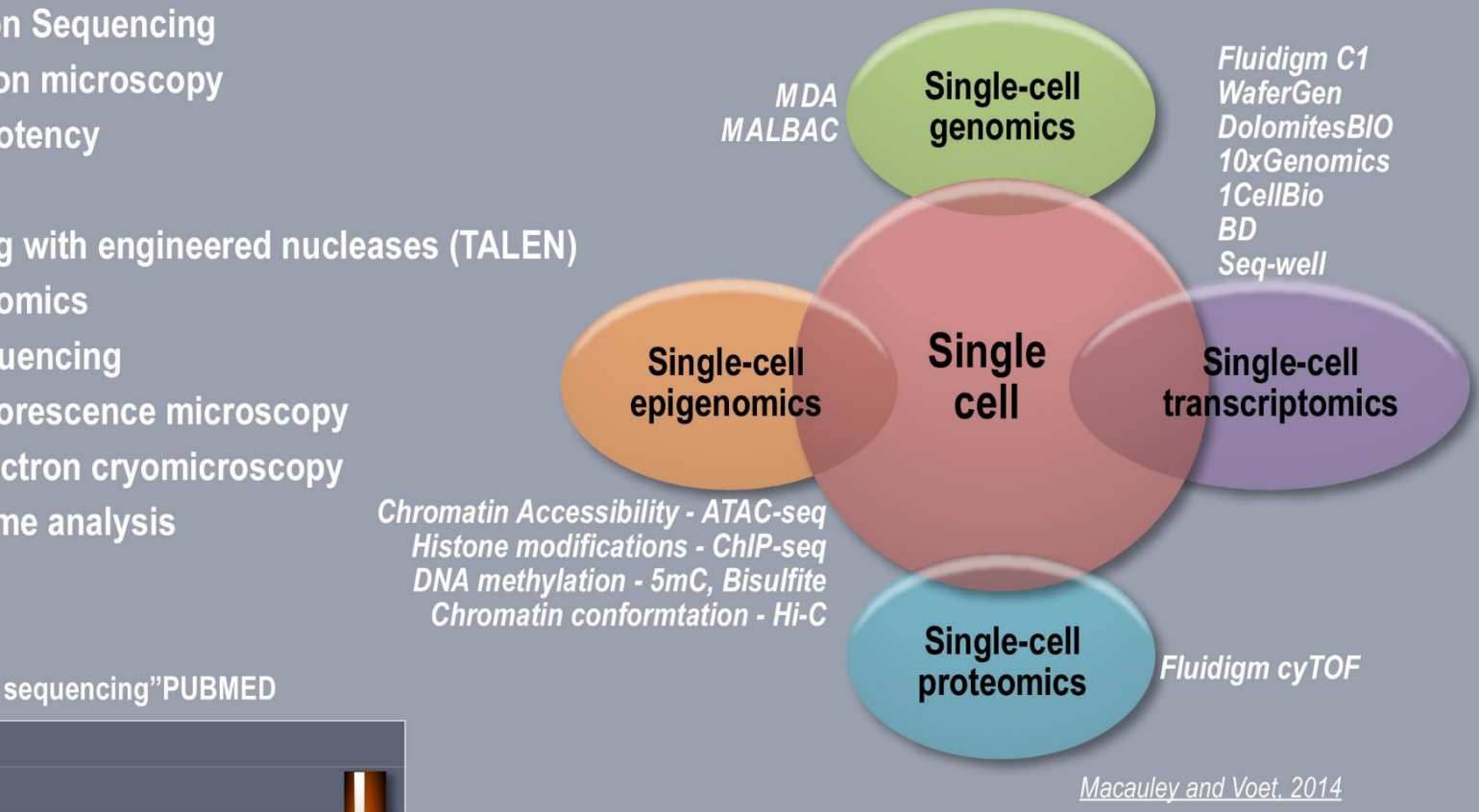
Singel Cell RNA-seq: Technologies and Experimental Approaches

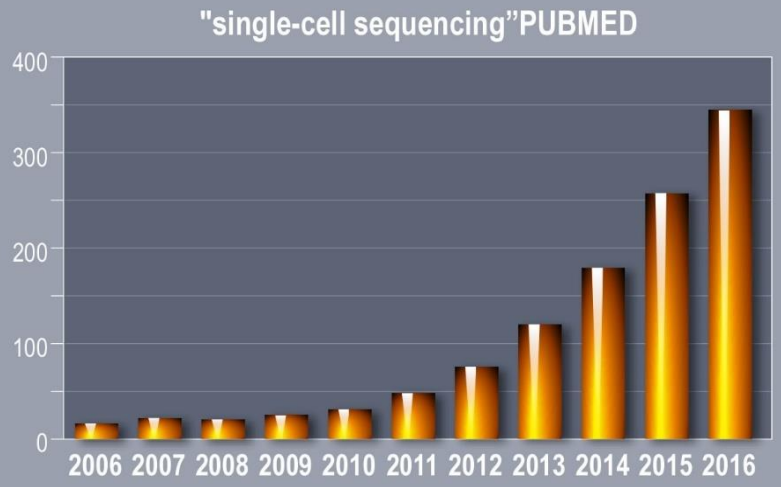
Kévin Lebrigand
UCAGenomix, Nice-Sophia-Antipolis
lebrigand@ipmc.cnrs.fr
@kevinlebrigand
Roscoff, 18 juin 2018

Methods of the year 2013 (Nature Methods)

2007: Next Generation Sequencing
2008: Super-resolution microscopy
2009: Induced pluripotency
2010: Optogenetics
2011: Genome editing with engineered nucleases (TALEN)
2012: Targeted proteomics
2013: Single cell sequencing
2014: Light-sheet fluorescence microscopy
2015: Cryo-EM or electron cryomicroscopy
2016: Epitranscriptome analysis

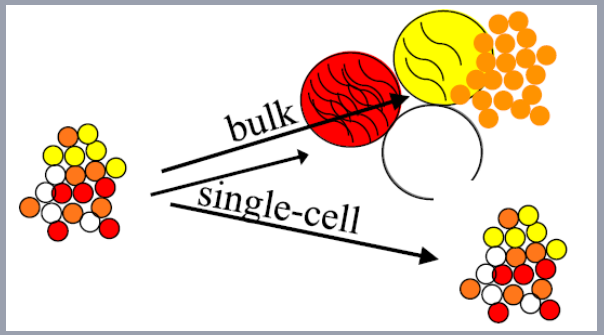


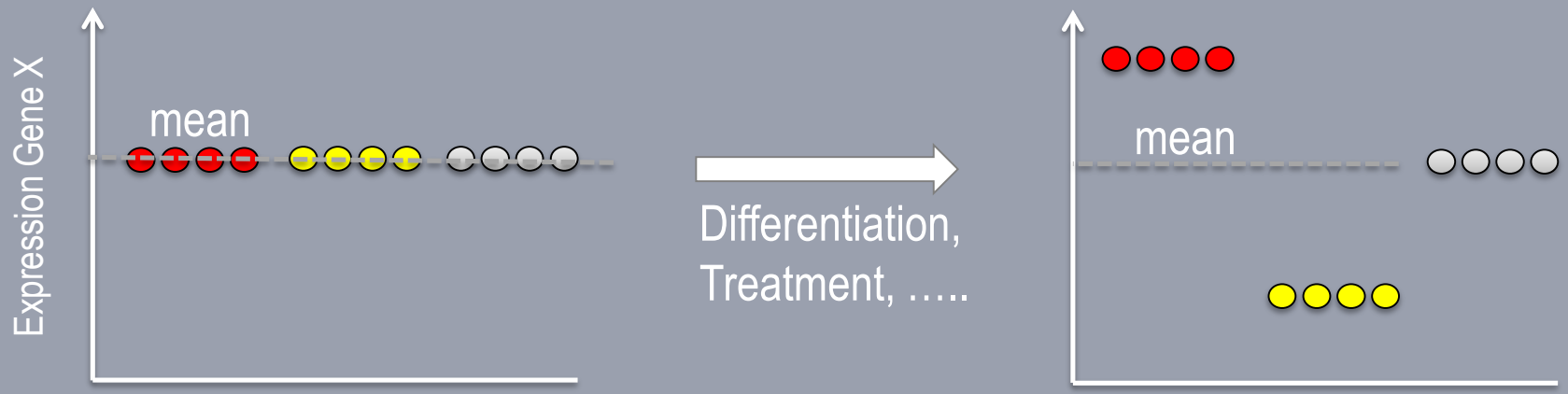
Macauley and Voet, 2014



Why single cell profiling?

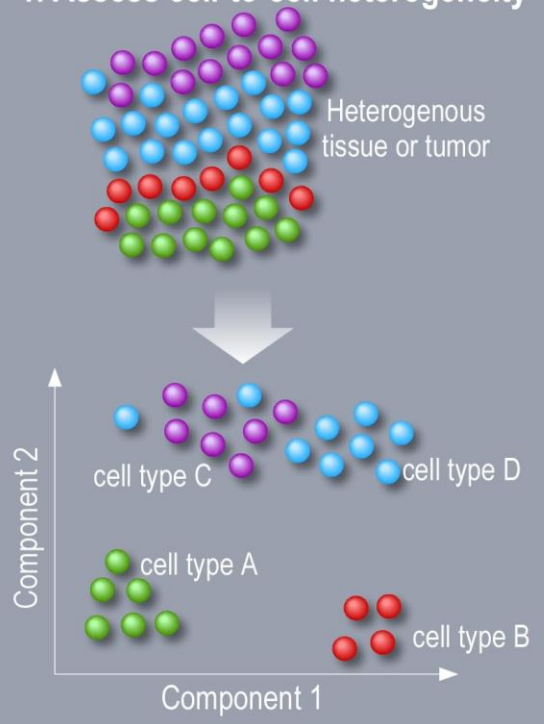
Stop measuring gene average as in bulk
Population sequencing yields average values
Changes in subpopulation might remain undetected



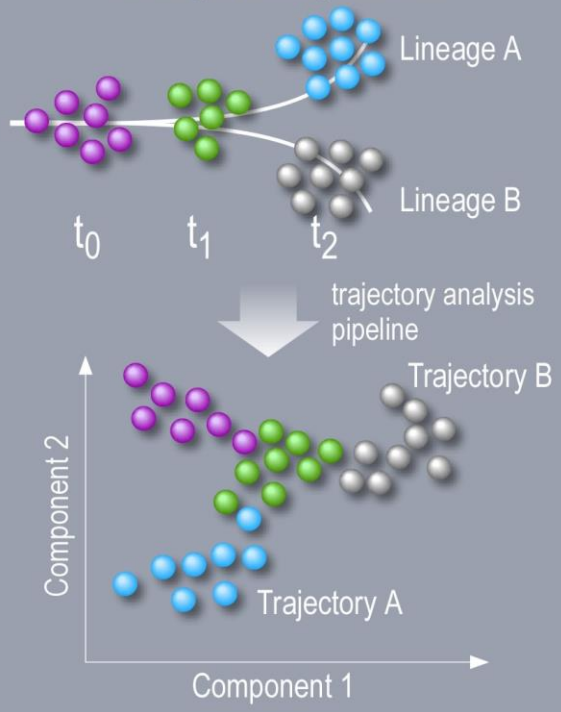


Why single cell profiling?

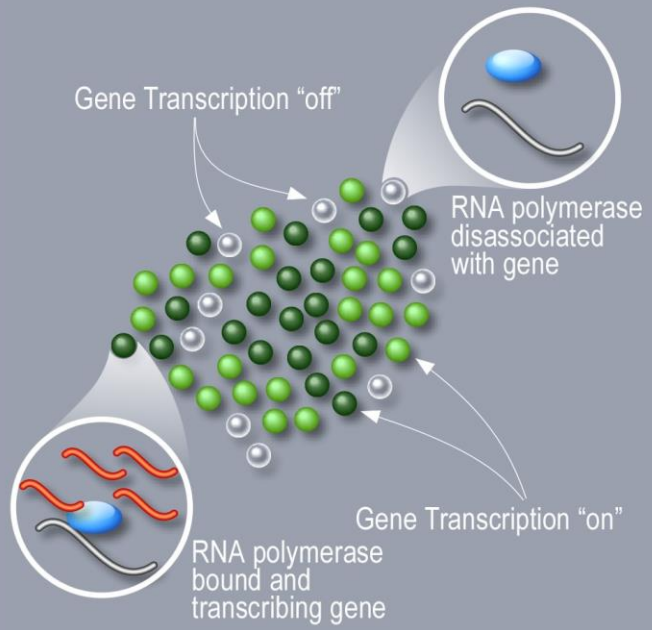
1. Assess cell-to-cell heterogeneity



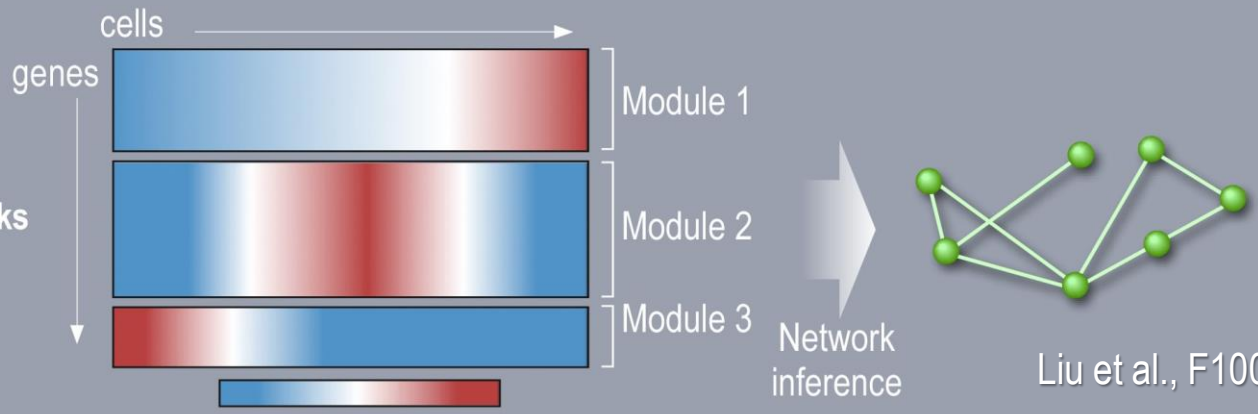
2. Map cell trajectories



3. Dissect transcriptional mechanics



4. Infer gene regulatory networks



Liu et al., F1000, 2016

Single cell analysis: the context

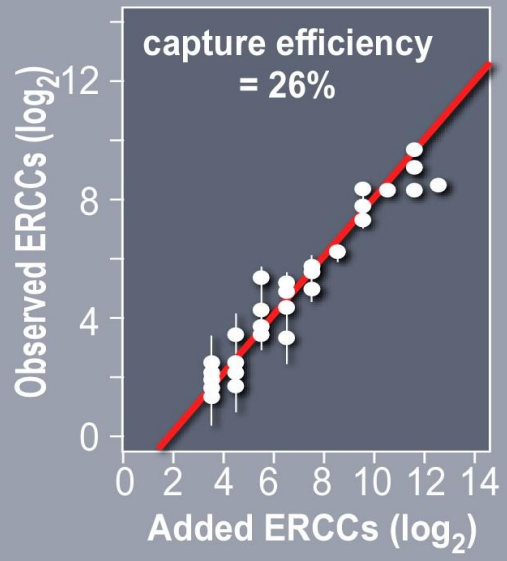
mRNA libraries	100 ng total RNA (10.000 cells)
Single cell mRNA libraries	10 pg total RNA (<<1pg mRNA)

- Highly efficient library preparation techniques
- Elimination of PCR amplification bias

- Spike-in ERCC molecules to evaluate yield and capture efficiency
- Use of Unique Molecular Identifiers to monitor the number of molecules:
  - ***Random multimers to reduce cloning biases during small RNAs profiling*** *(Jayaprakash, NAR, 2011)*
  - ***Detection of rare DNA mutations*** *(Ståhlberg et al, NAR, 2016)*
  - ***Improved accuracy of molecule counting***

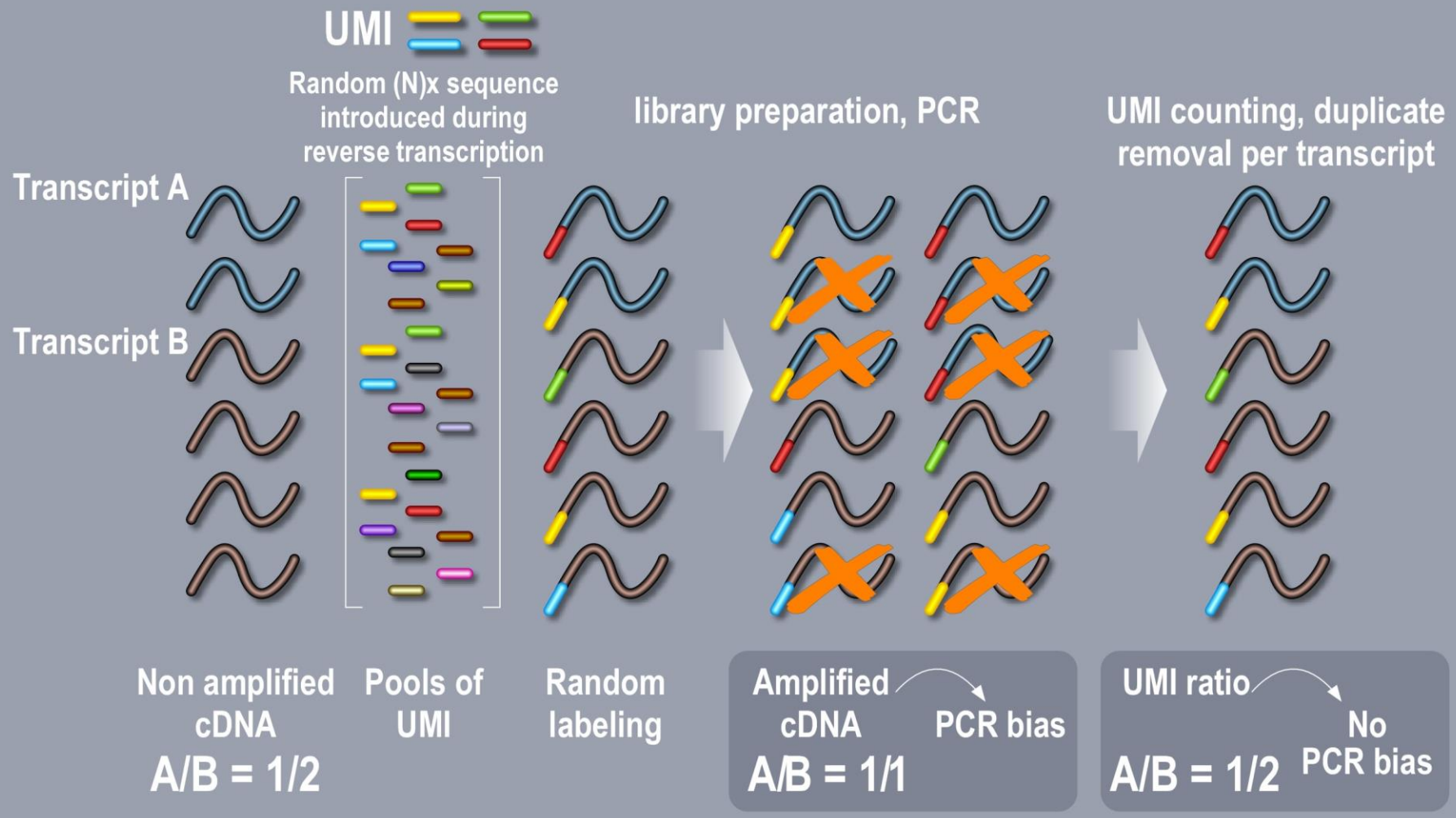
ERCC spike-ins

ERCC (Externals RNA Controls Consortium) spikes is 92 polyadenylated RNA molecules
- different sequences and lengths,
- no homology with known genome sequences,
- 2 mixes (mix1 and mix2) with the 92 sequences but in different amounts
- relative amounts of spikes in mix1 and mix2 from 1 to 10e6



Single Cell capture efficiency
Arguel et al., NAR (2016)

Unique Molecular Identifier (*Islam et al., Nature Methods, 2014*)



UMIs : *Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Meth 9, 72-74 (2012)*
UMIs for single cell transcriptome: *Islam, S. et al. Quantitative single-cell RNA-seq with UMI . Nat Methods 11, (2014).*

How much RNA does a typical mammalian cell contain?

https://www.qiagen.com/fr/resources/faq?id=06a192c2-e72d-42e8-9b40-3171e1eb4cb8&lang=en

The RNA content and RNA make up of a cell depend very much on its developmental stage and the type of cell. To estimate the approximate yield of RNA that can be expected from your starting material, we usually calculate that a typical mammalian cell contains **10-30 pg total RNA**.

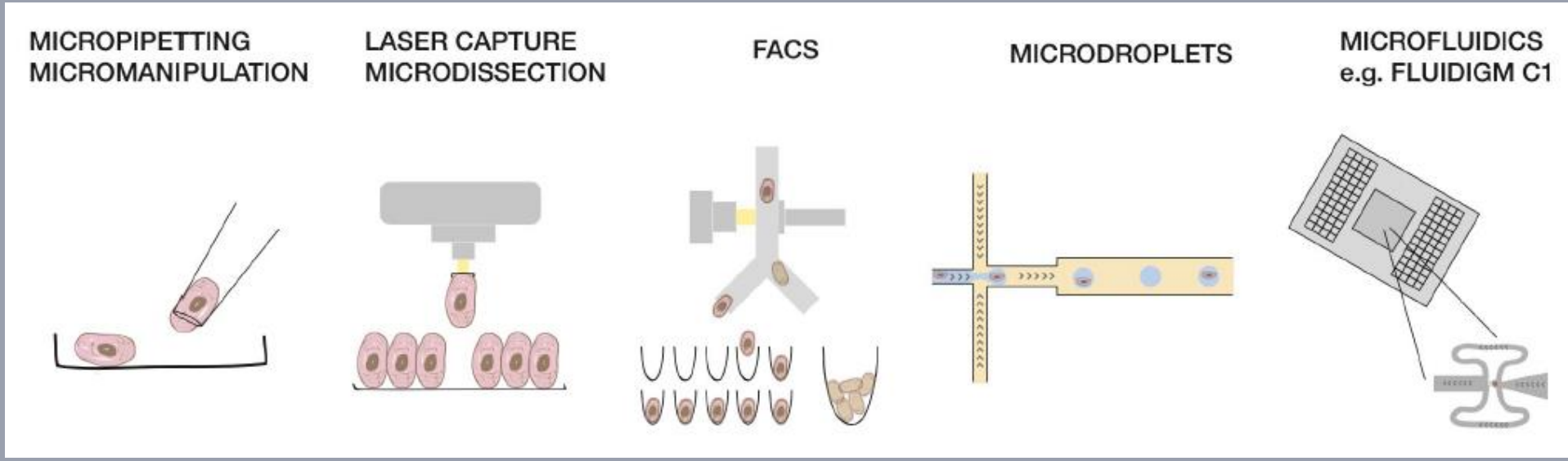
The majority of RNA molecules are tRNAs and rRNAs. mRNA accounts for only **1–5%** of the total cellular RNA although the actual amount depends on the cell type and physiological state. Approximately **360,000 mRNA** molecules are present in a single mammalian cell, made up of approximately 12,000 different transcripts with a typical length of around 2 kb. Some mRNAs comprise 3% of the mRNA pool whereas others account for less than 0.1%. These rare or low-abundance mRNAs may have a copy number of only 5-15 molecules per cell.

Average total RNA yields

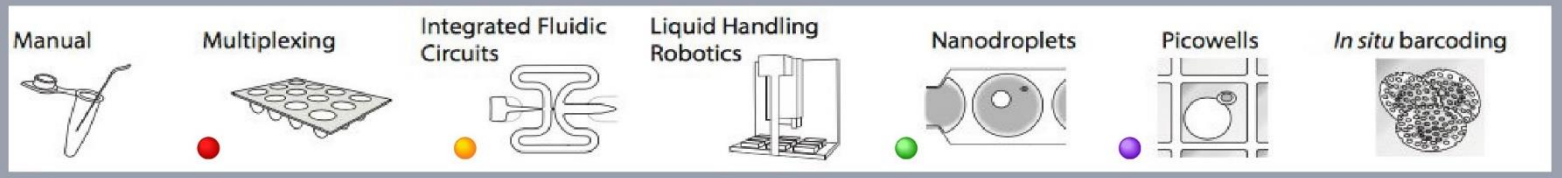
Primary cells (1×10^6 cells)	Total RNA (µg)
Dendritic cells, human	4
Hematopoietic progenitor cells ($CD34^+$), human	1
Fibroblasts, rat	5
PBMC	8

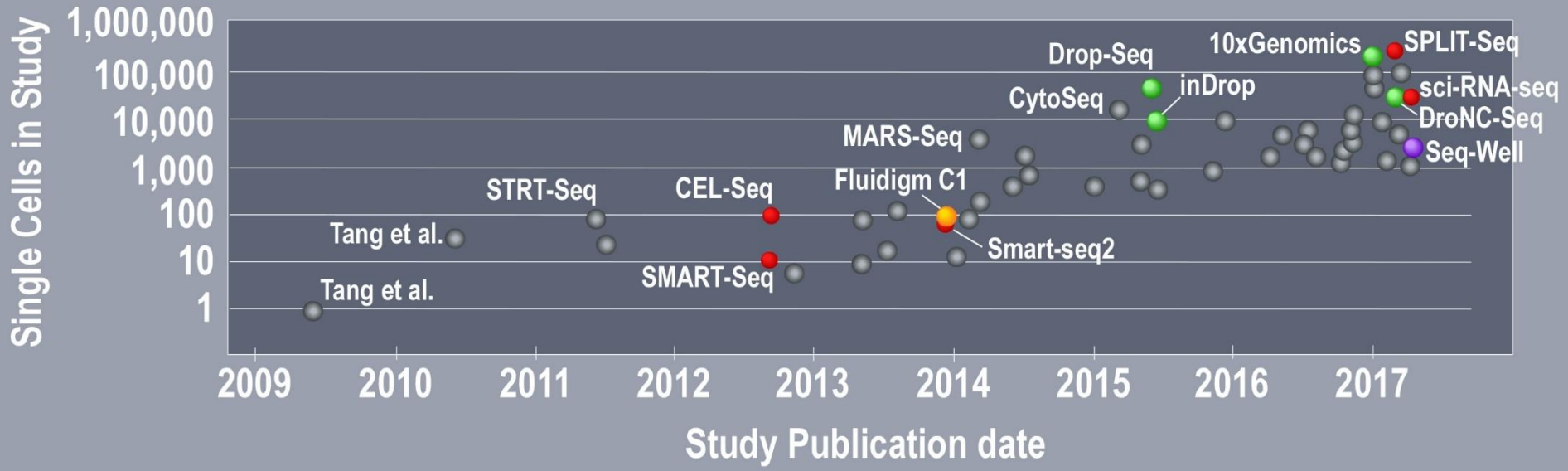
Cell lines (1×10^6 cells)	Total RNA (µg)
Colon carcinoma cells	30
HEK 293 cells	16
HeLa cells	32
HUV-EC-C	38
THP1 cells	16
U937 cells	12

To measure sequences in individual cells, need methods that capture one cell at a time

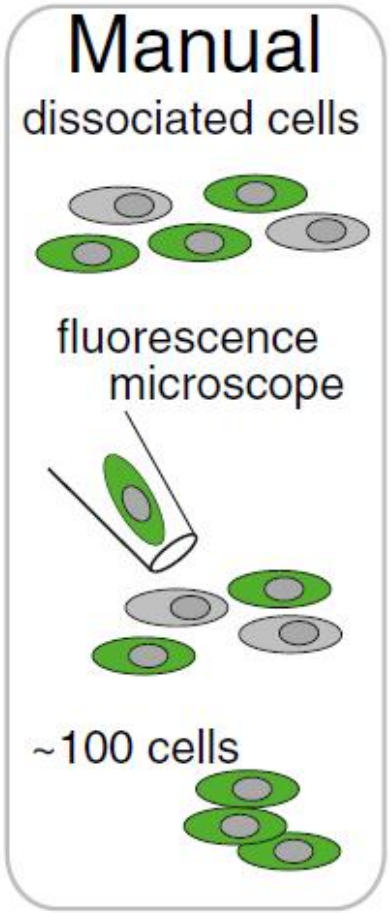


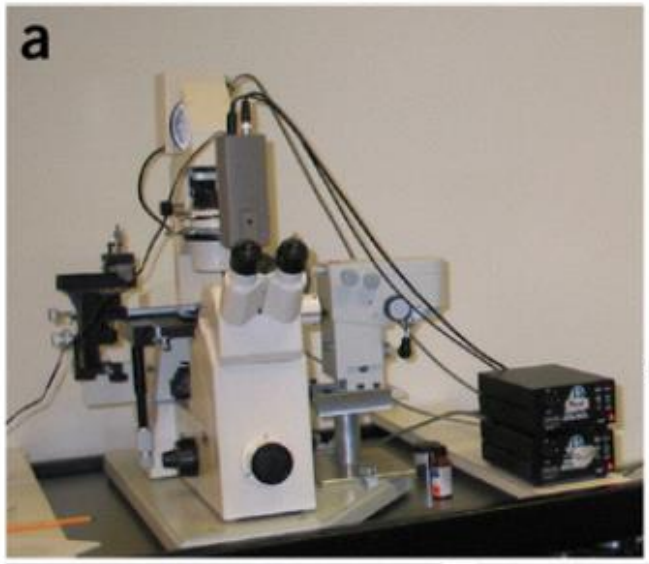
Scaling Single Cell Transcriptomics (*Svensson et al., Nature Methods, 2017*)

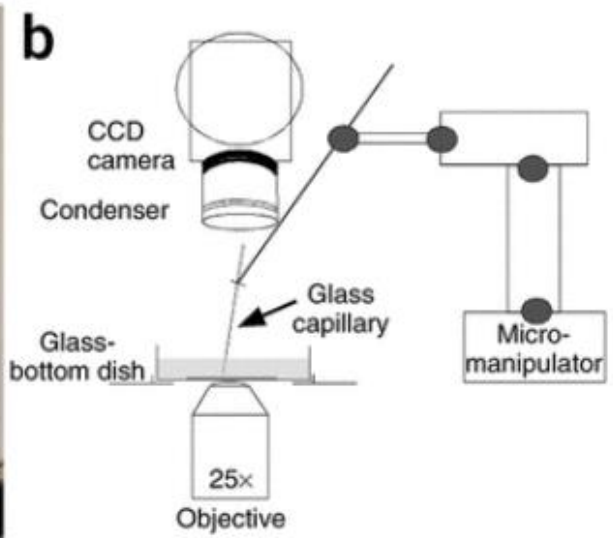


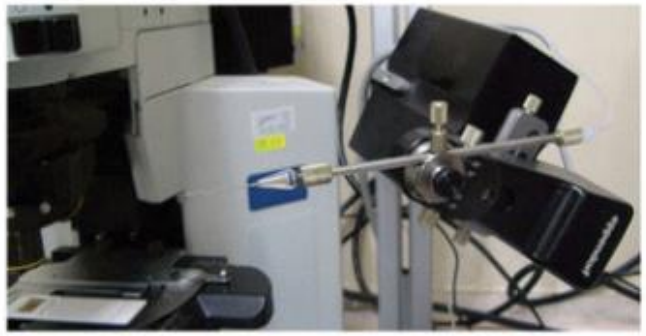


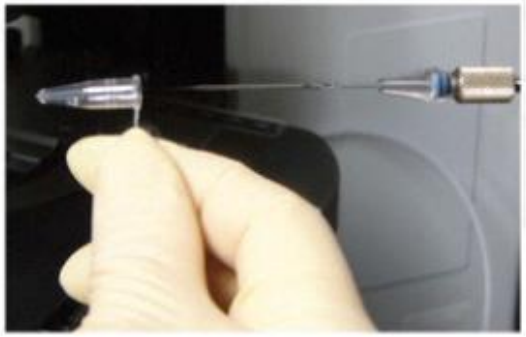
Manual Cell sorting





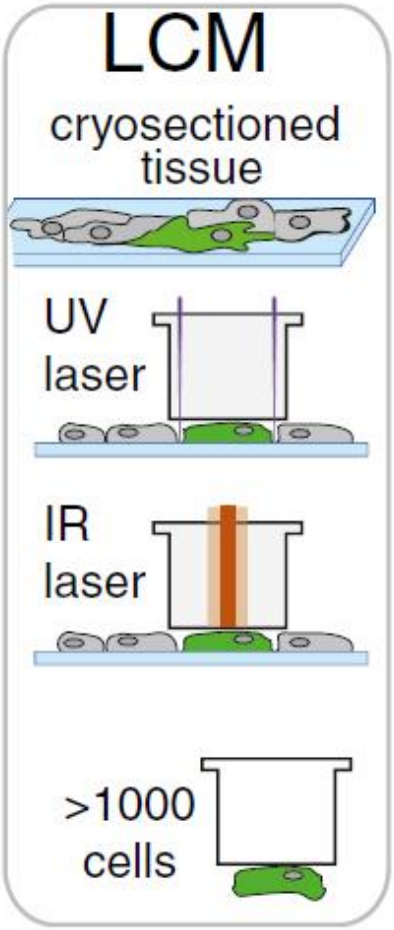


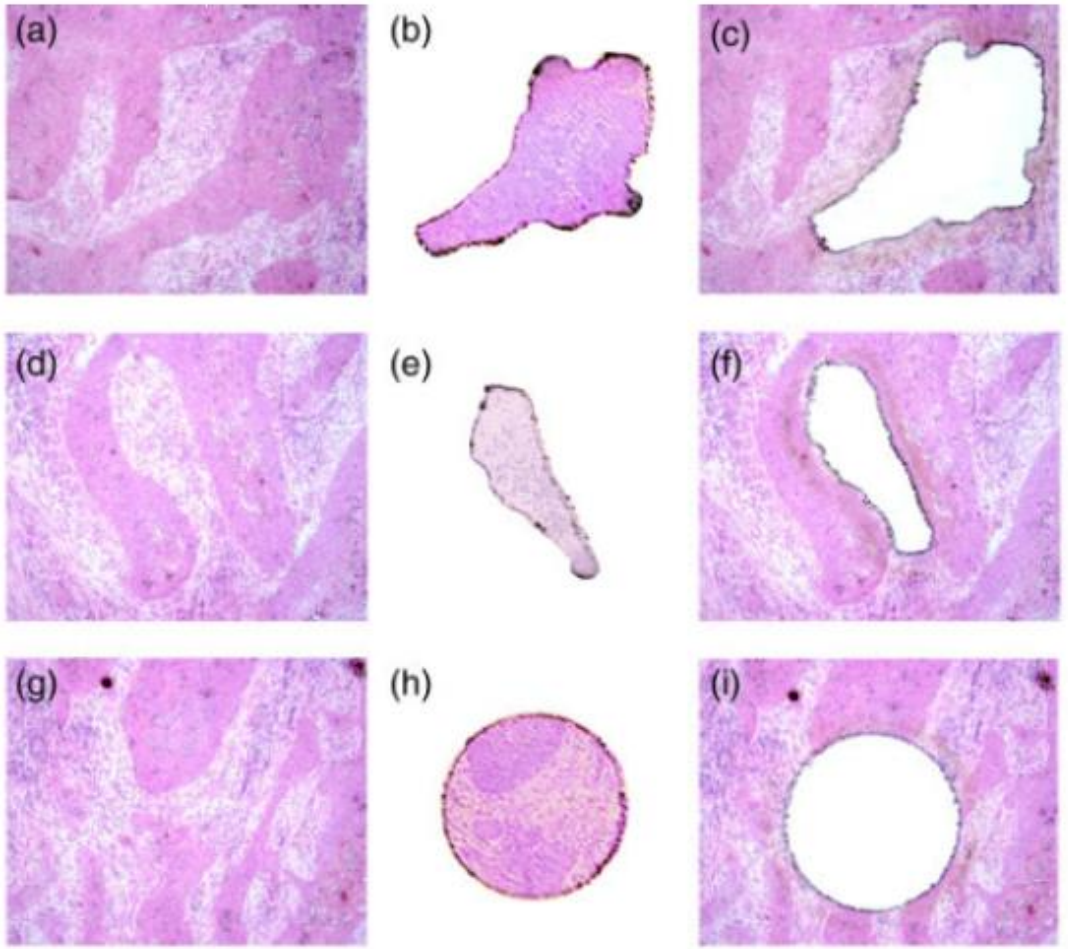




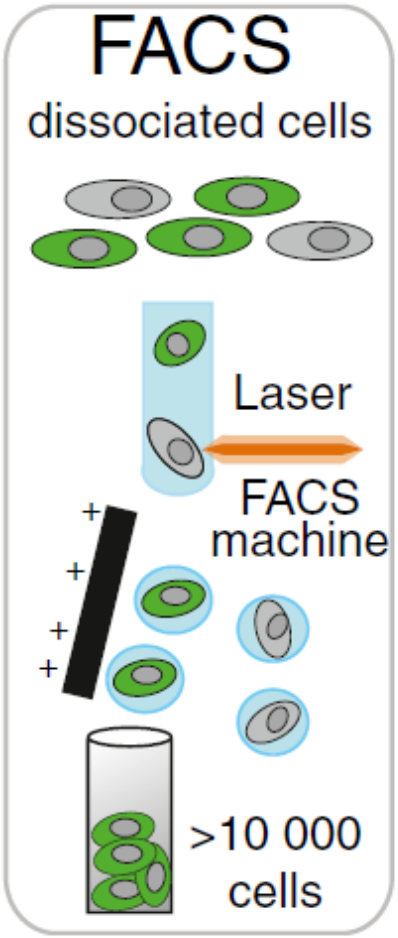
http://www.nature.com/nprot/journal/v6/n5/images_article/nprot.2011.322-F2.jpg

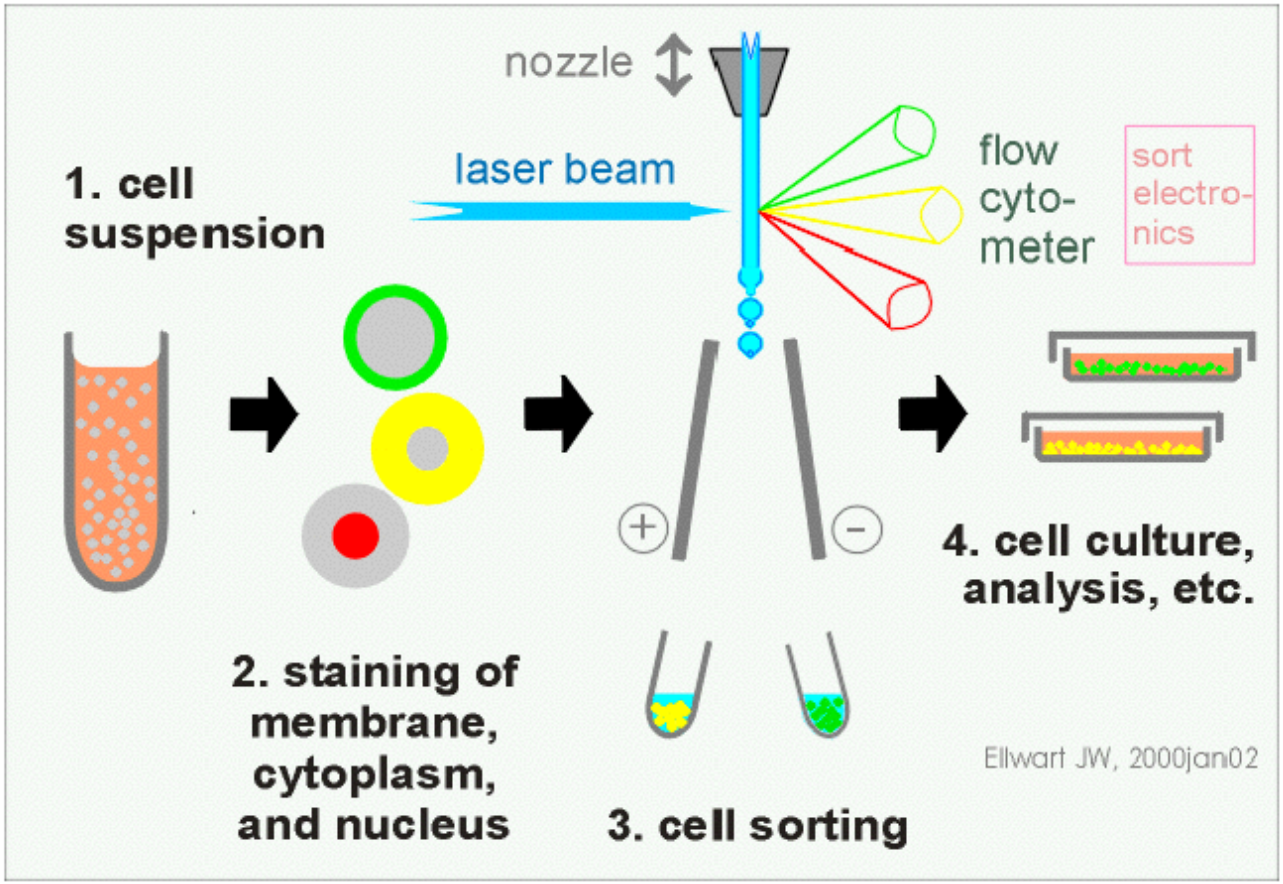
Laser Capture Microdissection





http://www.genomemedicine.com/content/figures/gm247-2-l.jpg



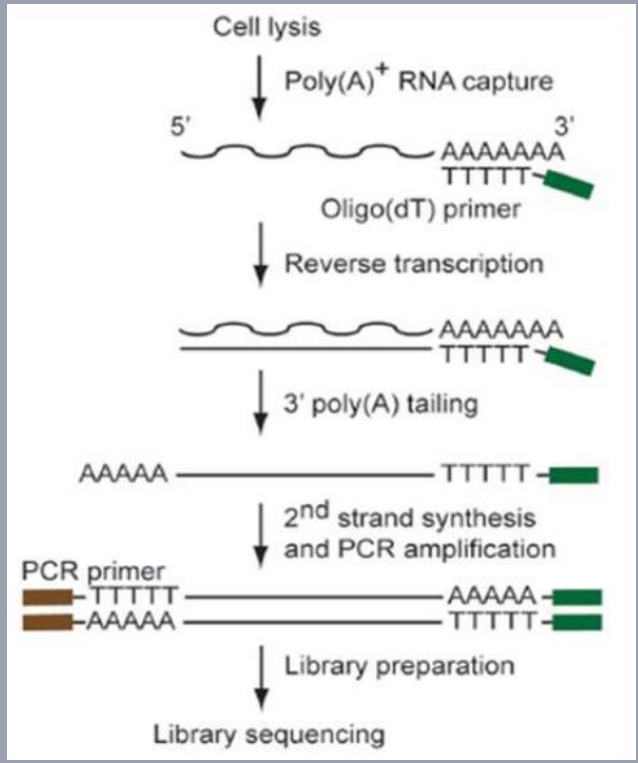


http://www.flowlab-childrens-harvard.com/yahoo_site_admin/assets/images/principle123.285181420_std.gif

Tang et al., 2009

Protocol

- Total RNA is isolated and fragmented,
- Converted to cDNA by using an **oligodT** primer with a specific anchor sequence,
- **Second strand synthesis using a polyT primer** with another anchor sequence,
- **PCR amplified** from primers against the two anchor sequences.



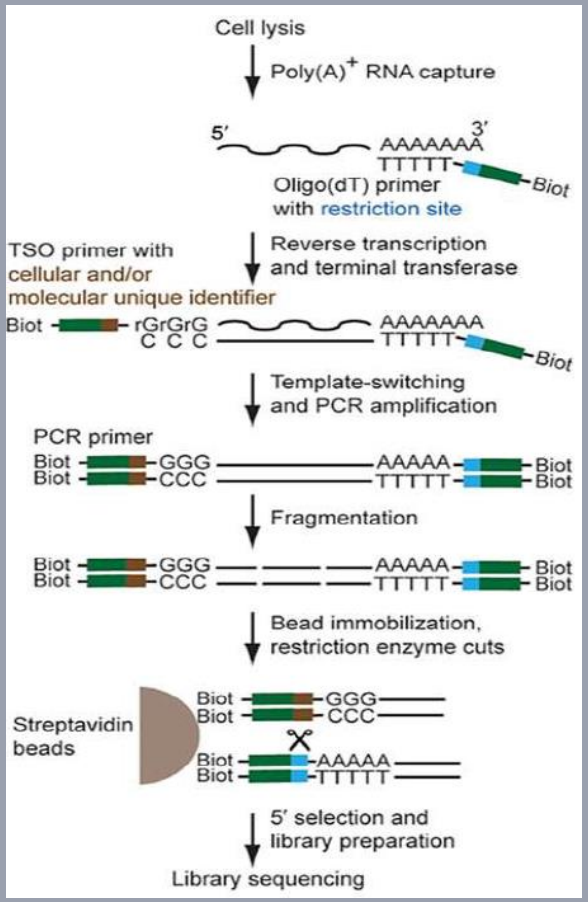
Drawback

- Premature termination of RT reduces transcript coverage at the 5' end
- Introduction of a polyA tail in addition to its own polyA sequence at the **3' end** of the input RNA causes a **loss of strand information** in the resulting double-stranded cDNA

STRT: Single cell tagged reverse transcription, Islam, Nat.Prot. (2012) @ Linnarsson's lab

Protocol

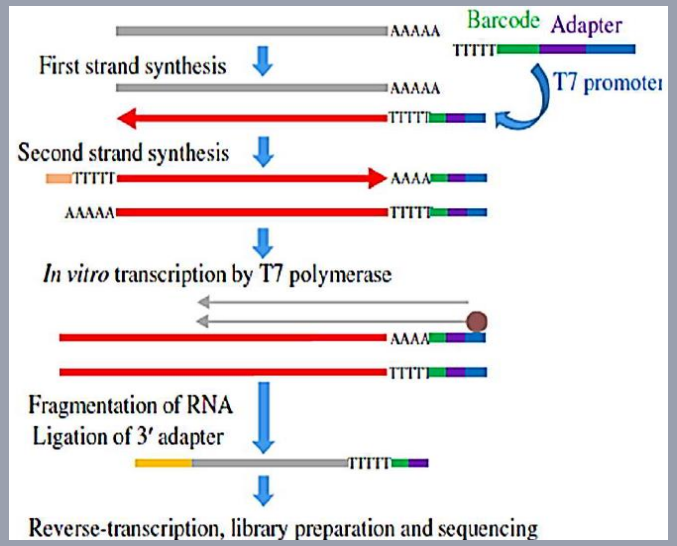
- based on **template switching**,
- **5' end cDNA tagged N5 UMI**,
- biotin is introduced at both the 3' and 5' ends via the use of biotinylated primers.
- enzymatic cleavage leads to the selection of only the 5' fragments for library construction.
- sequencing and analysis shows **5' read bias**



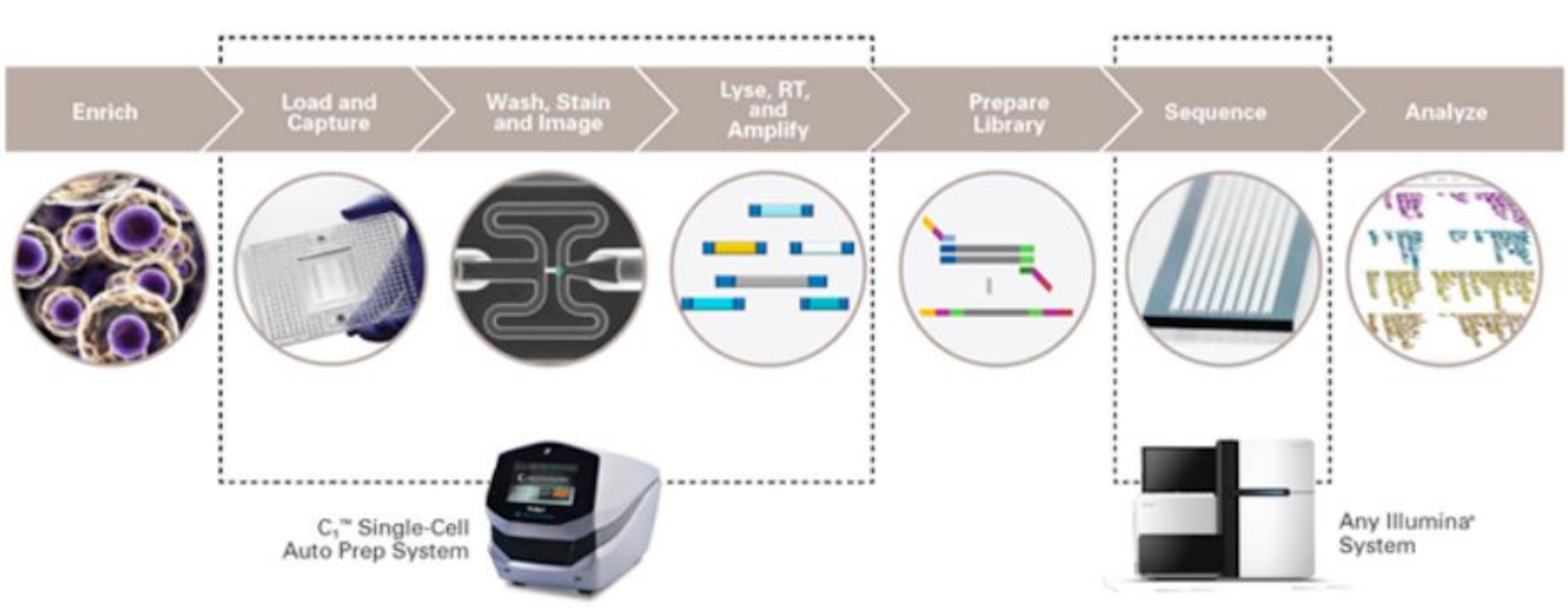
Cell expression by linear amplification and sequencing (CEL-seq, CEL-seq2)

Protocol

- OligodT primer containing the 5' Illumina adaptor, a cell barcode, and a T7promoter (**CEL-seq2 add UMI**),
- RT and second-strand synthesis,
- cDNA from all the cells is pooled and amplified **by in vitro transcription** from the T7 promoter,
- RNA fragmentation, Illumina adaptor ligated at 3' end
- RNA is reverse transcribed, library is prepared then sequencing,
- Sequencing of the **3´terminal fragments**



Microfluidics: Fluidigm C1



- Limiting factor is size of capture chambers (96 chips: 5-10µm, 10-17µm, >17µm)
- 800 cells chip (10-17 µm diameter cells)
- ScriptHub: protocols for running SMARTer, SmartSeq2, CEL-seq, STRT, …

openApp Chip for custom protocols development

SMART-Seq, Ramskold (2012), SMART-seq2, Picelli (2014) @ Sandberg's lab

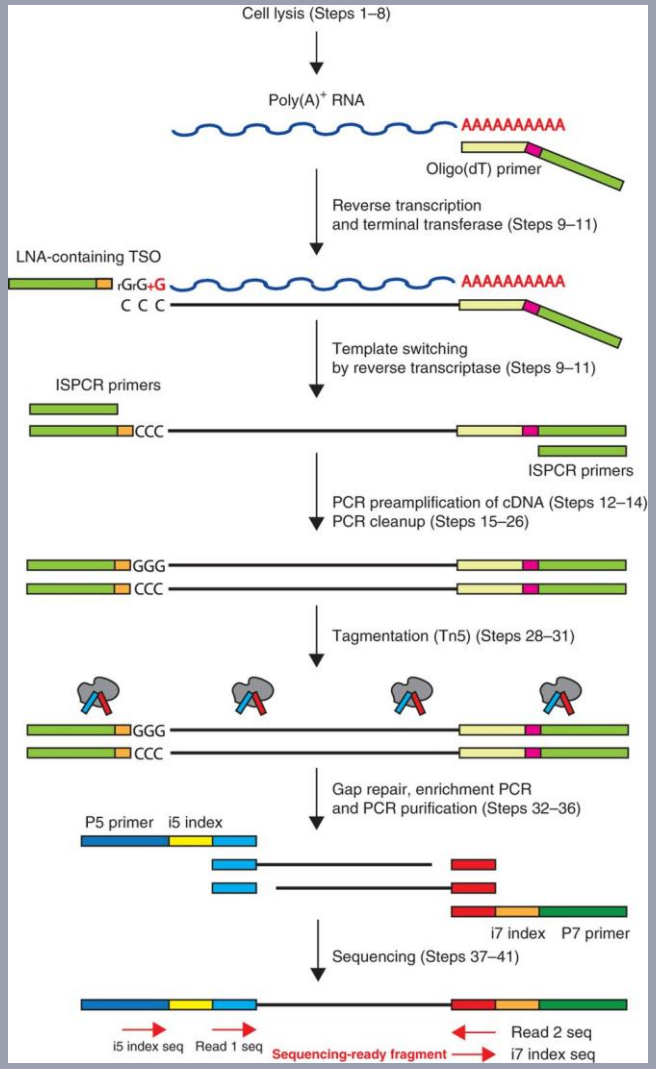
SMART= Switching Mechanism at the end of the 5´-end of the RNA Transcript

Protocol

- Based on template switching mechanism,
- Anchor a 5' universal seq. along with Locked nucleic acid by reverse transcription,
- cDNA is then **PCR amplified ,**
- Tagmentation is used to construct libraries,
- Generate **full transcript coverage**

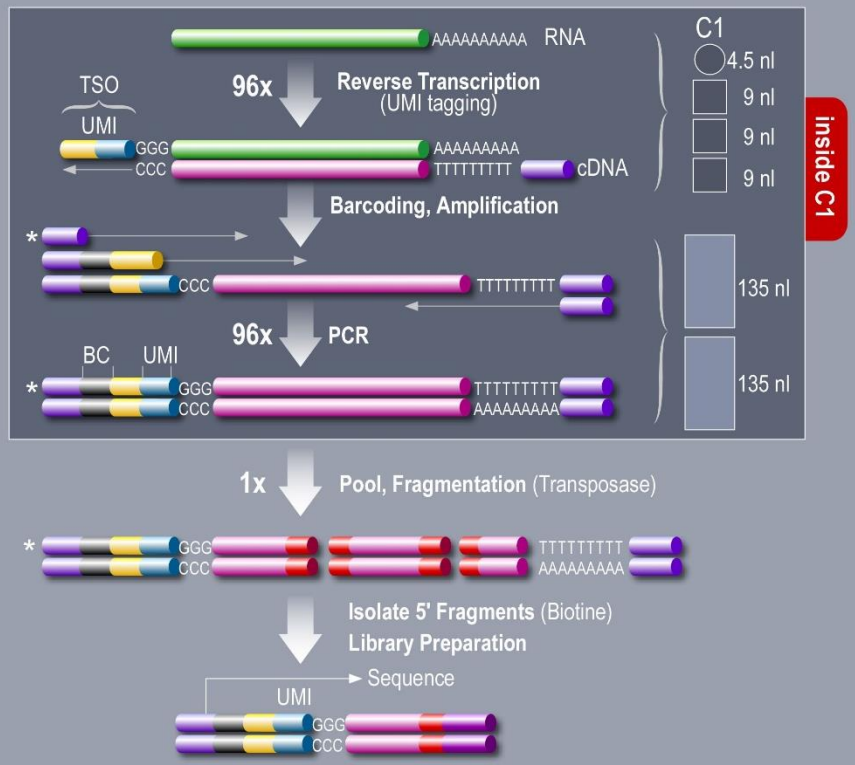
Drawback

- No UMIs



A cost effective 5' selective single cell transcriptome profiling approach

- UMIs inside Fluidigm C1, PCR bias removal (molecules couting)
- 60k transcripts in HEK (homogeneity of chemistry)
- Illumina and Ion Torrent sequencing (no paired-end required)
- Capture efficiency 26% (ERCC spike-ins)
- Sequences 5' end of transcripts → TSS identification



Arguel et al., Nucleic Acid Research, 2016

Now available on ScriptHub (Fluidigm)

Single-cell mRNA Seq with Integrated Barcoding mRNA Sequencing

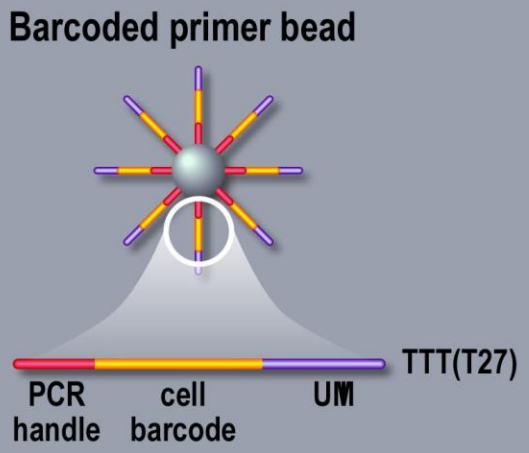
Arguel et al.
Pascal Barbry - UCA Genomix - IPMC, CNRS - University of Cote d'Azur

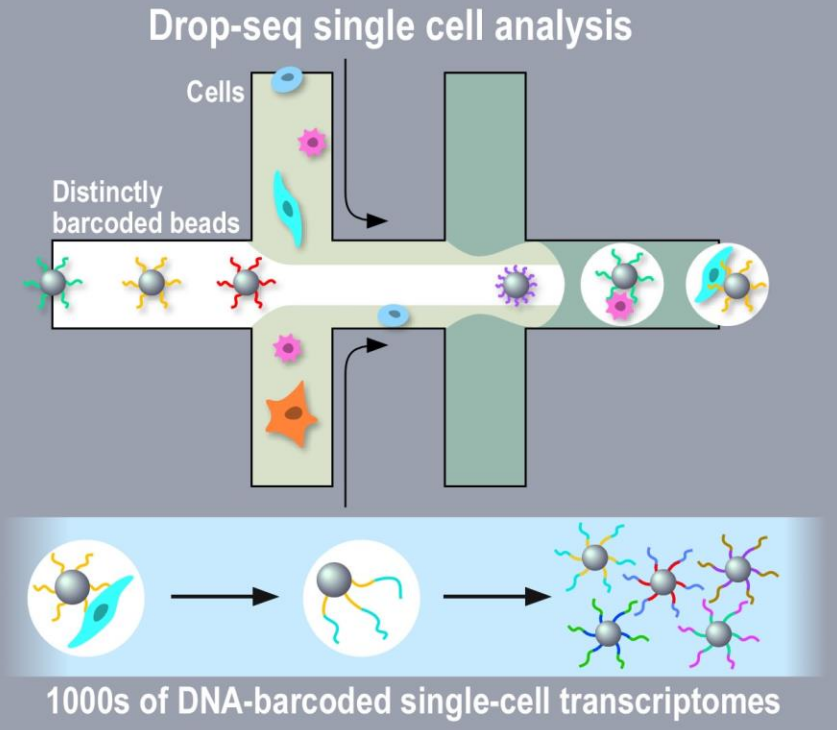
Drop-seq / inDROP: droplets-based scRNA-seq (*Mascosko et al., Klein et al., CELL, may 2015*)

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Evan Z. Macosko[1, 2, 3], Anindita Basu[4, 5], Rahul Satija[4, 6, 7], James Nemesh[1, 2, 3], Karthik Shekhar[4], Melissa Goldman[1, 2], Itay Tirosh[4], Allison R. Bialas[8], Nolan Kamitaki[1, 2, 3], Emily M. Martersteck[9], John J. Trombetta[4], David A. Weitz[5, 10], Joshua R. Sanes[9], Alex K. Shalek[4, 11, 12], Aviv Regev[4, 13, 14], Steven A. McCarroll[1, 2, 3]

- droplets encapsulation of cells and barcoded beads
- 3' selective single cell RNA-seq
- 12bp cell barcode and 8bp UMI
- capture efficiency 12.5%
- 44,808 mouse retinal cells
- identification of 39 differents cell types

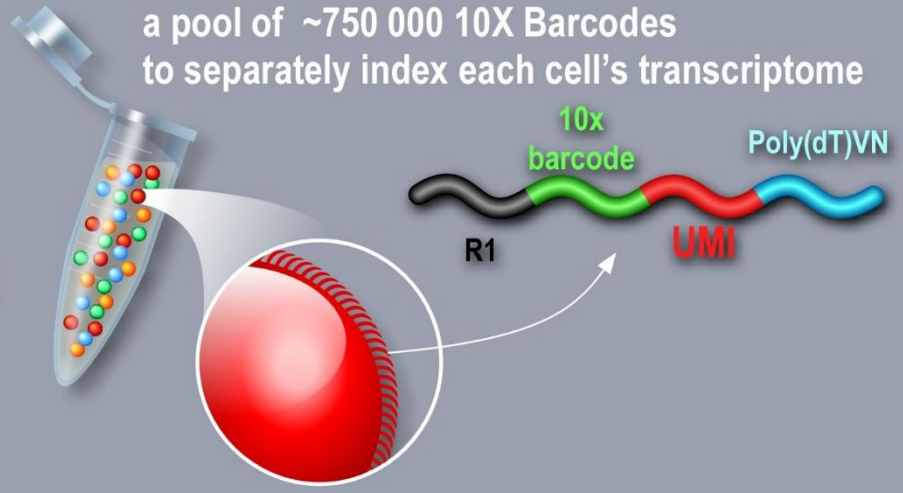




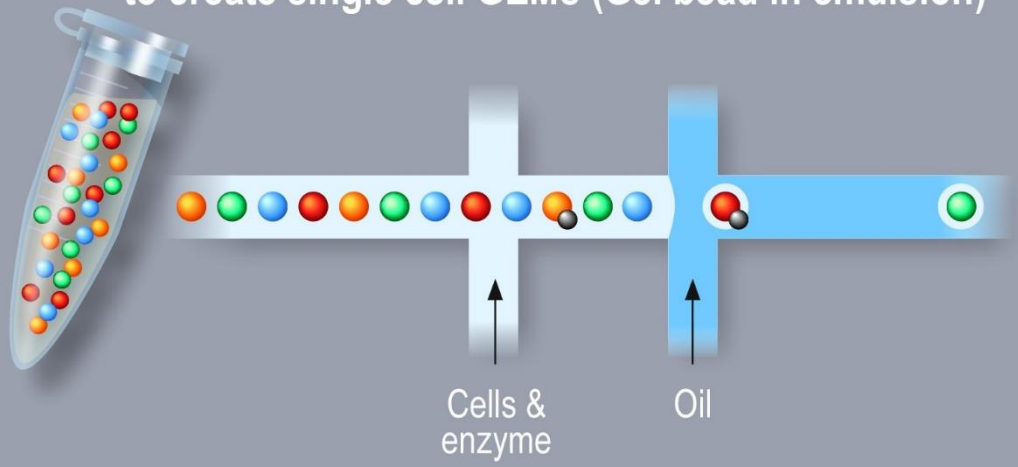
10X Genomics Chromium

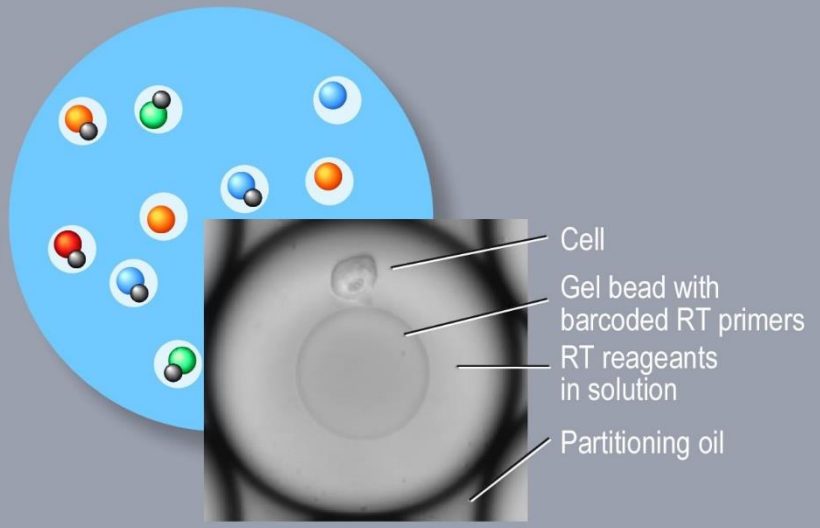
Single-use microfluidics chip

10X GemCode Technology samples
a pool of ~750 000 10X Barcodes
to separately index each cell's transcriptome



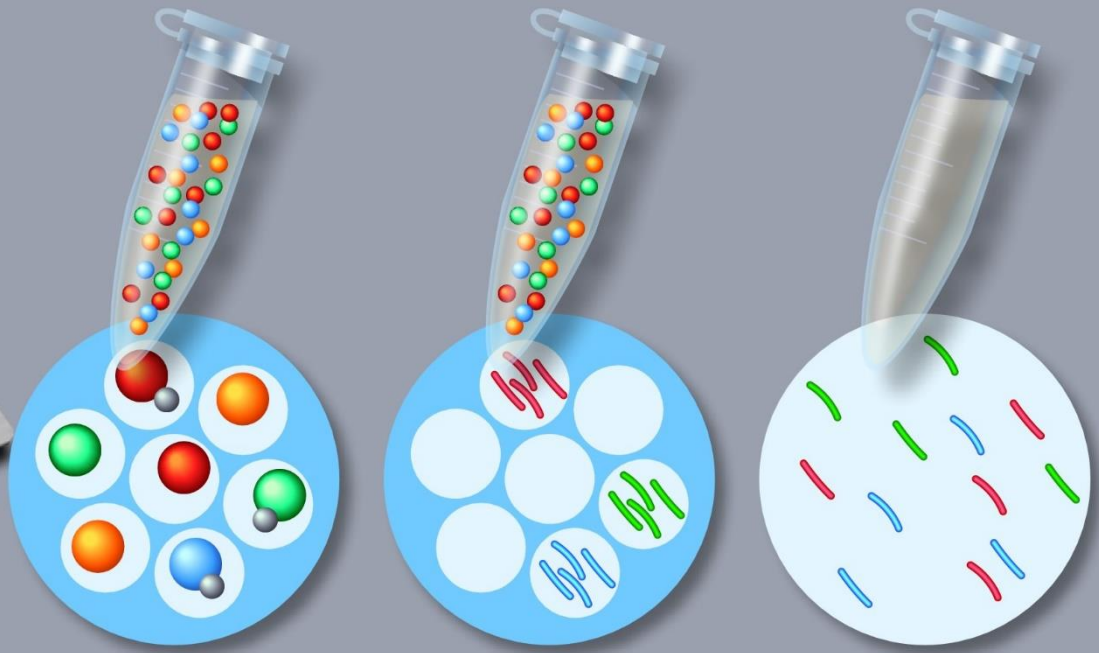
10X Barcoded Gel Beads are mixed with cells, enzyme and oil
to create single cell GEMs (Gel bead in emulsion)

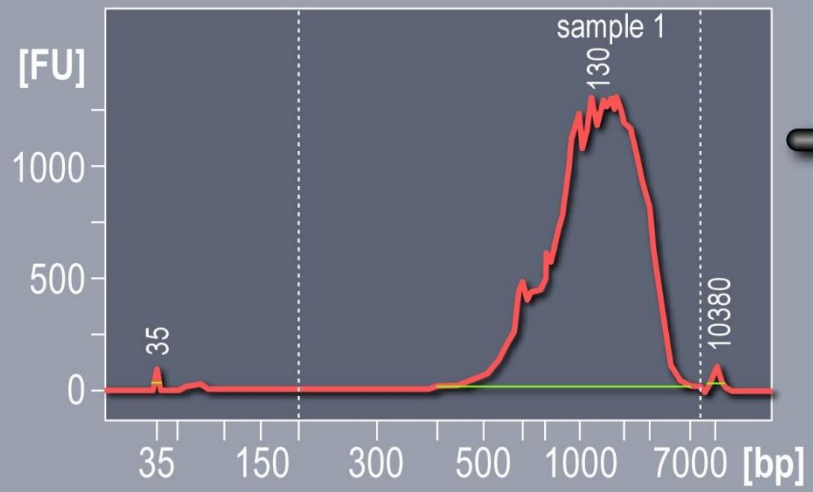




GEMs recovery and cDNA amplification

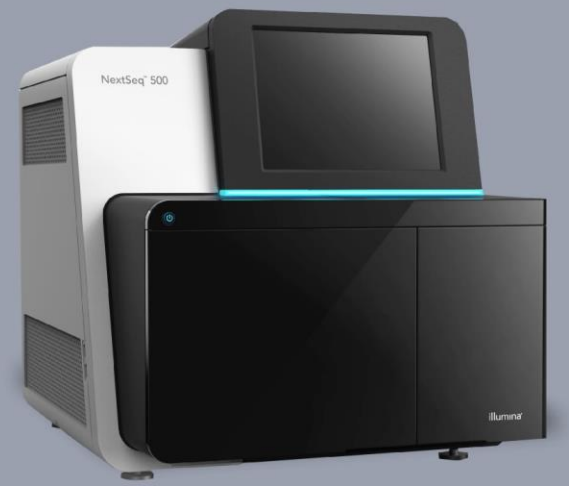
GEM recovery RT
cDNA Amplification
QC and Quantification

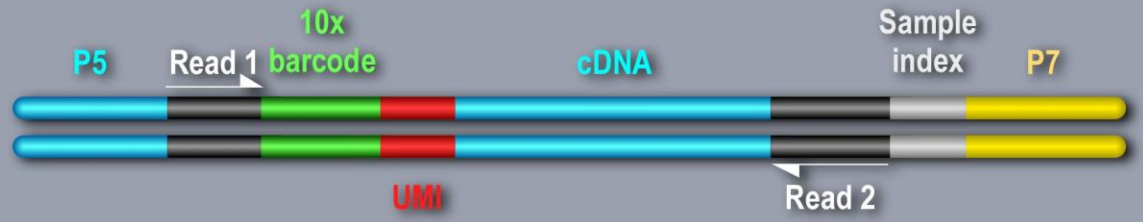




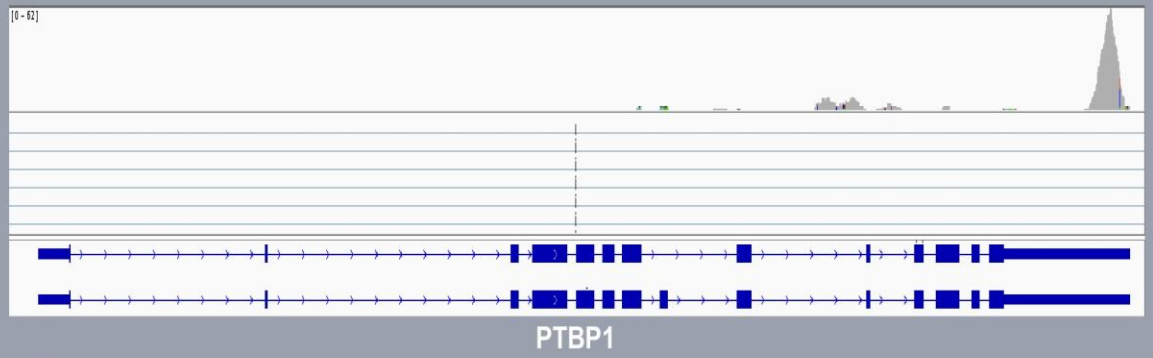
Library construction and Illumina sequencing

Fragment, End Repair and A-tailing
Adaptor Ligation
Sample Index PCR
QC and qPCR quantification





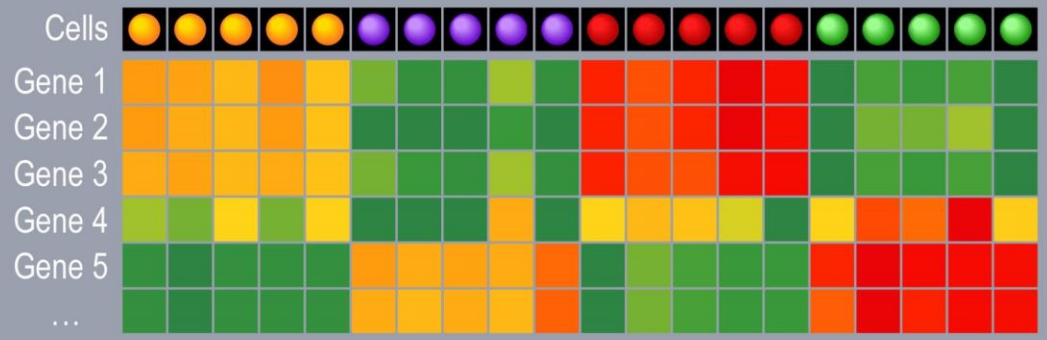
Sequencing read	Number of cycles
Read 1	26 cycles
i7 index	8 cycles
i5 index	0 cycles
Read 2	57 cycles



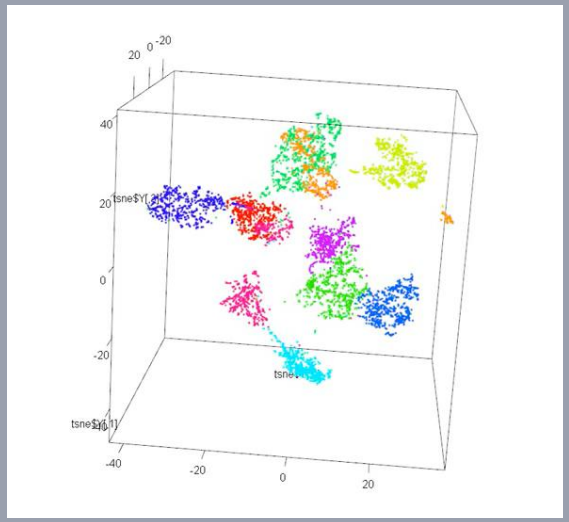
➡ Production of a 3p gene counting expression matrix

Single Cell statistical analysis and classification

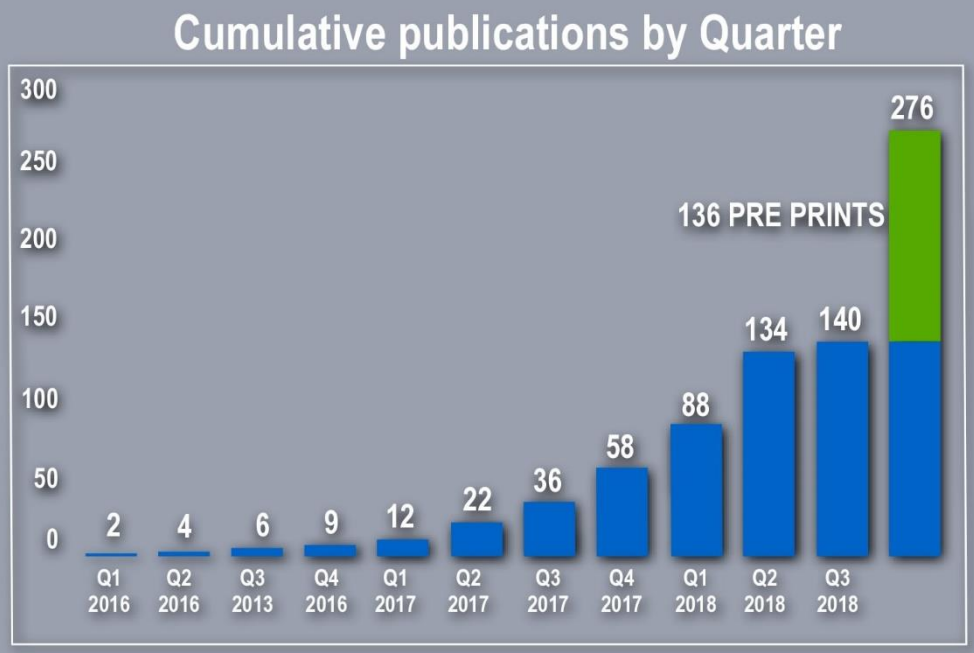
- Normalization to the Median UMI counts per cell (scaling factor)
- Matrix is log-transformed, centered and scaled per-gene (mean=0, SD=1)
- PCA analysis based on the most variables 1.000 genes
- t-SNE analysis based on 1st 10 components of the PCA-projected matrix
- k-means clustering (K=2..10) on 1st 10 components of the PCA-projected matrix
- Maximum of silhouette score as default K
- Differential expression analysis between clusters (genes markers)



3p gene counting expression matrix



10xGenomics Chromium publications



Technology which have a broad impact on many diverse fields of biology, including microbiology, neurobiology, development, tissue mosaicism, immunology, and cancer research

Experimental approaches comparison

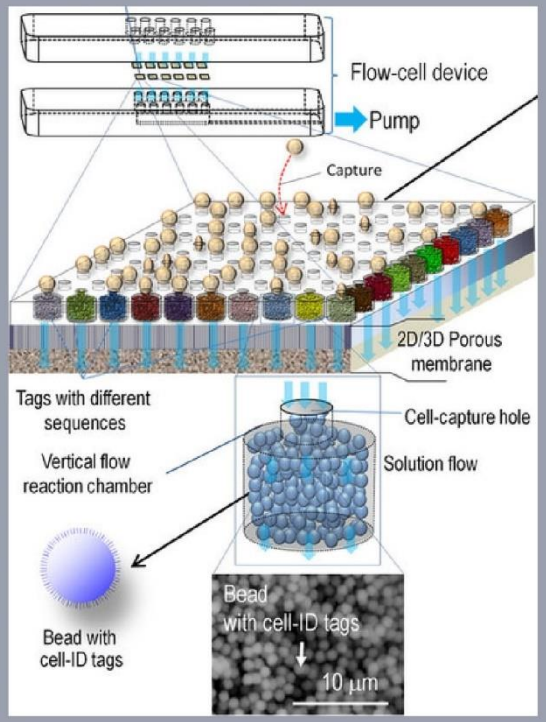
Lab (Ref) Name of approach	Region sequenced	UMI		Capture efficiency (ERCC)	Amplification	Barcoding	Device	Sequencer
		Design	Counting / error correction					
This approach	5'	N4H4	Unique (a) / edit distance (b)	26%	PCR	PCR pre-fragmentation	C1	Illumina
Linnarsson (10) STRT-seq	5'	N5	Unique start (c) / percentile (d)	48% (e)	PCR	Tagmentation *custom barcoded transposons*	C1	Illumina
Linnarsson (7) STRT-seq			Unique start (c) / percentile + (df	22% (b)				
Amit (13) Mars-seq	Close to 3'	N4	edit distance (b)	2 – 3%	Isothermal	RT	Plates	Illumina
McCarol (16) Drop-seq	Close to 3'	N8	edit distance (b)	12.8%	PCR	RT	Microdroplets	Illumina
Kirschner (17) InDrop	Close to 3'	N6	(g)	7.1%	Isothermal	RT	Microdroplets	Illumina
Yanai (15) Cel-seq	Close to 3'	No	none	6%	Isothermal	RT	Plates	Illumina
Yanai (15) Cel-seq2	Close to 3'	N6	none	22% in C1	Isothermal	RT	Plates, C1	Illumina
Smartseq2	Internal (a)	No	–	n.d.	PCR	–	Tubes	Illumina

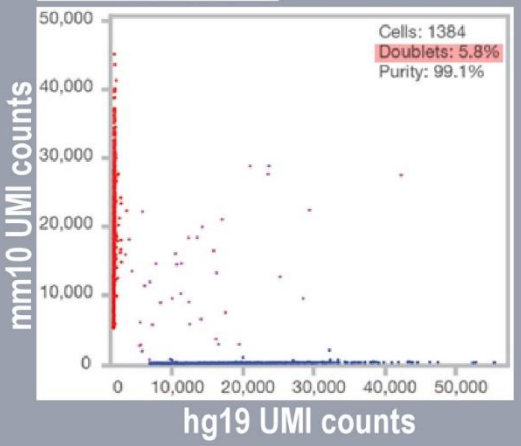
Arguel et al., NAR, 2016

Alternative options for high throughput single cell RNA-seq

Companies systems

- **Dolomite Bio:** Drop-seq setup
- **Hitachi:** Vertical Flow Array Chips (VFACs)
- **1CellBIO:** Kirschner's lab (*Klein et al.*) In-drop startup (Isothermal amplification)
- **Bio-Rad:** Illumina Bio-Rad SureCell™ WTA 3' Library

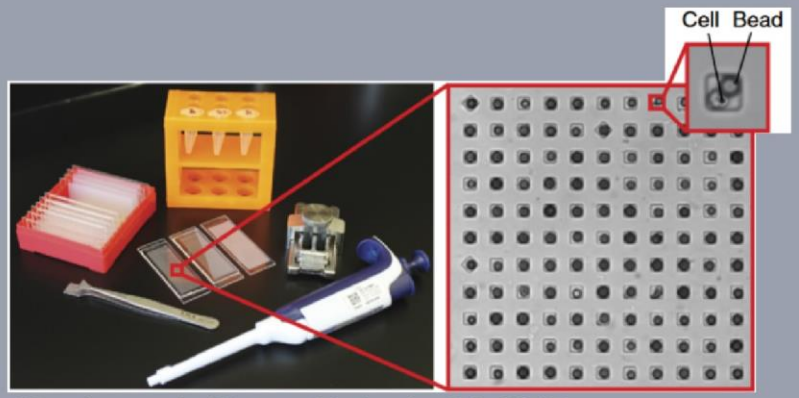




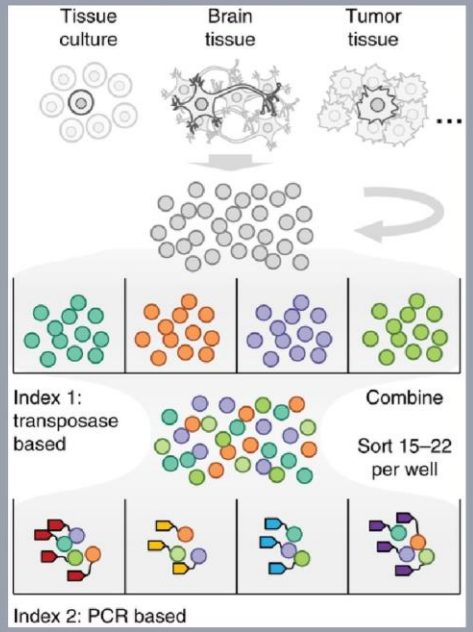
Alternative options for high throughput single cell RNA-seq

Wetlab protocol

- **Seq-Well:** portable and low cost scRNAseq in subnanoliter wells
- **SCi-seq:** combinatorial indexing (rounds of barcoding of pools of cells)



Gierahn et al., Nature Methods, feb.2017



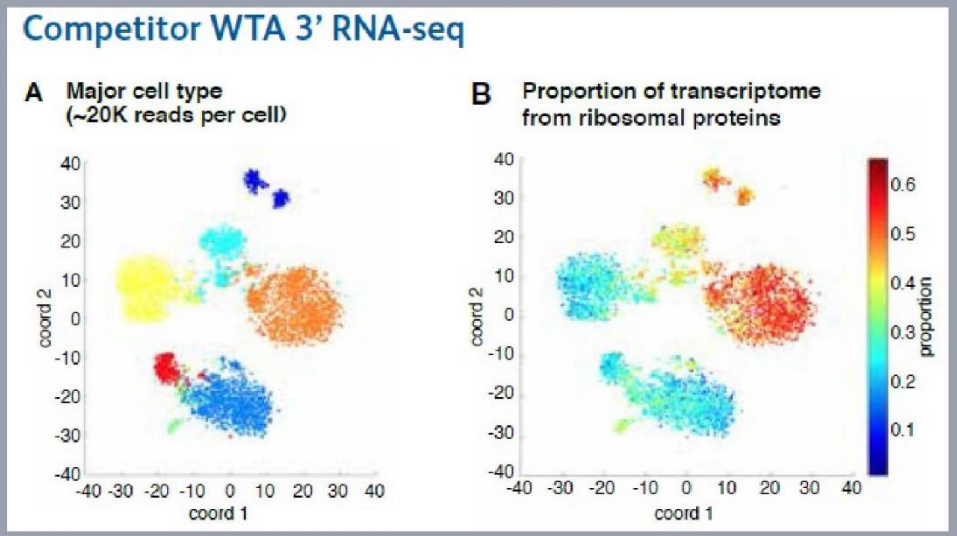
Vitak et al., Nature Methods, jan.2017

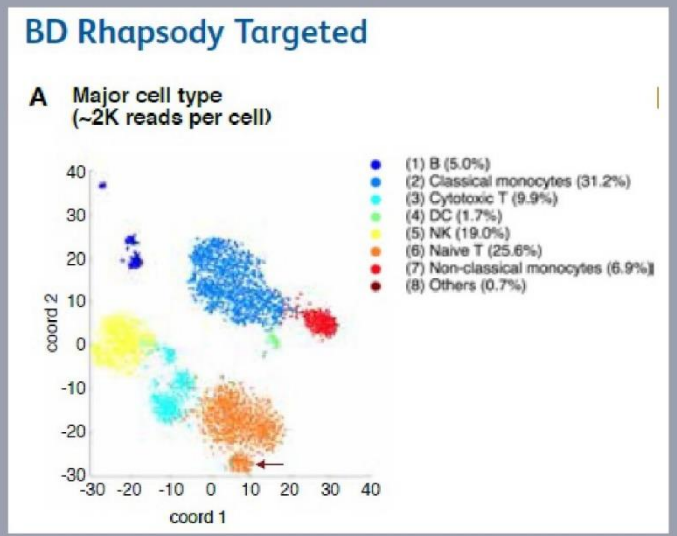
BD Rhapsody Single Cell Analysis System

Analyze 100's of genes across tens of 1000's of single cells

- microwells platform with barcoded beads and UMIs,
- 15.000 cells per sample,
- 99.4% count purity, minimal crosstalks between microwells,
- doublets rate close to 0% for 1.000 cells, under 5% for 15.000 cell,

- targeted assays with standard or custom gene panels:
  decrease cost of sequencing (2k reads/ cell <=> 20k reads/ cells for 10x ?)



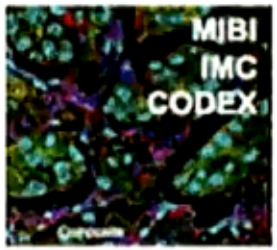


Approaches for Spatial Genomics

Direct Measurement

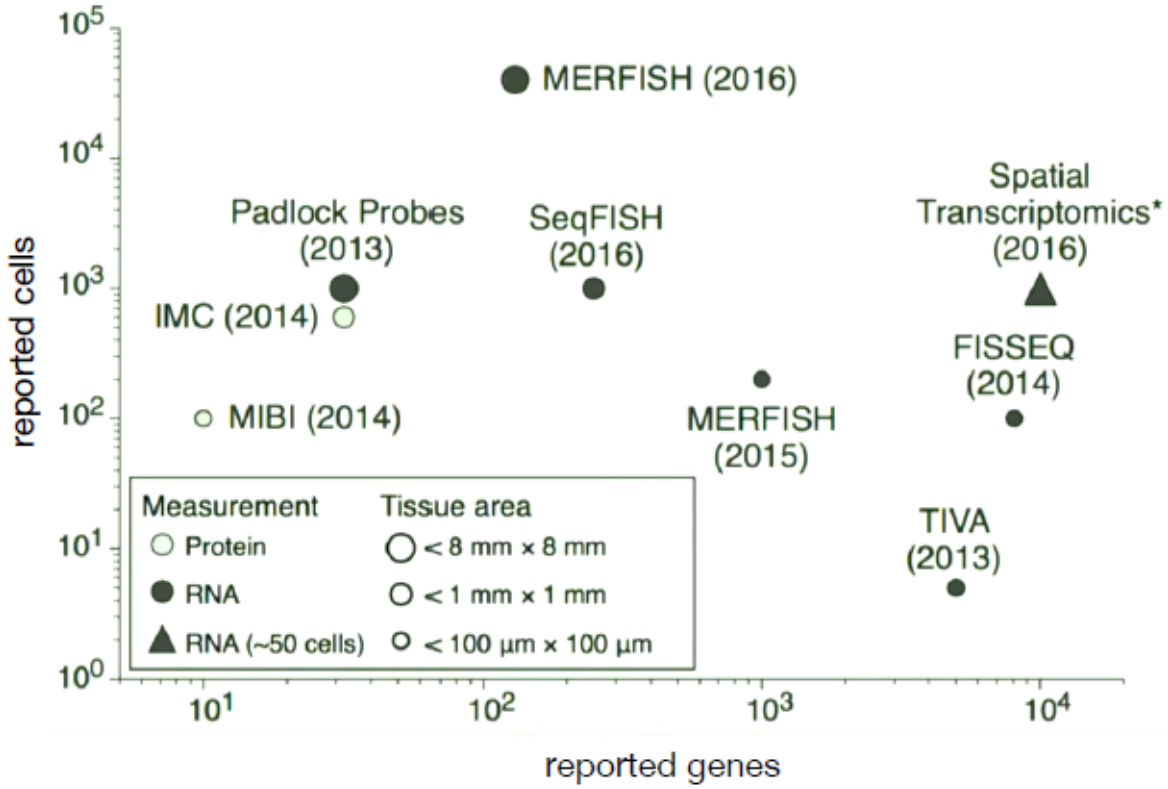
multiplex RNA fISH

in situ multi-plex protein localization



In situ sequencing

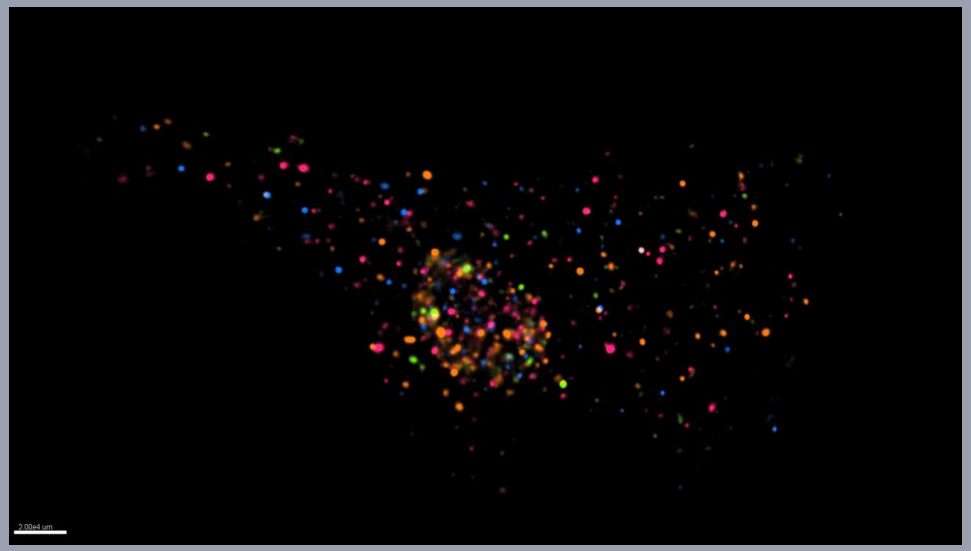
Spatial Technologies



FISSEQ, Lee et al. Science, 2014

- **Sequencing in situ (FISSEQ):** yields precise information on both cellular and subcellular localisation of transcripts.
  Lee at al., Science, Mar.2014

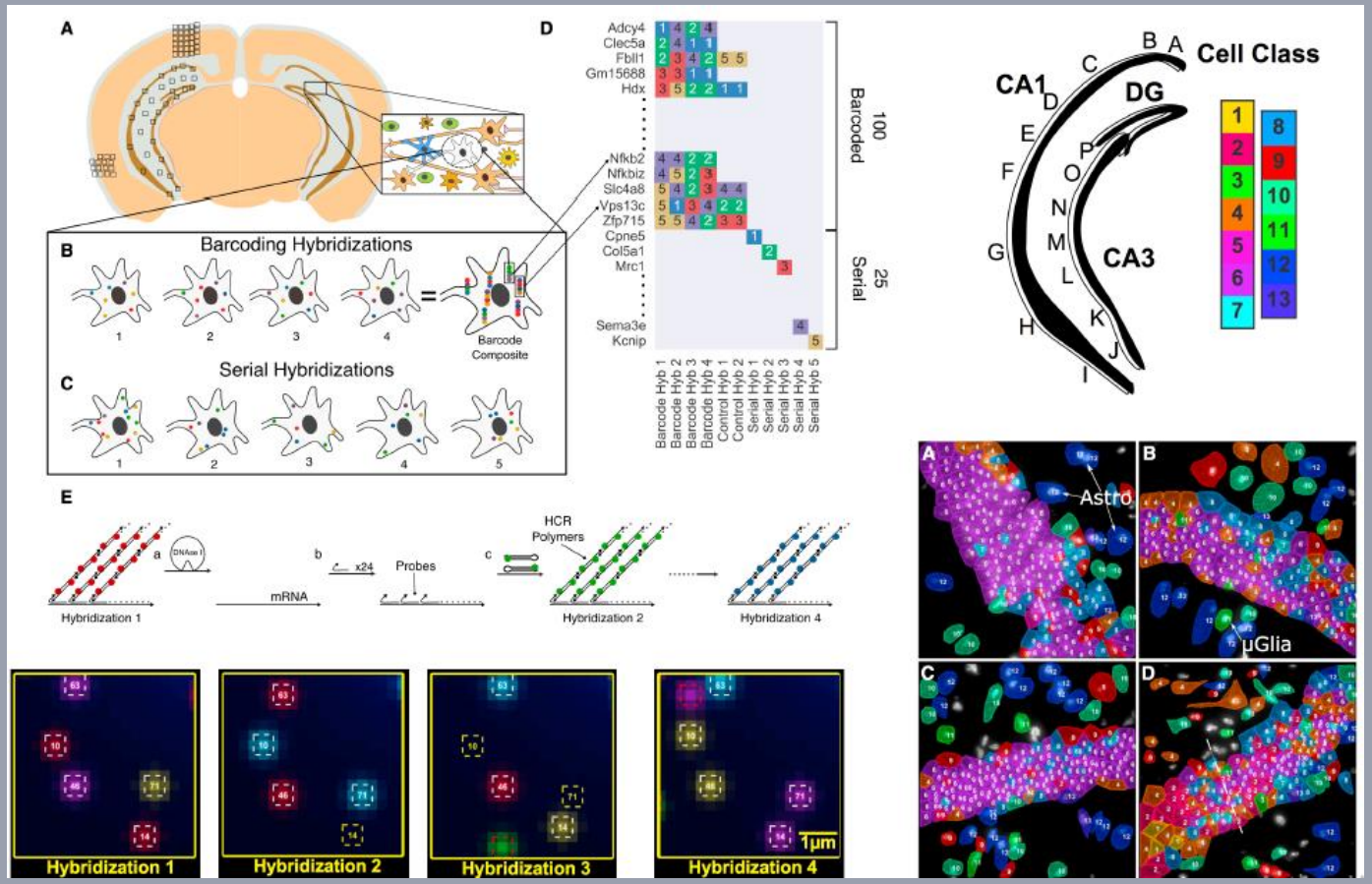
Cross-linked 3D RNA-seq library in cells



Based on **SOLID** chemistry and **sequencing by ligation**, you can get info on cellular and sub-cellular localisation of transcripts
2 days experiment for reads of 25 bases long

Amplified seqFISH, Shah et al. Neuron, 2016

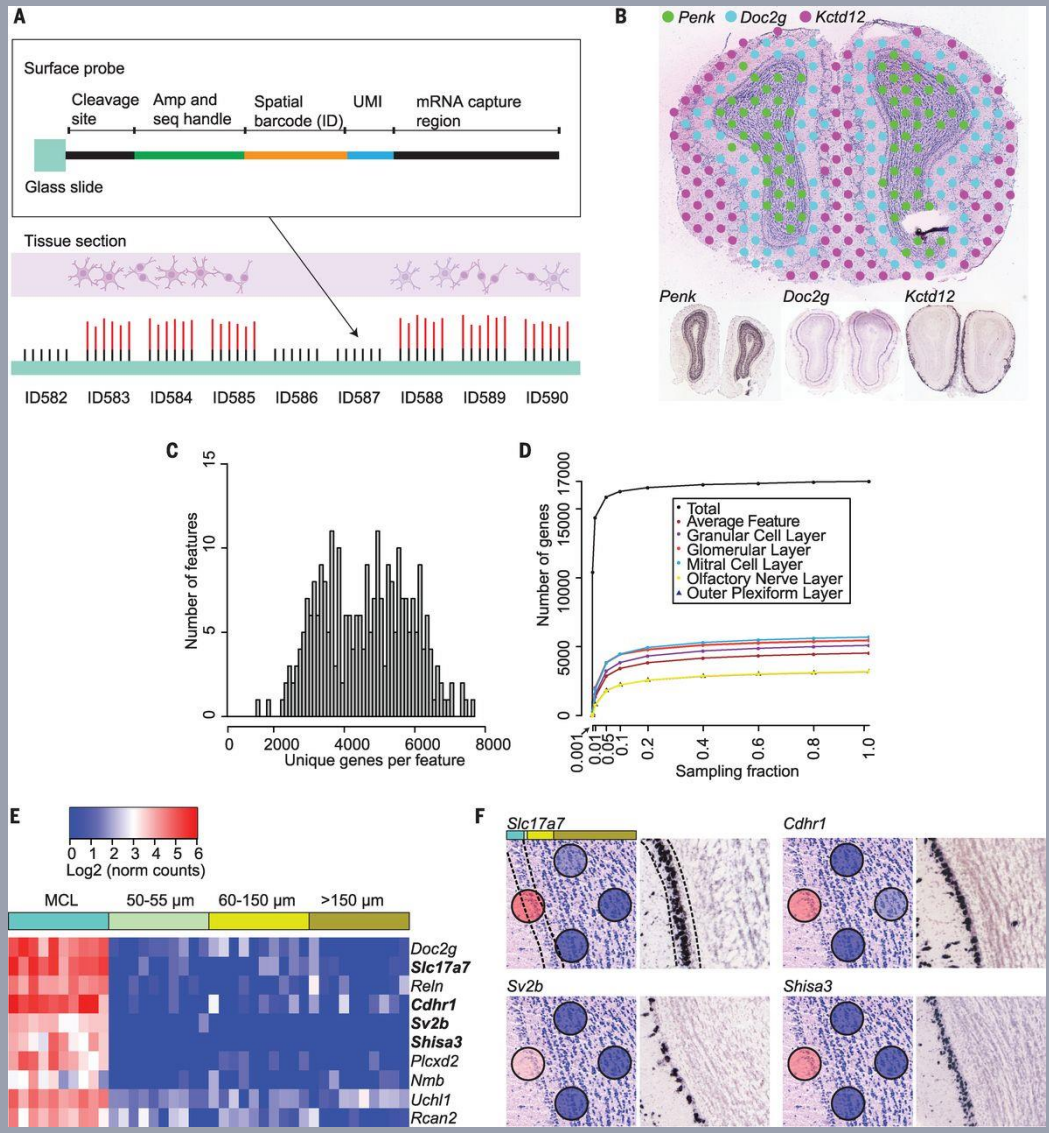
- **Highly multiplexed in-situ hybridization (Amplified seqFISH)**
  in situ 3D multiplexed imaging method



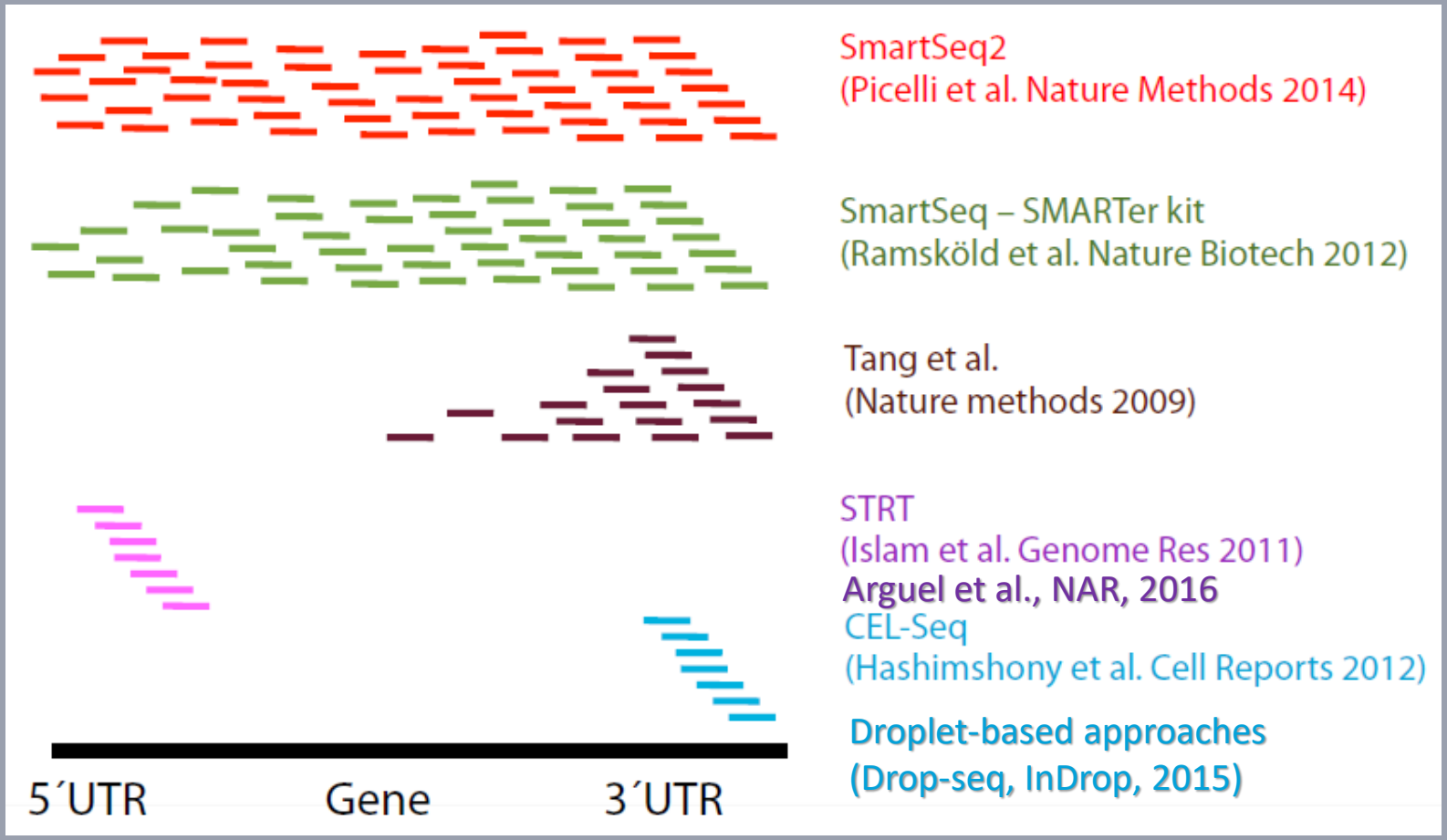
4 fluorophores and 4 rounds of hyridization can discriminate 4^4 genes = 256 genes

Spatial transcriptomics, Stahl et al. Science, 2016

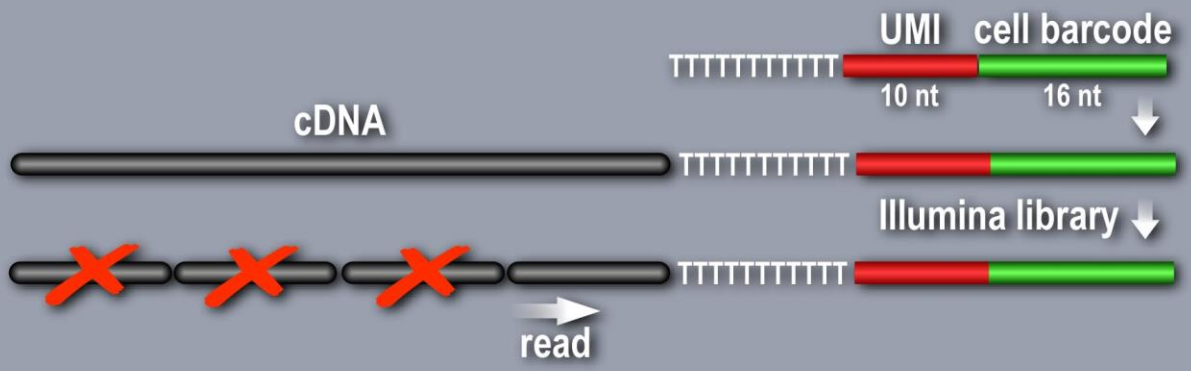
- **Spatial RNA-seq with position-indexed RT primer arrays on slides (spatial transcriptomics)**
  200 million oligos, 1007 features (100µm), ~5.000 genes per feature.
  Stahl et al., Science, Aug.2016



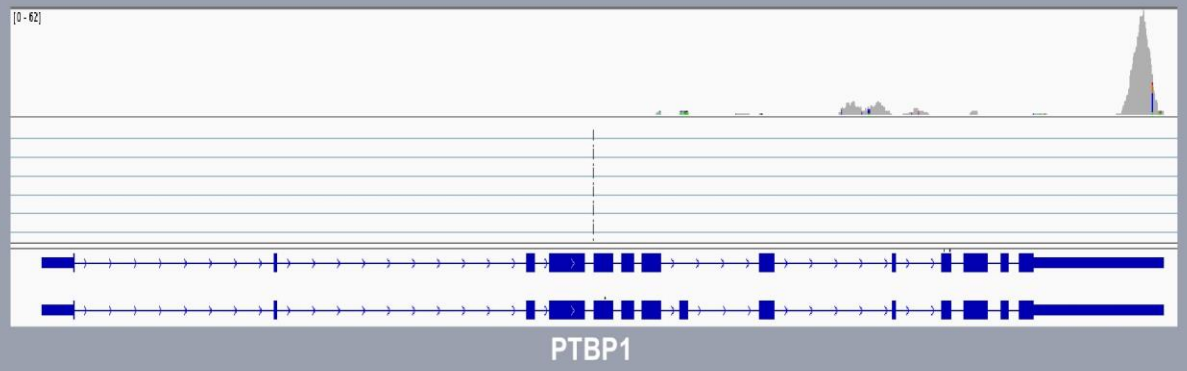
Single Cell RNA-seq methods



Short read single cell RNA-seq



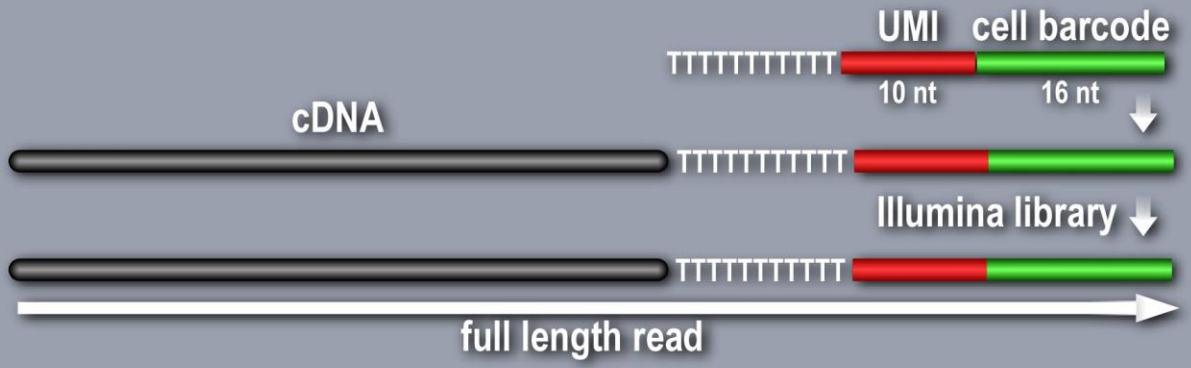
Reverse transcription, library preparation



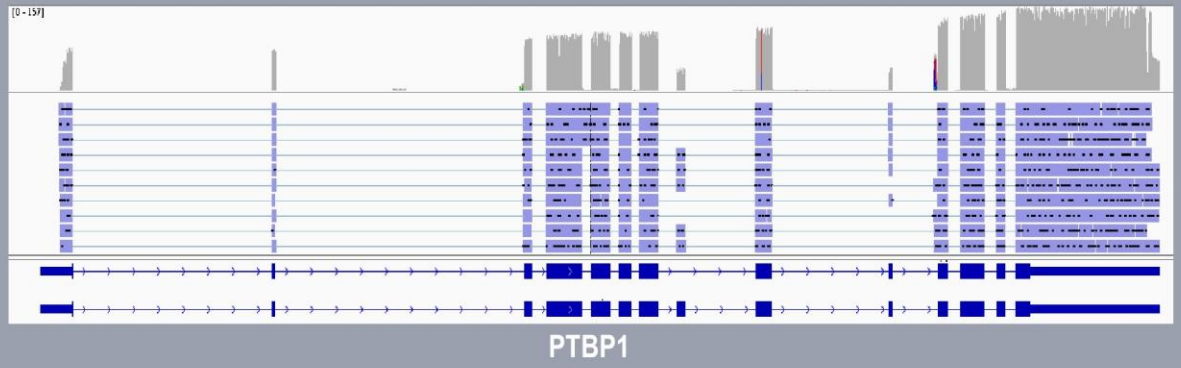
Illumina sequencing yields just a short read close to 3'

Informations on splicing, fusions, SNPs, editing, imprinting are lost

Long read single cell RNA-seq



Reverse transcription, library preparation



Informations on splicing, fusions, SNPs, editing, imprinting are preserved

February 2017, Linnarsson's lab

Karlsson and Linnarsson *BMC Genomics* (2017) 18:126
DOI 10.1186/s12864-017-3528-6

BMC Genomics

RESEARCH ARTICLE Open Access

Single-cell mRNA isoform diversity in the mouse brain

Kasper Karlsson[1] and Sten Linnarsson[2*]

Abstract

Background: Alternative mRNA isoform usage is an important source of protein diversity in mammalian cells. This phenomenon has been extensively studied in bulk tissues, however, it remains unclear how this diversity is reflected in single cells.

Results: Here we use long-read sequencing technology combined with unique molecular identifiers (UMIs) to reveal patterns of alternative full-length isoform expression in single cells from the mouse brain. We found a surprising amount of isoform diversity, even after applying a conservative definition of what constitutes an isoform. Genes tend to have one or a few isoforms highly expressed and a larger number of isoforms expressed at a low level. However, for many genes, nearly every sequenced mRNA molecule was unique, and many events affected coding regions suggesting previously unknown protein diversity in single cells. Exon junctions in coding regions were less prone to splicing errors than those in non-coding regions, indicating purifying selection on splice donor and acceptor efficiency.

Conclusions: Our findings indicate that mRNA isoform diversity is an important source of biological variability also in single cells.

Keywords: Alternative isoform usage, Single-cell RNA sequencing, STRT, PacBio, Long read sequencing, UMI, Oligodendrocytes

Feb.2017:

We selected six single cells for which cDNA was available from an earlier experiment … was used for PacBio sequencing. The cDNA had been produced with the STRT method adapted to the Fluidigm C1 instrument for single cell RNA sequencing

July 2017, Vollmer's lab

Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells

Ashley Byrne, Anna E. Beaudin, Hugh E. Olsen, Miten Jain, Charles Cole, Theron Palmer, Rebecca M. DuBois, E. Camilla Forsberg, Mark Akeson & Christopher Vollmers

Nature Communications **8,**
Article number: 16027 (2017)
doi:10.1038/ncomms16027

Received: 24 April 2017
Accepted: 23 May 2017
Published: 19 July 2017

Understanding gene regulation and function requires a genome-wide method capable of capturing both gene expression levels and isoform diversity at the single-cell level. Short-read RNAseq is limited in its ability to resolve complex isoforms because it fails to sequence full-length cDNA copies of RNA molecules. Here, we investigate whether RNAseq using the long-read single-molecule Oxford Nanopore MinION sequencer is able to identify and quantify complex isoforms without sacrificing accurate gene expression quantification. After benchmarking our approach, we analyse individual murine B1a cells using a custom multiplexing strategy. We identify thousands of unannotated transcription start and end sites, as well as hundreds of alternative splicing events in these B1a cells. We also identify hundreds of genes expressed across B1a cells that display multiple complex isoforms, including several B cell-specific surface receptors. Our results show that we can identify and quantify complex isoforms at the single cell level.

July 2017:

To test this, we used our ONT RNAseq approach to analyse seven individual mouse B1a cells and compared it with the standard Illumina RNAseq approach. To this end, we FACS-sorted single B1a cells into individual wells containing lysis buffer and amplified cDNA from each individual cell using a modified Smartseq2

Juin 2018:

R2C2: Improving nanopore read accuracy enables the sequencing of highly-multiplexed full-length single-cell cDNA , Volden et al., Vollmers's lab, bioRxiv, 2018

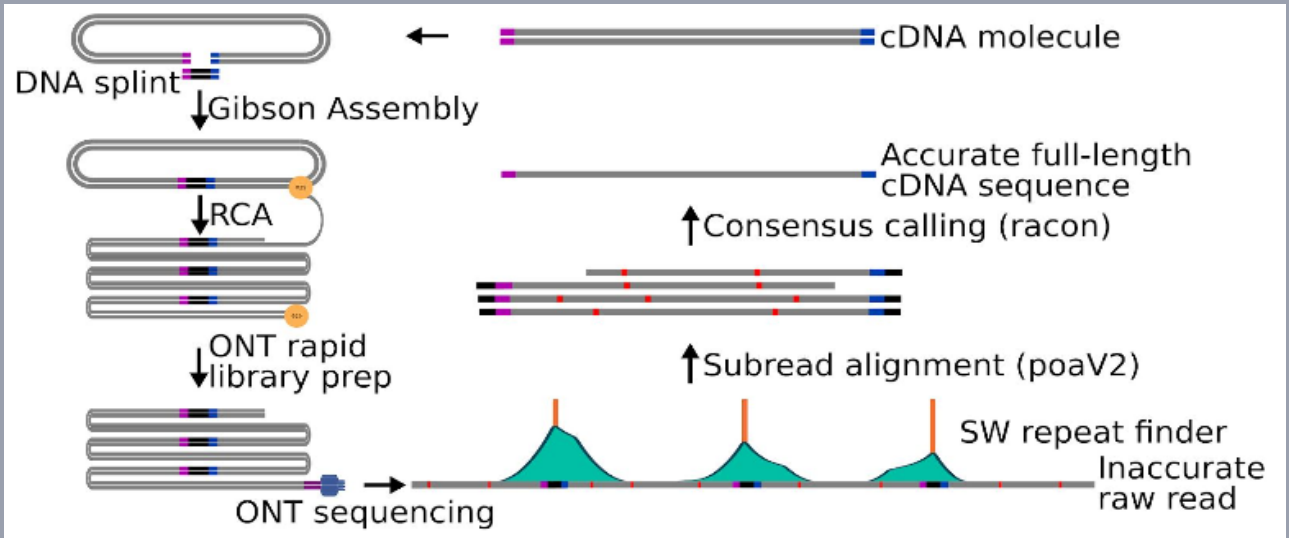


Fig. 1: R2C2 method overview. A) *cDNA is circularized using Gibson Assembly, amplified using RCA, and sequenced using the ONT MinION. The resulting raw reads are split into subreads containing full-length or partial cDNA sequences, which are combined into an accurate consensus sequences using our C3POa workflow which relies on a custom algorithm to detect DNA splints as well as poaV2 and racon.*

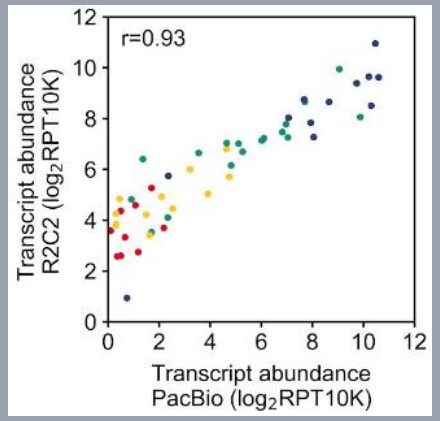
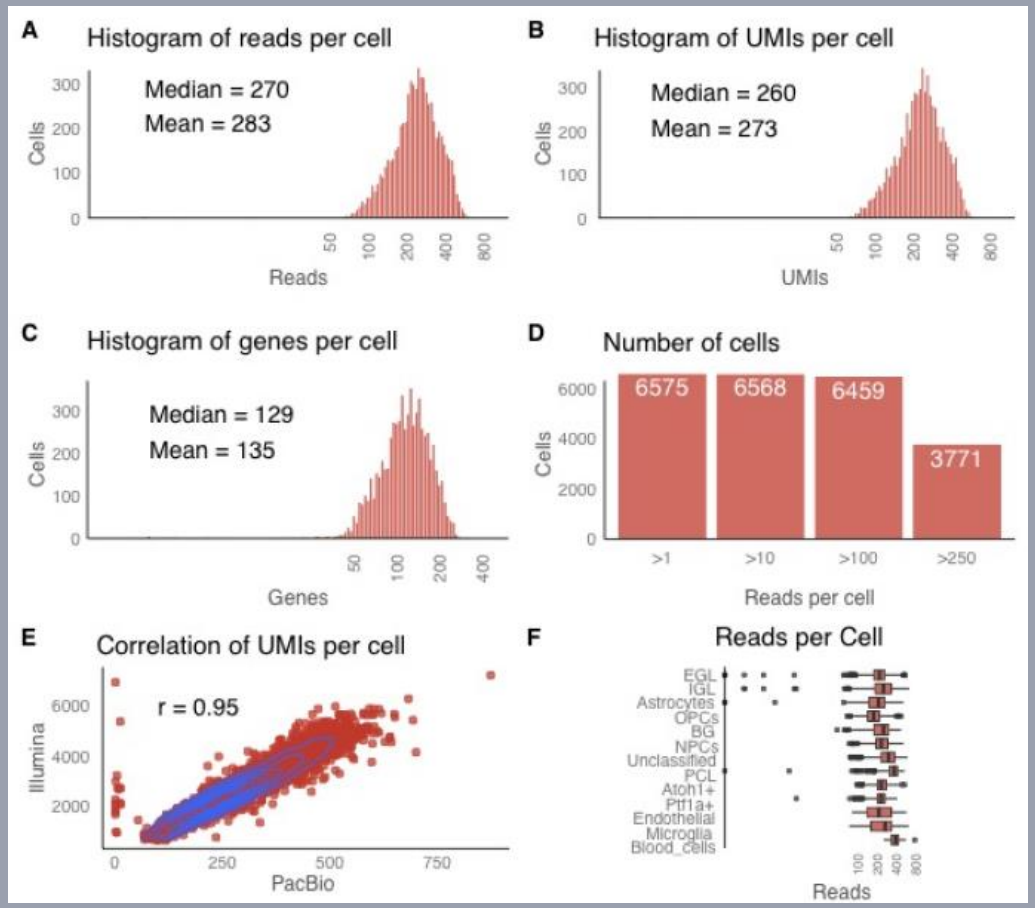


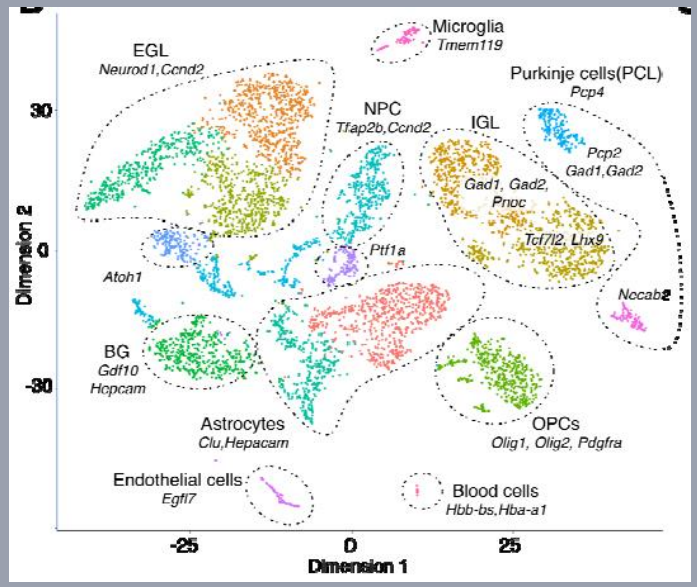
Plate of 96 B cells, Smart-seq2 modified protocol with Rolling Circle Amplification (RCA) and sequence with 1D R9.5 ONT flowcells. Good correlation with PacBio profiling, R2C2 can be easily adapted to any RNAseq library preparation protocol (10x, smart-seq2, drop-seq)

October 2018, Tilger's Lab, Nature Biotechnology, ScISOr-Seq

Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells

Ishaan Gupta, Paul G Collier, Bettina Haase, Ahmed Mahfouz, Anoushka Joglekar, Taylor Floyd, Frank Koopmans, Ben Barres, August B Smit, Steven A Sloan, Wenjie Luo, Olivier Fedrigo, M Elizabeth Ross & Hagen U Tilgner





Oct.2018:

6.627 mouse P1 cerebellum cells

10xGenomics sample

23 SMRT PacBIo flowcells

5.2M reads generated

260 UMIs per cell out of the

3.875 UMIs identified in Illumina.

Options for full length single cell transcriptome sequencing

PacBio Sequel

Oxford Nanopore Technology

+++ higher accuracy	+++ high throughput (PromethION: > 60 * 10^6 reads)
--- low throughput (< 400.000 reads / SMRT)	--- lower accuracy

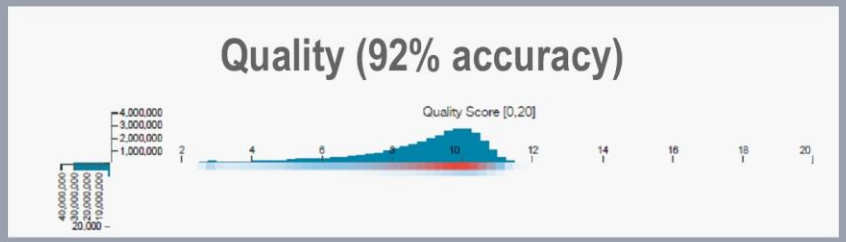
2 challenges to tackle

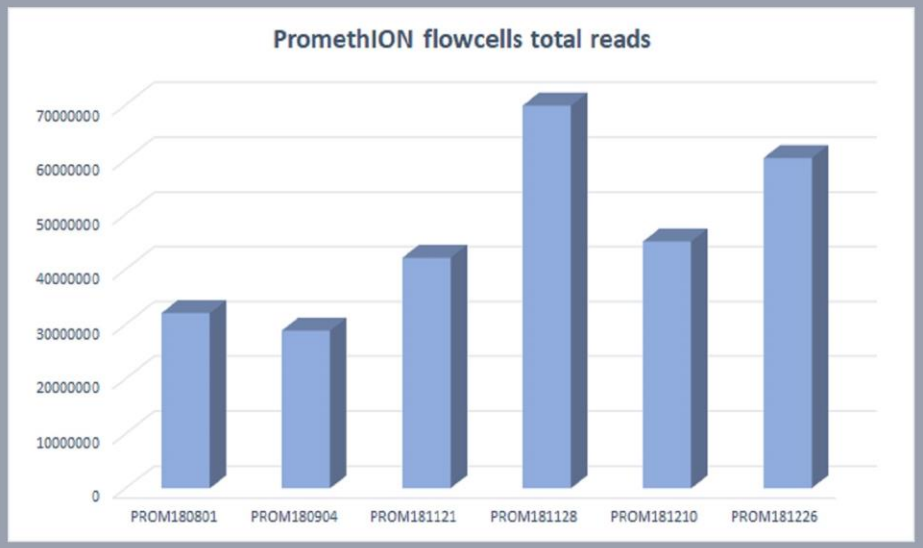
(1) get enough reads to sequence all molecules (50k reads / cell)

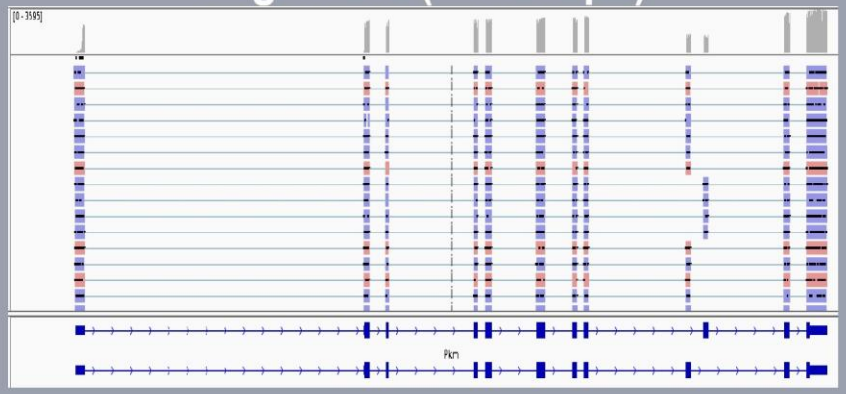
(2) high accuracy for cell barcode and UMI identification

Challenge 1: get enought reads to sequence all molecules

PromethION flowcells





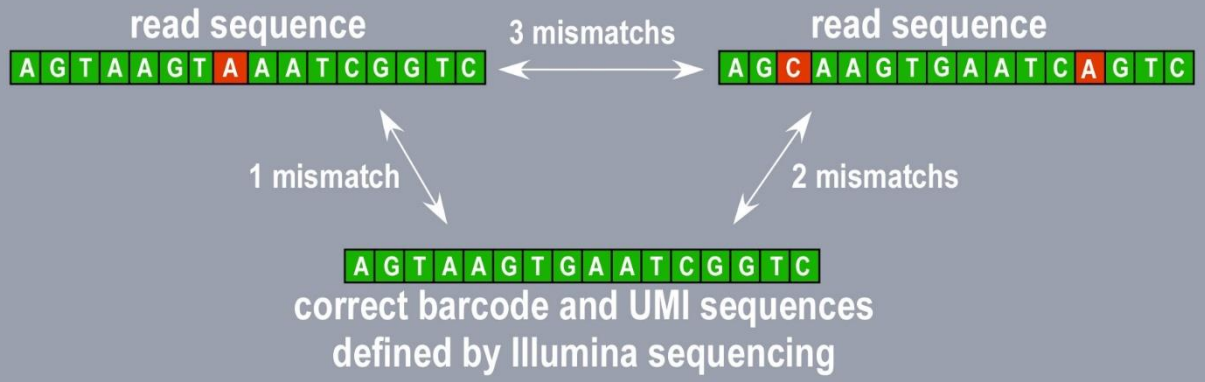


Challenge 2: cell barcode and UMI identification in Nanopore reads

- Cell barcodes (16nt) are randomly selected from a pool of a million barcodes
- UMIs (10nt) are totally random sequences

@d6e7c3a3-2ec4-4ead-b876-f24622460270 sampleid=951_nonSized read=10 ch=55 start_time=2018-11-21T13:23:07Z
CGCTGAGCTATCTTGCCCTGAGGGCCCATACCTAGCTCACTGTCAGTCTTGTTCCATCCTGTCCTGAGGGCCCCCACTCTGTCTCCTCTGCTCTTTCTAATAAAACAGCAG
TTGCACC**AAAAAAAAAAAAAAAAAAAAAA**TTAGCAGCACGAATACGGTAGC**AGATCGGAAG**AGCGTCGTGGTATCAGCACATGTAAATCGAACGAAGTGGCAATTGAT
...
@132849bf-70fe-4518-88e6-1ed3f4bdf4fc sampleid=951_nonSized read=12 ch=974 start_time=2018-11-21T13:23:07Z
CTACACCTGAGCCGAGCACTGGCCTGCGGGTCTCTAAGGACCTCCACTAATCTGGATCGCTGGAATTCACTGGTCGCCCTGGAGACAGAAAATTATTTGTAAGGATTGAT
GAAATAAAATACCTGC**AAAAAAAAAAAAAAAGAAAAA**TCTGGTTCATTCAATTGTAATGA**AGATCGGAAG**AGGCCGGTAACAGCACATGTAACTGAACGAAGTACAATTGA
...

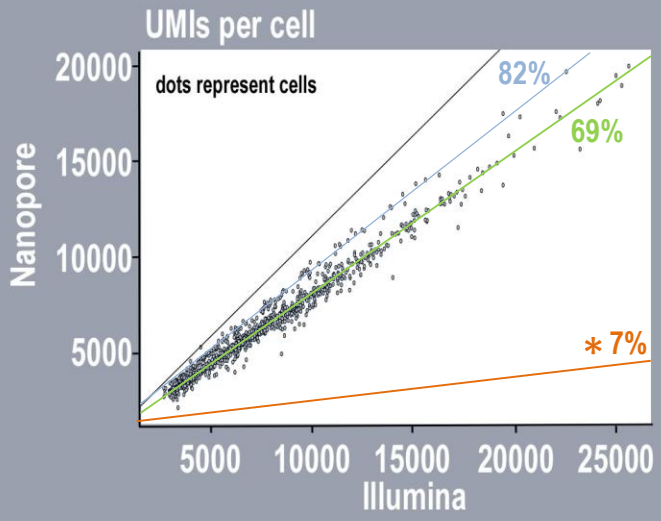
UMI and barcode assignment is challenging and error prone without knowledge of correct UMI and BC sequences

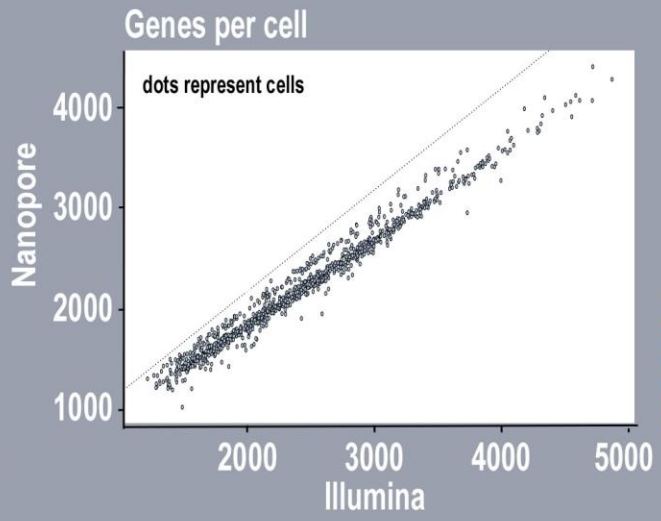


Computational challenge

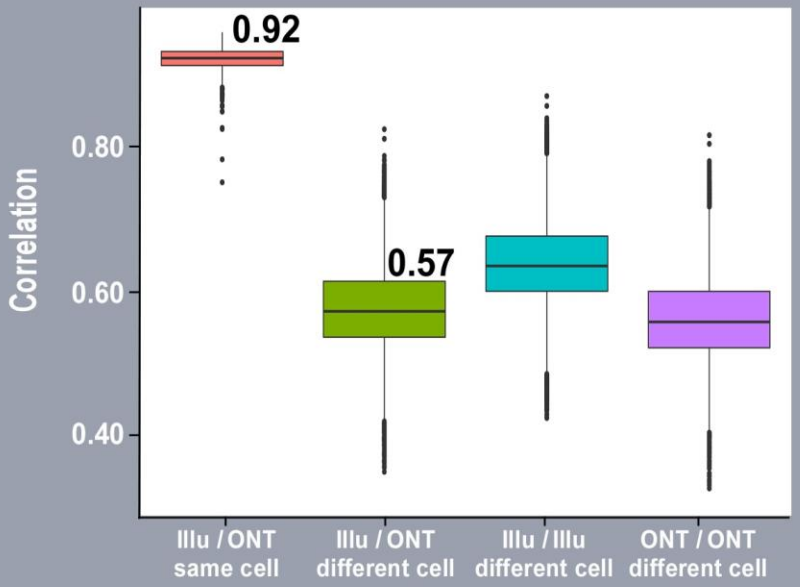
- Allowing 3 mutations in a 16 nt. barcode means generation of > 2 million variants per barcode
- For a 50 million read Promethion run > 10^14 barcode variants need to be generated
- Use exclusively CPU efficient bitshift and bitwise “and” or “or” operations to generate mutations.

Illumina / Nanopore correlation (190 + 951 cells)





* Gupta et al. Nature Biotechnology (2018)
6.627 cells mouse cerebelum post-natal day (P1), 5.2M PacBio reads (23 SMRT cells)
260 UMIs / cell, 129 Genes / cell

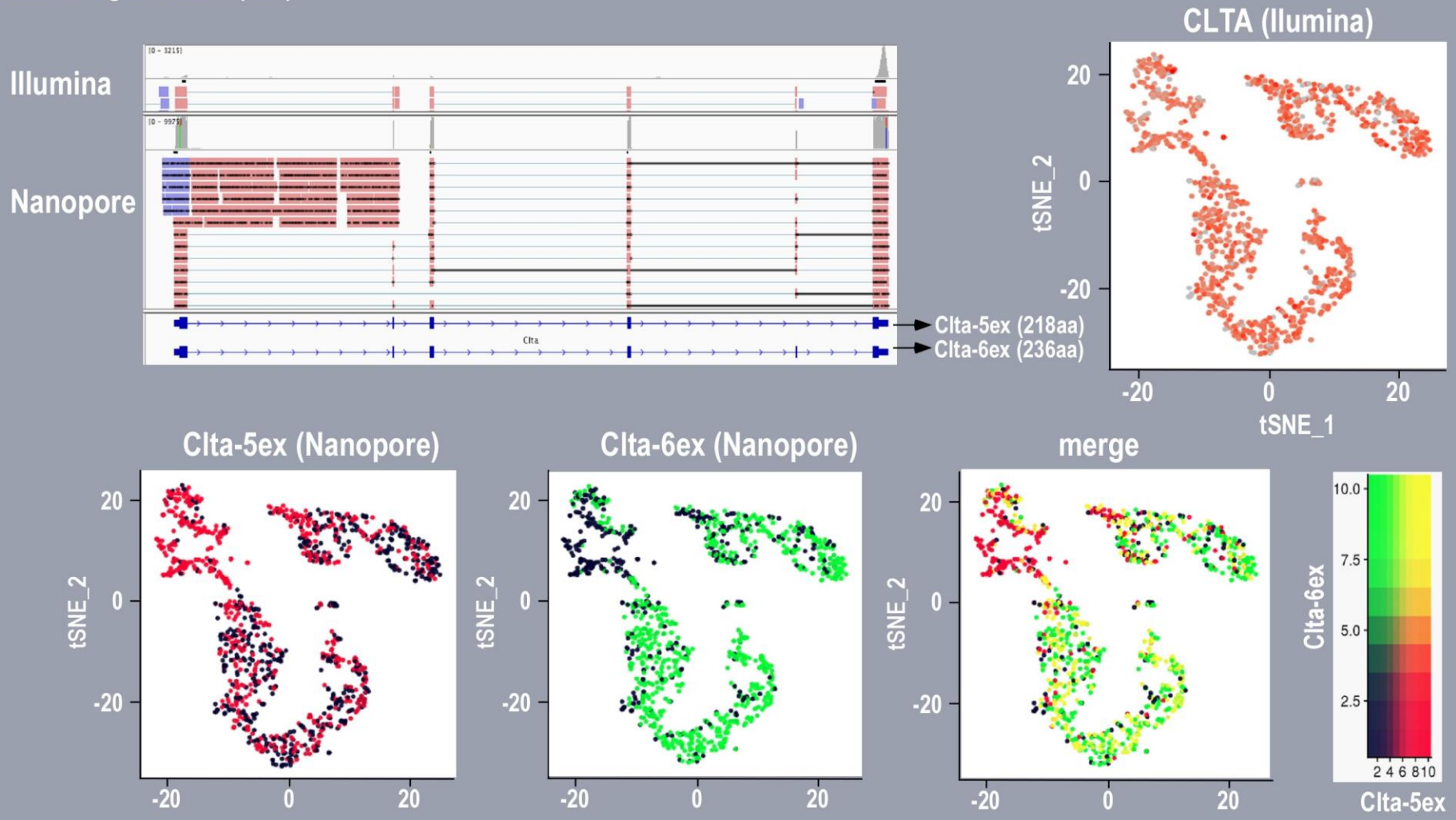


- Illumina and Nanopore data correlate well

Dataset	190c	951c	Gupta et al.
Flowcells	1	5	23
Reads	35M	246M	5.2M
Depth	82%	69%	7%

Long read sequencing reveals diversity

Clathrin Light Chain A (Clta)

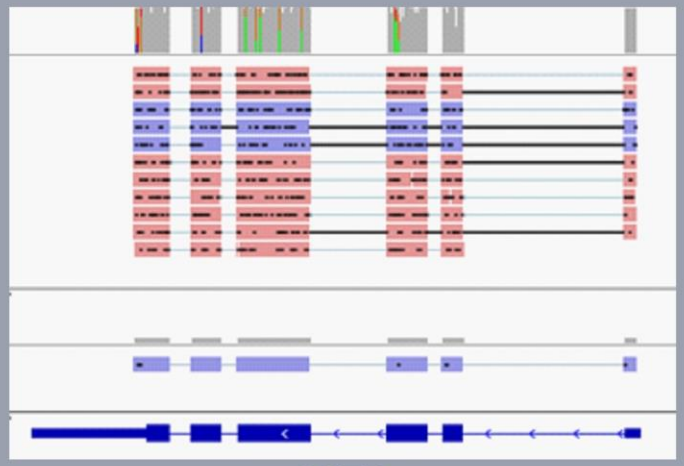


Multiple reads for the same molecule (UMI) allow errors correction

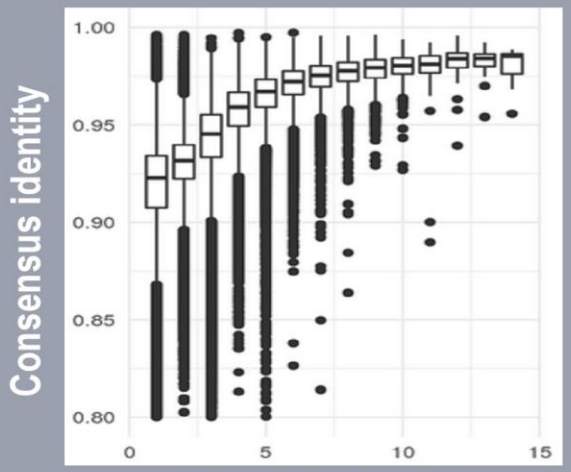
- Multiple alignment of reads to define consensus sequence (if >10 reads, best 10 reads used)
- racon polishing using all reads

Reads from the same ACTB molecule 92% identity each

Consensus sequence 99.2% identity



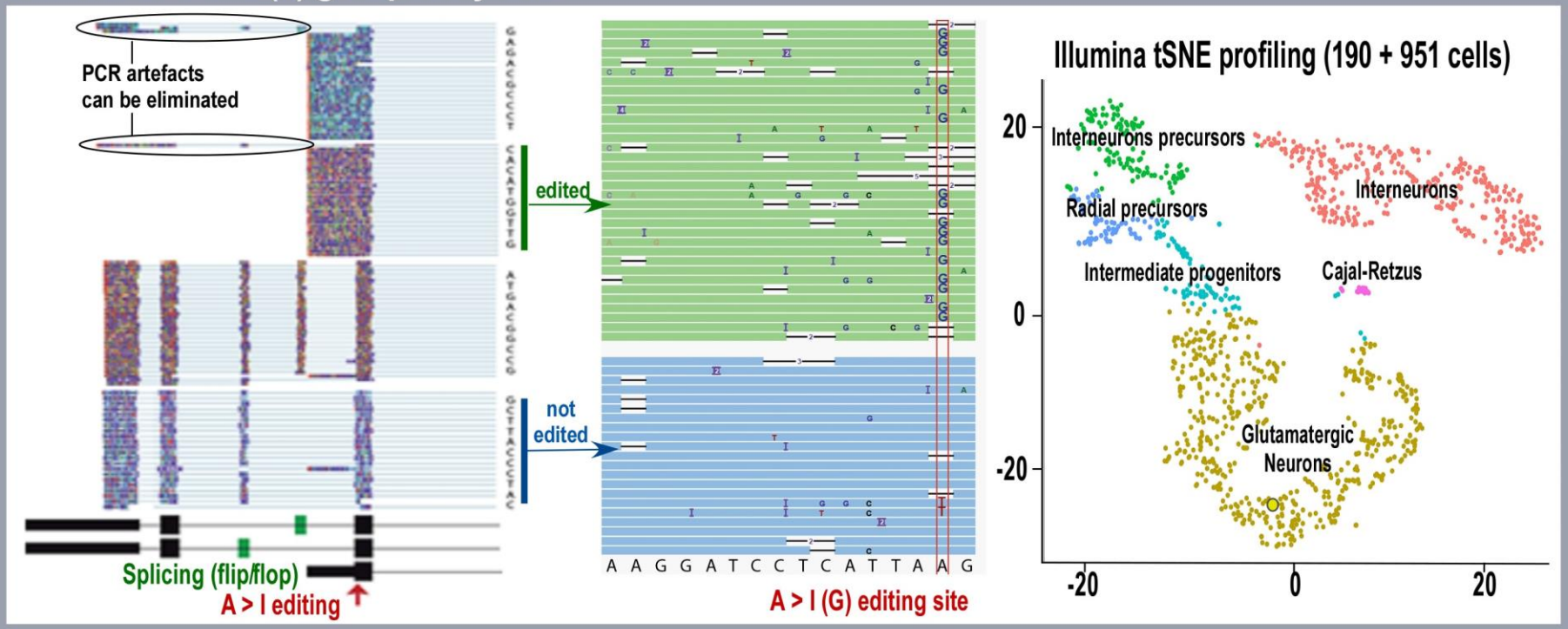
ACTB



- Analysis of raw signal instead of basecalled reads should yield better accuracy
- SNP calling for well sequenced molecules is possible

Enrichment of transcripts of interest - targeted sequencing

Data for one cell (●) grouped by UMI



Unique Molecular Identifier allow

- elimination of PCR artefacts,
- generation of a consensus sequence per molecule,
- SNPs, editing, imprinting calling is possible

Internal RT priming in certain long RNAs

Some weird cDNAs generated by priming on internal poly(A) rich regions