Single-cell sequencing

Jie He

2020-11-10

The Core of Biology Is All About One Cell

Forward Approaching

The Nobel Prize in Physiology or Medicine 2004

Dr. Richard Axel

Dr. Linda Buck

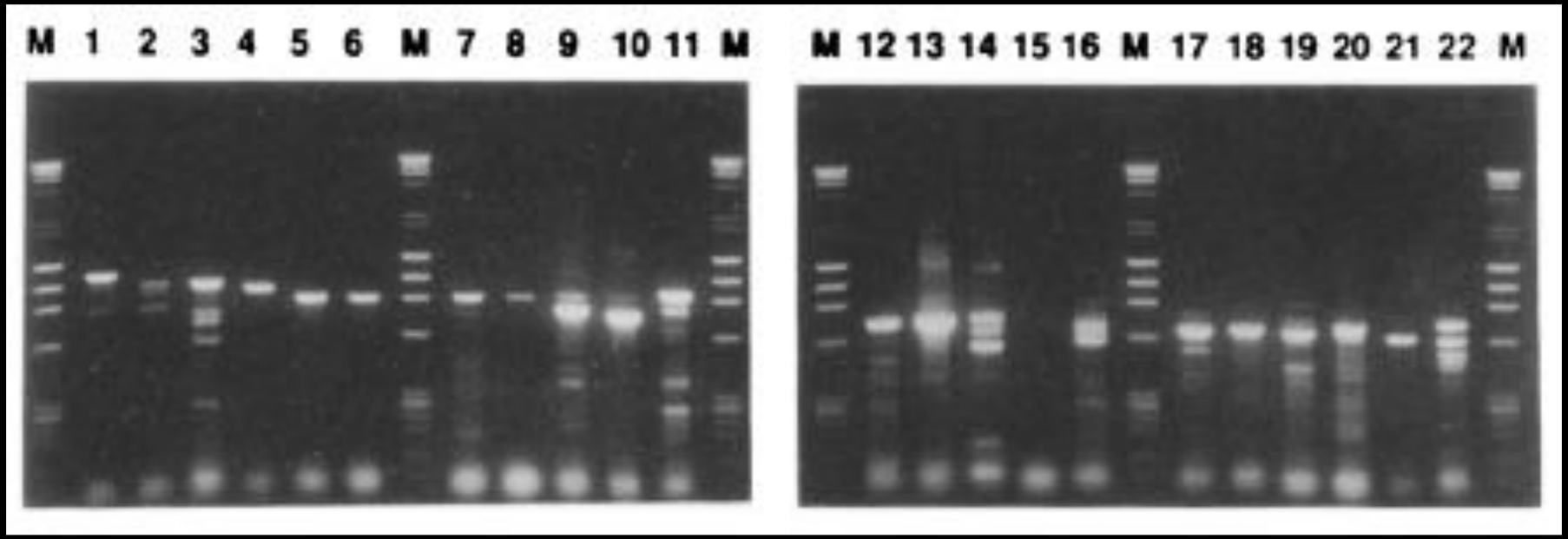
Hypothesis

1. The odorant receptors are likely to belong to the superfamily of receptor proteins that transduce intracellular signals by coupling to GTP-binding proteins.
2. Odorant receptors themselves should exhibit significant diversity
3. The expression of the odorant receptors should be restricted to the olfactory epithelium.

Buck et al. 1991

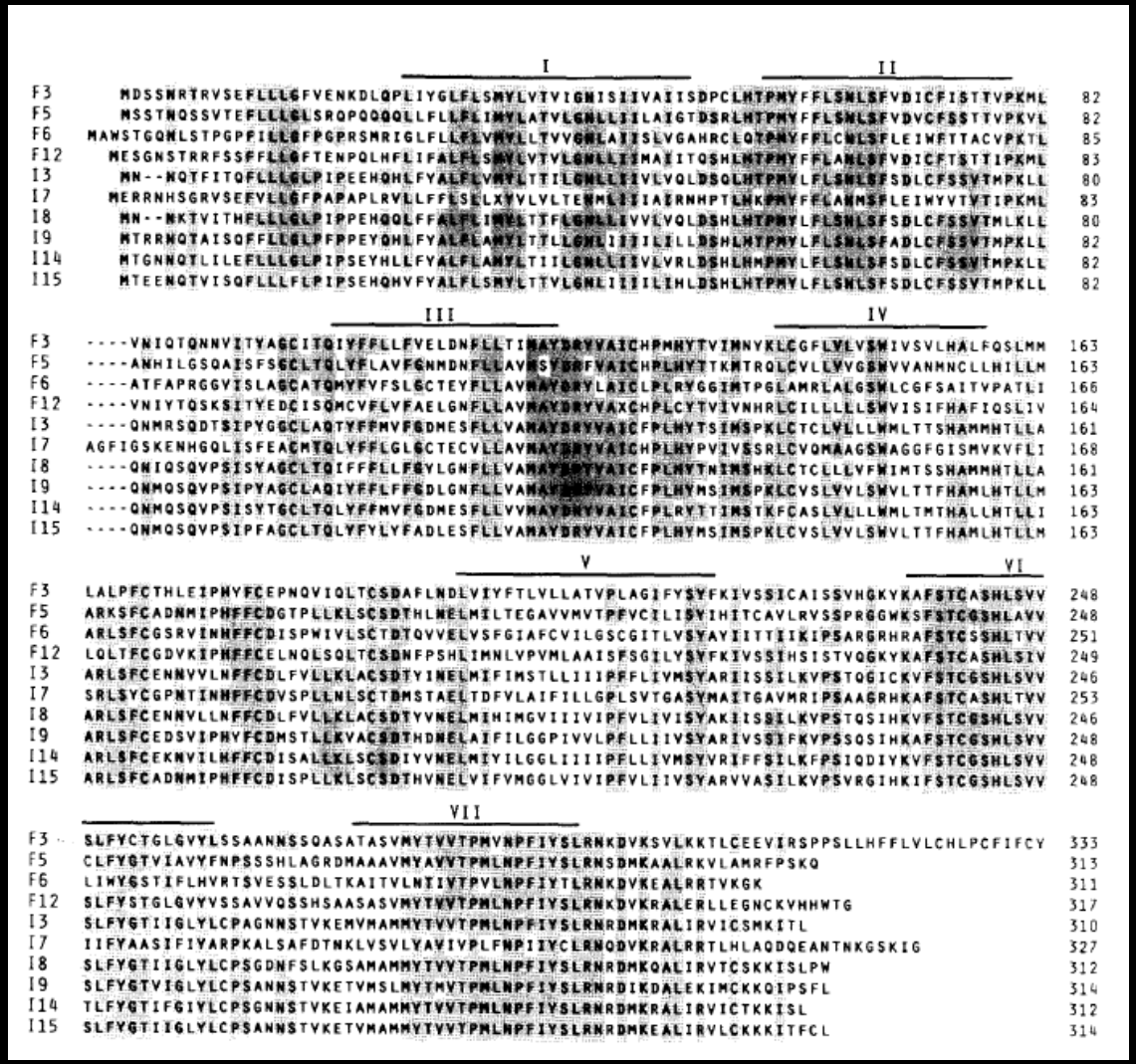
Strategy

Degenerative primers that could anneal to conserved regions of G protein-coupled seven transmembrane domain receptor genes



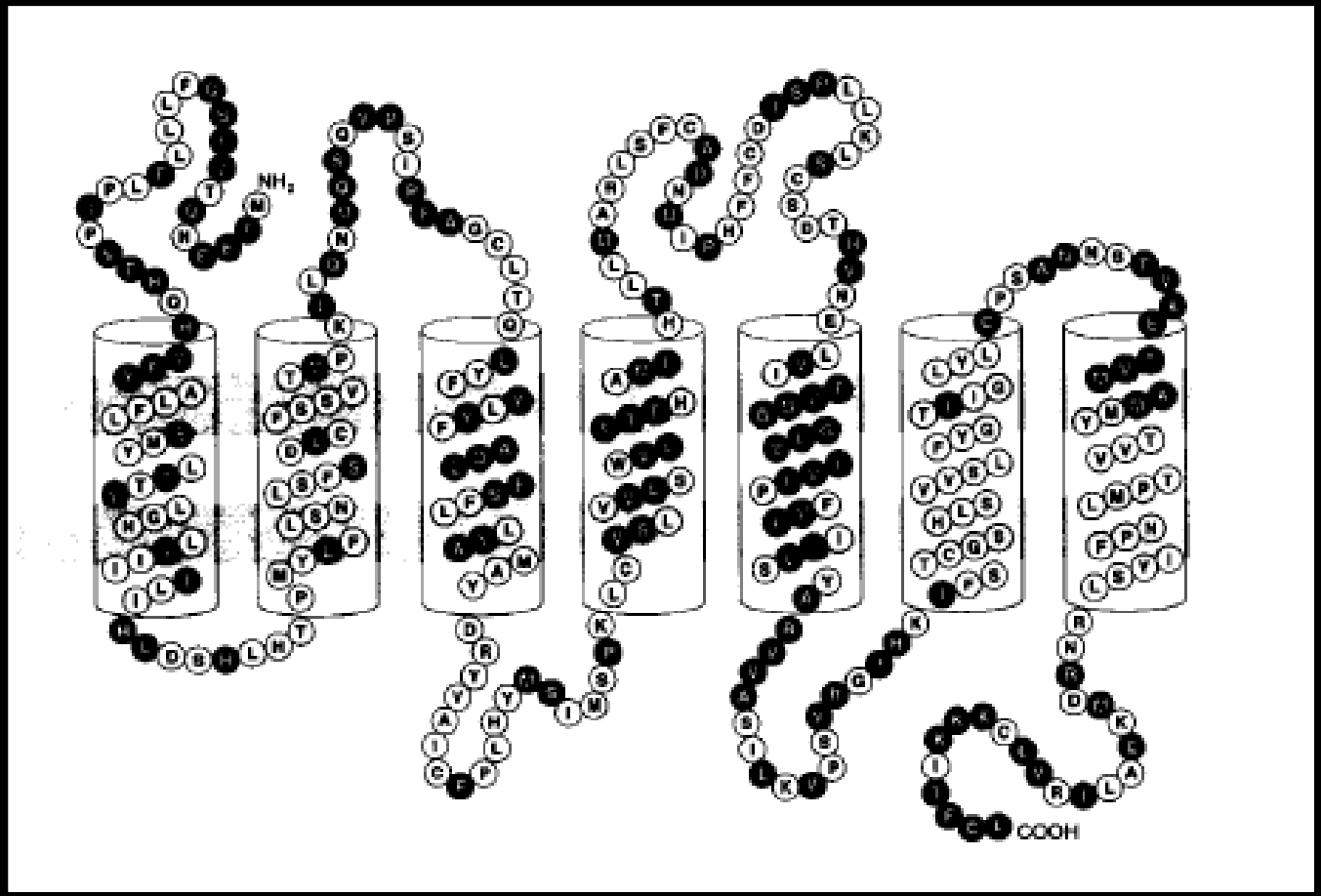
Buck et al. 1991

Identification of a new family of GPCR



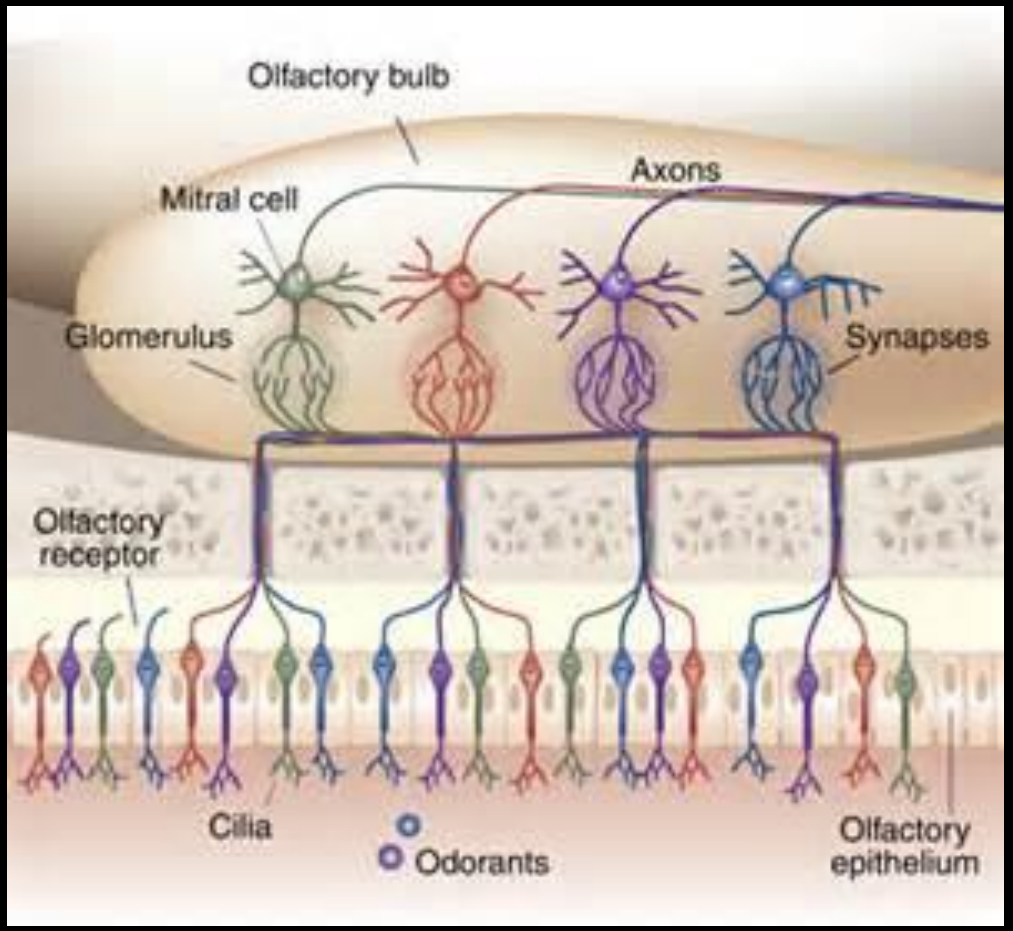
Buck et al. 1991

Olfactory sensory receptors



Buck et al. 1991

Olfactory circuits

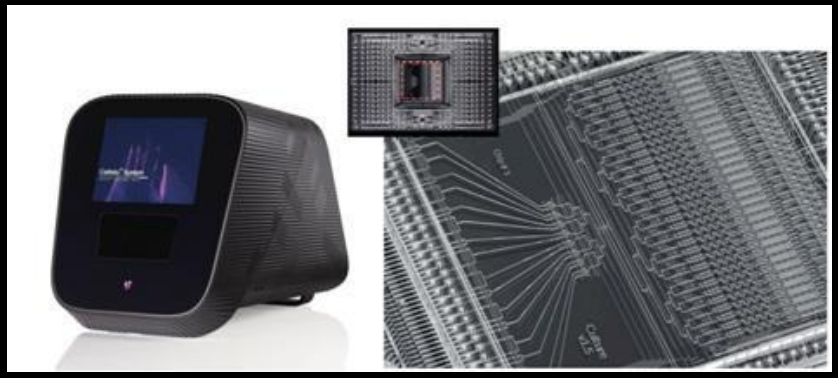


Isolating Single Cells for Sequencing

FACS:
Morphology;
Chemical indicator for certain cellular property;
Genetic labelling;
Antibodies against cell surface markers;

Micromanupulation:

Microfluidic devices: (Fluidigm C1)



Sequencing the genome of individual cells

The reveals of somatic mutations and allows the investigation of clonal dynamics

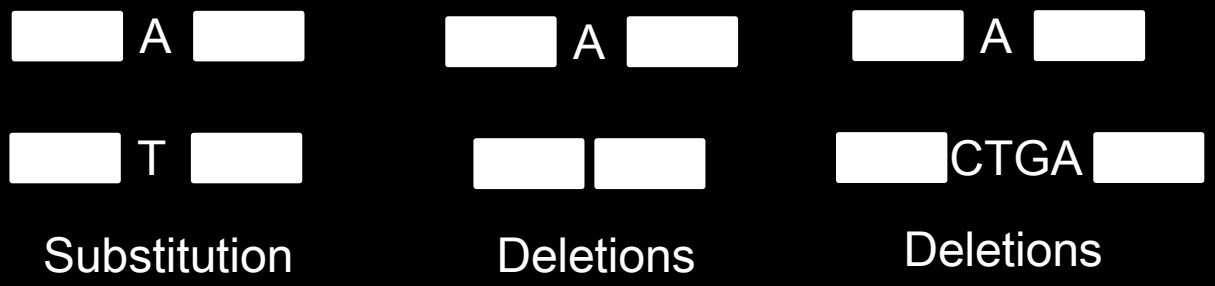
Tumor evolution inferred by single-cell sequencing

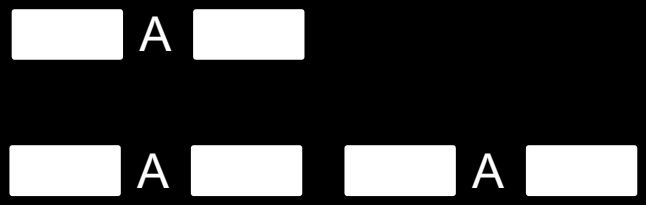
Navin et al 2011

Whole-Genome Amplification (WGA) Techniques

Germline and somatic genome mutations:

substitutions, insertions and deletions, copy number variations





Whole-Genome Amplification (WGA) Techniques:

Polymerase chain reaction (PCR)

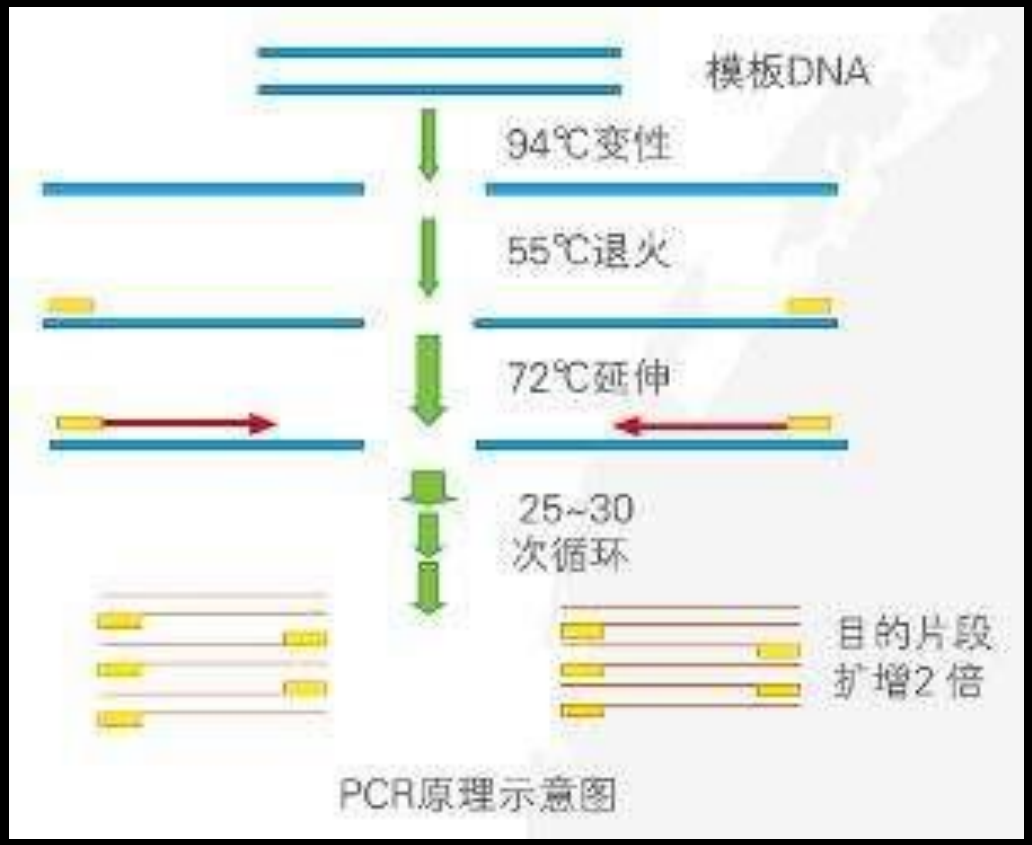
Multiple displacement amplification (MDA)

PCR&MDA (MALBAC)

Polymerase chain reaction (PCR)

Random or nonrandom primers

DNA polymerase



Technical artifacts (PCR)

1. Biased amplification of sequence rich in cytosine and guanosine (GC-bias)
2. Preferential allelic amplification
3. Chimeric DNA molecules

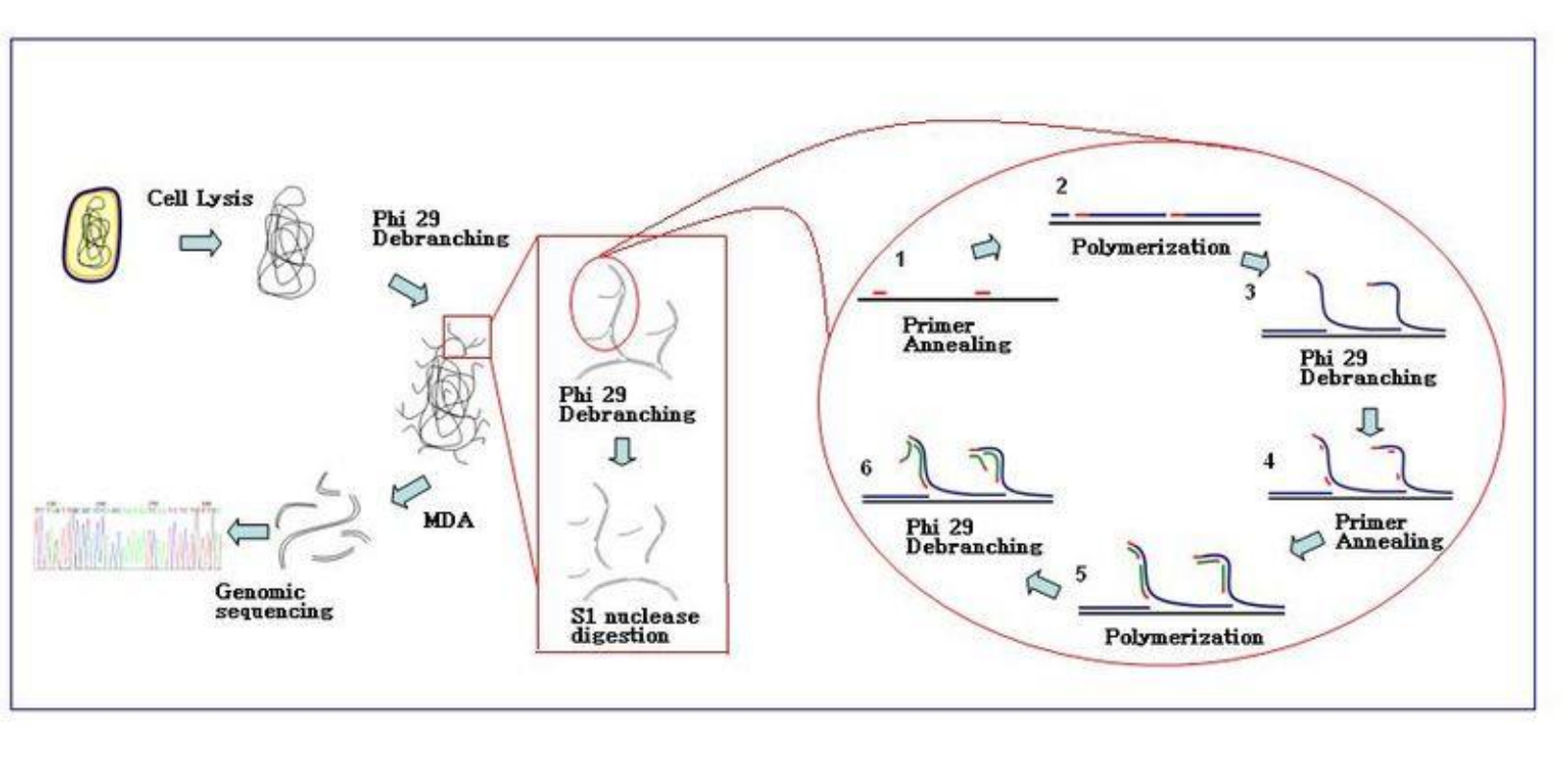
In general, random primed PCR-based methods achieve a highly uniform amplification but yield only sparse coverage of the genome

Multiple displacement amplification (MDA)

A non-PCR based DNA amplification

Random hexamer primers

A high fidelity enzyme: Φ29 DNA polymerase



General procedures for MDA

Sample preparation: Samples are collected and diluted in the appropriate reaction buffer (Ca2+ and Mg2+ free). Cell are lysed with alkaline buffer.

Condition: The MDA reaction with $\Phi 29$ polymerase is carried out at 30 C, which takes 2.5-3 hours

End of reaction: Inactivate enzymes at 65 C before collection of the amplified DNA products

DNA products can be purified with commercial purification kit.

Advantages of MDA

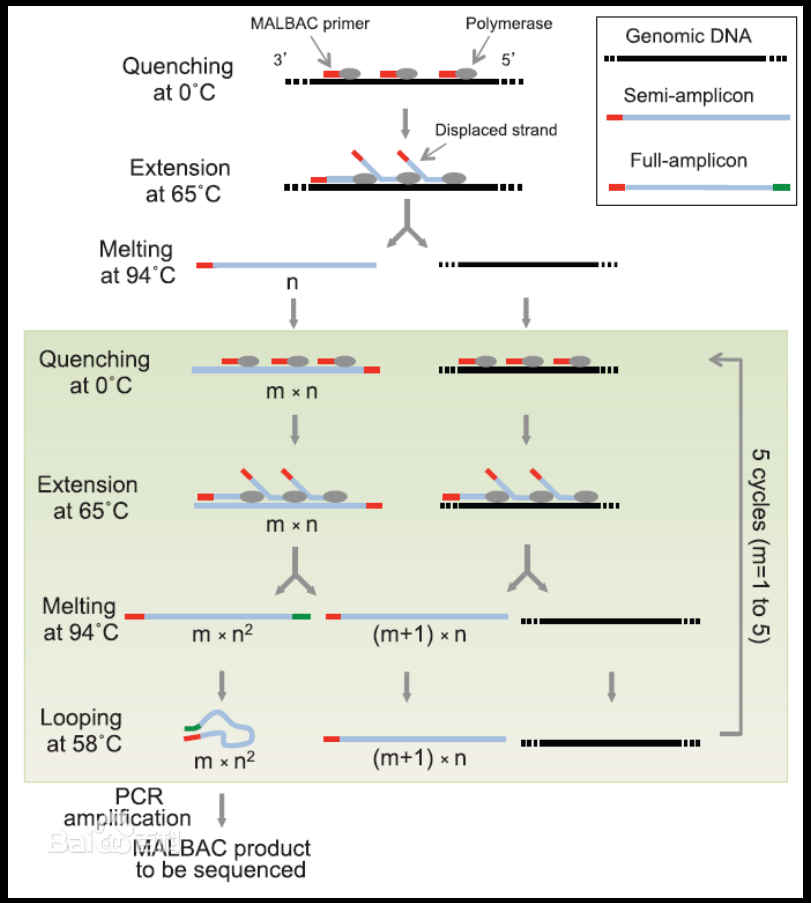
1. Better genome coverage
2. Larger size products (70kb) with a lower error frequency
3. If >70 kb, use Bst DNA polymerase

Technical artifacts (MDA)

1. Allelic dropout
2. Preferential amplification
3. Primer-primer interactions

Multiple annealing and looping-based amplification cycles (MALBAC)

1. Pre-amplies DNA using MDA and generates amplicons with complementary ends
2. This complementary induces loop formation and prevents the amplicon from being used as a template during subsequent cycles to attain close-to-linear amplification
3. After five cycles of pre-amplification, the material is amplified exponentially by PCR



1. Yielding 93 % genome coverage
2. Showing higher detection efficiency for SNPs and CNVs

Reduction of the reaction volume (nano-liter reaction wells)

Micro-well displacement amplifcation system (MIDAS)

Yielding an extremely low error rate (4 x 10-9)

Analysis of Single-Cell Genome Sequencing Data

1. Inspect the read quality and trim low-quality bases and remaining adaptor sequences at the end of the reads

2. If the remaining read is too short, reads should be discarded in order to avoid erroneous mapping.

3. Removal of PCR duplicates

4. Mapping (Obtaining a file with sequencing reads is mapping to a reference genome USSC genome browser and Ensembl)

5. Reads that map to more than a single locus should be discarded or counted with reduced uniform weight for each locus

6. To determine genomic mutants

Analysis of Single-Cell Genome Sequencing Data

1. GC bias

2. Preferential allelic amplification

3. Random sequencing errors represent another source of uncertainty for SNP detection.

4. Cell-cycle phase

Single-Cell RNA Sequencing (scRNAseq)

Analysis of Single-Cell RNA Sequencing (scRNA-Seq)

The main problem with any of these methods is the presence of amplification bias, which can distort the relative abundances of mRNAs from different genes.

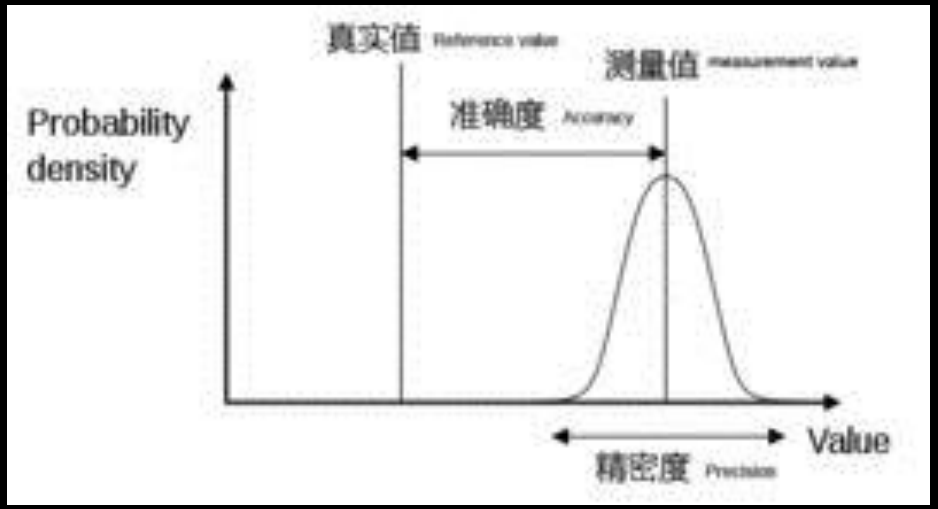
Key Variables

Sensitivity

Accuracy

Precision

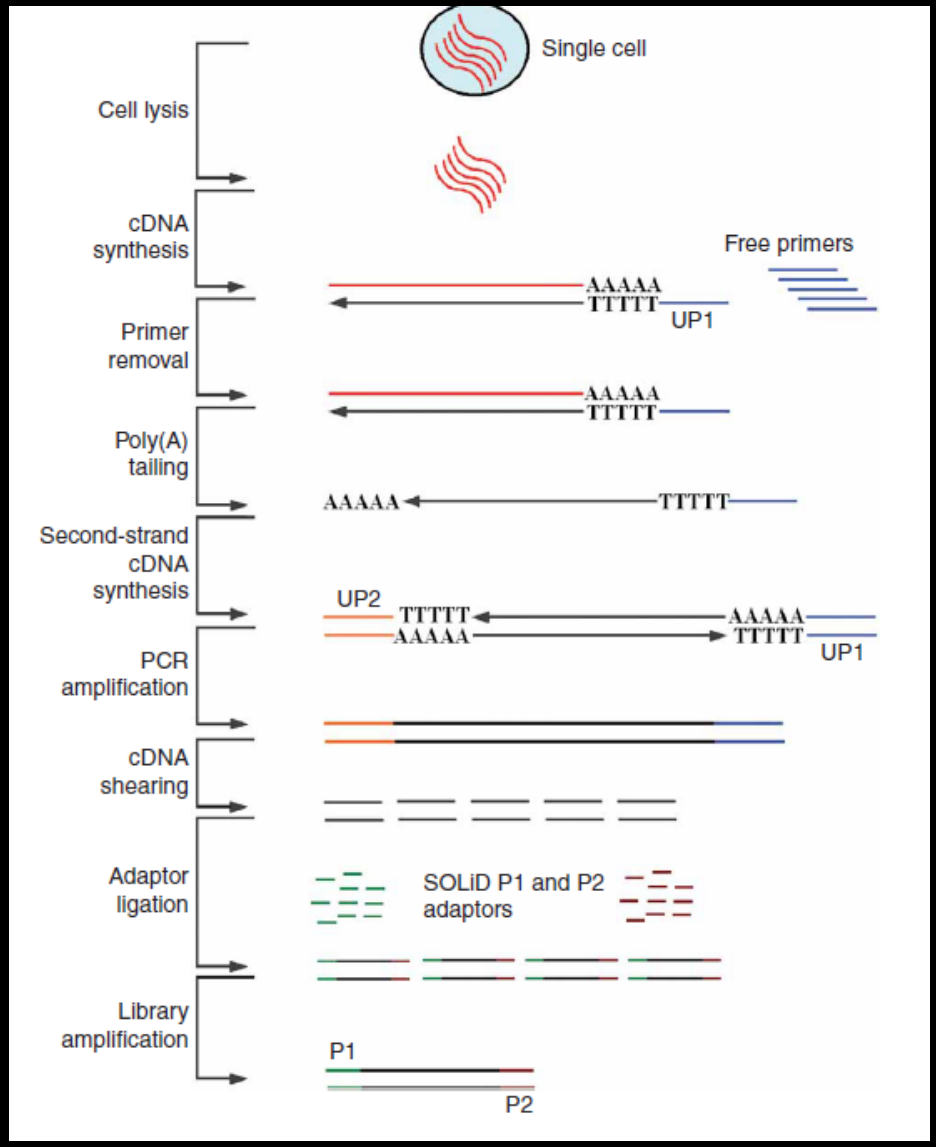
Cost



Main Challenge

Sensitivity & Amplification Bias

The first protocol for single-cell sequencing



TANG et al. 2009
TANG et al. 2010

CEL-Seq 1/2 (Cell expression by linear amplification and sequencing)

MARS-seq (Massively parallel RNA single-cell sequencing)

SCRB-seq (single-cell RNA barcoding and sequencing)

Smart-seq 1/2 (Switching mechanism at 5’ end of the RNA transcript)

Drop-seq/In-Drop

Design I: UMI (Unique molecular identifiers)

4 to 10 random nucleotides to serve as a random barcode for each mRNA molecules

Allow for the distinction between original molecules and amplification duplicates that derive from the cDNA or library amplification

It has been shown that counting UMIs instead of reads lead to a 2-fold reduction of technical noise.

It is important to consider UMI if gene expression variability is the goal

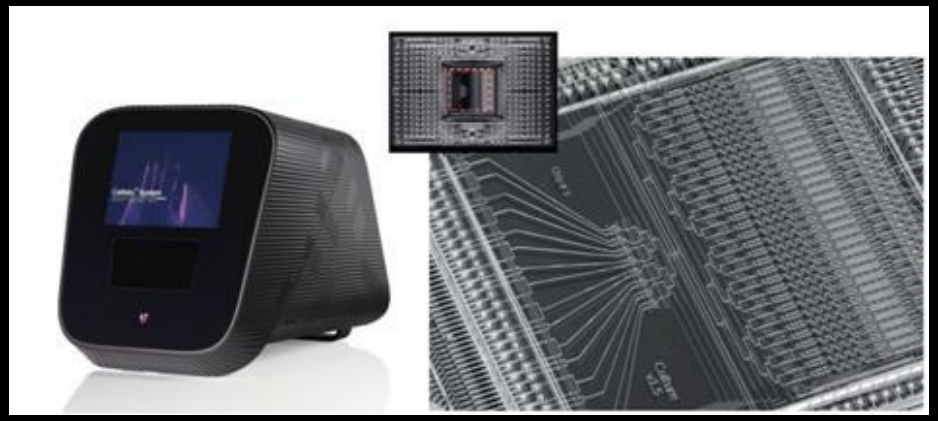
Design III: BC (Barcodes for Cells)

Batch processing

Design III: Nano-liters vs. micro-liters

scRNA-seq in the small volume , such as Fluidigm C1 outperforms the ones in microliter volumes.

Hashimashony et al., 2016;
Wu et al., 2014

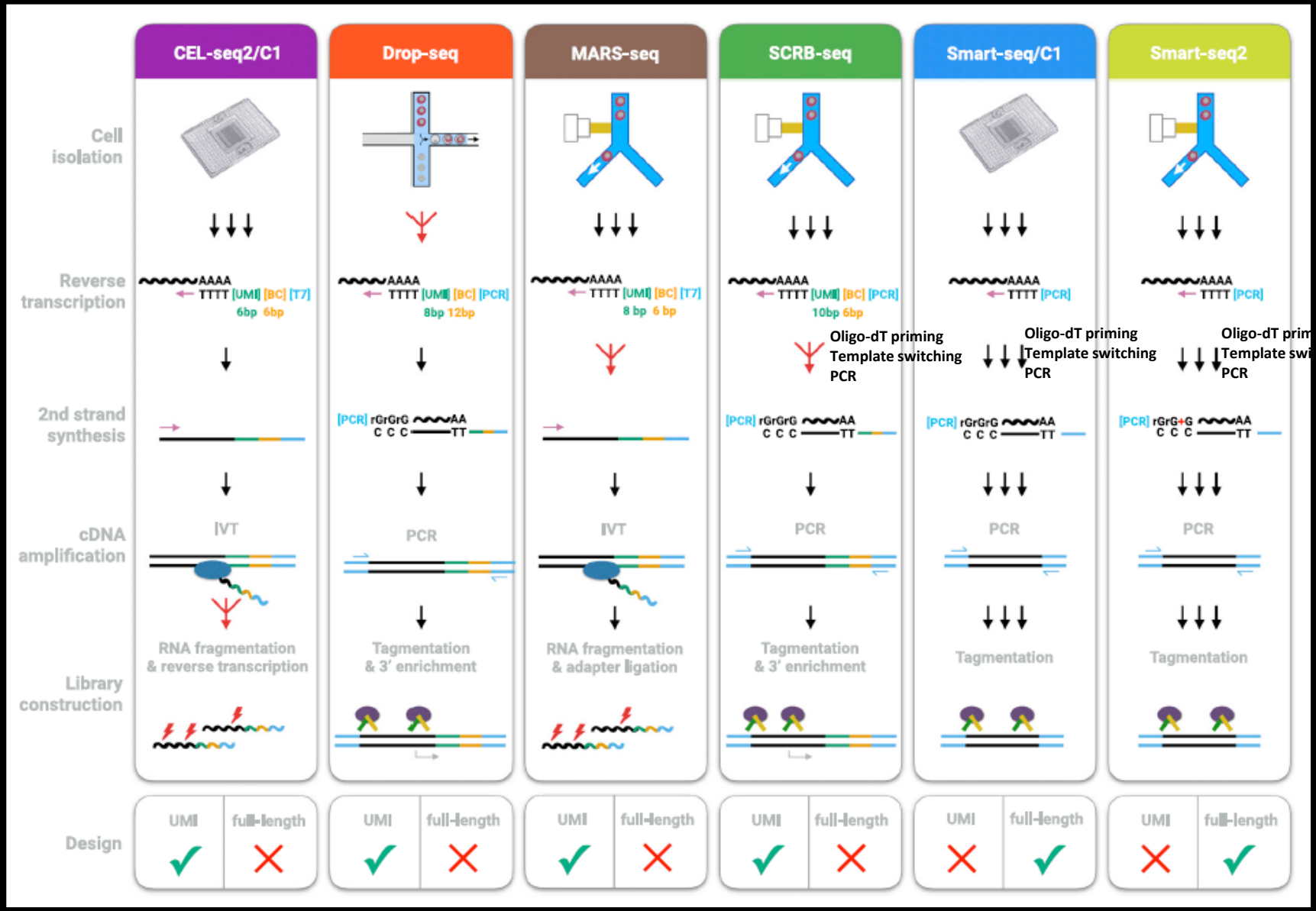


Fluidigm's C1

Design IV: Quantifying sensitivity

The use of external spike-in RNA of known concentration

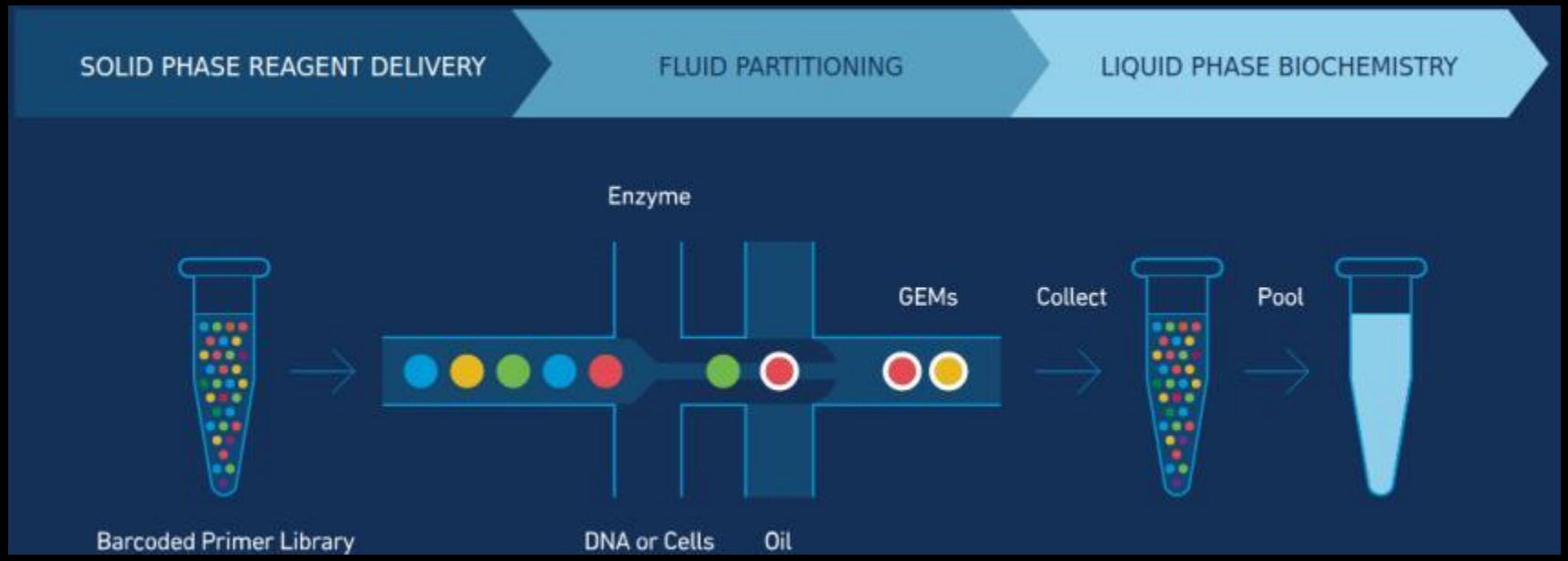
The spike-in concentration should be chosen such that spike-in RNA contributes 1-5 % of the number of mRNA molecules



~100 ~1-10k ~100-1000 ~100-1000 ~100 ~100 (Cells)

Christoph et al., 2017

Droplet-based microfluidic methods


SOLID PHASE REAGENT DELIVERY
FLUID PARTITIONING
LIQUID PHASE BIOCHEMISTRY
Enzyme
GEMs
Collect
Pool
Barcoded Primer Library
DNA or Cells
Oil

Processing of scRNA-Seq Data

The number of cells

(Describing profiling the cell composition of a sample with high sensitivity)

Gene numbers per cells

(Describing gene abundance per cell)

The sequencing complexity

(Describing sequencing each single cell with sufficient sequencing depth)

Processing of scRNA-Seq Data

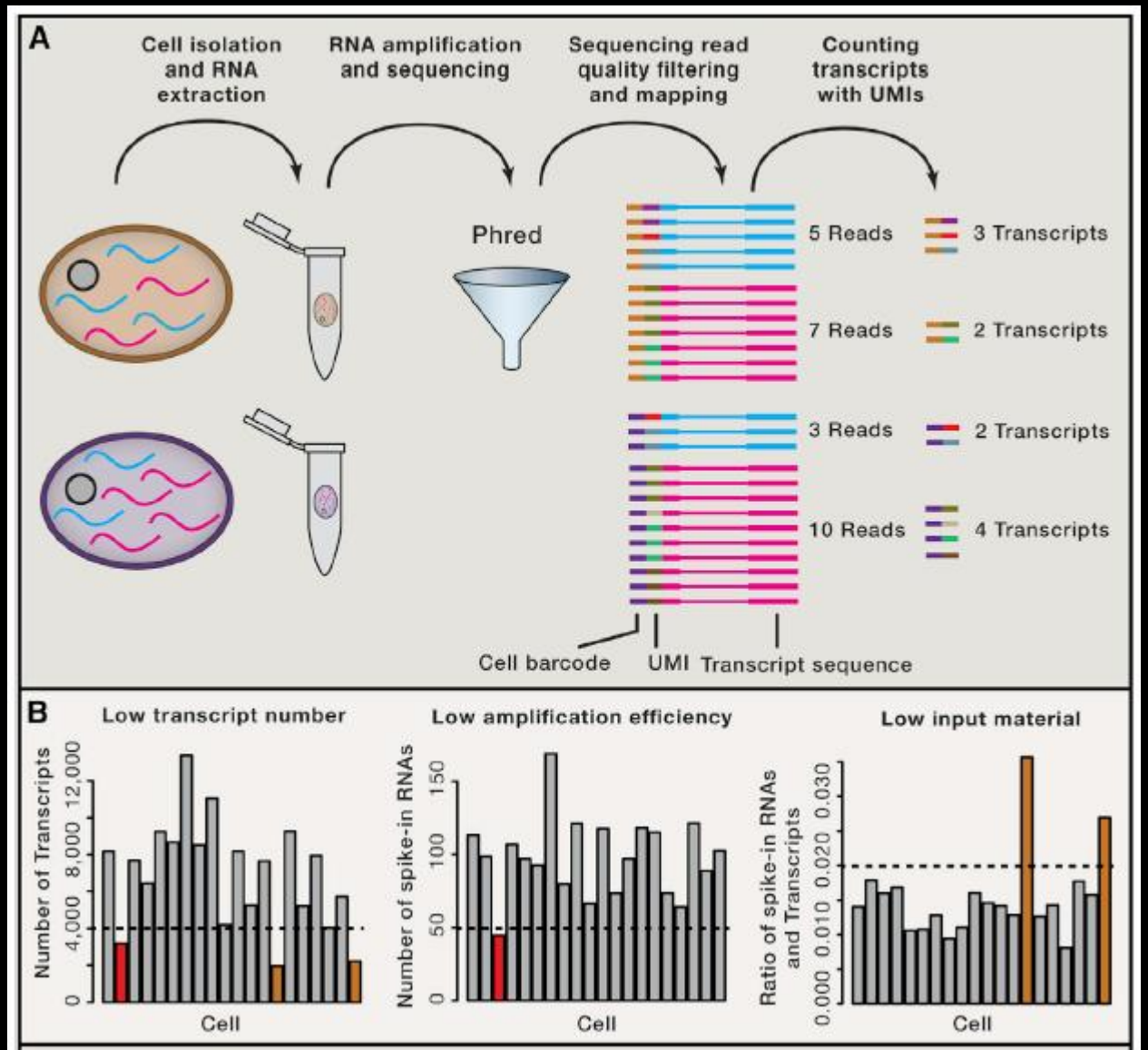
Preprocessing and read mapping

Fastqc: permit a quality analysis of the sequenced library
Bwa: trimming of low-quality bases from the end of the reads

For the mapping, available tools for bulk RNA-seq can be used
Merge all isoforms of a given gene into a so-called gene locus and quantify the expression of these gene loci

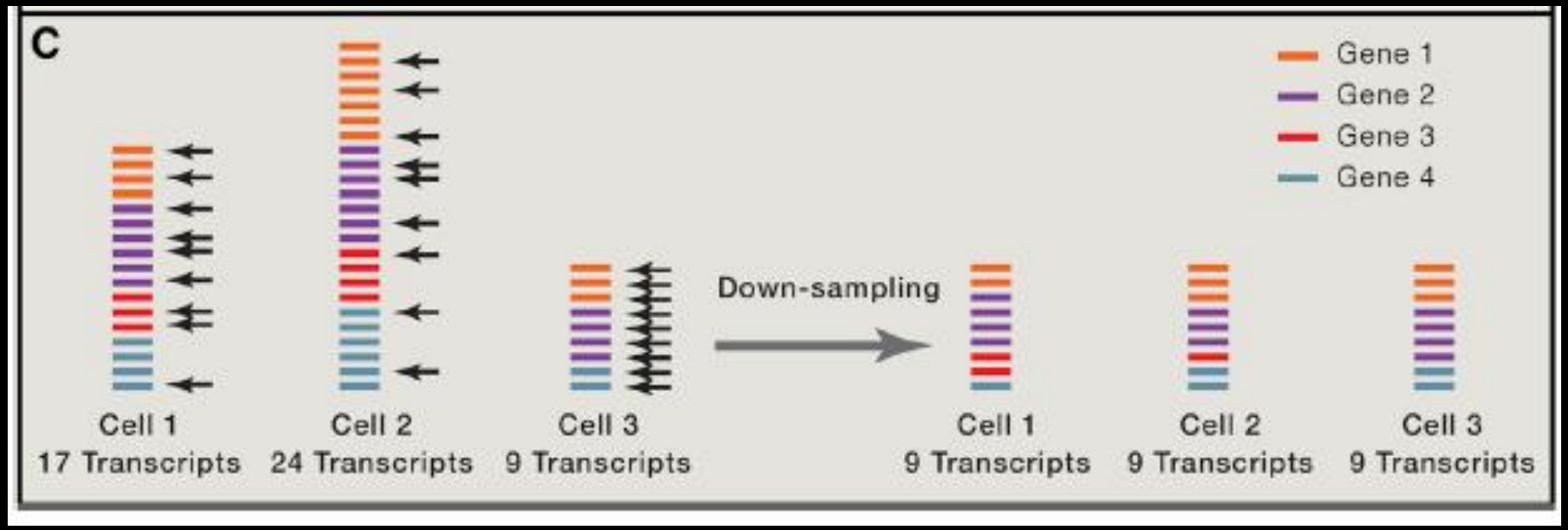
Processing of scRNA-Seq Data

Expression Quantification and Filtering



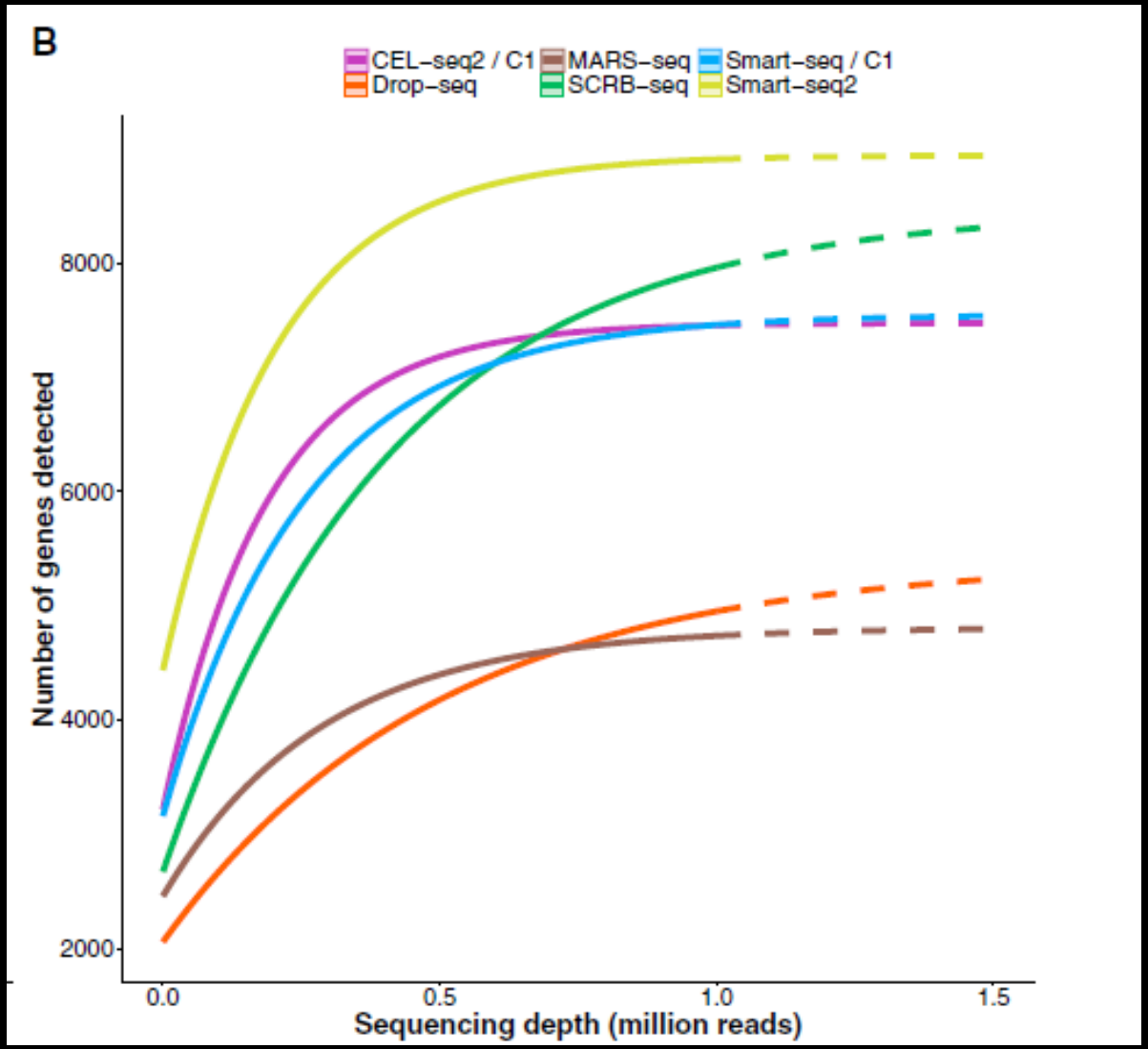
Processing of scRNA-Seq Data

Data Normalization

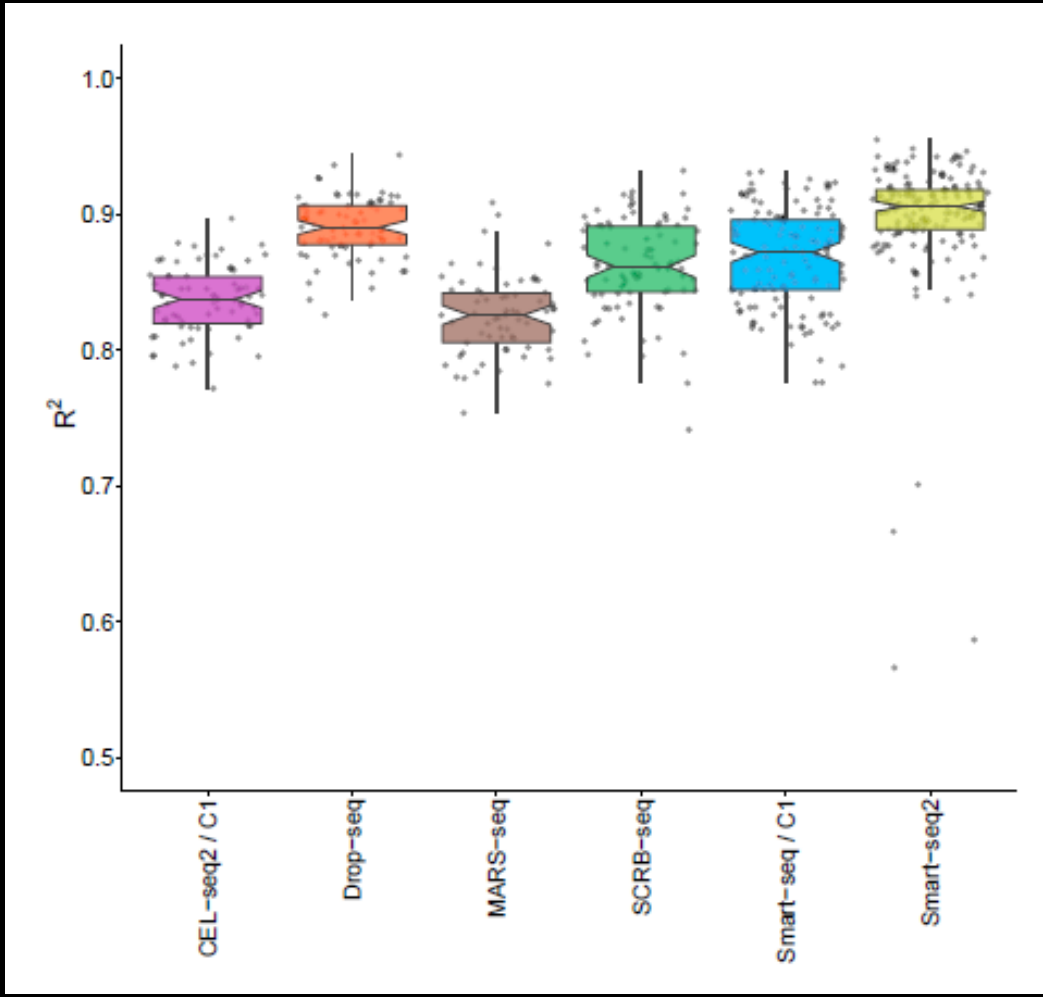


Subsampling of the same number of transcripts from each cell (Down-sampling)

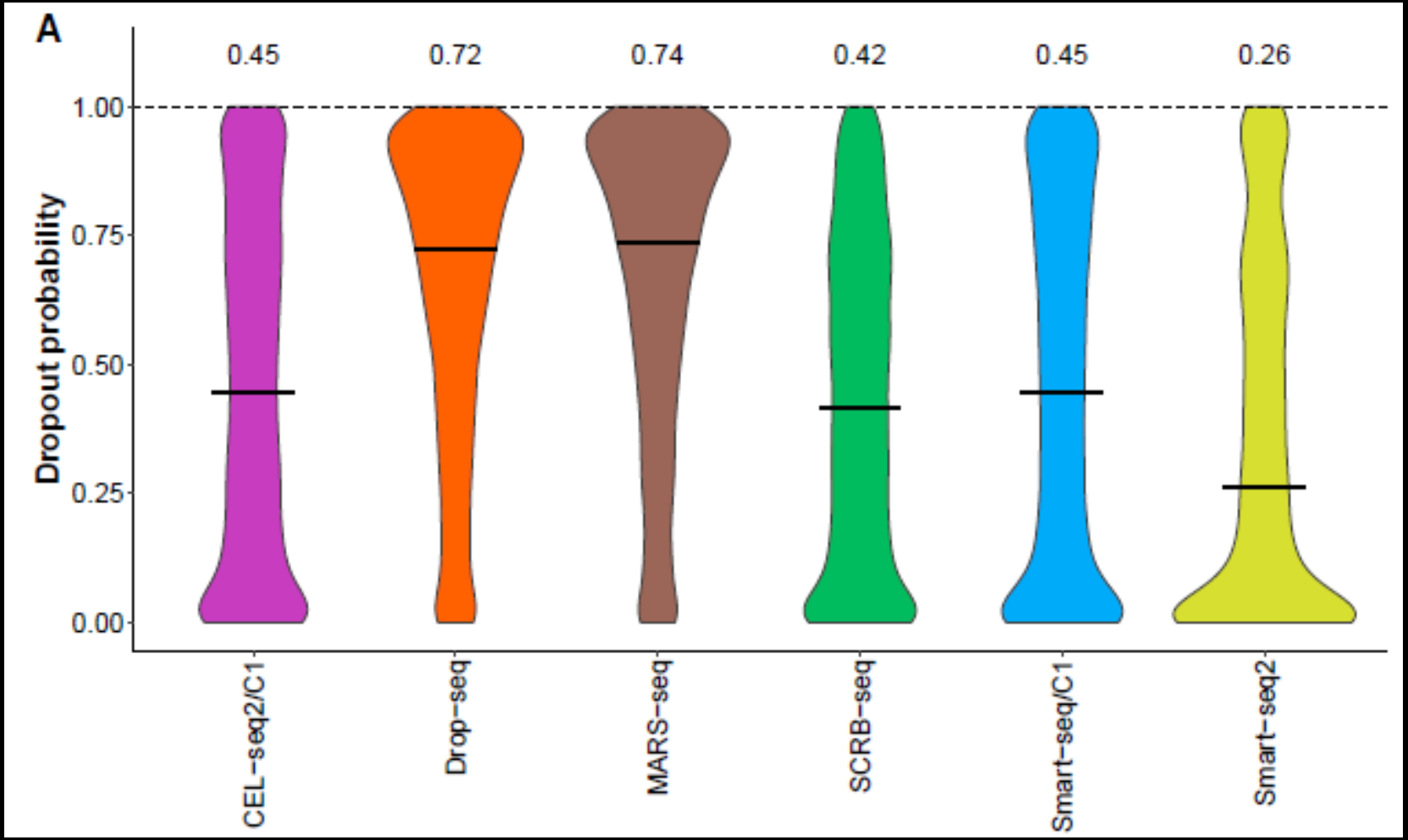
Sensitivity



Accuracy


1.0
0.9
0.8
0.7
0.6
0.5
R^2
CEL−seq2 / C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq / C1
Smart−seq2

Precision


A
0.45
0.72
0.74
0.42
0.45
0.26
1.00
0.75
0.50
0.25
0.00
Dropout probability
CEL−seq2/C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq/C1
Smart−seq2

Cost

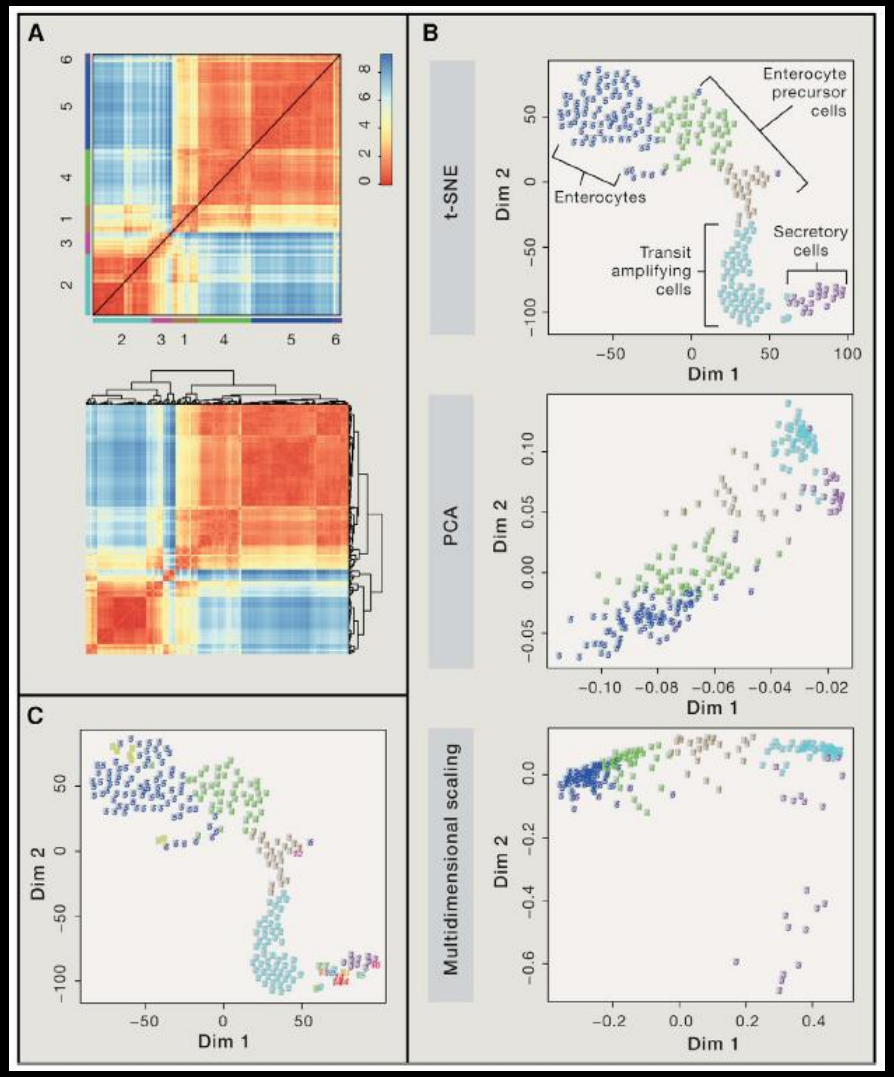
Method	CEL-seq2/C1	Drop-seq	MARS-seq	SCRB-seq	Smart-seq/C1	Smart-seq2
Single-cell isolation	automated in the C1 system	droplets	FACS	FACS	automated in the C1 system	FACS
ERCC spike-ins	yes	no	yes	yes	yes	yes
UMI	6 bp	8 bp	8 bp	10 bp	no	no
Full-length coverage	no	no	no	no	yes	yes
1st strand synthesis	oligo-dT	oligo-dT	oligo-dT	oligo-dT	oligo-dT	oligo-dT
2nd strand synthesis	RNAseH / DNA Pol	template switching	RNAseH / DNA Pol	template switching	template switching	template switching
Amplification	IVT	PCR	IVT	PCR	PCR	PCR
Imaging of cells possible	yes	no	no	no	yes	no
Protocol usable for bulk	yes	no	yes	yes	yes	yes
Sequencing	paired-end	paired-end	paired-end	paired-end	single-end	single-end
Library cost /cell	~9.5€	~0.1€	~1.3€	~2€	~25€	~3/30*

Table S1 (related to Figure 2): Overview of single-cell RNA-seq methods.

* in-house produced Tn5 / commercial Tn5

Biological Insights from scRNA-Seq

Identification of cell types



Biological Insights from scRNA-Seq

Identification of marker genes

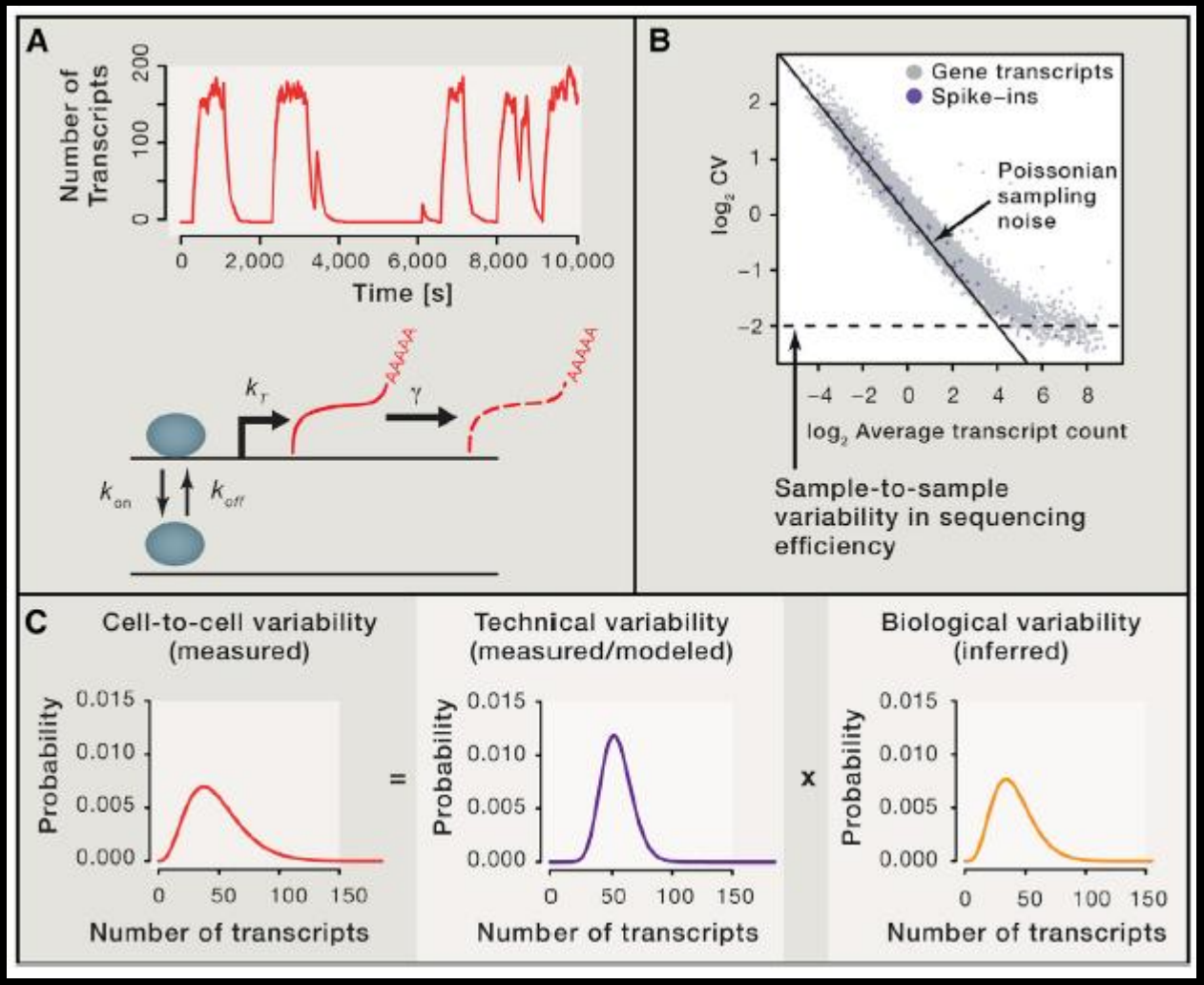
Biological Insights from scRNA-Seq

Inference of differentiation Dynamics

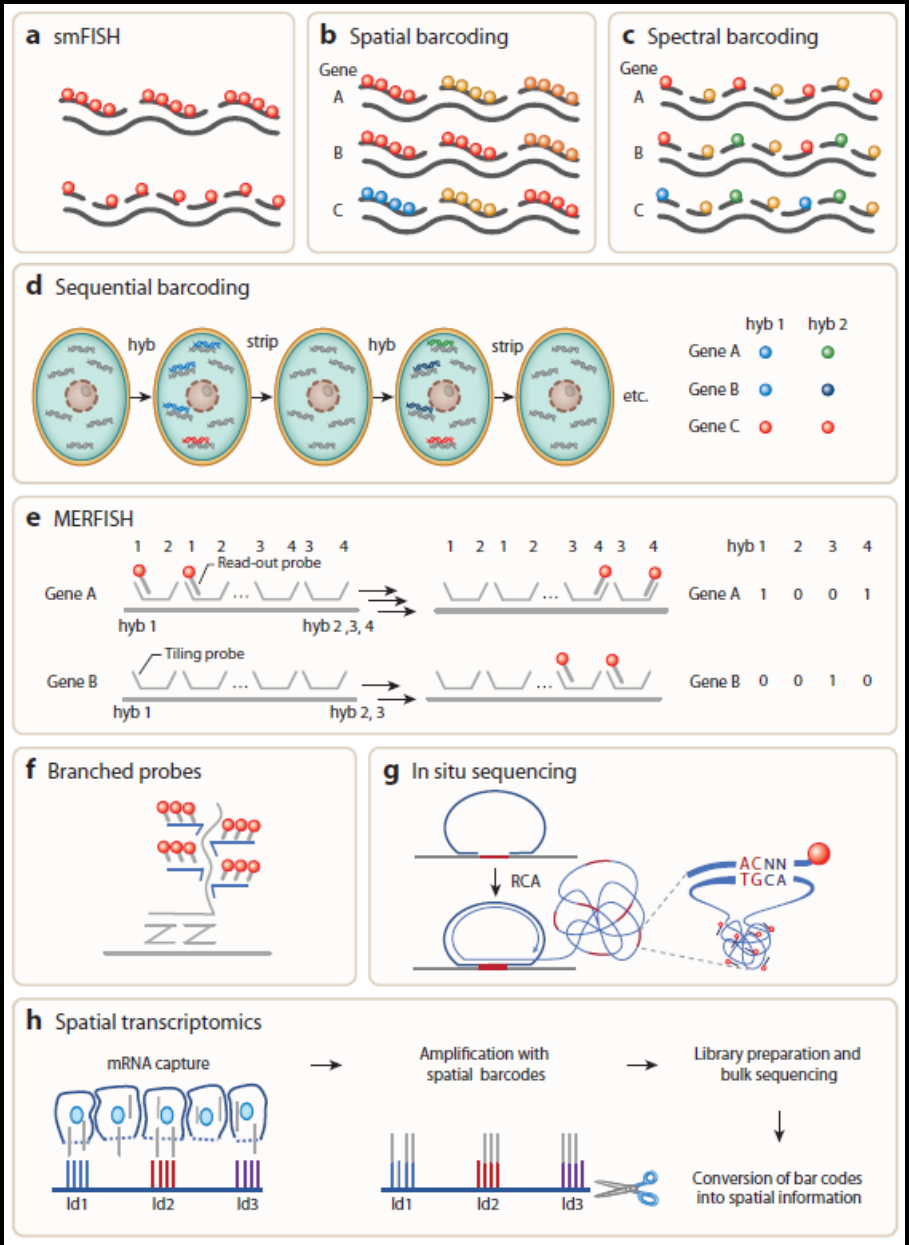
Generally, if a sample is analyzed that contains all differentiation stages of a given cell lineage, a pseudo-temporal ordering of single-cell transcriptomes can be inferred.

Biological Insights from scRNA-Seq

Measuring Gene Expression Noise

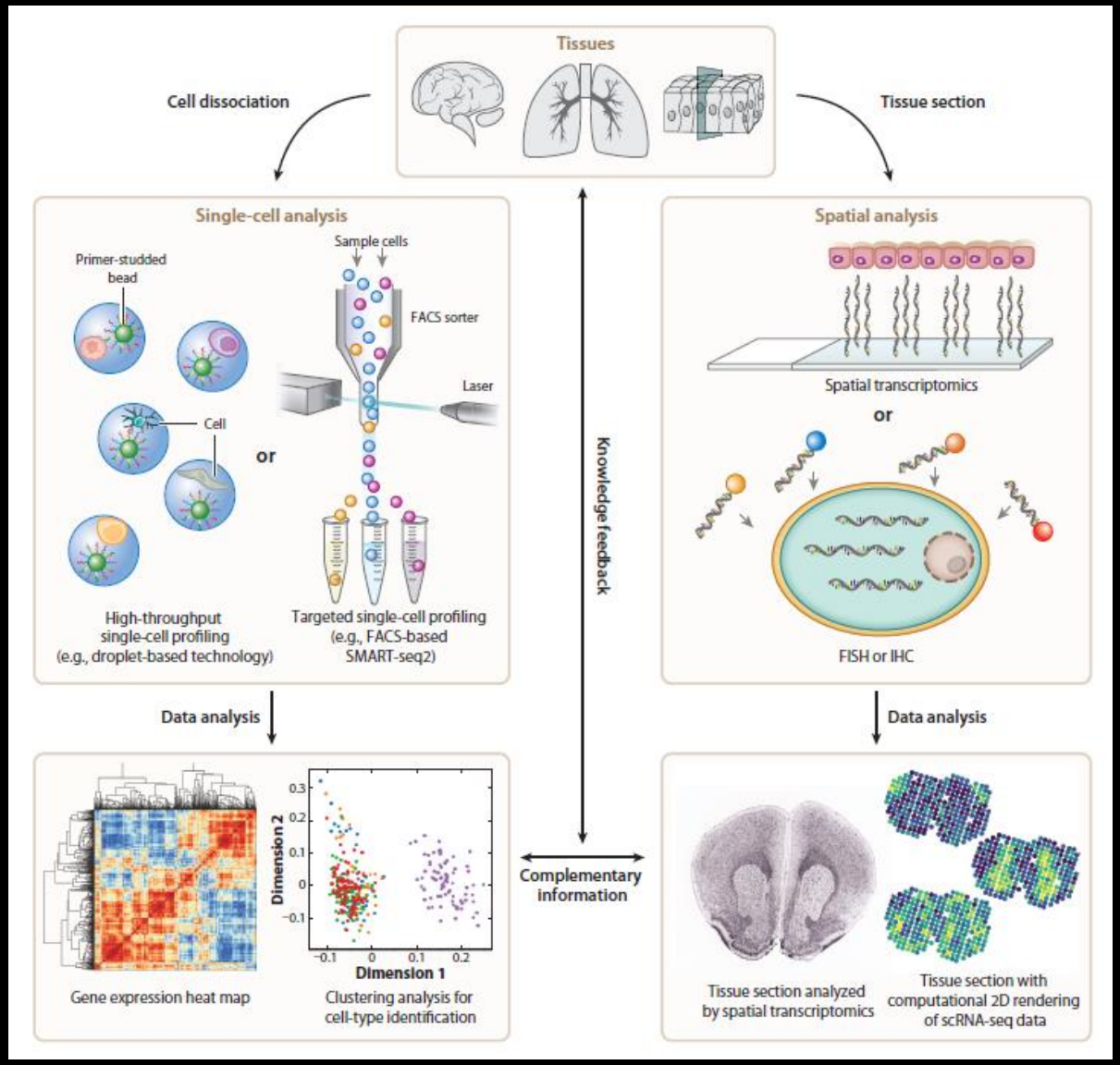


Gene expression assays that retain spatial information



Chen et al., 2018

Hypothetical future workflow



Chen et al., 2018

Thanks for your attentions