Transcriptomics / NGS lab methods
2025
Bastien Mangeat (GECF)

FYI = means for sure not in exam

Technological toolbox & application in transcriptomics

- nucleic acids QC
  - Nanodrop & qubit
  - capillary electrophoresis

- NGS
  - Historical perspective
  - Illumina short reads description, strengths, pitfalls…
  - Long reads (Oxford Nanopore, PacBio)
  - RNA-seq
    - mRNA-seq methods
    - Whole transcriptome methods
    - New developments (3’ end…)
    - Single cells RNA-seq

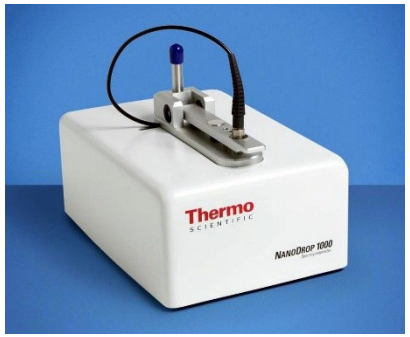
Transcriptomics, genomics, epigenomics:

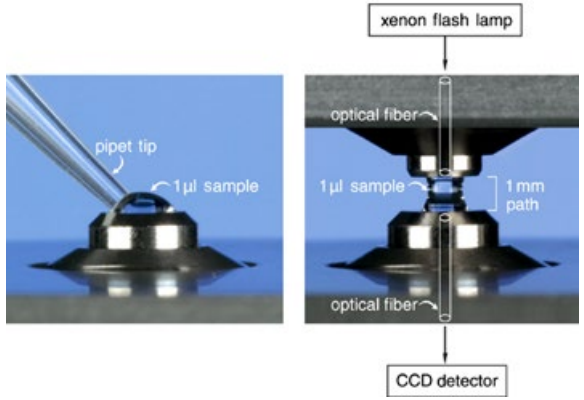
Fields of studies
And
Groups of methods
both intimally linked (parallel development)

Nucleic acids quantification

Spectrophotometer (nanodrop)

- Old technique.
- Molecules absorb light at specific wavelength (DNA, RNA, proteins, lipids…).
- Measures absorbance -> calculate concentration.





To distinguish RNA, DNA, ssDNA → shapes of curves
Also used to assess purity:

- 260/280 ratio 1.8 = pure dsDNA
- 260/280 ratio 2= pure RNA
- 260/230> 2 for pure RNA or DNA
- 320nm = 0

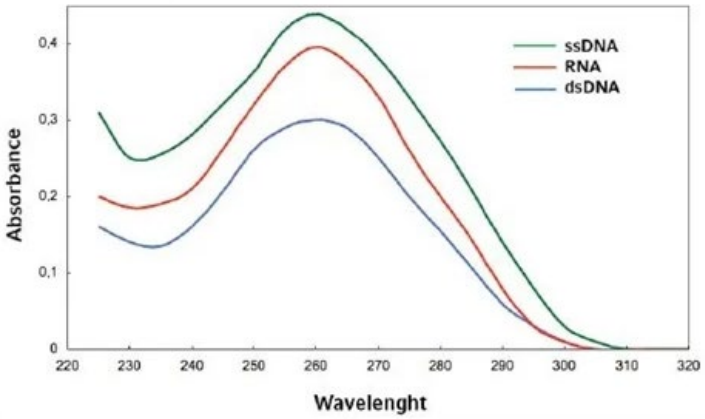
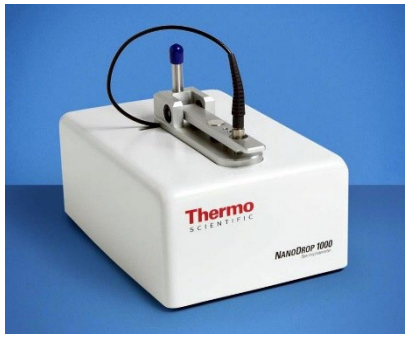


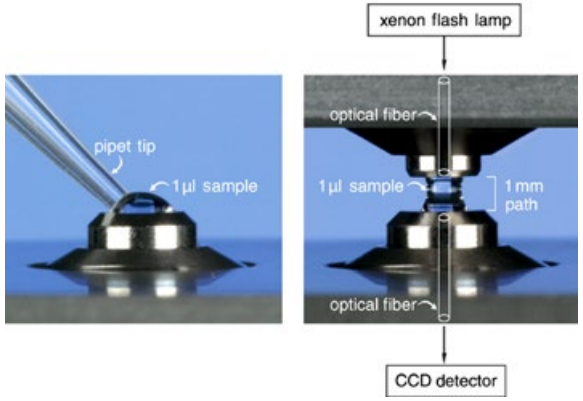
Image: Eppendorf

Nucleic acids quantification

Spectrophotometer (nanodrop)

- Old technique.
- Molecules absorb light at specific wavelength (DNA, RNA, proteins, lipids…).
- Measures absorbance -> calculate concentration.



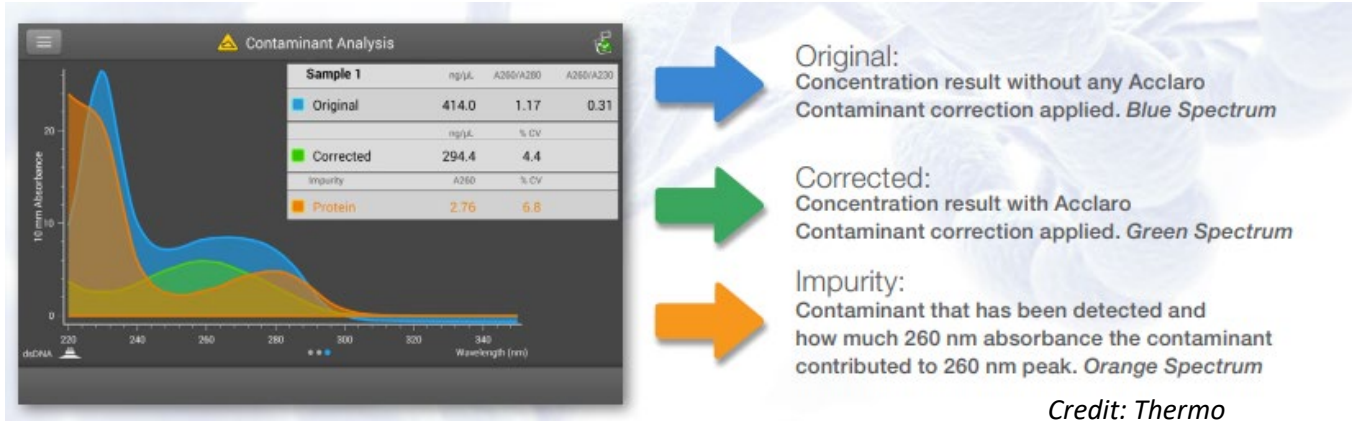


Problems:

1. Cannot truly discriminate RNA from DNA when mixed
2. Contaminating solvants can absorb at same wavelengths -> OK-enough at high concentration, but can lead to huge overestimation at low concentration (e.g: 50 ng/ul instead of 1 ng/ul).
3. Sensitivity: 1ng/ul → Not sensitive-enough for new genomics applications

Nucleic acids quantification

If using nanodrop, use latest “nanodrop One” version, → identifies/corrects for most frequent contaminants.



Credit: Thermo

DNA contaminated with proteins

Nucleic acids quantification

Fluorescence-based methods (Qubit, Quant-It picogreen, ribogreen…)

- Intercalating dyes → fluorescent upon binding
- Very accurate (no fake signal from most solvants/contaminants)
- Very specific (discriminate RNA/DNA/ssDNA)
- Very sensitive: ˜0.1 ng/ul for DNA, 2 ng/ul for RNA
- Downsides: requires pipetting, not free, standard curve (hence errors, use pos control).

Nucleic acids quantification

Recommendation of GECF:

- <50 ng/ul → always fluorescence-based
- >50 ng/ul, nanodrop can be used, specially for large series

Nucleic acids QC

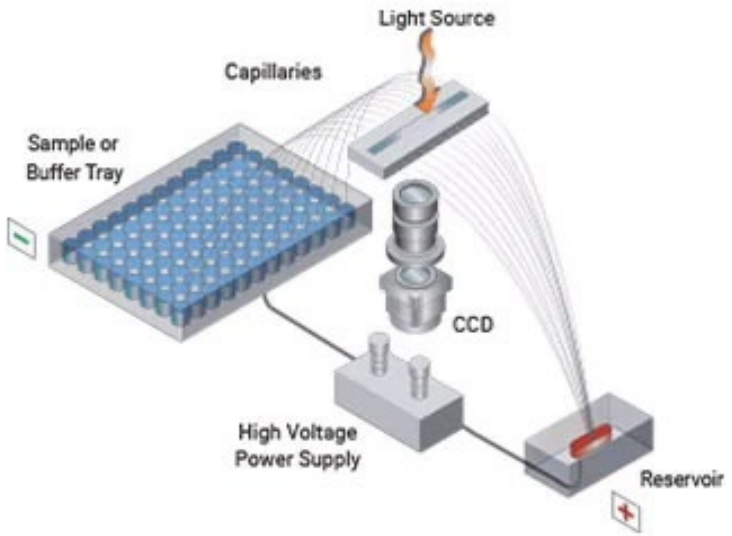
High sensitivity capillary electrophoresis for RNA and DNA samples

Nucleic acids QC

High sensitivity capillary electrophoresis for RNA and DNA samples

Applications: QC of RNAs, cDNAs, genomic DNA, NGS libraries...

- Very sensitive both for RNA and DNA (0.1 ng/ul)
- Fluorescence detection when reaches camera -> size



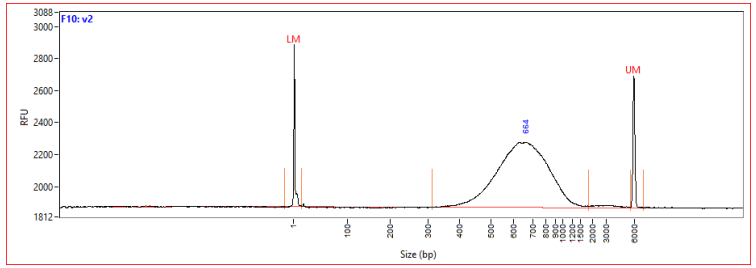
Nucleic acids QC

High sensitivity capillary electrophoresis for RNA and DNA samples

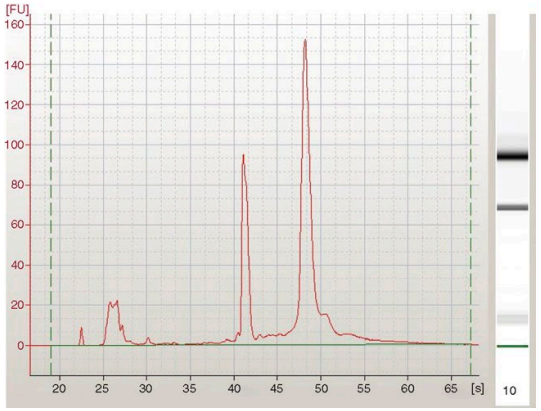
Applications: QC of RNAs, cDNAs, genomic DNA, NGS libraries...

- Very sensitive both for RNA and DNA (0.1 ng/ul)
- Fluorescence detection when reaches camera -> size

Fragment Analyzer

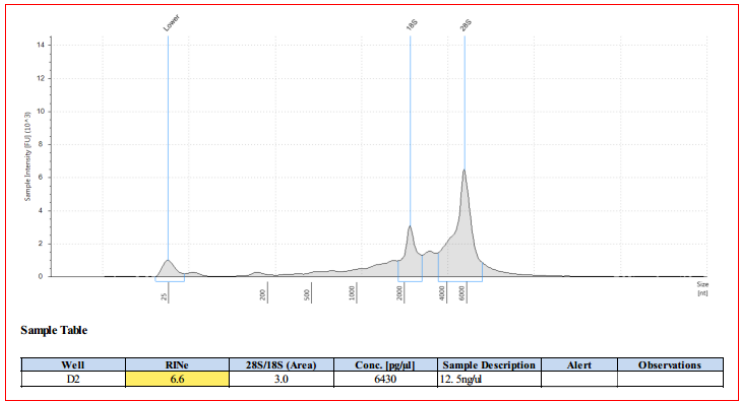


BioAnalyzer



TapeStation 4200

Not true capillary, less resolution, but faster.



Sample Table

Well	RINe	28S/18S (Area)	Conc. [pg/µl]	Sample Description	Alert	Observations
D2	6.6	3.0	6430	12. 5ng/ul		

Nucleic acids QC

High sensitivity capillary electrophoresis for RNA and DNA samples

Typical profile of a good quality RNA

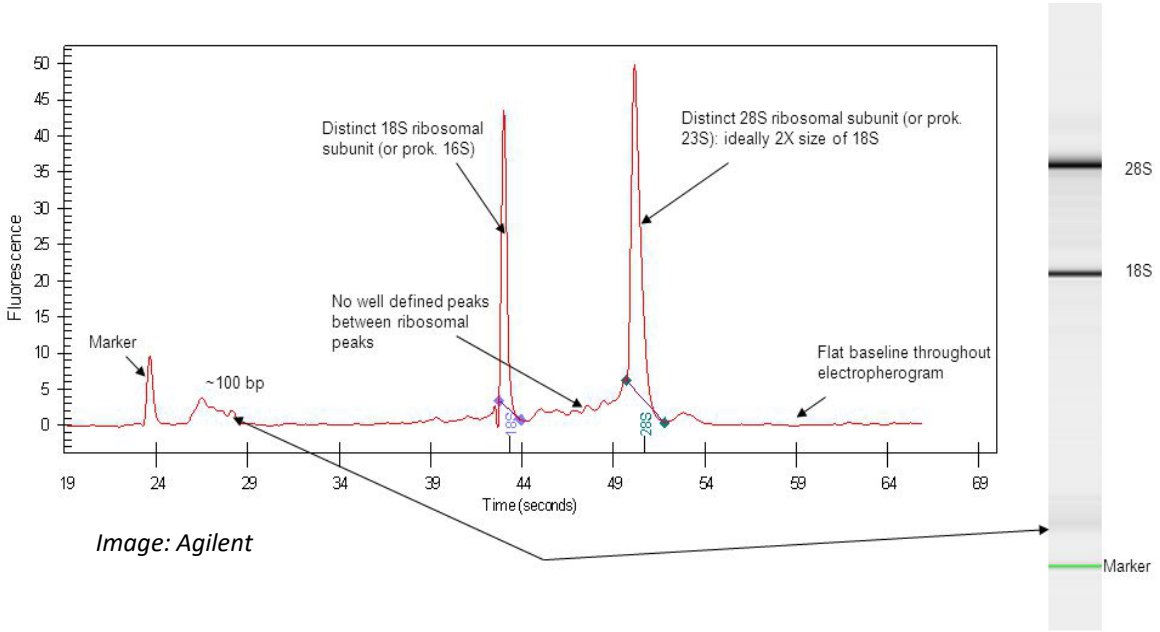
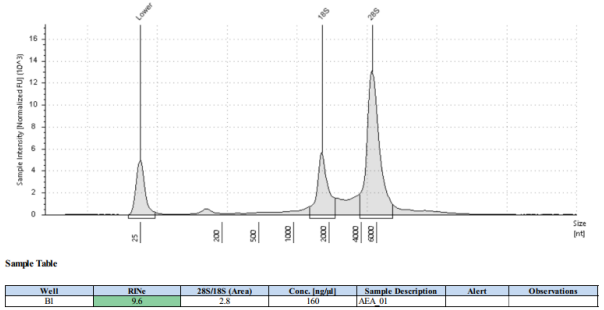


Image: Agilent

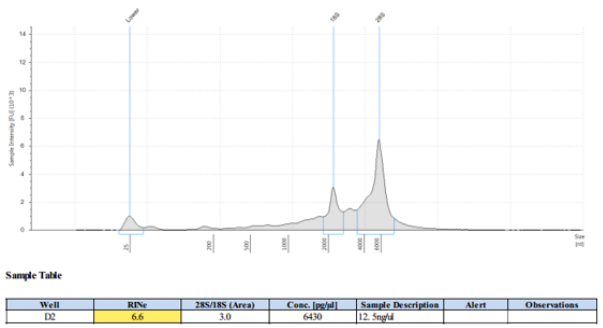
- mRNAs too low amount (2-5%)→ rRNA (80%) as surrogate
  → RIN (RQN) score (profiles of 28S and 18S)
- RIN (RQN) : 1-10, 10 is best, above 7 is acceptable



Sample Table

Well	RINe	28S/18S (Area)	Conc. [pg/µl]	Sample Description	Alert	Observations
B1	9.6	2.8	160	AEA_01		

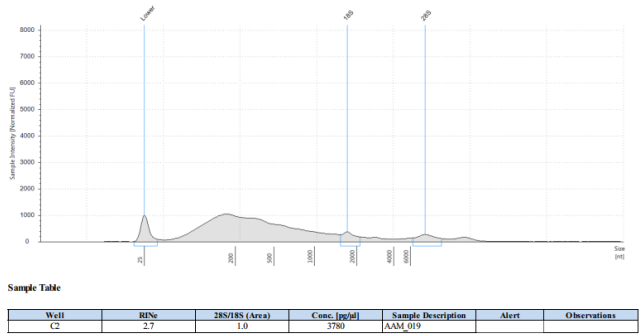
intact



Sample Table

Well	RINe	28S/18S (Area)	Conc. [pg/µl]	Sample Description	Alert	Observations
D2	6.6	3.0	6430	12. 5ng/µl		

intermediate



Sample Table

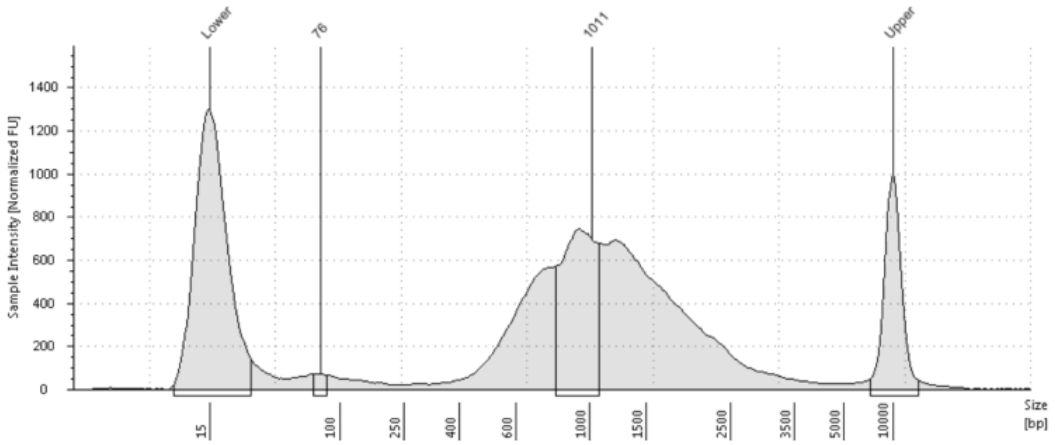
Well	RINe	28S/18S (Area)	Conc. [pg/µl]	Sample Description	Alert	Observations
C2	2.7	1.0	3780	AAM_019		

degraded

Nucleic acids QC

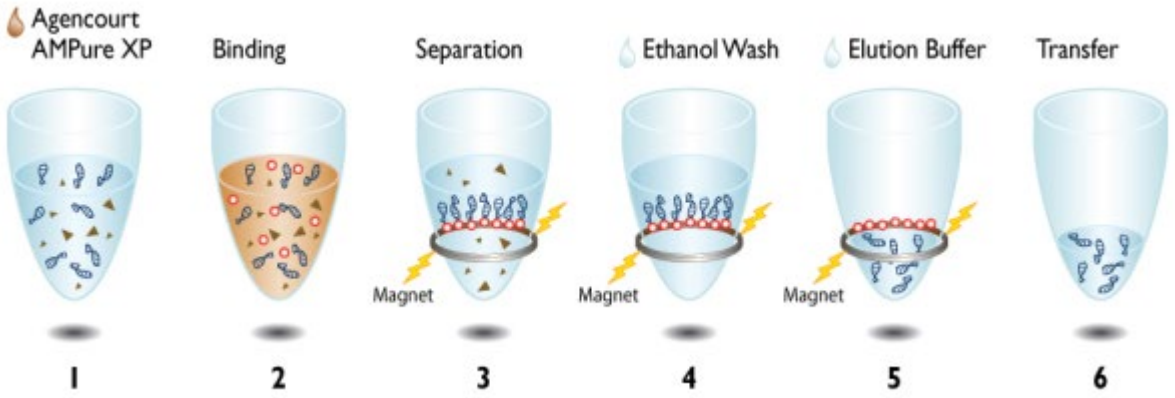
High sensitivity capillary electrophoresis for RNA and DNA samples

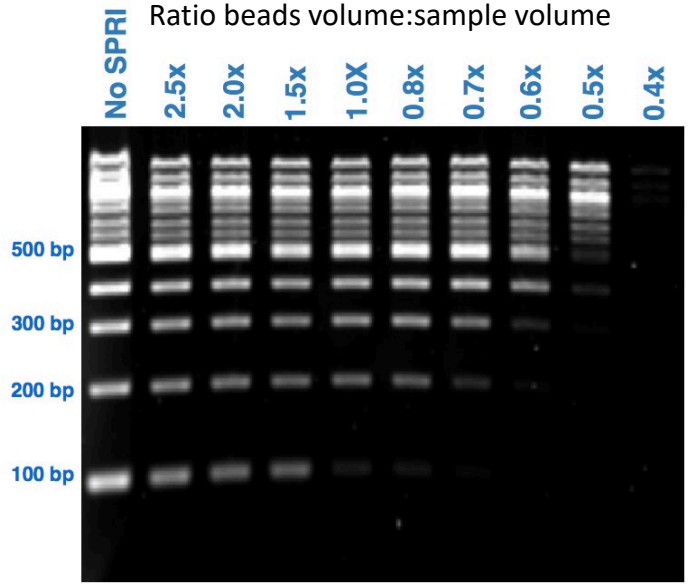
Profile of a good quality cDNA (dsDNA form)



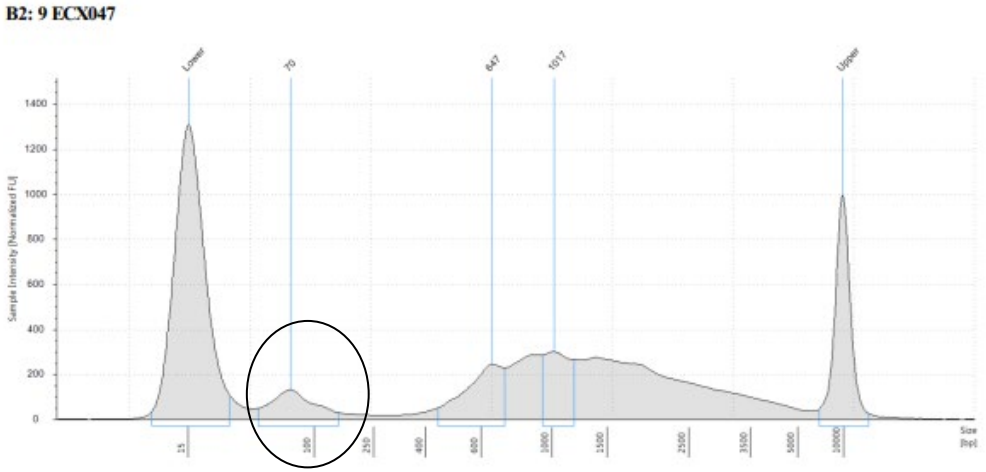
Smear from 500-2’000nt (rRNA excluded by cDNA creations protocols)

Nucleic acids size selection (AMPure beads)

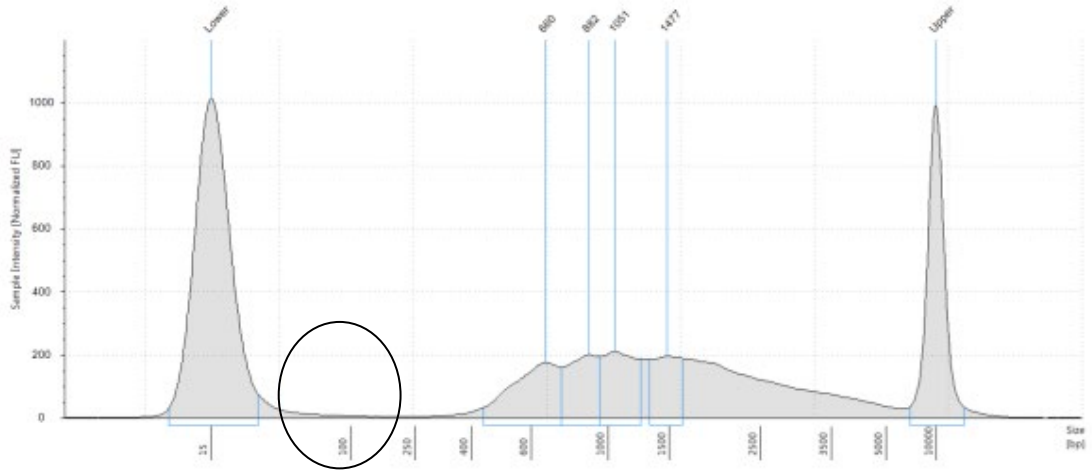




The longer the DNA molecule the higher the affinity for the beads →
if low amount of beads, only longer fragments are retained



→



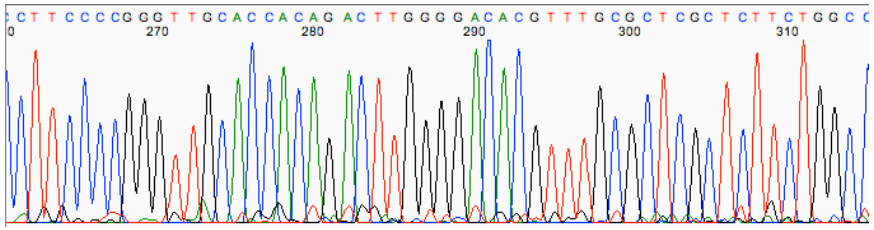
High Throughput sequencing (NGS)

High Throughput sequencing (NGS)

Historical perspective

Sequencing:

- Sanger sequencing: 1 DNA fragment per reaction, used for 1st human genome sequencing



- Around 2005: “next generation sequencing” → millions sequencings done in parallel (Illumina, Roche/454, SoliD, Ion Torrent)
- Only Illumina remained on the market (+ Ion Torrent in diagnostics)
- Now, competition is again active, with Element Biosciences, Ultima Genomics, MGI...

FYI

High Throughput sequencing (Illumina)

Typical NGS Services

MiSeq

- Low yield
- Only specific applications (amplicons sequencing, bacterial genomes)

NextSeq

- Medium yield
- Broad range of use

NovaSeq

- Up to ultra-high yield
- Any application (WGS in particular)

High Throughput sequencing (Illumina)

Clustering (bridge amplification) and sequencing run

Clustering= serves to amplify signal during imaging

Input in the sequencer:
NGS libraries (dsDNA).

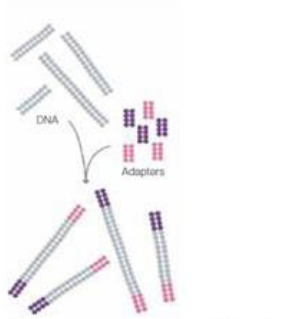


Figure 1
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

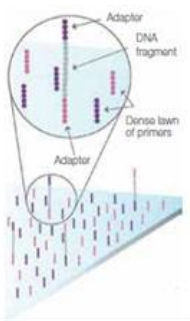


Figure 2
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

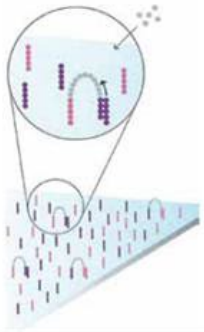


Figure 3
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

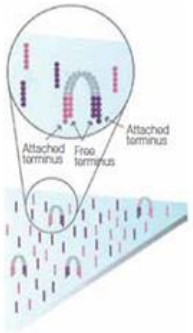


Figure 4
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

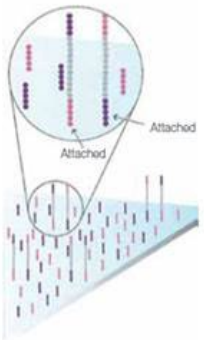


Figure 5
Denaturation leaves single-stranded templates anchored to the substrate.

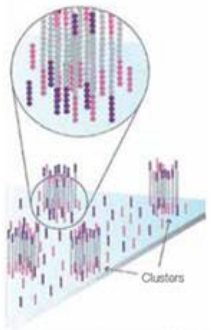


Figure 6
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

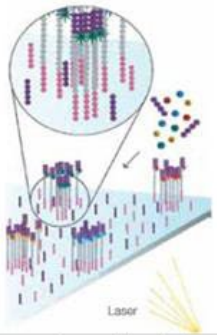


Figure 7
The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

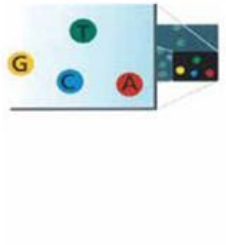
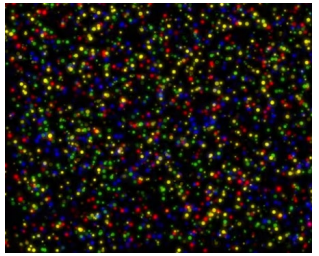


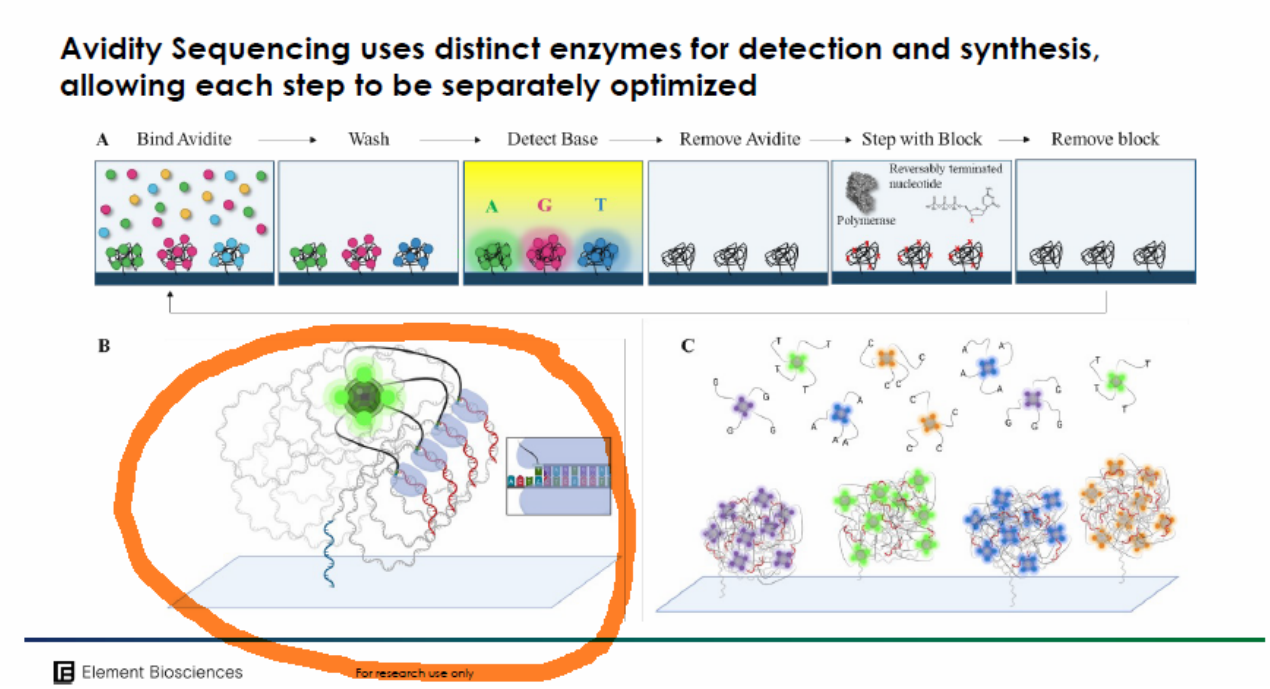
Figure 8
After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.



Each dot = one cluster

High Throughput sequencing (NEW: Aviti)

- Rolling circle amplification → polonies
- fluorescently labelled oligos bind → washed away (sequencing by binding)→ no scar→ low errors
- Why cheaper? Less reagents usage, 2 reasons:
  - Very low non-specific binding surface
  - “Avidites”:



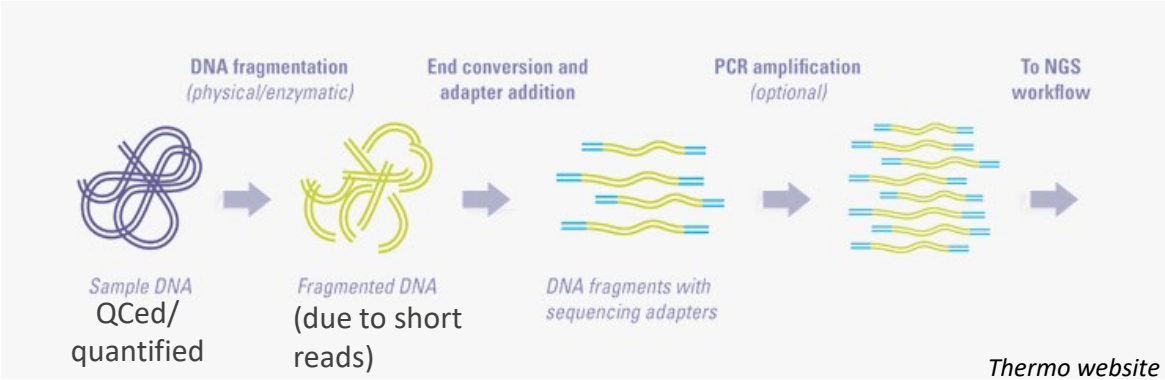
High Throughput sequencing (Illumina)

clustering and sequencing run

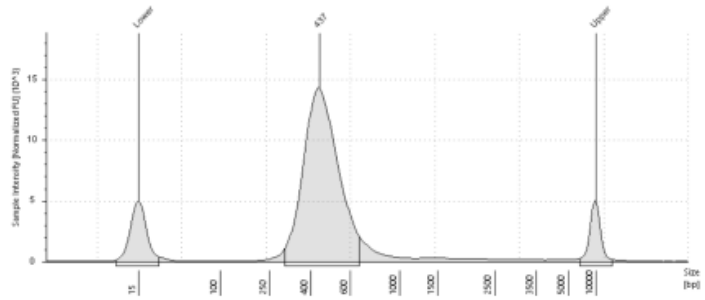
- Microscope requires strong fluorescent signal → clusters → out of sync molecules within the cluster → rapid decrease quality over fragment length → only small DNA fragments, aka "short reads" sequencing (50-300nt).
- Throughput from 1 mio to 25'000 mio reads (1 human genome at 30x coverage is ˜400mio reads)

High Throughput sequencing (Illumina)

Library prep



Thermo website



- 2 modes: Single-end (single read, SR, e.g. SR75), or paired-end (PE, e.g. PE150)
- Index: for multiplexing/pooling samples on the flow cell

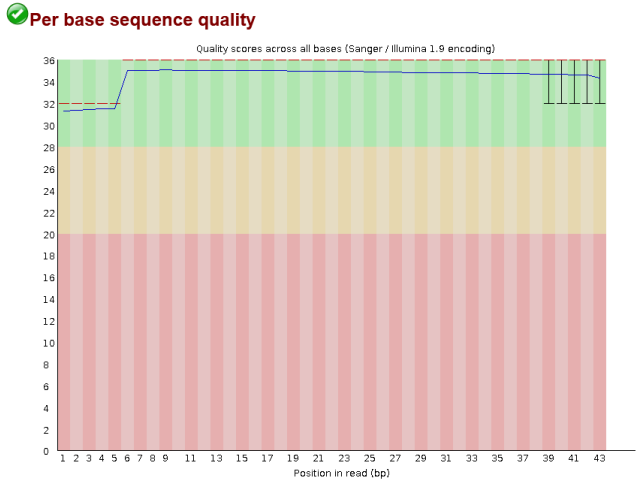
→ Files obtained = .fastq

High Throughput sequencing (Illumina)

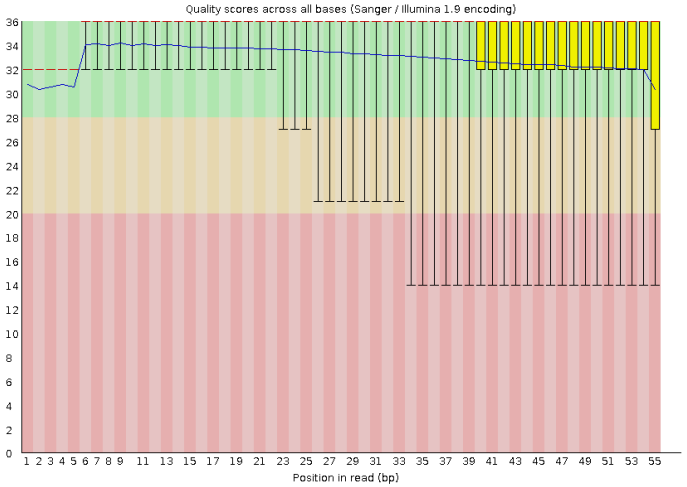
QC with FastQC

Q30 metrics (quality estimation)

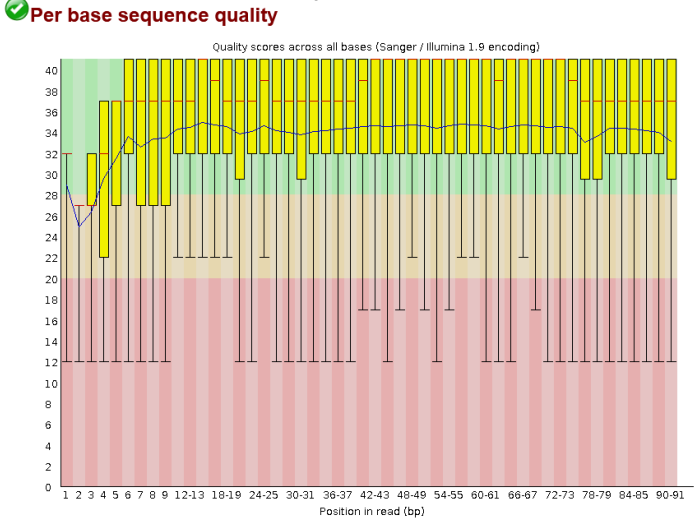
perfect



good



˜suboptimal (but frequent on older instruments)

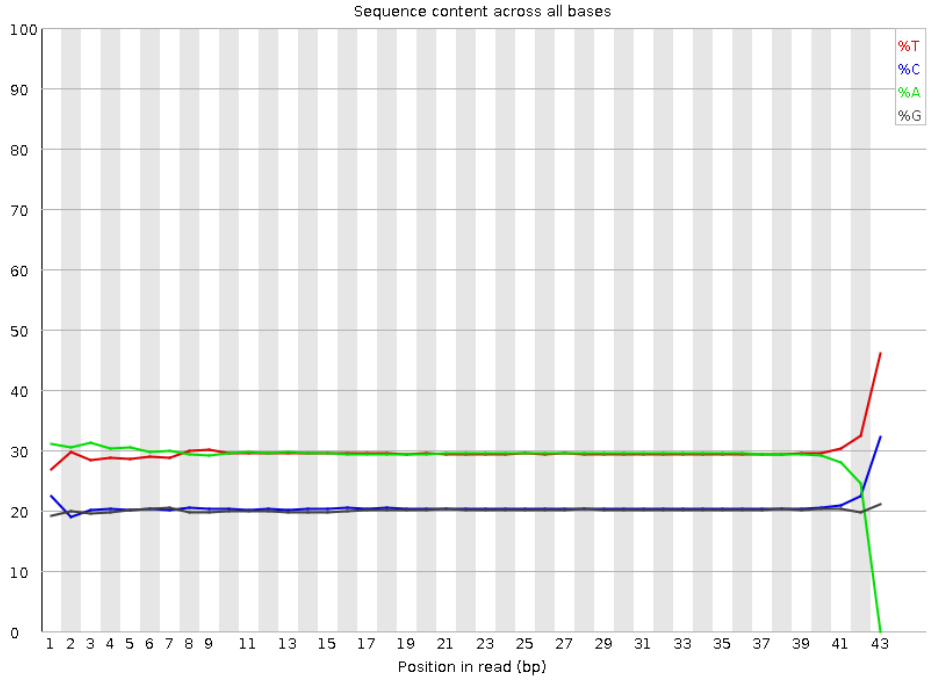


The real metrics for read quality is % error rate (spiking of a known PhiX genome library in each run): typical <1%

High Throughput sequencing (Illumina)

QC with FastQC

Per base sequence content



GC content (%)		
	genome (can vary amongst chromosomes)	polyA+ transcriptome
Saccharomyces cerevisiae	38	
Homo sapiens	41	46-48
Mus musculus	42	48-50
Pseudomonas aeruginosa	66	

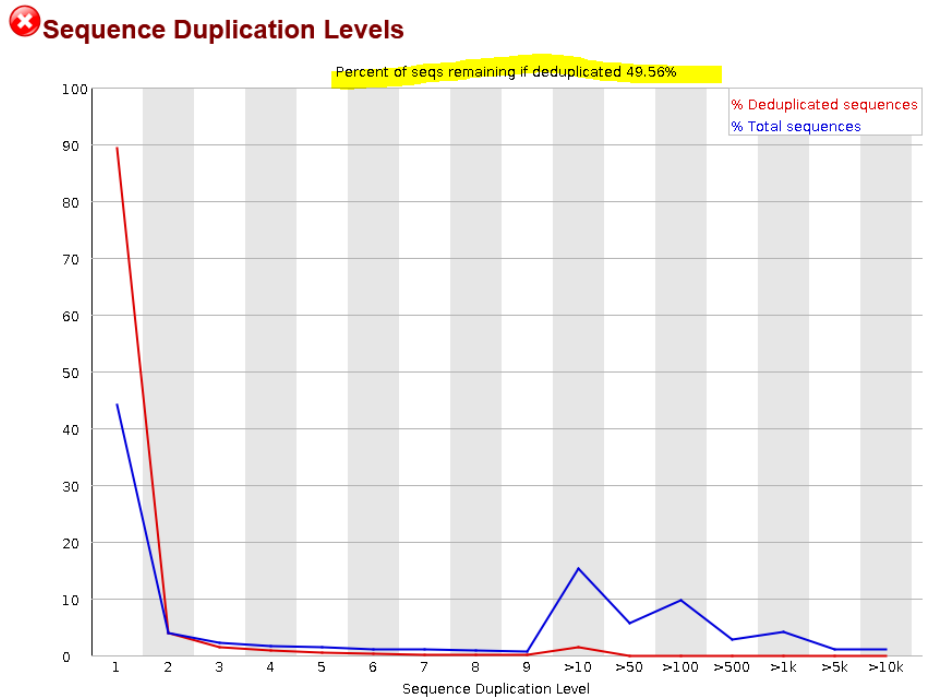
GECF internal

% base content can indicate issues, but no universal “good values”

FYI

High Throughput sequencing (Illumina)

QC with FastQC



Fragment duplication (=mostly PCR duplicates) rate is important but vary a lot depending on applications

High Throughput sequencing (Illumina)

Short-reads sequencers conclusion

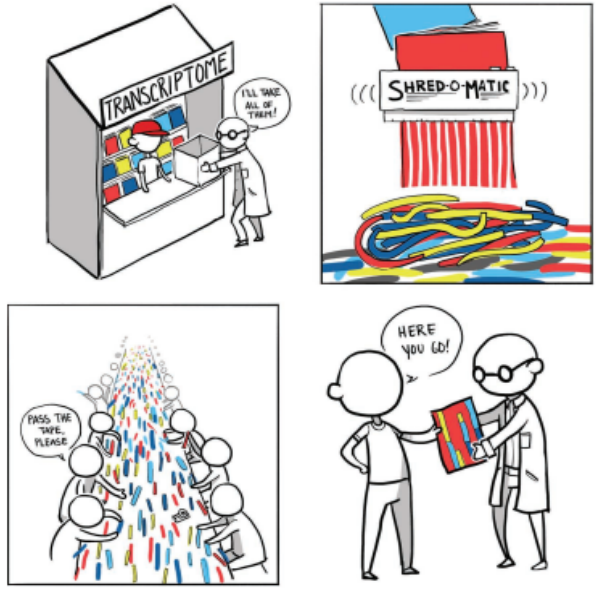
Strength:

- Very high throughput
- Low error rate
- Flexible and very well established (hundreds of library prep protocols)

BUT:

- Still quite expensive
- Short reads only (max 600nt)→ poor for isoforms/alternative splicing, structural variants…

→ Long reads sequencing: PacBio and Oxford Nanopore



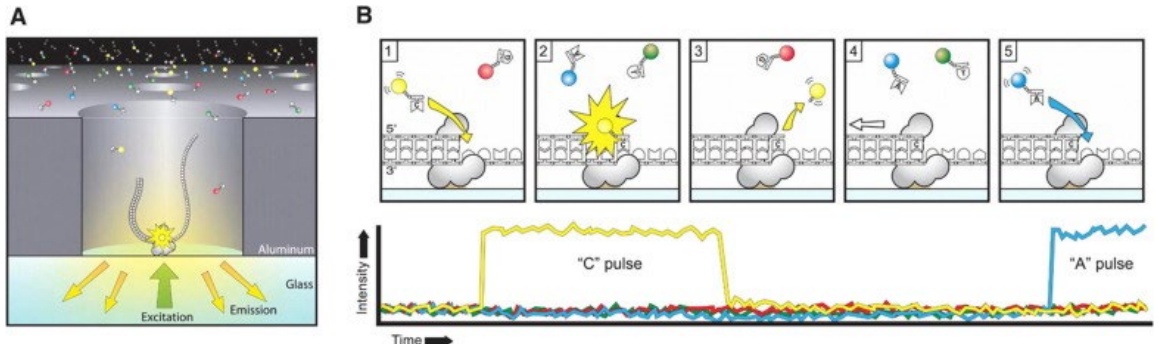
Korf, Nature Meth., 2013

High Throughput sequencing

Long Reads Sequencing

Single-molecule sequencing (no clusters) → no signal alteration over length of fragment → Mb long reads
But faint signal more prone to background error rate)

PacBio:



Circular library → Multiple sequencing rounds → lower error rate.

Oxford Nanopore (MinION, Flongle):

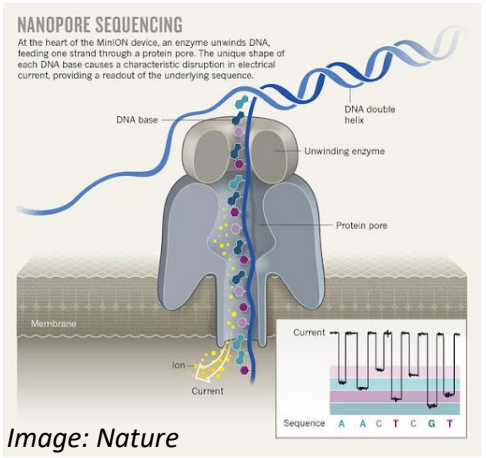


Image: Nature

- Also possible to detect DNA methylation, and to directly sequence RNA
- In the field!

Both: too low throughput, and too high error rate → complementary to short reads sequencing

FYI

High Throughput sequencing
Gene Expression analysis (RNA-seq)

High Throughput sequencing
Gene Expression analysis (RNA-seq)

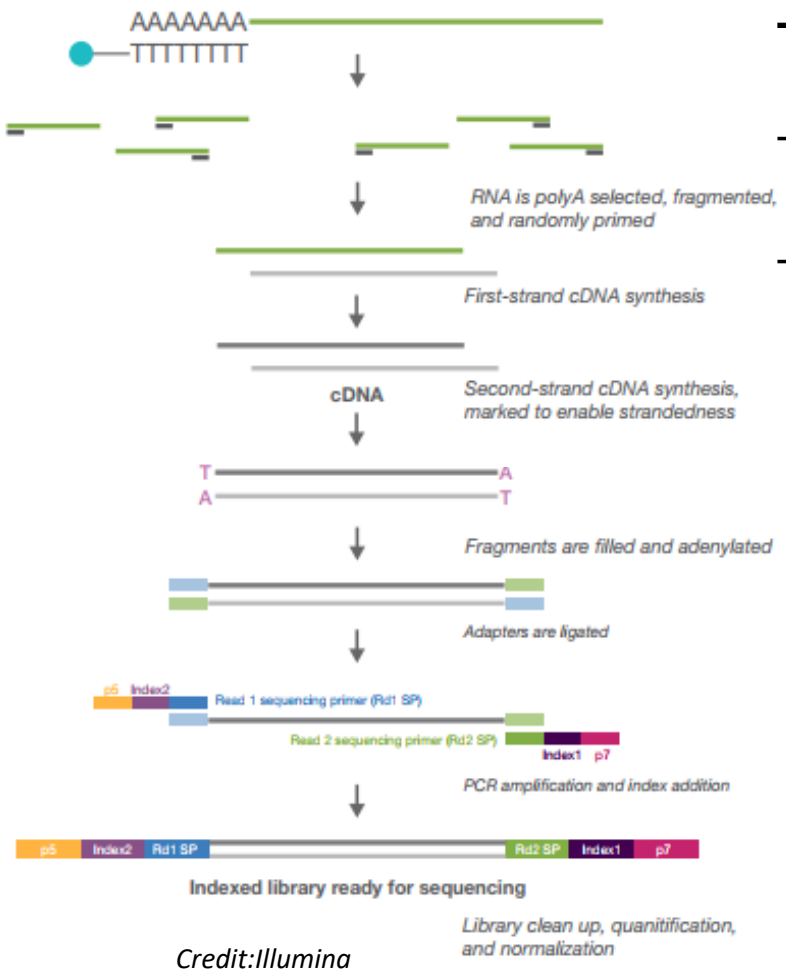
Workflow: RNAs → libraries → sequencing → mapping to transcriptome

Main applications:

- **quantification** of RNAs/mRNAs (counting mapped reads):
  → only done with short reads (since needs very high depth)
- "**sequencing**" *per se* (*de novo* transcriptome, isoforms…)
  → best done with long reads sequencing

High Throughput sequencing
Gene Expression analysis (RNA-seq)

<u>mRNA-seq</u> library prep ("coding transcriptome") (RIN>7, eukaryotes only)



Credit:Illumina

→ mRNA capture by oligodT beads (discard rRNA…)(explains requirement for good RIN)

→ heat fragmentation (short reads requirement)

→ reverse transcription (random primers)

→ dsDNA (2nd strand synthesis) + strand marking

→ adapters ligation (for PCR, flow cell binding and sequencing primer binding sites)

→ PCR (to increase amount)

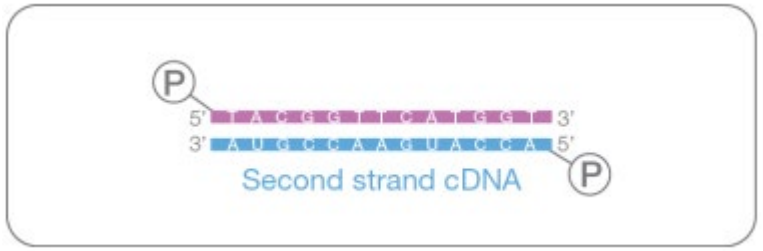
High Throughput sequencing
Gene Expression analysis (RNA-seq)

2nd strand synthesis

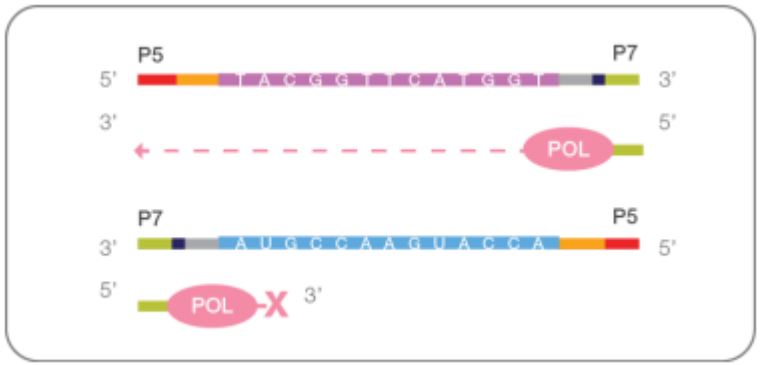
Not needed for qPCR (since 2 primers), but needed for transcriptome-wide library prep!

RNase H + E. Coli Polymerase 1

Strand-specificity («stranded protocols»)

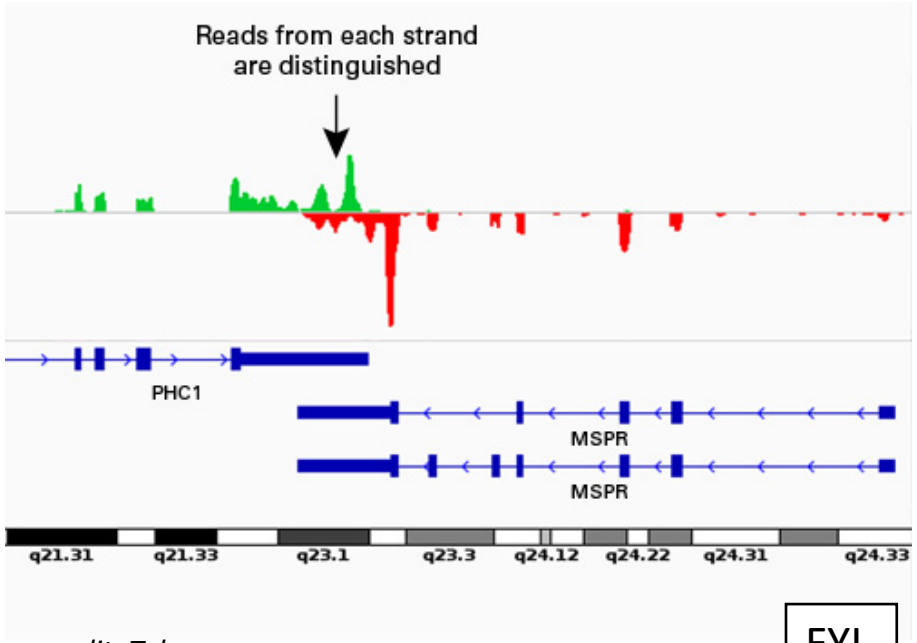


dUTP incorporation in 2nd strand



2nd strand is not amplified during PCR

credit: Illumina

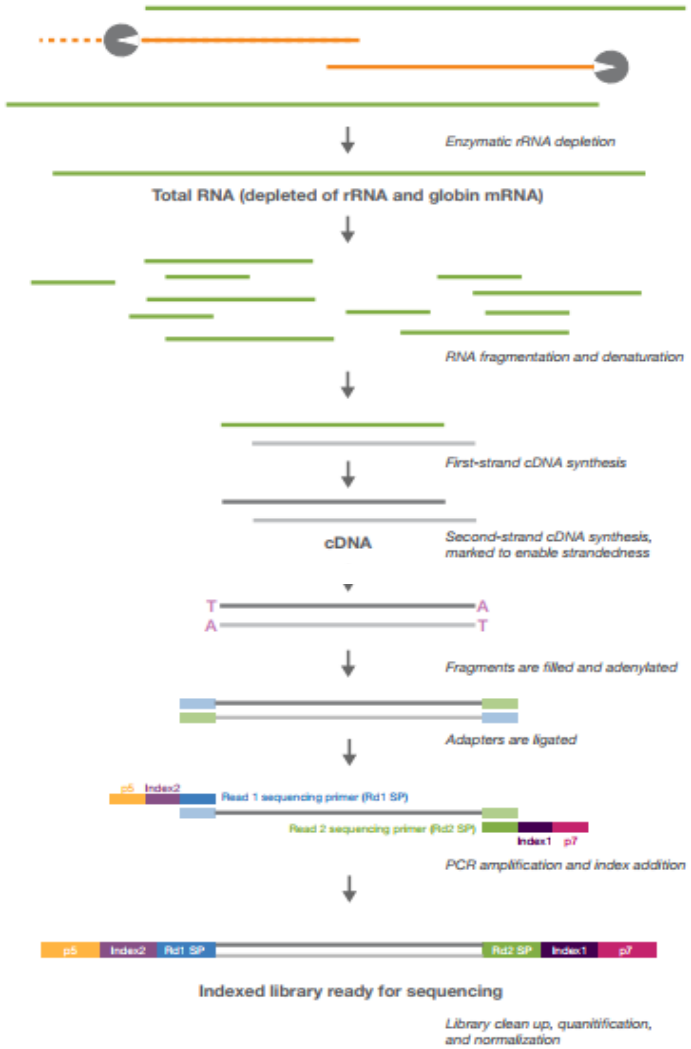


credit: Takara

FYI

High Throughput sequencing
Gene Expression analysis (RNA-seq)

“total” RNA-seq (whole transcriptome RNA-seq) (˜any RIN, bacteria OK)



Credit:Illumina

→ rRNA depletion by rRNA-targeting oligos (ribodepletion)

→ heat fragmentation (short reads requirement)

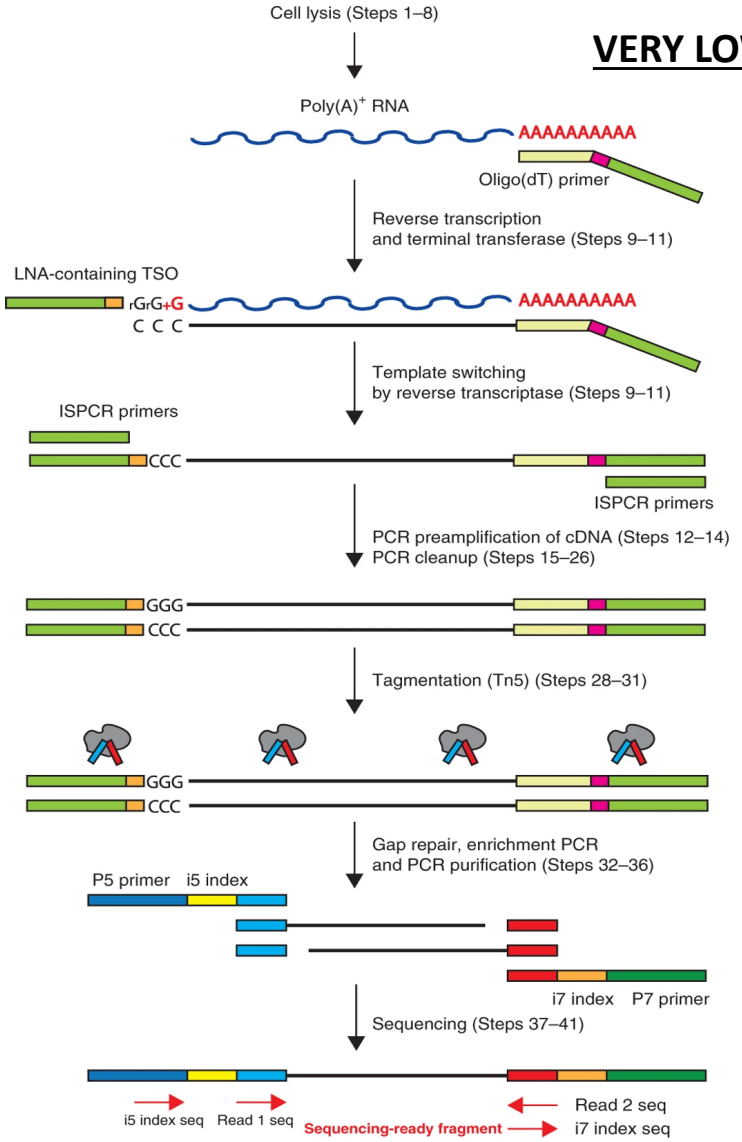
→ reverse transcription (random primers)

→ 2nd strand synthesis → dsDNA

→ adapters ligation (for PCR, flow cell binding and sequencing primer binding sites)

→ PCR (to amplify)

High Throughput sequencing
Gene Expression analysis (RNA-seq)



Sandberg Nature 2014, FYI

<u>VERY LOW AMOUNT mRNA-seq</u> library prep (RIN>7, “smart-seq”)

-> RT adds a few CCCs at the end of cDNA → GGG-containing oligo annealed (“TSO”) → <u>“template switch”</u>

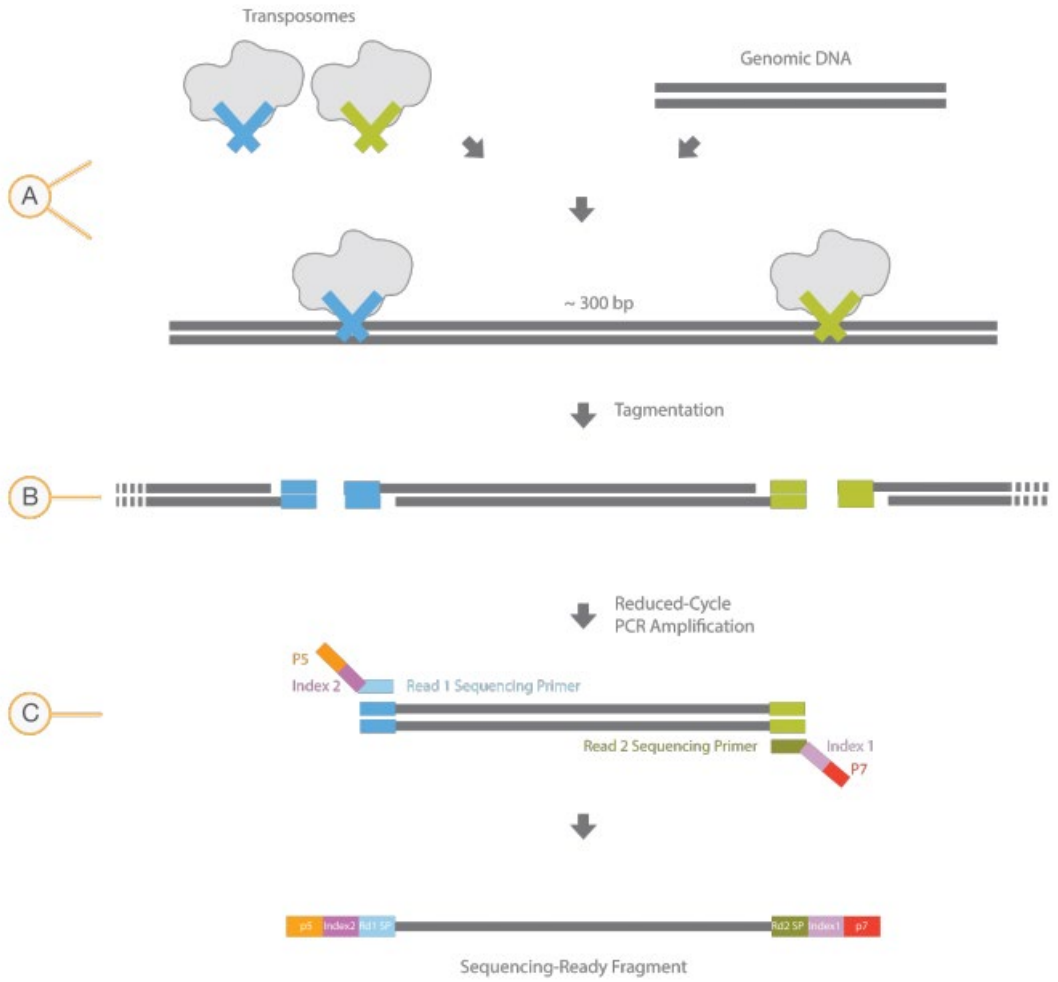
→ known sequence now on both sides (in only 1 step) → no need for “2nd strand synthesis”

→ fragmentation + partial adapters addition by tagmentation (less steps, less loss)

FYI

High Throughput sequencing
Gene Expression analysis (RNA-seq)

«tagmentation» by transposase from Tn5 transposon

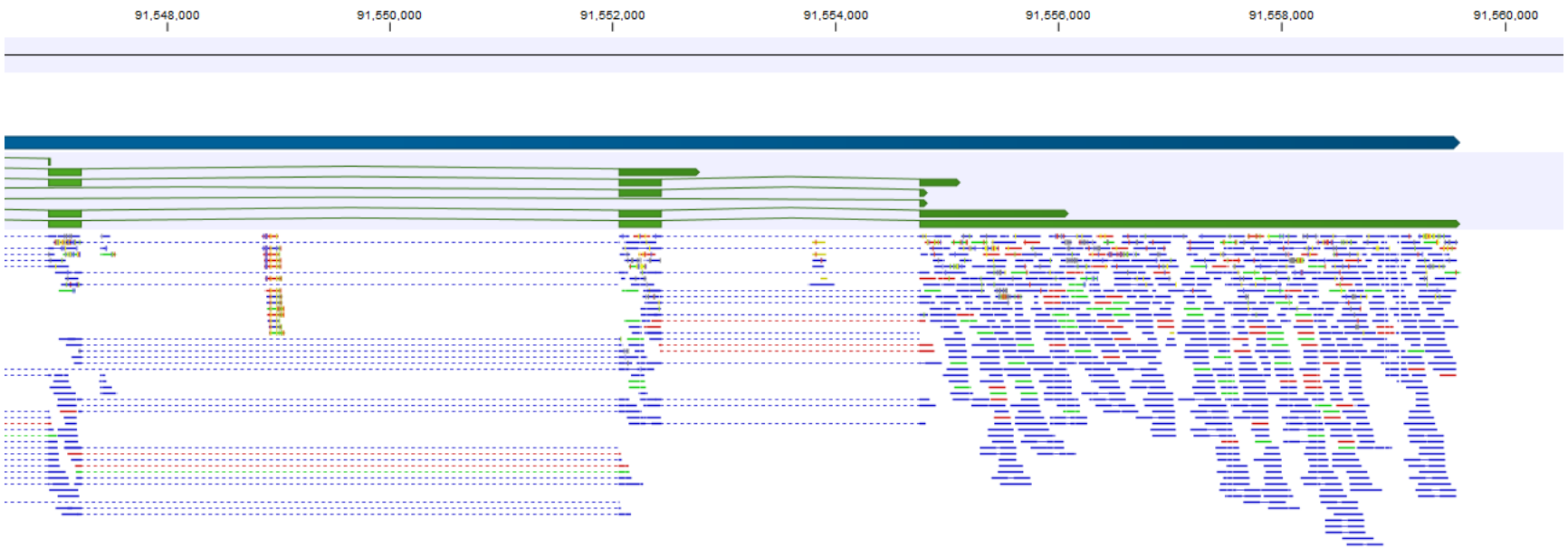


Tn5 transposase:
- Mutated hyperactive enzyme
-19nt recognition sequence
(«mosaic ends»)

From: Illumina

High Throughput sequencing
Gene Expression analysis (RNA-seq)

Mapping and data analysis



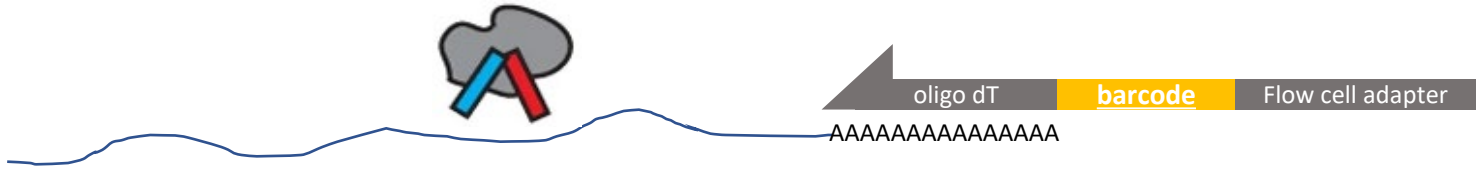
→ 15-50 mio reads/sample → mapping → counting
→ normalization & expression quantification, e.g. for mRNA lengths & mio reads (rpkm) → differential expression (multiple pipelines, see bioinfo part)

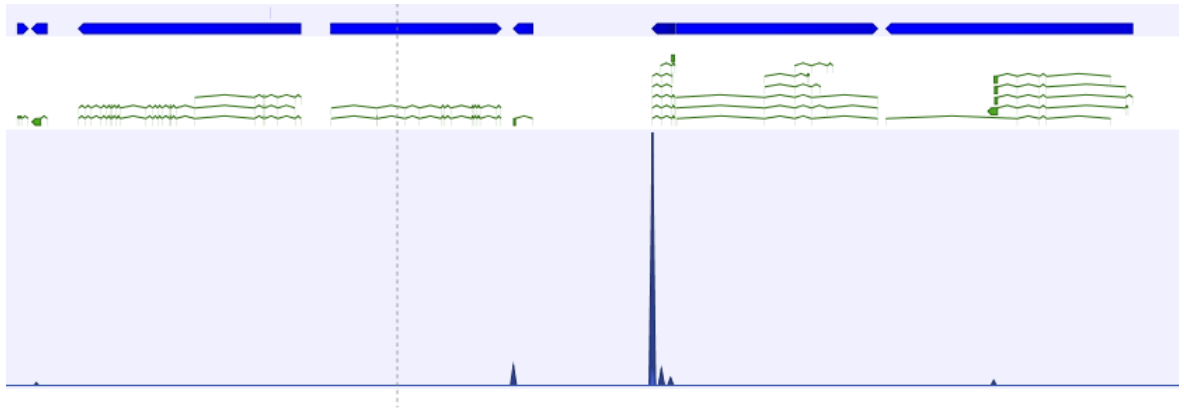
High Throughput sequencing
Gene Expression analysis (RNA-seq)

Recent development: 3'end sequencing

Only 3' end of cDNA kept

→ Barcode added at beginning (RT) → early multiplexing/pooling → streamlined protocol
→ no need for mRNA length normalization.
→ less reads/sample needed (5mio/sample, cheaper).
→ … but misses isoforms…





Mapping on 3'end

Alpern *et al. Genome Biology* (2019) 20:71
https://doi.org/10.1186/s13059-019-1671-x

Genome Biology

METHOD Open Access

BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing

Daniel Alpern[1,2†], Vincent Gardeux[1,2†], Julie Russeil[1], Bastien Mangeat[3], Antonio C. A. Meireles-Filho[1,2], Romane Breysse[1], David Hacker[4] and Bart Deplancke[1,2*]

Recent EPFL method

High Throughput sequencing
Gene Expression analysis (RNA-seq)
Conclusions

Strength:
- Exhaustive
- Extremely broad linear/dynamic range
- Very sensitive if enough sequencing depth
- Many protocols exist: any quantity/quality of starting RNA

Pitfalls:
- expensive for many samples
- For a defined set of lowly expressed genes-> qPCR may be as good
- Much longer turnaround time than qPCR
- (Data analysis less straightforward →more possibilities of bias than qPCR)

Gene expression studies

Experimental design

Which technology is best suited?:

- **qPCR?:** Low price, very fast data, when only a few genes of interest, lots of samples
- **RNA-seq?:** comprehensive analysis of coding transcriptome, intermediate number of samples, weeks before data, well-established
- **3'-end mRNA-seq?** comprehensive analysis of coding transcriptome, "cheap", high number of samples, weeks before data

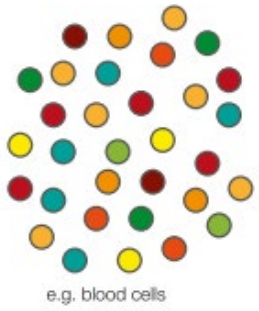
SINGLE CELLS RNA-seq

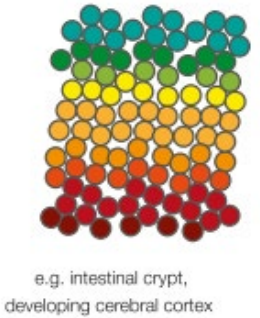
Single Cells

Why single-cell?

Single cell transcriptomics

- More resolution on the studied system





Teichmann, Mol Cell, 2015

Single Cells

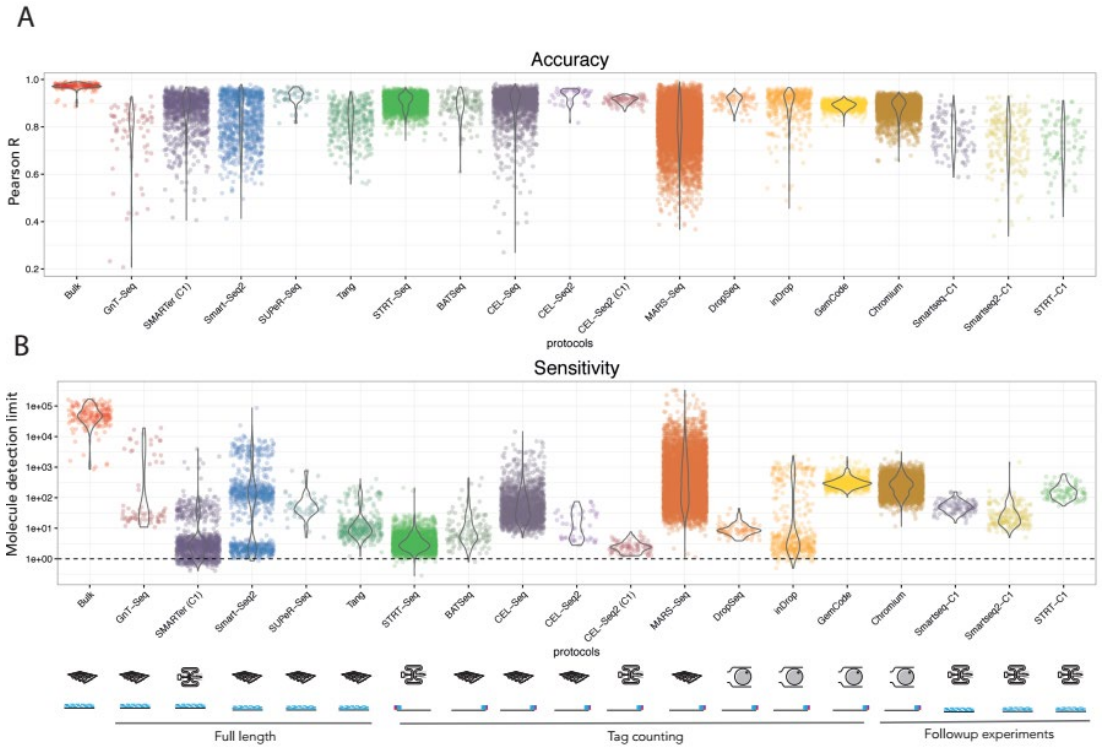
Why single-cell?

Single cell transcriptomics

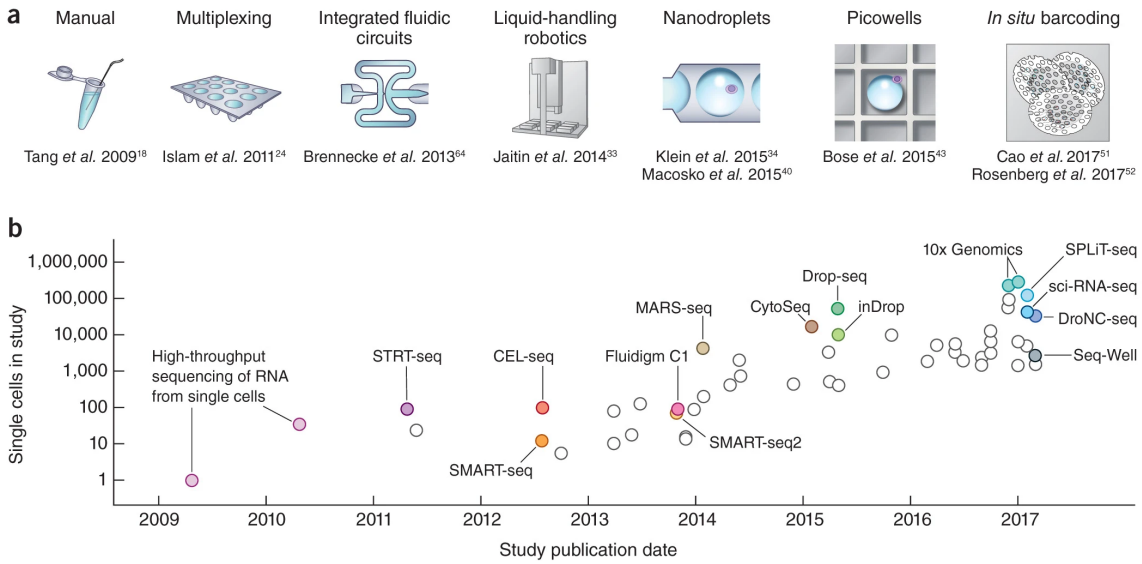
- More resolution on the studied system
- Rare cell types (unsupervised, no prior knowledge needed)… what is a cell type?
- Cell to cell heterogeneity (normal tissues, tumors)
- Developmental process (intermediate cell states, transitions)
- Easier to define gene regulatory networks (easier correlations)

Single Cells scRNA-seq

Historical perspective: explosion of scRNA-seq methods since 10 years



Teichmann, Molec. Cell, 2015



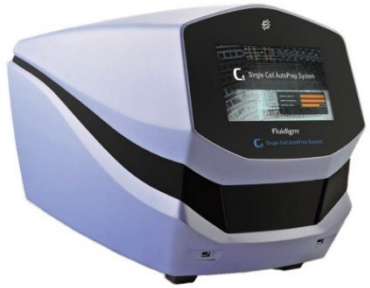
Svensson V et al. (2018) Nature Protocols 13: 599–604. DOI: 10.1038/nprot.2017.149.

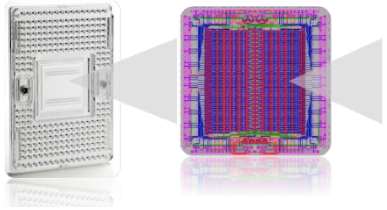
FYI

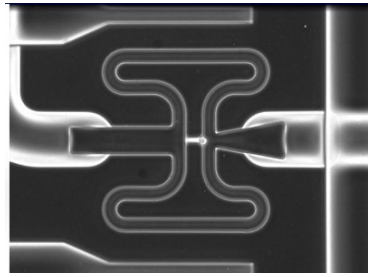
Single Cells scRNA-seq

Historical perspective: C1 system (*Fluidigm)*

Microfluidics-based single-cells capture and processing







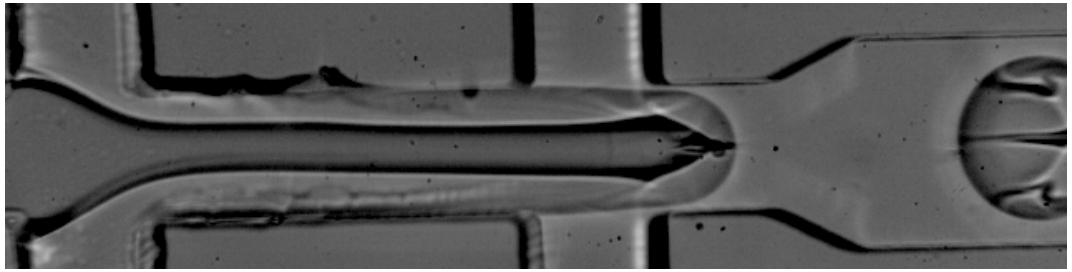
A captured cell

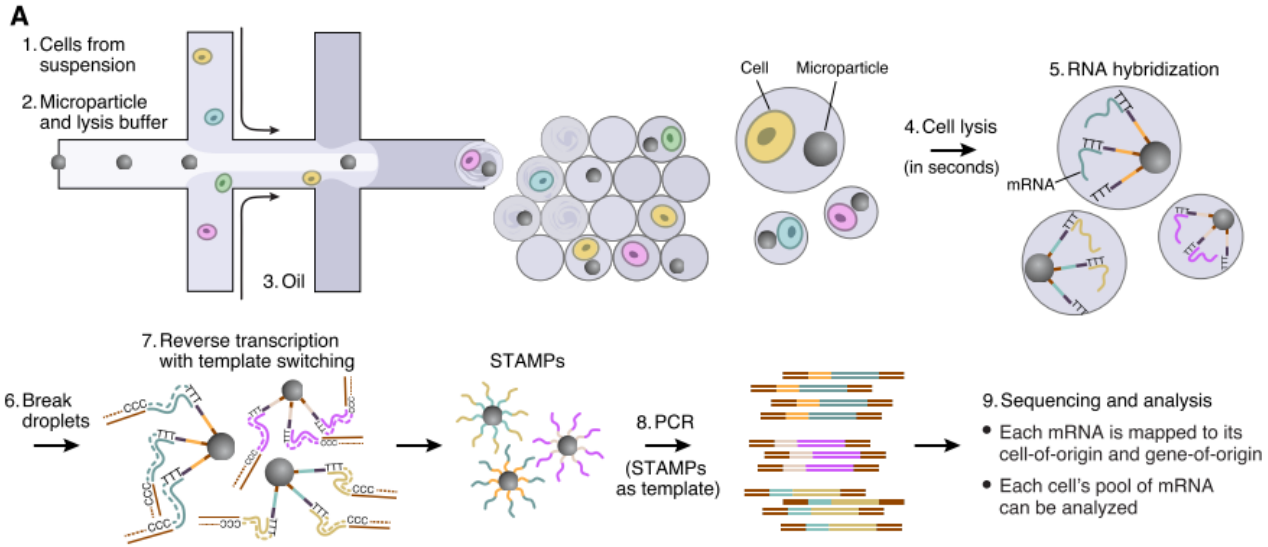
- Full-length (Smart-seq)
- <u>Only 96 cells</u>

FYI

Single Cells scRNA-seq

Historical perspective: Higher throughput → DROP-seq





Barcoded cDNA pooling

Macosko, Cell, 2014

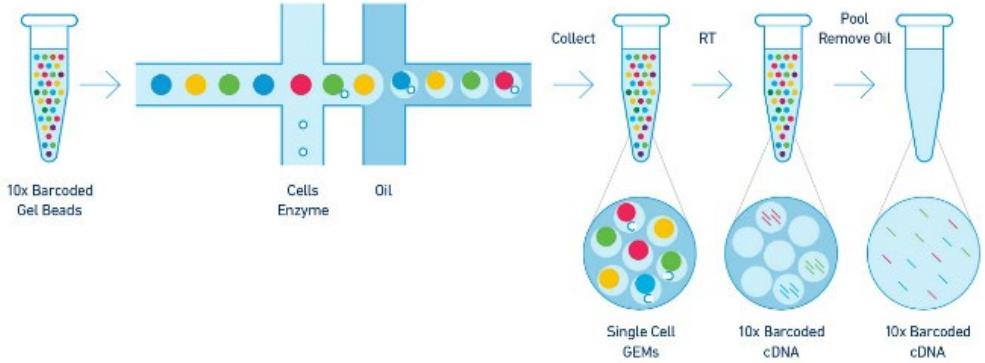
bead

oligo dT | Cell barcode | Flow cell adapter

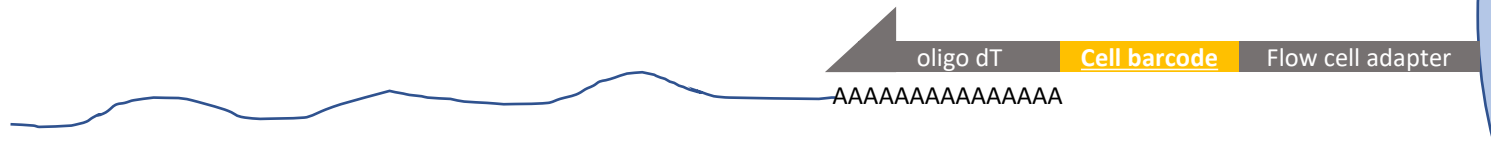
AAAAAAAAAAAAAAA

Single Cells scRNA-seq

10X Genomics Chromium

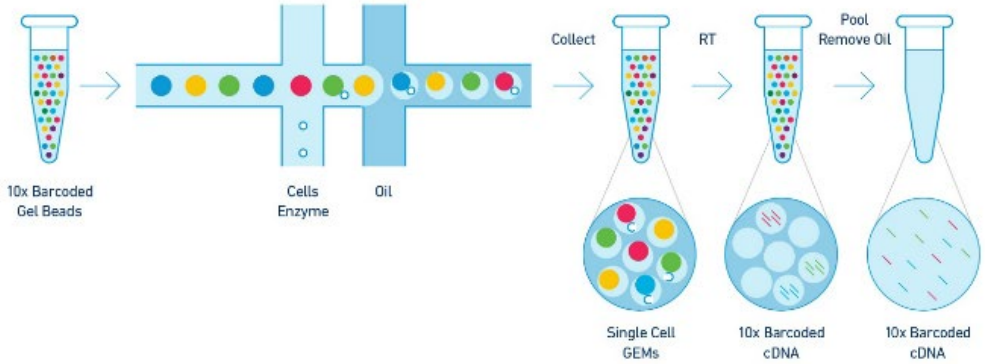


- Gel beads that dissolve in droplet
- Strengths:
  - High throughput (500 – 30’000 cells)
  - Any cell size up to 60um (if larger: use nuclei)
- Sensitivity 1’000-5’000 genes (won’t get rare transcripts)
- Nuclei when true single-cells suspension not possible (neurons)
- Can get TCR/Ig sequence in parallel -> clonotypes / “immune cells profiling”
- Downsides:
  - Cannot image cells
  - Not best-in-class sensitivity (bad for lowly expressed genes)

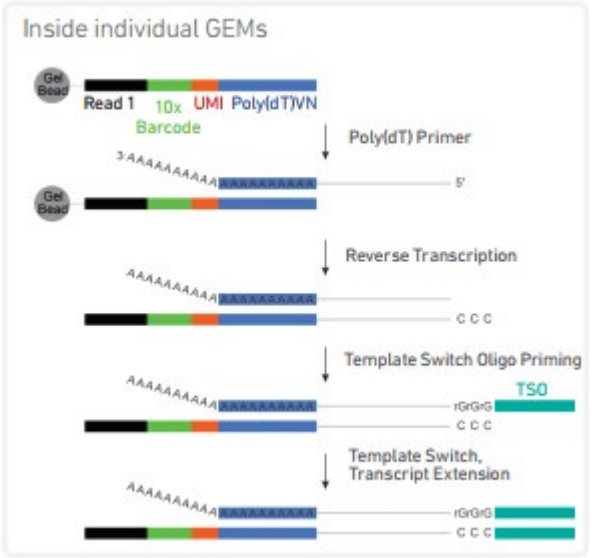


Single Cells scRNA-seq

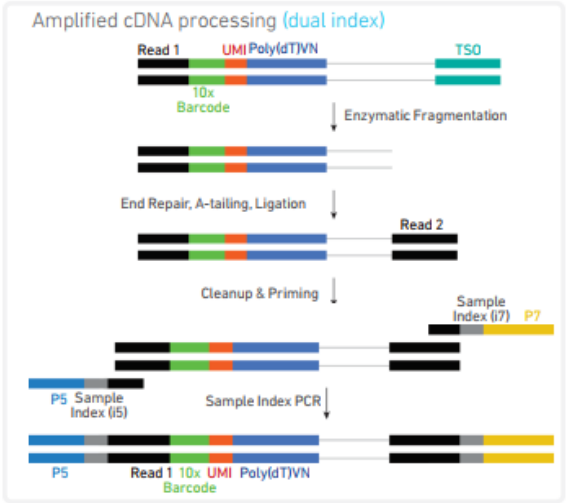
10X Genomics Chromium



cDNA creation similar to low-amount RNA-seq (TSO), with cell barcode in addition (“10x barcode”):



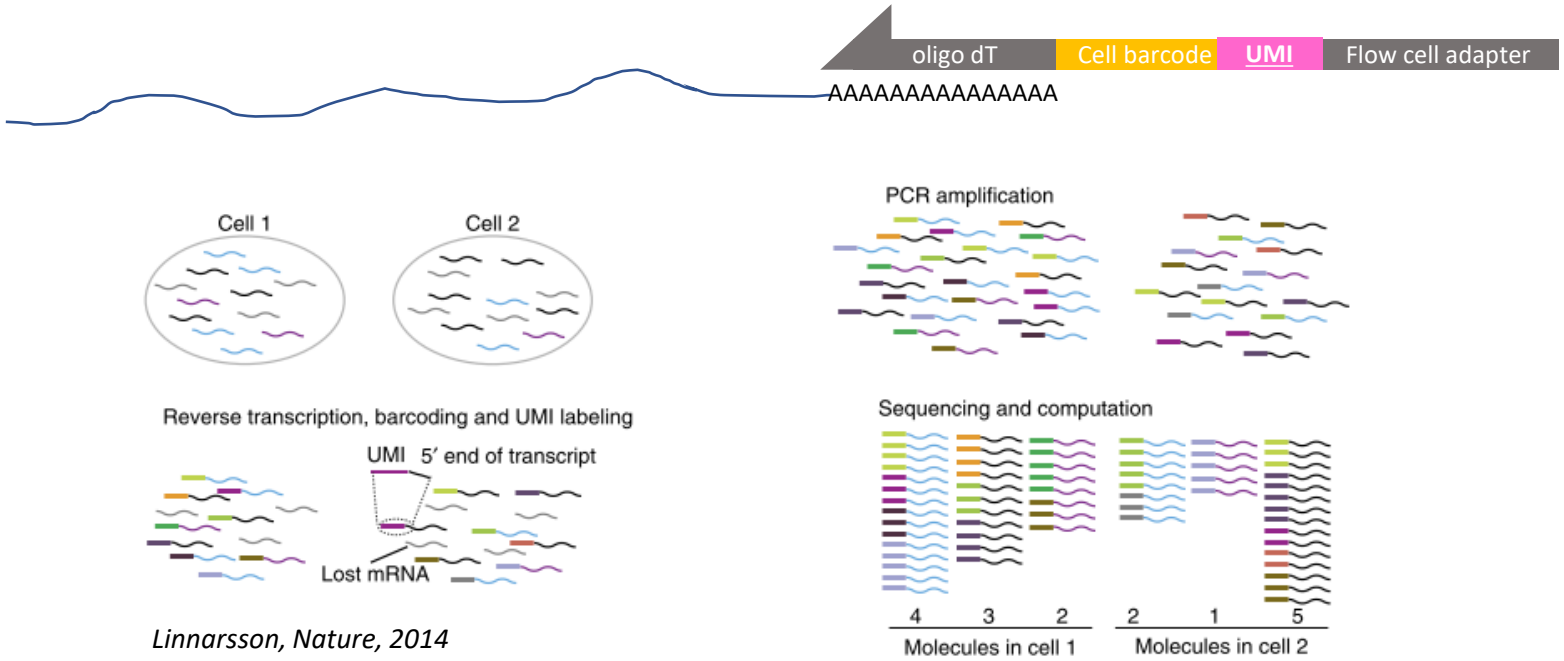
3’-end sequencing (mRNA read sequence needs to be “attached” to the cell barcode, so only 3’ end can be sequenced)



Single Cells scRNA-seq

10X Genomics Chromium
UMI

Very low starting amounts → PCR bias → solution: **UMI (unique molecular identifiers)**



Linnarsson, Nature, 2014

Single Cells scRNA-seq

10X Genomics Chromium

CellRanger user friendly reports, with QC and warnings

Estimated Number of Cells
3,168

Mean Reads per Cell	Median Genes per Cell
82,105	2,815

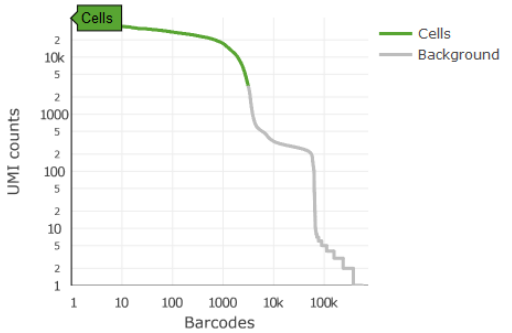
Sequencing

Number of Reads	260,111,191
Valid Barcodes	98.1%
Sequencing Saturation	66.1%
Q30 Bases in Barcode	97.3%
Q30 Bases in RNA Read	86.7%
Q30 Bases in Sample Index	94.9%
Q30 Bases in UMI	96.6%

Mapping

Reads Mapped to Genome	90.9%
Reads Mapped Confidently to Genome	85.4%
Reads Mapped Confidently to Intergenic Regions	3.0%
Reads Mapped Confidently to Intronic Regions	10.6%
Reads Mapped Confidently to Exonic Regions	71.8%
Reads Mapped Confidently to Transcriptome	69.8%
Reads Mapped Antisense to Gene	1.0%

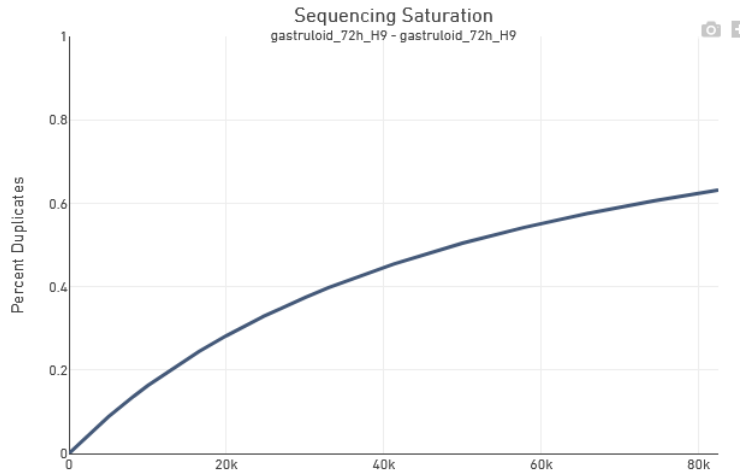
Cells

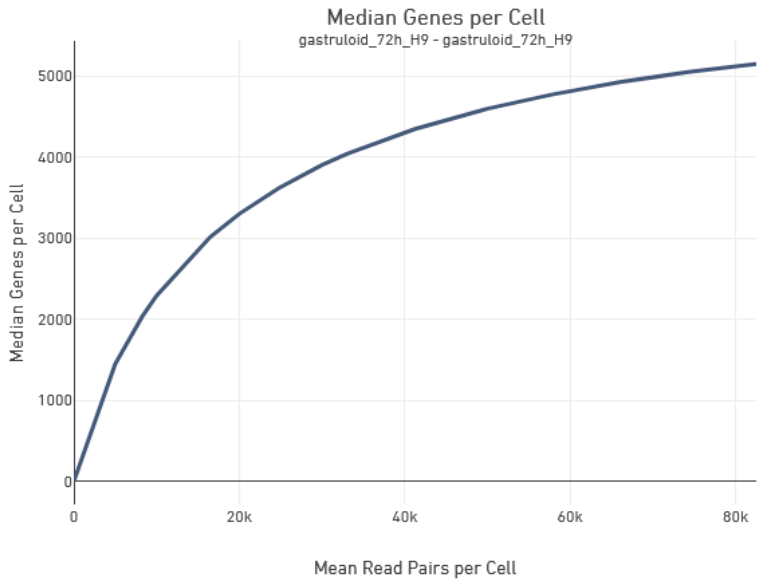


Estimated Number of Cells	3,168
Fraction Reads in Cells	73.0%
Mean Reads per Cell	82,105
Median Genes per Cell	2,815
Total Genes Detected	19,275
Median UMI Counts per Cell	12,221

Sample

Name	NXT0268_CellRang_P1
Description	
Transcriptome	mm10
Chemistry	Single Cell 3' v2
Cell Ranger Version	2.1.0





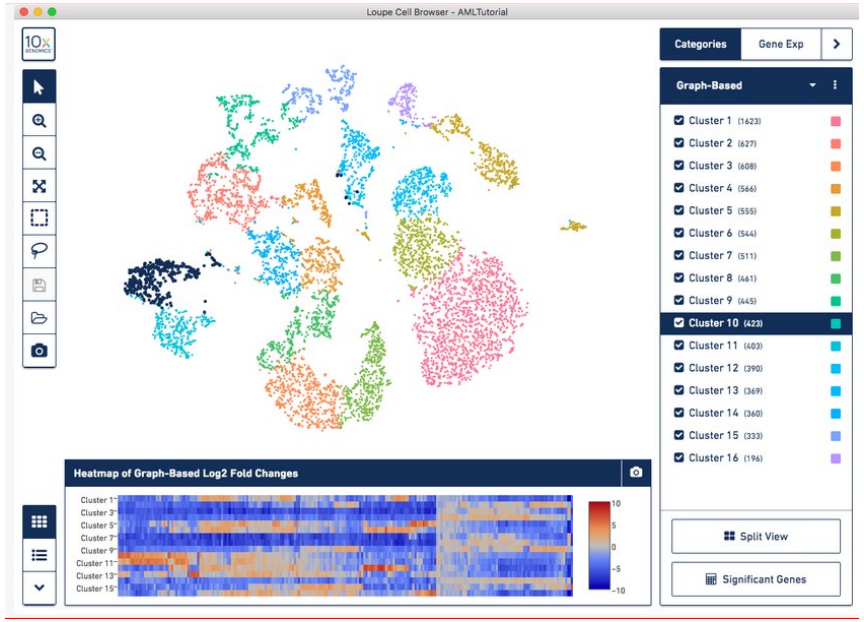
FYI

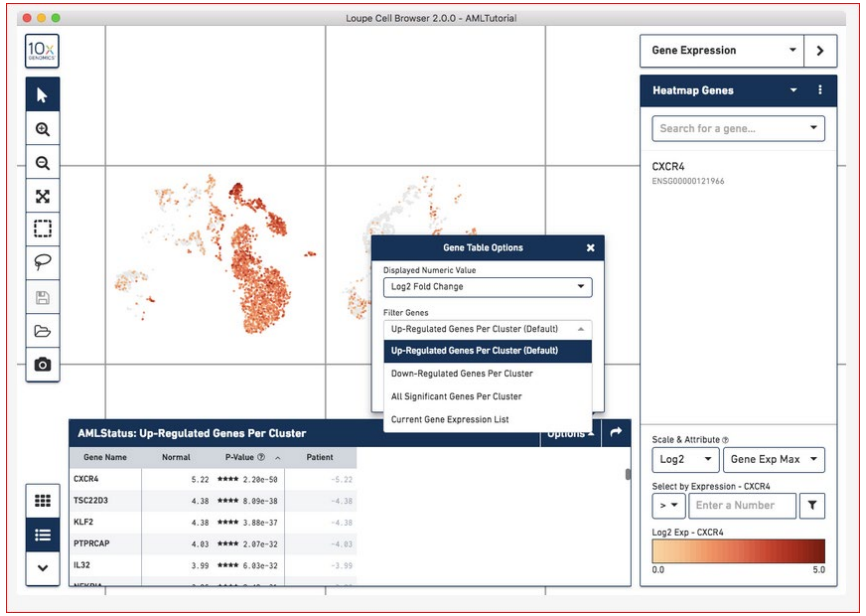
Single Cells scRNA-seq

10X Genomics Chromium

Loupe browser:

- user friendly
- clustering (rare populations)
- differential gene expression





FYI

Single Cells scRNA-seq

10X Genomics Chromium
Issues?

Typical issues (any single cell method):

- **Doublets** → fake subpopulations (specially if data have only 2-3 big clusters)? Cause?
  - Poor dissociation?
  - Or too many cells loaded?

Nb of Recovered Cells	Multiplet Rate (%)
500	~0.4%
1 000	~0.8%
2 000	~1.6%
3 000	~2.3%
4 000	~3.1%
5 000	~3.9%
6 000	~4.6%
7 000	~5.4%
8 000	~6.1%
9 000	~6.9%
10 000	~7.6%

  - Fake populations from **dying cells**? Cell cycle stage?... (look for mitochondrial gene reads for suffering)
  - Sub-population of tiny cells, or **debris**?
  - **dropout** (lowly expressed genes)