	RNase H and DNA pol I	TSO	TSO	RNase H and DNA pol I	TSO
Full-length cDNA synthesis?	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
Barcode addition	Library PCR with barcoded primers	Barcoded RT primers	Barcoded TSOs	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers and library PCR with barcoded primers	Ligation of barcoded RT primers
Pooling before library?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Library amplification	PCR	In vitro transcription	PCR	PCR	In vitro transcription	PCR	In vitro transcription	PCR	PCR	PCR	PCR
Gene coverage	Full-length	3′	5′	3′	3′	3′	3′	3′	3′	3′	3′
Number of cells per assay											

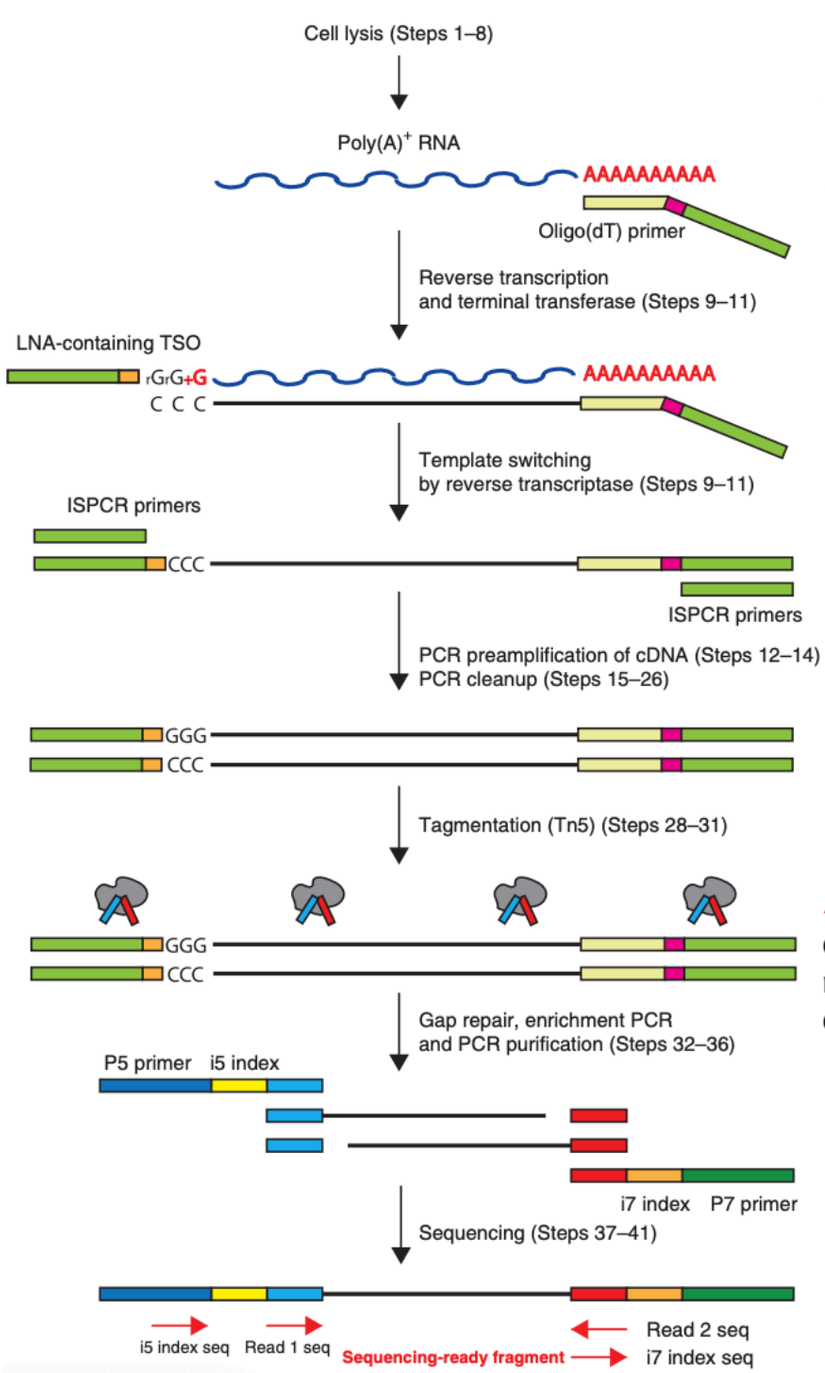
Chen et al. 2018. DOI:10.1146/annurev-biodatasci-080917-013452

Physical separation of cells in wells



Papalexi and Satija, 2017

Details of SMART-Seq2



▲ **CRITICAL STEP** All the experiments must be performed under a UV-sterilized hood with laminar flow, and all the surfaces must be free from RNase to prevent degradation of RNA and from DNA to prevent cross-contamination from previous samples. The hood must be used only for single-cell experiments up to (but excluding) the cDNA amplification step (Step 12). An ideal scenario would be to place the hood in a separate room with a positive air pressure to prevent any contaminants from being carried inside, where they might affect the experiments. The room should be equipped with a garmenting area in which the user changes into a fresh disposable lab coat, hair net, dust mask, shoe covers and vinyl gloves (powder-free).

▲ **CRITICAL STEP** Thaw all the reagents in advance and assemble the RT mix while performing denaturation (Step 7) to minimize bias.

▲ **CRITICAL STEP** The number of PCR cycles depends on the input amount of RNA. We typically use 18 cycles for single eukaryotic cells to obtain ~1–30 ng of amplified cDNA. The number of cycles can be increased for smaller cells (with less RNA content) or lowered for large cells (with more RNA).

▲ **CRITICAL STEP** The number of cycles depends on the amount of DNA used for tagmentation. If we are starting from 100 pg of amplified cDNA, we usually perform 12 PCR cycles. The optimal number of cycles depends on the sample and the experiment. It may be helpful to run a range of cycles to determine the best conditions. Above are cycling guidelines on the basis of the input DNA used for tagmentation.

40+ step protocol; 2 days of work

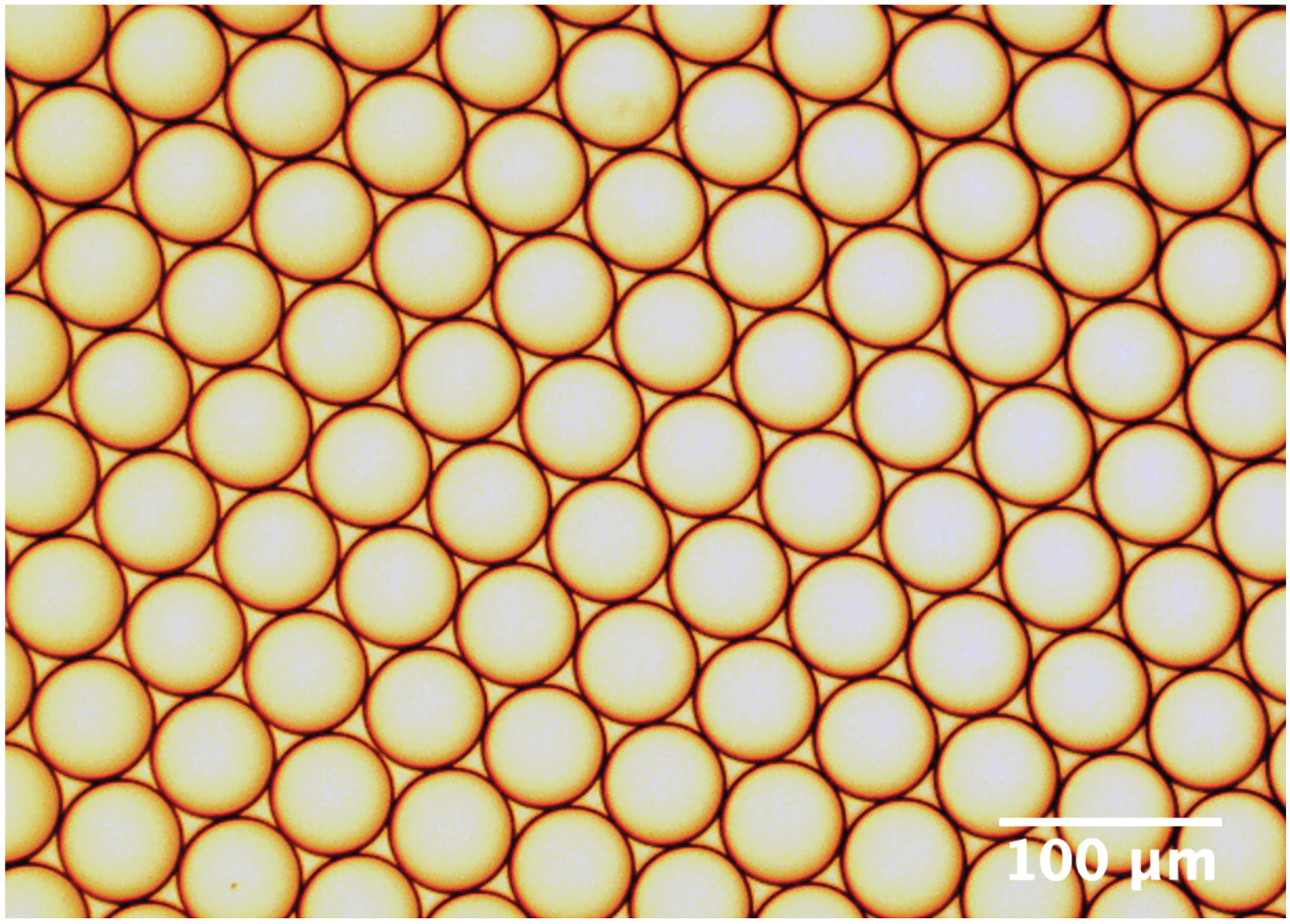
Picelli et al., 2014

Microfluidic methods

- Macosko et al., **Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets**, 2015.
- Klein et al., **Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells**, 2015.

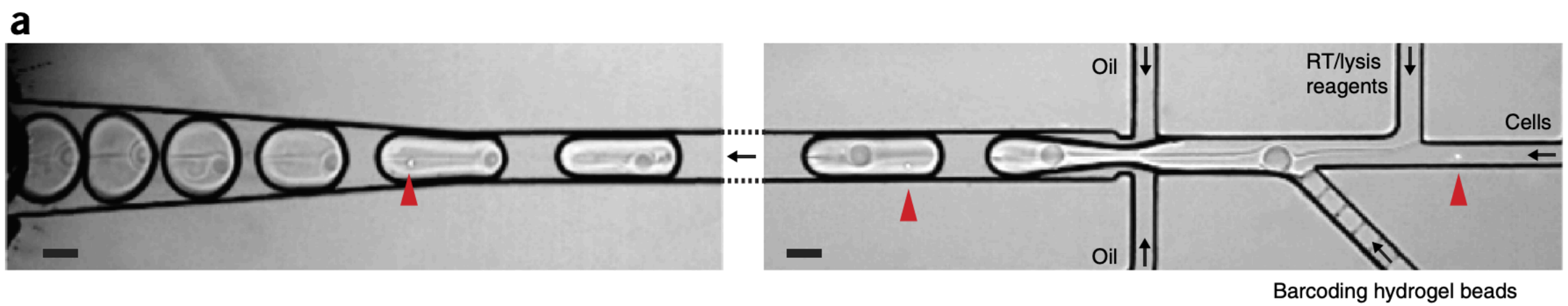
- Song, Chen, Ismagilov, Reactions in droplets in microfluidic channels, 2006.
- Guo, Rotem, Heyman and Weitz, Droplet microfluidics for high-throughput biological assays, 2012.
-

Emulsions



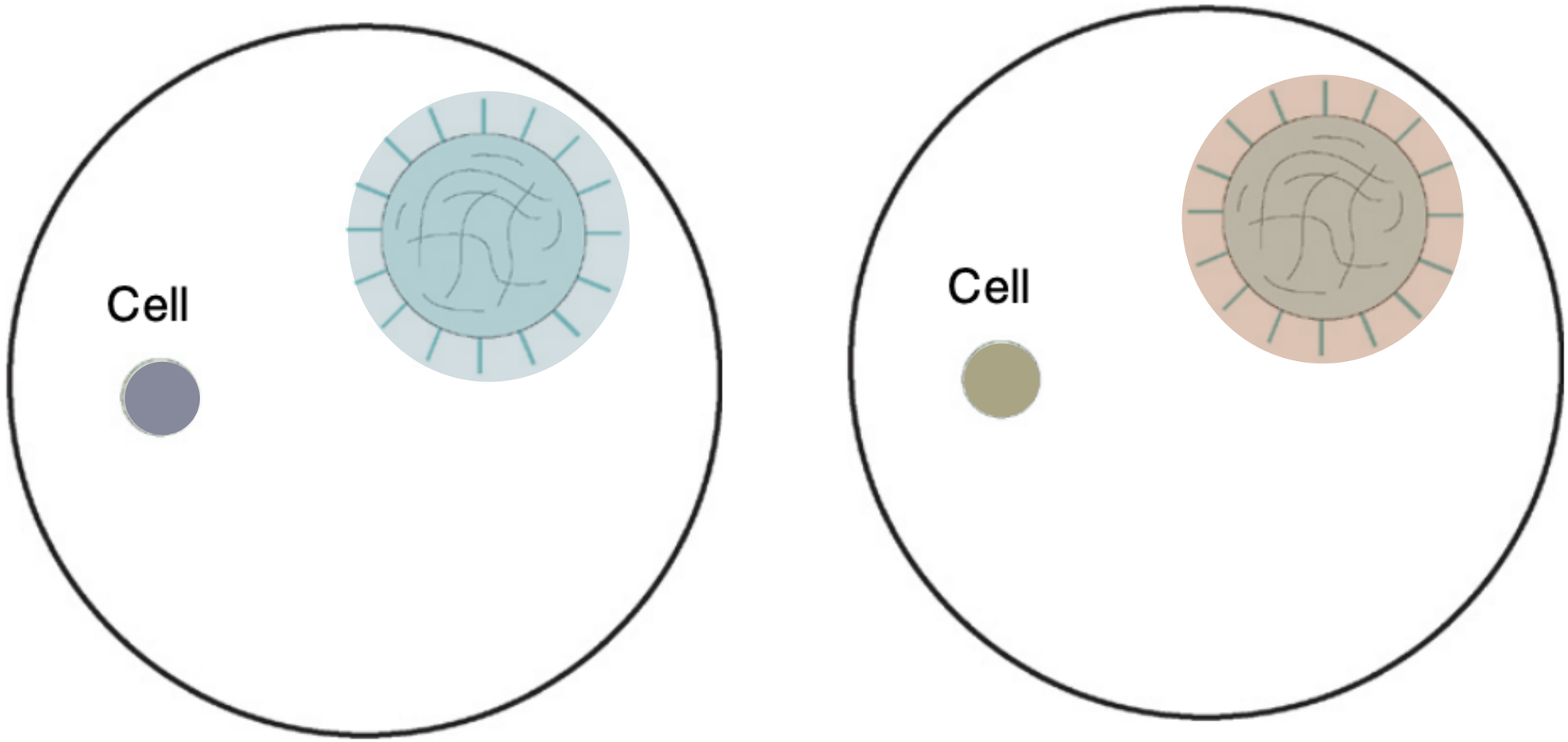
Booeshaghi et al. 2019. DOI:10.1101/521096

The inDrops approach



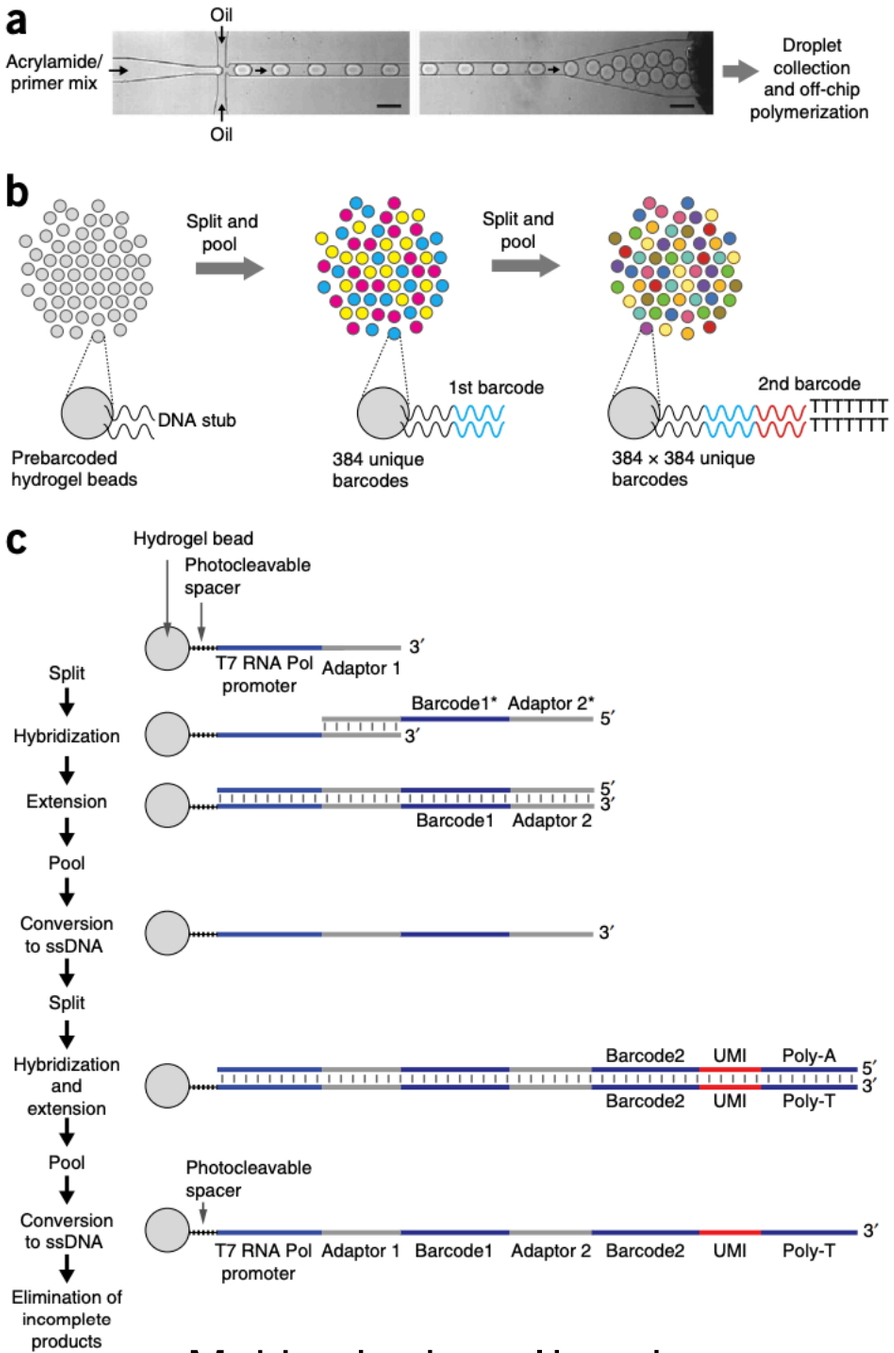
Zilionis et al. 2016. DOI:10.1038/nprot.2016.154

Beads, cells and droplets

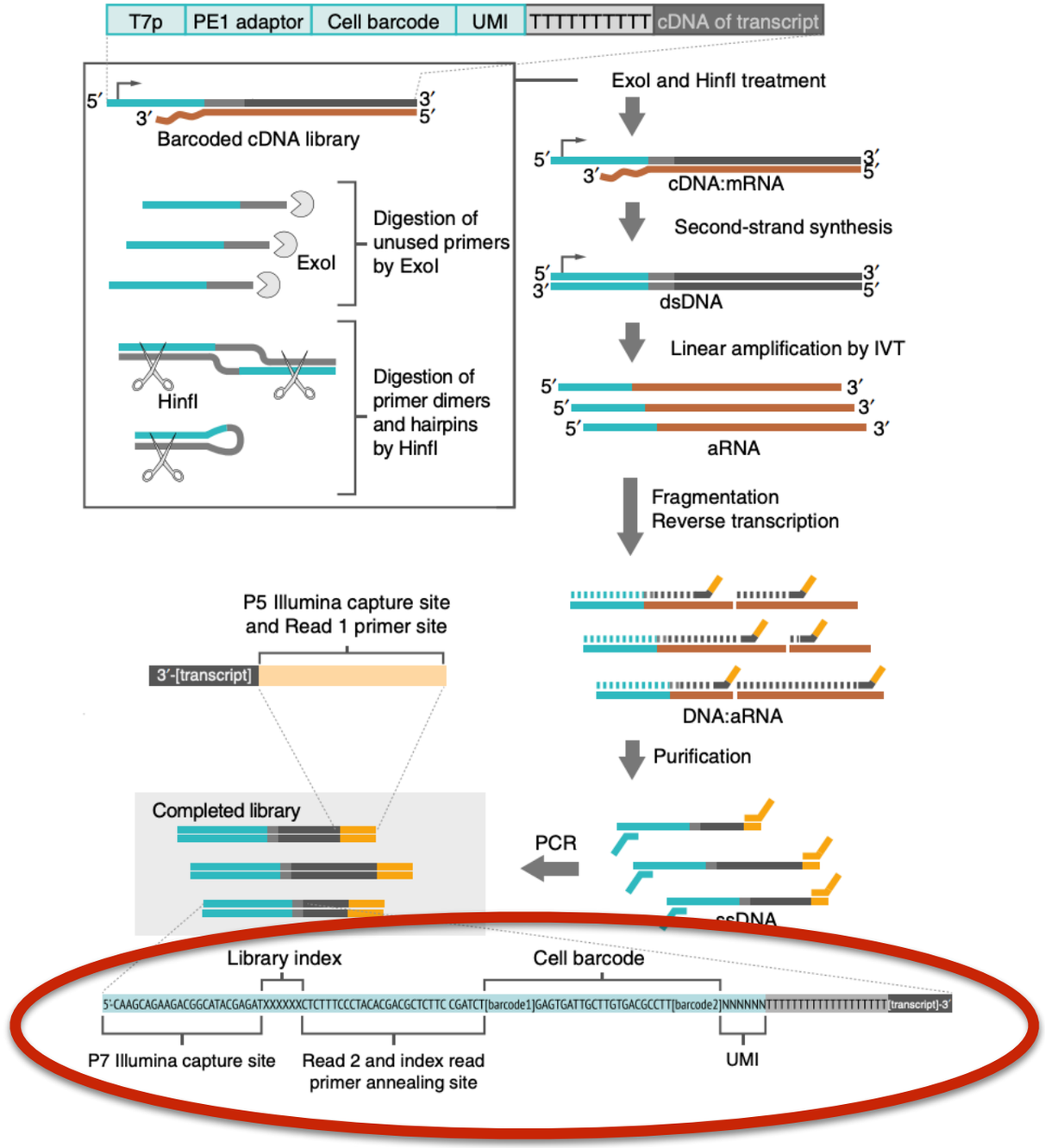


Zilionis et al. 2016. DOI:10.1038/nprot.2016.154

Details of the inDrops protocol



Making hydrogel beads



Library preparation

Zilionis et al. 2016. DOI:10.1038/nprot.2016.154

The growth of RNA-seq

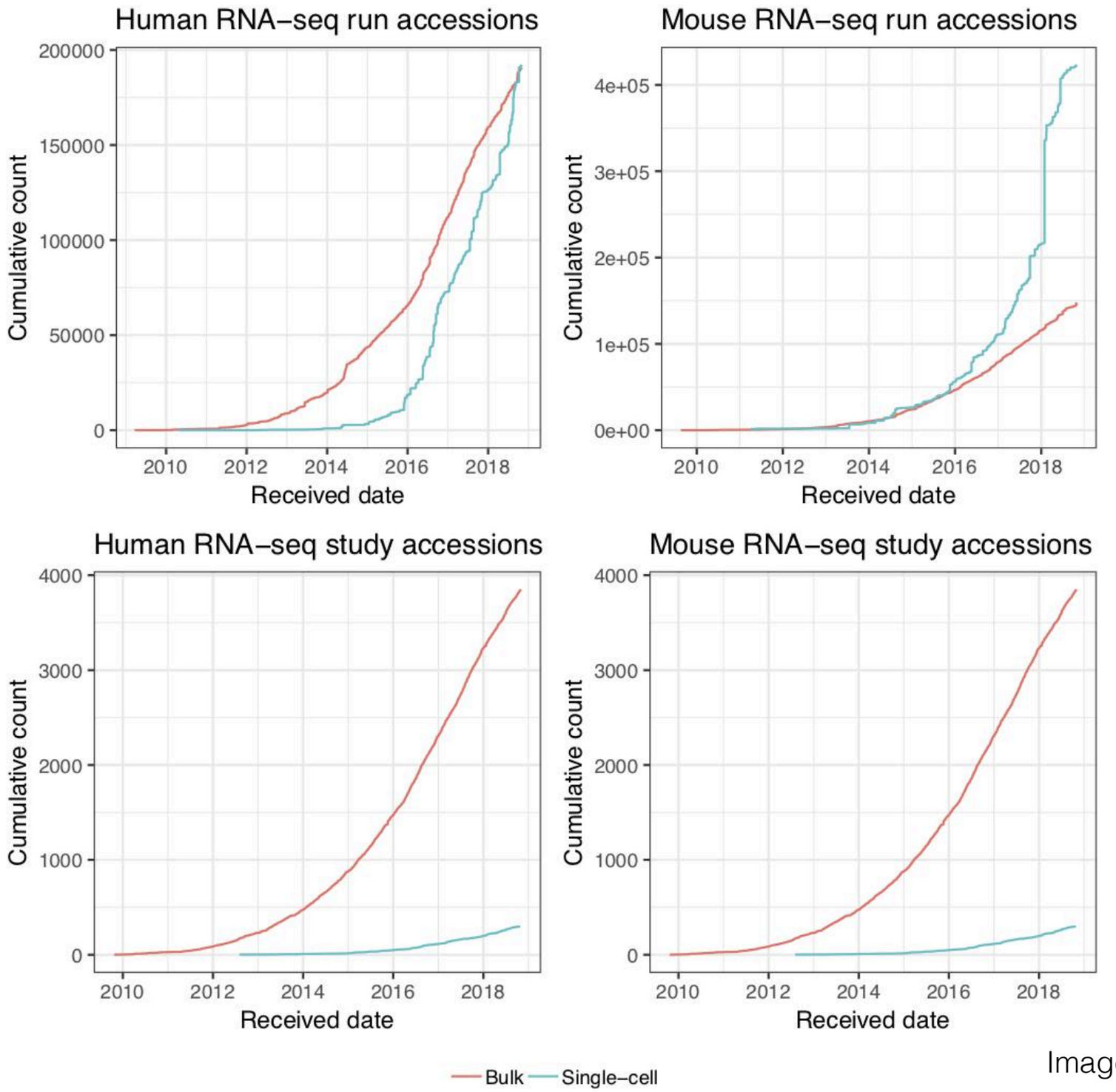
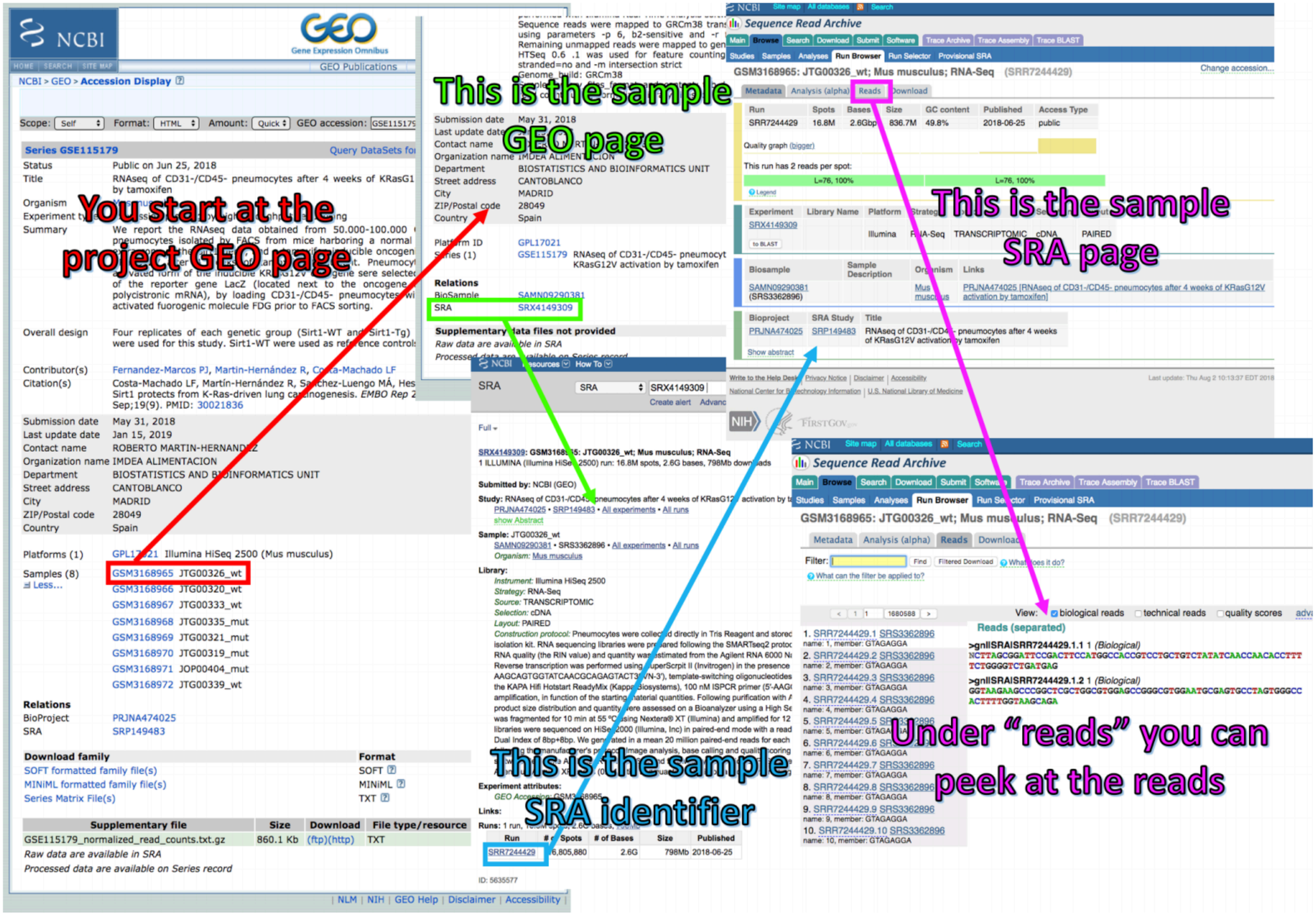


Image credit: Ben Langmead

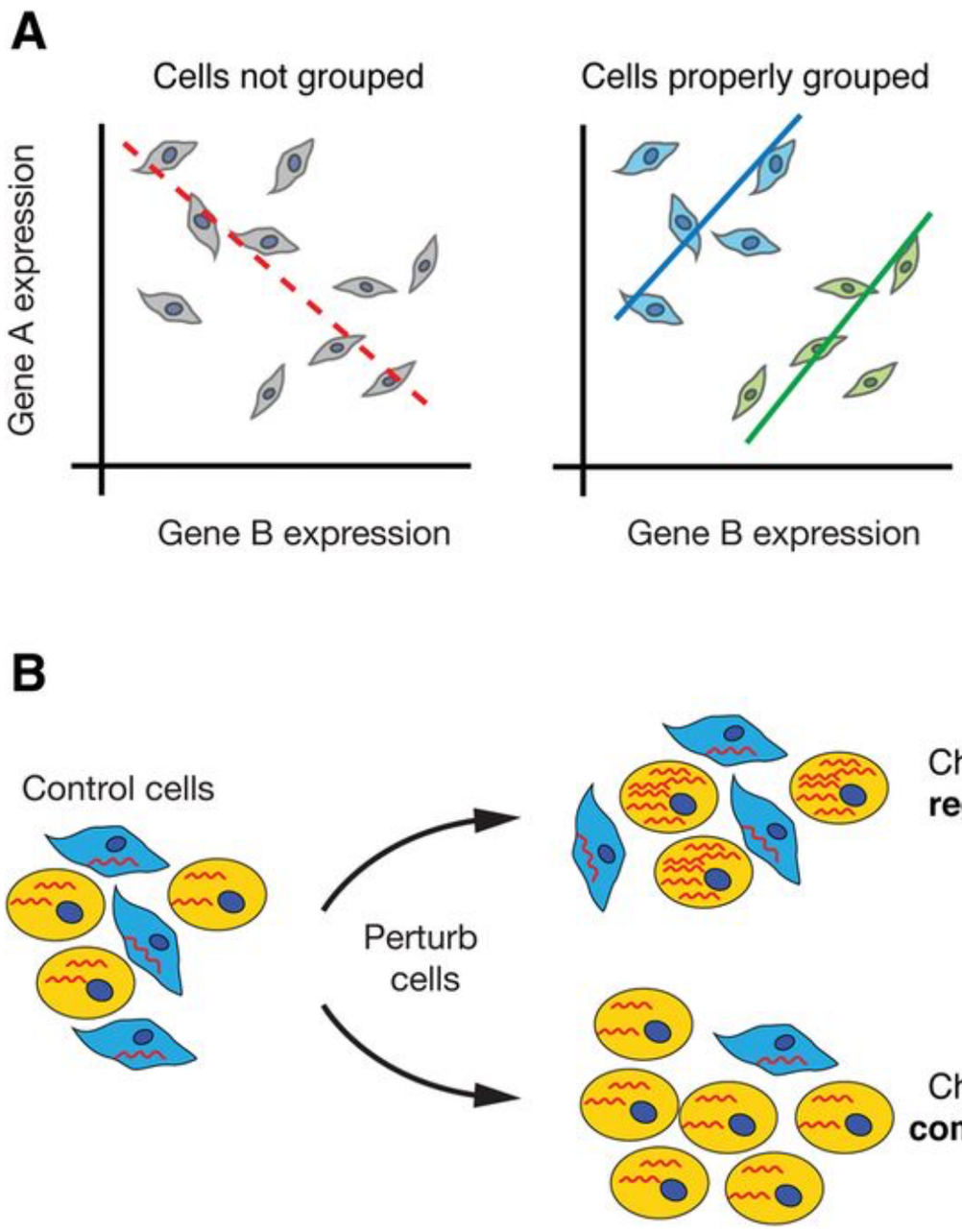
Navigating to the data



Components of an *analysis* (primary)?

- Barcode error correction (assignment of reads to cells)
- UMI collapsing (identification of reads with molecules)
- **cDNA alignment (association of reads to transcripts)**
- Gene counting (production of a cells x genes count matrix)

Components of an *analysis* (secondary)



cDNA alignment based analysis

read alignment information

read 1	GGGTTGTACCC
read 2	ATGTGATCC
read 3	CCGTTG
read 4	GAAAGGGTTG
read 5	CACAGGTGTGG

Alignment based analysis

read alignment information

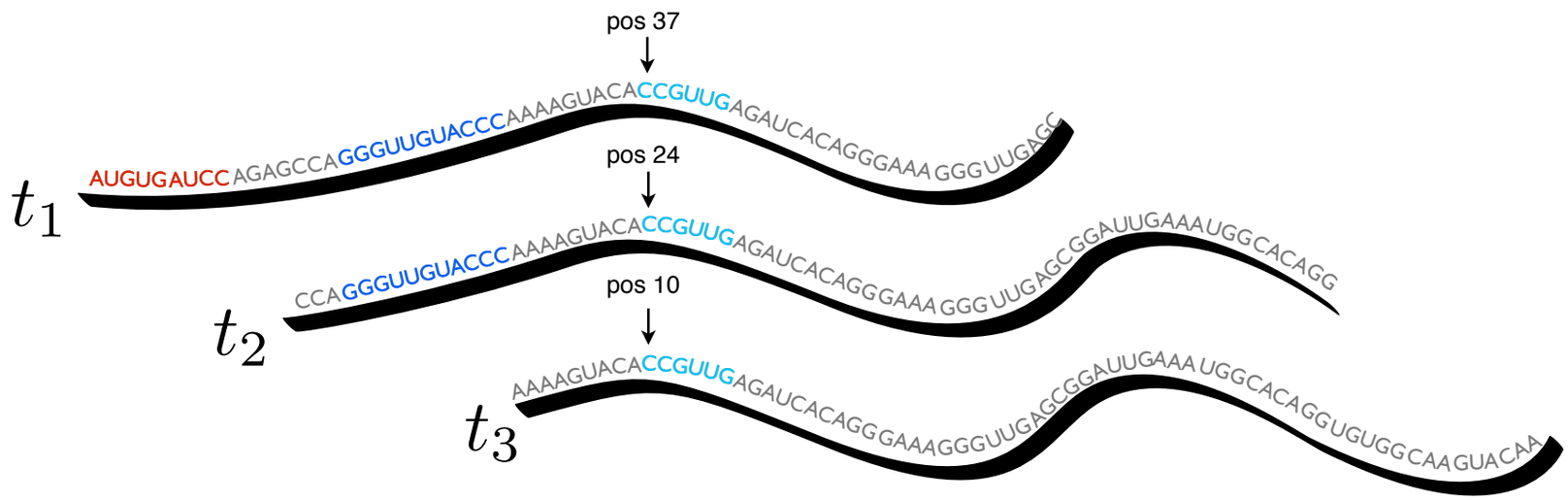
read 1	GGGTTGTACCC	t_1 @position 17, t_2 @position 4
read 2	ATGTGATCC	
read 3	CCGTTG	
read 4	GAAAGGGTTG	
read 5	CACAGGTGTGG	

Alignment based analysis

read alignment information

read 1	GGGTTGTACCC	t_1 @position 17, t_2 @position 4
read 2	ATGTGATCC	t_1 @position 1
read 3	CCGTTG	
read 4	GAAAGGGTTG	
read 5	CACAGGTGTGG	

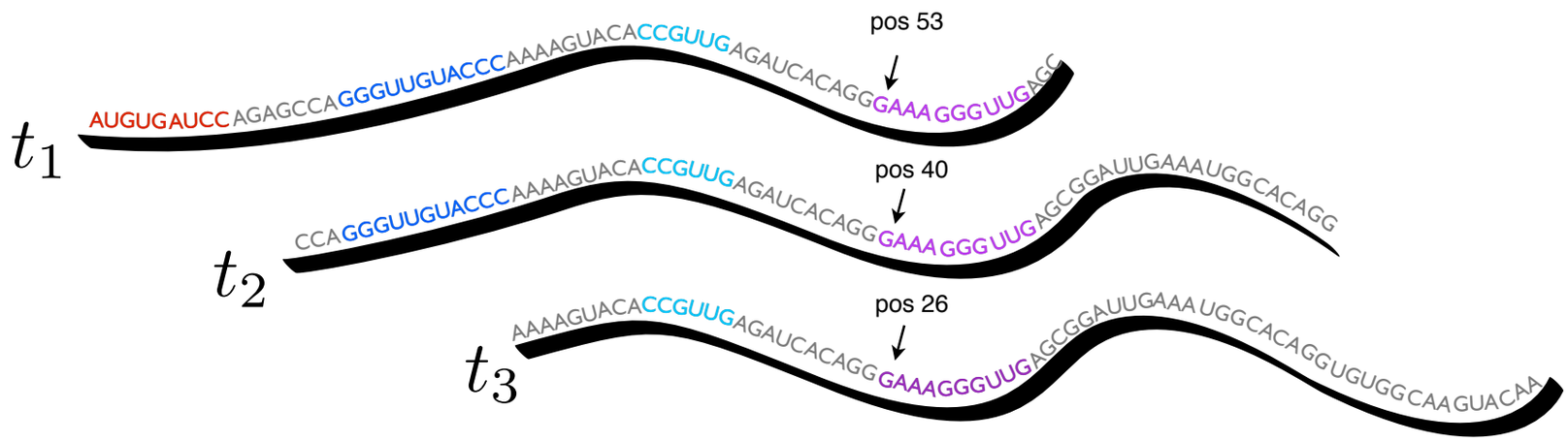
Alignment based analysis



read alignment information

read 1	GGGTTGTACCC	t_1 @position 17, t_2 @position 4
read 2	ATGTGATCC	t_1 @position 1
read 3	CCGTTG	t_1 @position 37, t_2 @position 24, t_3 @position 10
read 4	GAAAGGGTTG	
read 5	CACAGGTGTGG	

Alignment based analysis



read alignment information

read 1	GGGTTGTACCC	t_1 @position 17, t_2 @position 4
read 2	ATGTGATCC	t_1 @position 1
read 3	CCGTTG	t_1 @position 37, t_2 @position 24, t_3 @position 10
read 4	GAAAGGGTTG	t_1 @position 53, t_2 @position 40, t_3 @position 26
read 5	CACAGGTGTGG	

Alignment based analysis

read alignment information

read 1	GGGTTGTACCC	t_1 @position 17, t_2 @position 4
read 2	ATGTGATCC	t_1 @position 1
read 3	CCGTTG	t_1 @position 37, t_2 @position 24, t_3 @position 10
read 4	GAAAGGGTTG	t_1 @position 53, t_2 @position 40, t_3 @position 26
read 5	CACAGGTGTGG	t_3 @position 51

Alignment based analysis

read alignment information

read 1	GGGTTGTACCC	t_1 @position 17, t_2 @position 4
read 2	ATGTGATCC	t_1 @position 1
read 3	CCGTTG	t_1 @position 37, t_2 @position 24, t_3 @position 10
read 4	GAAAGGGTTG	t_1 @position 53, t_2 @position 40, t_3 @position 26
read 5	CACAGGTGTGG	t_3 @position 51

- Even "ultra-fast alignment" is still pretty slow
- Alignments contain information that we don't usually care about.

The kallisto mantra

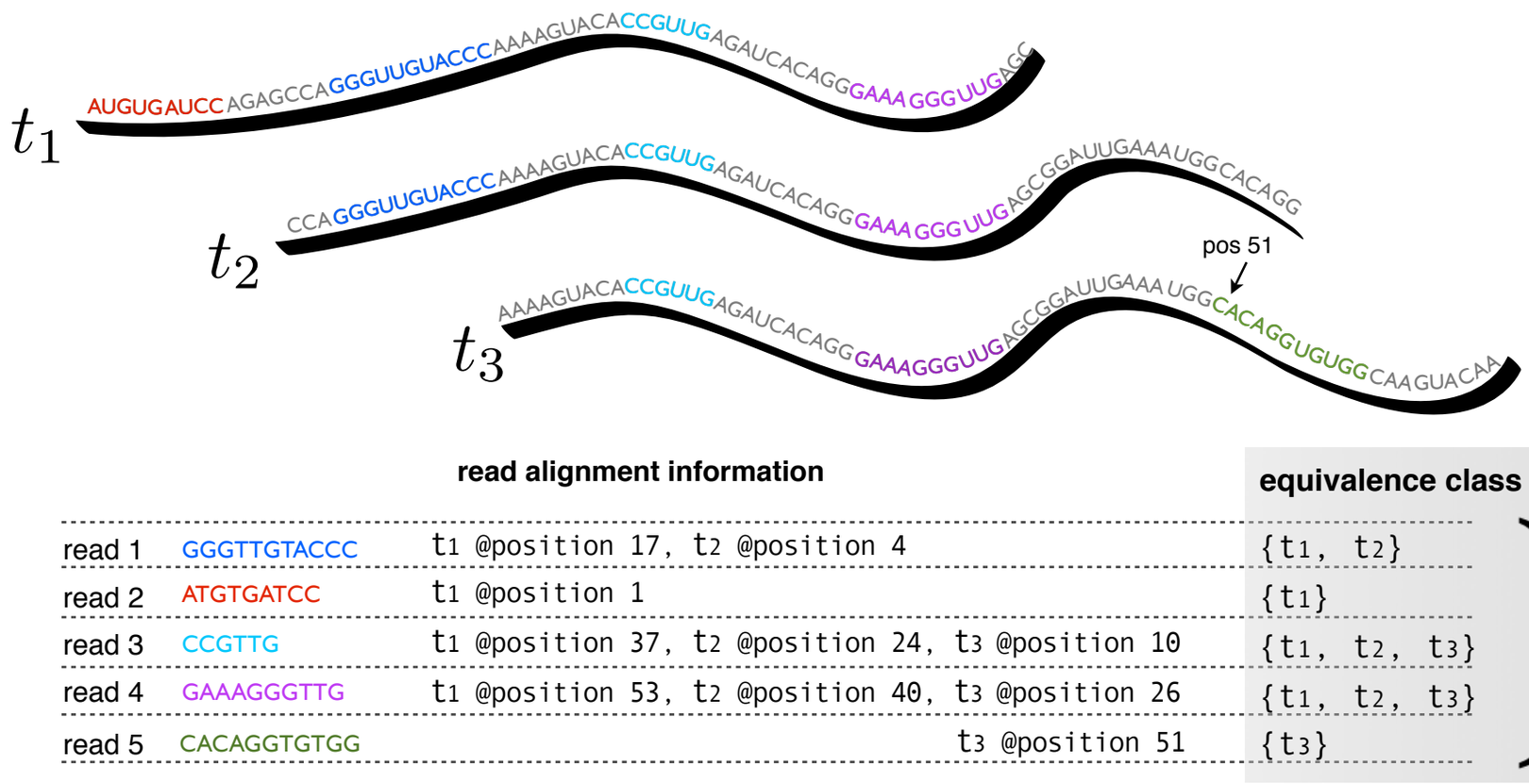
read alignment information

read 1	GGGTTGTACCC	t_1 @position 17, t_2 @position 4
read 2	ATGTGATCC	t_1 @position 1
read 3	CCGTTG	t_1 @position 37, t_2 @position 24, t_3 @position 10
read 4	GAAAGGGTTG	t_1 @position 53, t_2 @position 40, t_3 @position 26
read 5	CACAGGTGTGG	t_3 @position 51

Do as much as you can, with as little as you can.

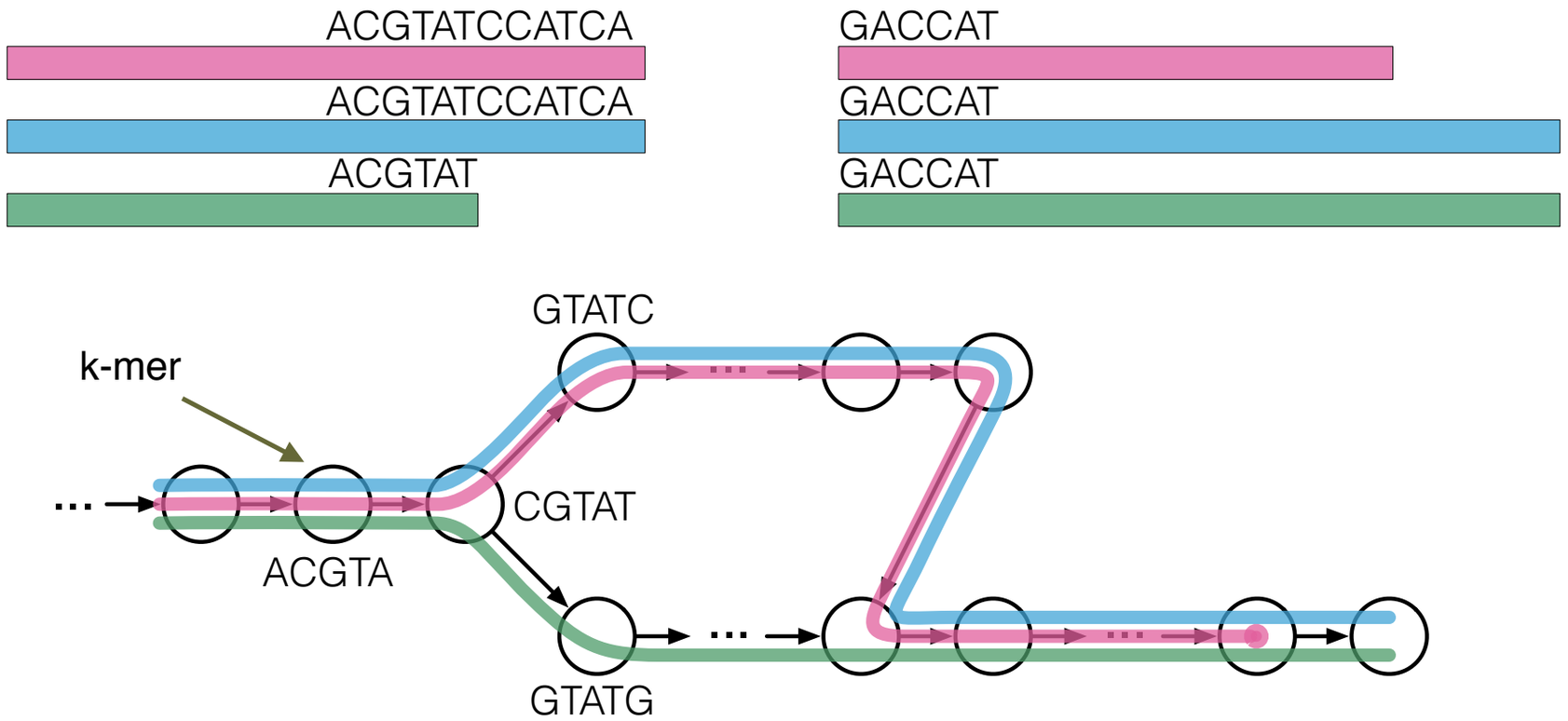
Bray et al. 2016

The kallisto mantra



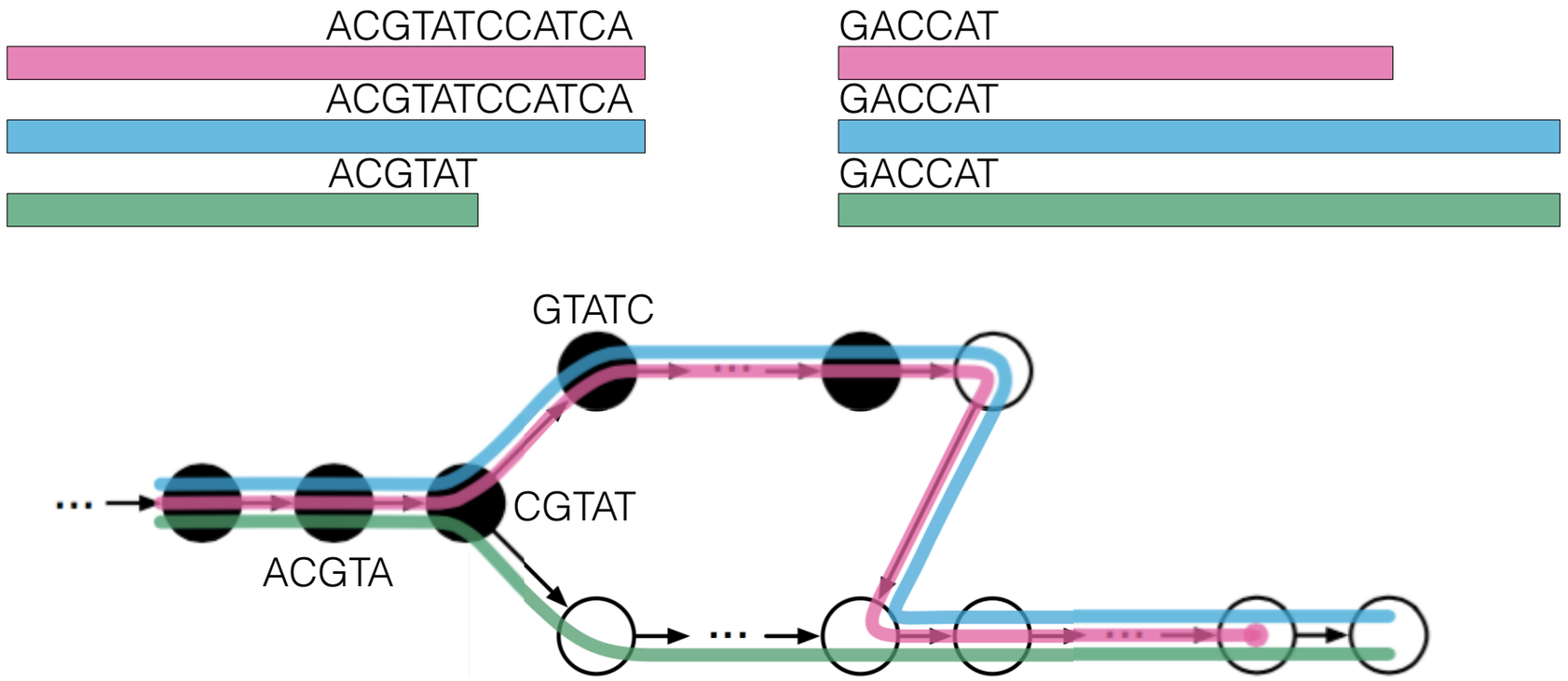
- for computing transcript abundances, the set of transcripts a read is compatible with tells you almost everything about it
- **idea**: let's compute that directly rather than a basepair-level alignment that has more information than we need

How kallisto computes pseudoaligments



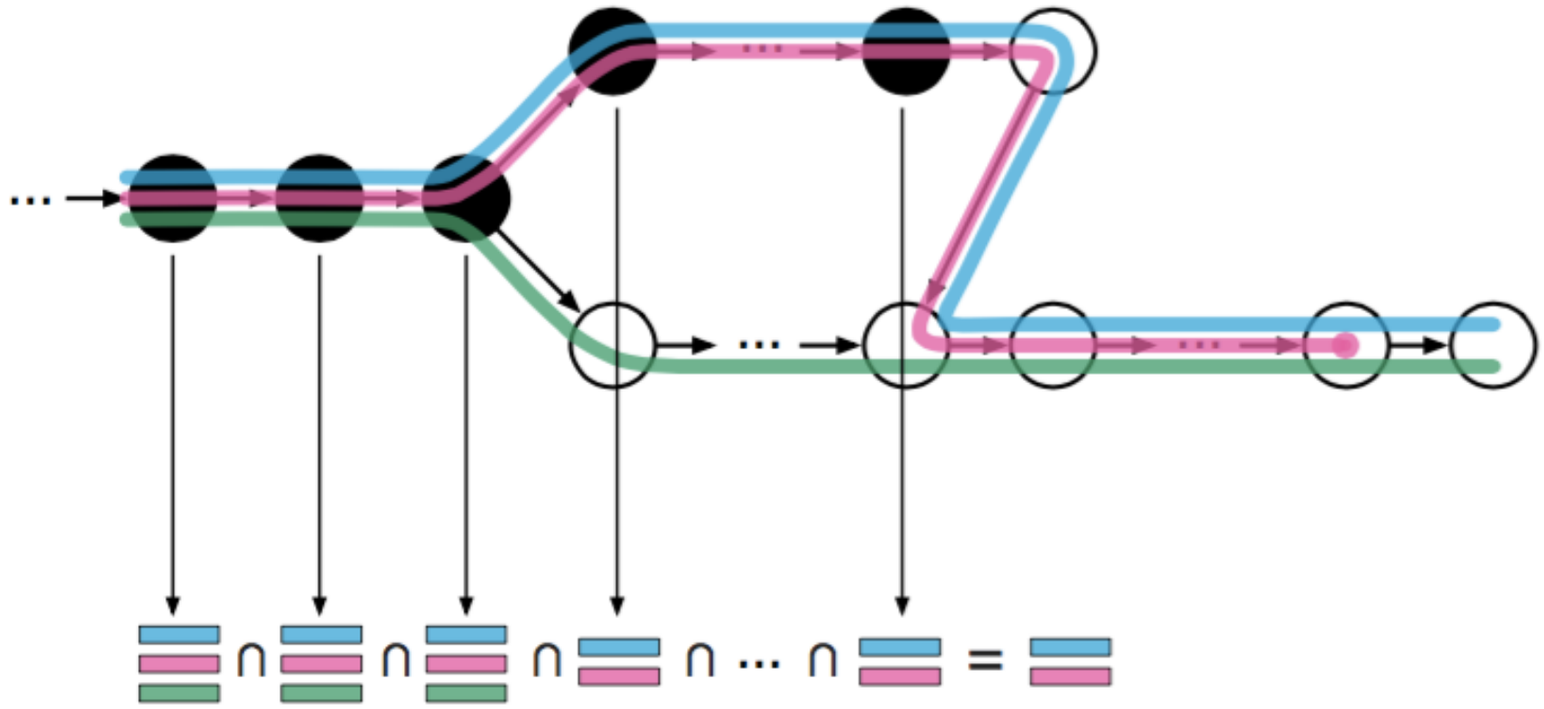
- Given our reference transcriptome, we first construct its *target de Bruijn Graph (T-DBG)*
- This encodes the transcript sequences but also provides information about how they overlap with each other
- Only has to be done *once* per transcriptome (and is fast)

How kallisto computes pseudoaligments



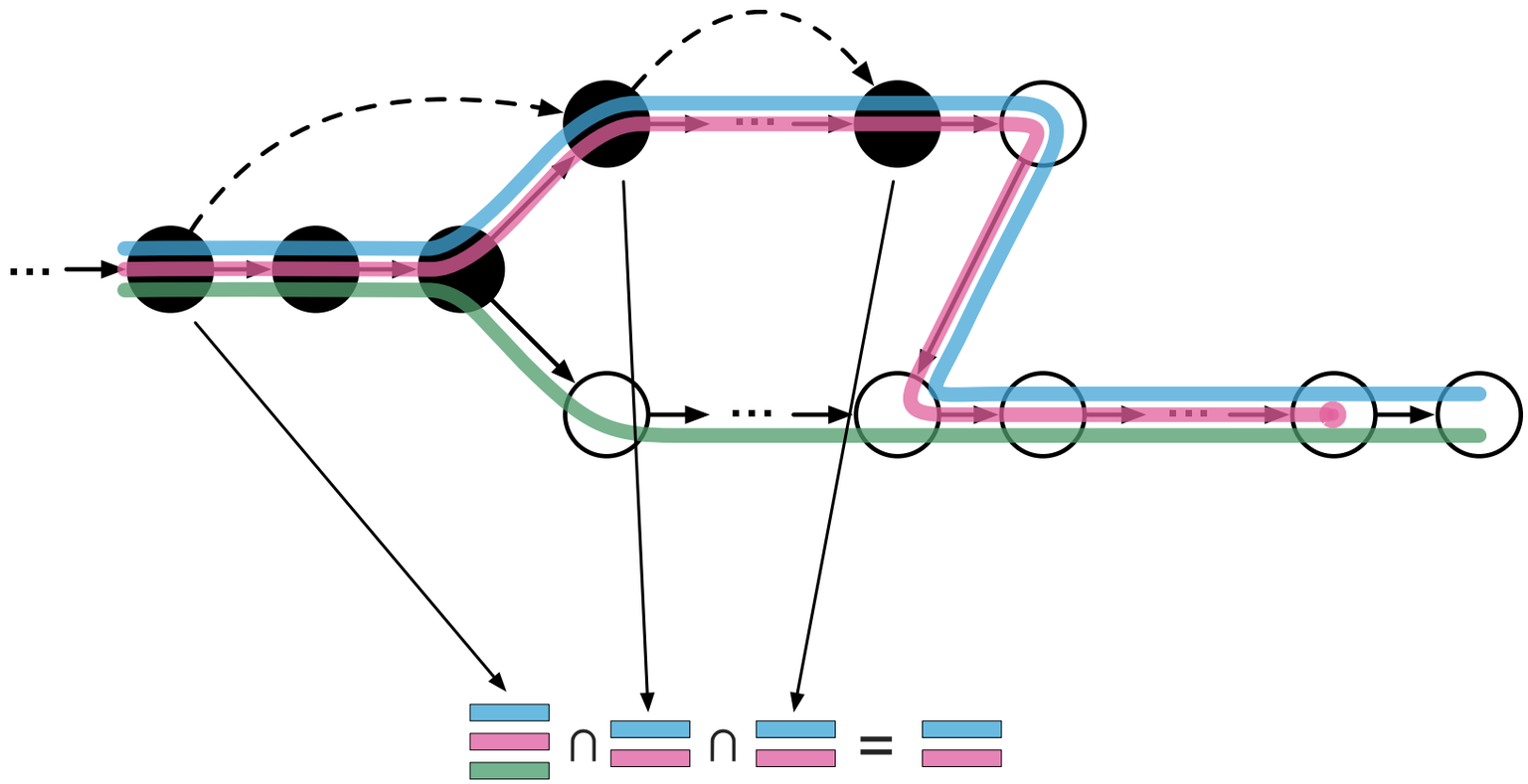
- Given a read, finding its constitutive *k*-mers in the T-DBG gives you information about where the read could have come from
- This can be done *very* fast
- **But individual *k*-mers might be more ambiguous than the read as a whole**

How kallisto computes pseudoaligments



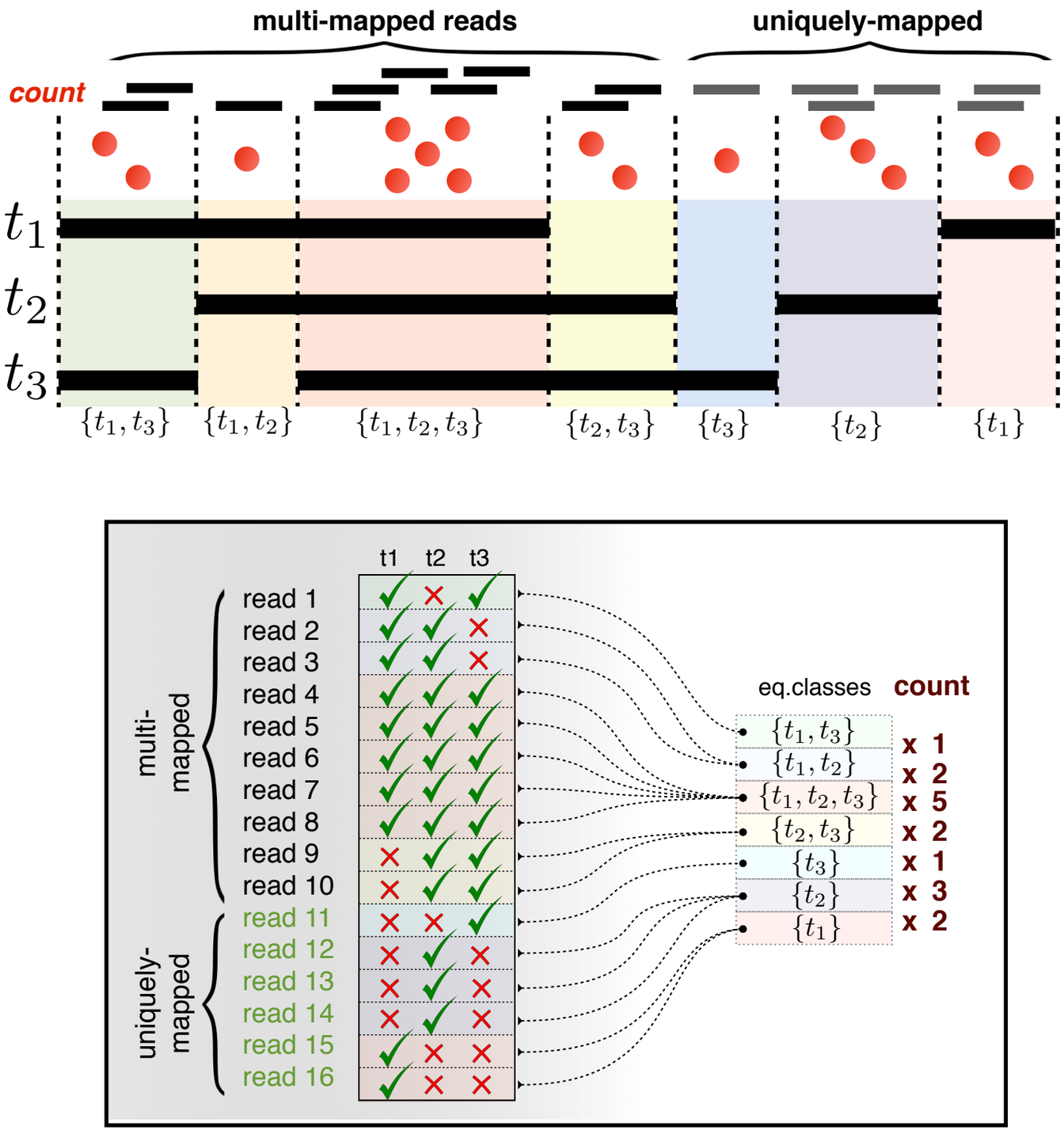
- Combining information across the k-mers can recover lost information
- For each k-mer we have the set of transcripts it could have come from. Intersecting them gives the set of transcripts that *all* k-mers could have come from
- It's possible for their combination to have information equivalent to the entire read, even if no single k-mer does by itself

How kallisto computes pseudoaligments

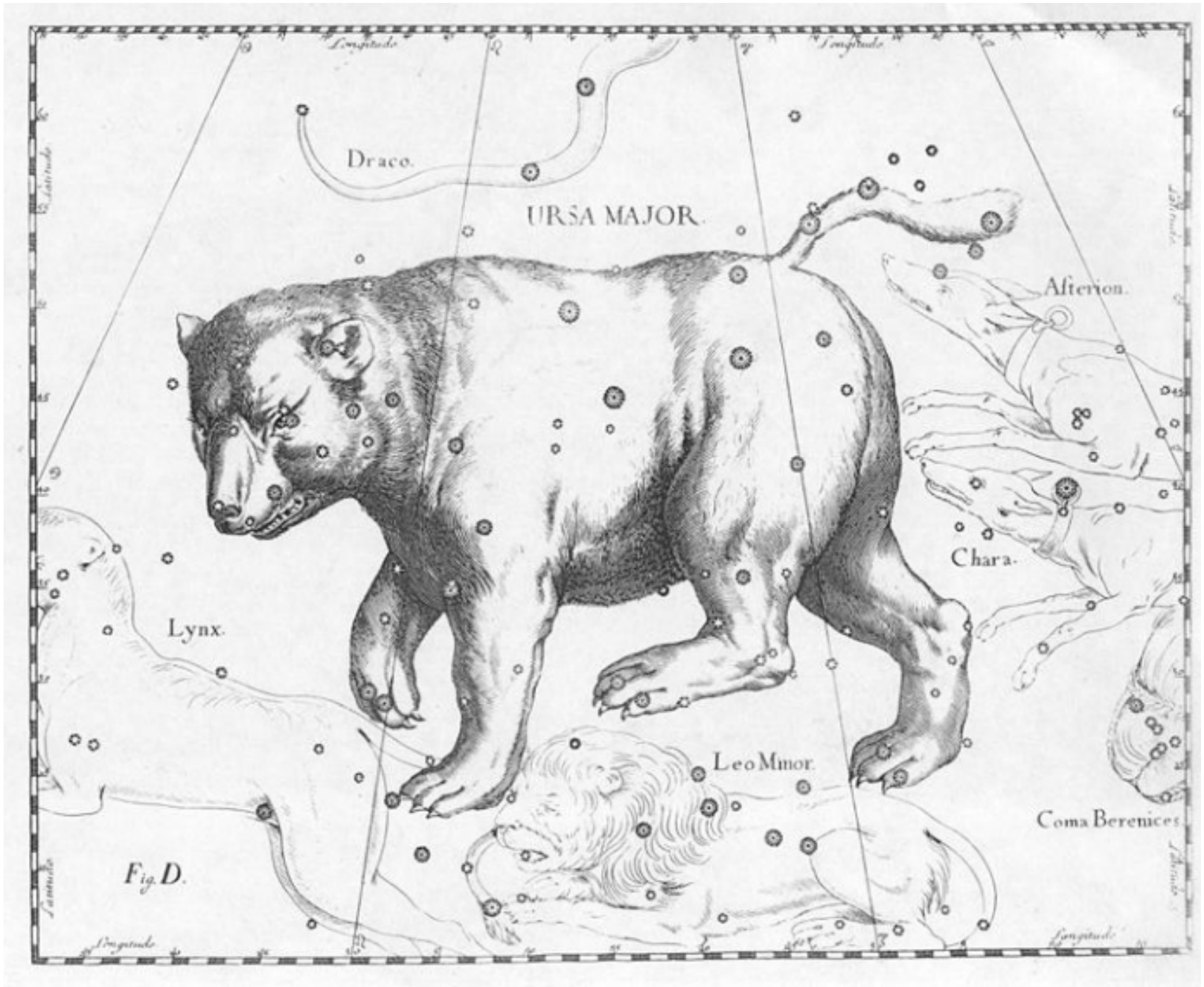


- Knowing the T-DBG, we can predict ahead of time which k-mers will be potentially interesting
- By only processing those k-mers, kallisto runs ~8 times faster

Transcript compatibility counts

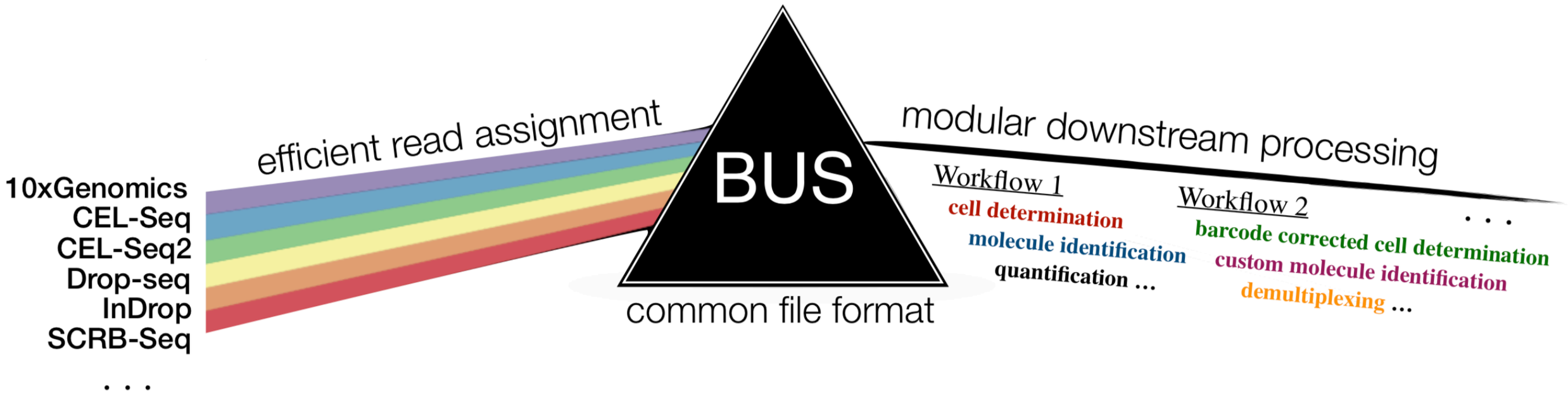


pachterlab.github.io/kallisto/



A new format

- A new format which decouples technology dependencies from algorithm choices.



- We call this format **B**arcode, **U**MI, **S**et (BUS) format.

Background to this workshop

- **Vasilis Ntranos** created a python kallisto based workflow for scRNA-seq
- **Páll Melsted** implemented a number of new features in kallisto to optimize the software for processing of sc-RNA-seq
- **Tina Wang** worked with **Lynn Yi** to implement this workflow in C++.
- During the course of development we (the **Pachter group**) revisited the current approach to building scRNA-seq workflows.
- **Valentine Svensson**, developer of umis (kallisto based software for processing different technologies) asked to what extent a naïve workflow would be sufficient for most purposes.
- **Vasilis Ntranos**, **Páll Melsted** and **Lior Pachter** identified BUS as a suitable checkpoint for modularizing workflows.
- **Páll Melsted** implemented the kallisto bus command, wrote bustools and created a naïve processing notebook.
- **Fan Gao**, **Eduardo Beltrame**, **Lambda** and **Jase Gehring** started developing BUS notebooks.
- **Jase Gehring** helped to greatly simplify the workflow.