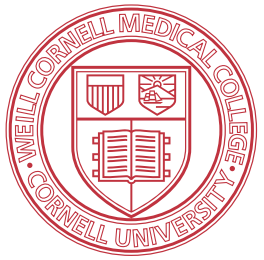
Differential gene expression analysis using RNA-seq

Applied Bioinformatics Core, November 2019



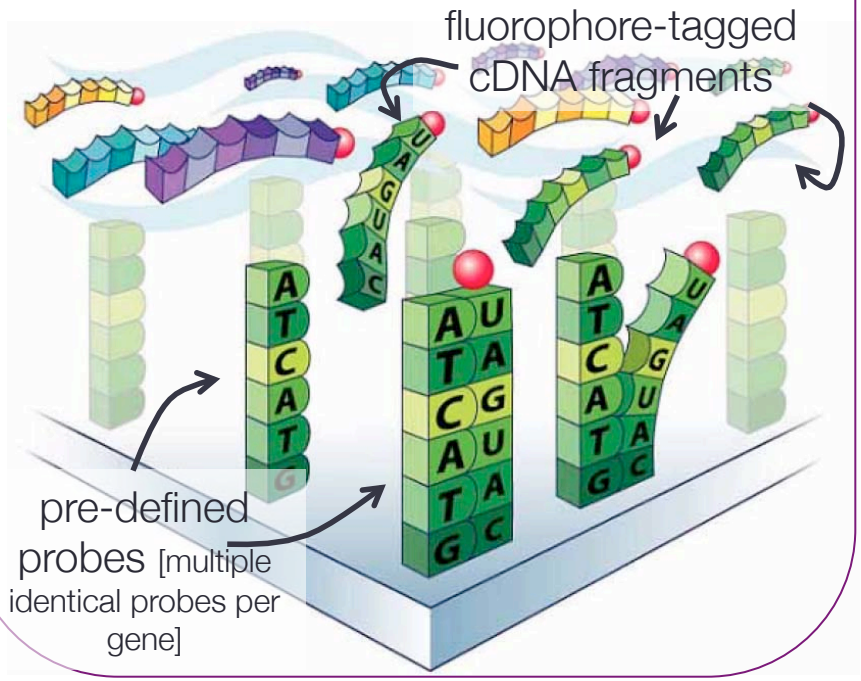
Friederike Dündar with Luce Skrabanek & Paul Zumbo

Day 1: Obtaining RNA-seq data

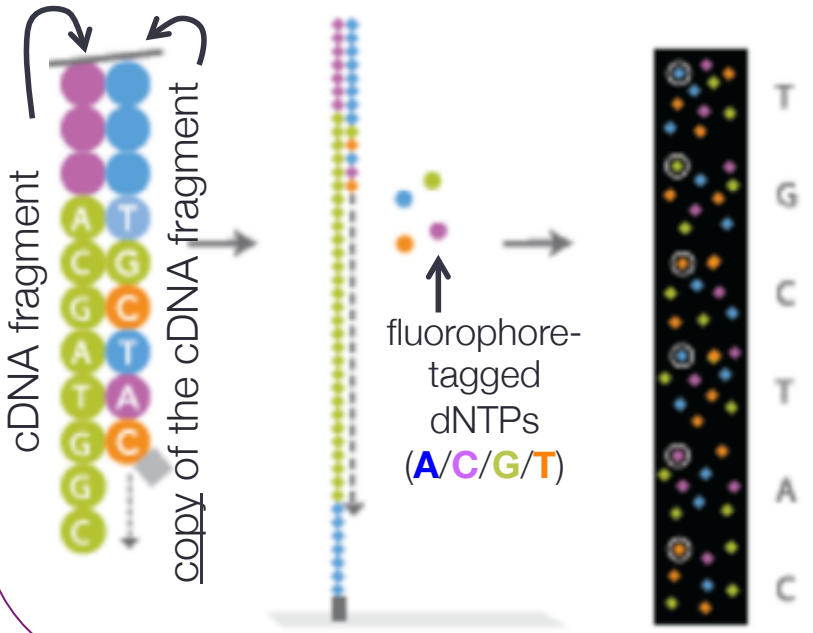
1. Introduction
2. Library preparation
3. Illumina-based sequencing
4. Raw data

Quantifying gene expression of thousands of genes:
Microarrays vs. “RNA” sequencing

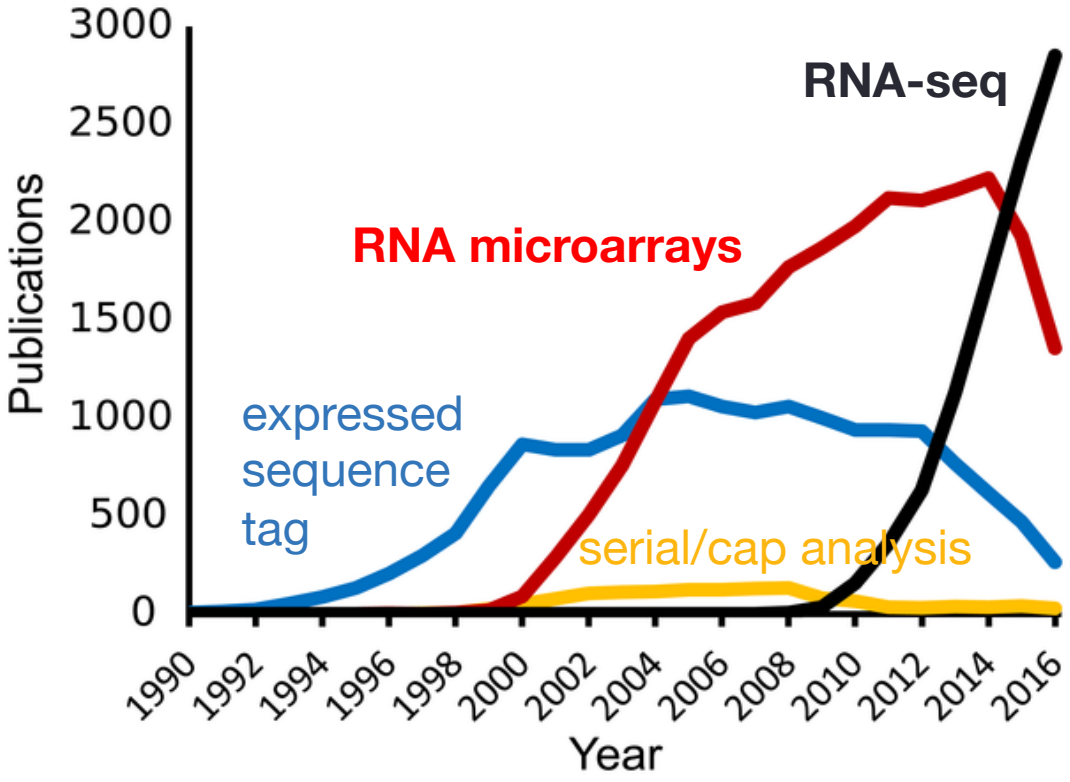
- fixed probes that capture DNA of **pre-defined sequences**
- the **read-out is fluorescence intensity** based on how many probes of a given sequence were able to **hybridize** with cDNA fragments



- **any cDNA fragment** can be sequenced
- the **read-out are the base identities** (= the actual cDNA sequence) inferred via DNA polymerase-based **amplification** of the captured fragments using labelled dNTPs



RNA-seq's rise goes hand-in-hand with progress of **high-throughput DNA sequencing**



- 1980s: Sanger sequencing, i.e. **one (1) DNA fragment** at a time
- 2000s: massively parallelized sequencing allows for **millions of DNA fragments** to be sequenced simultaneously

- short cDNA fragments (250-1,000bp), even shorter reads (= sequence info) (50-300bp)
  - requires **clonal amplification**
  - higher **error rates** than Sanger

"2nd generation" sequencing

Fig. from Lowe et al. (2017) Transcriptomics technologies. PLOS Comp Bio. doi: 10.1371/journal.pcbi.1005457

Evolution of HTS machines

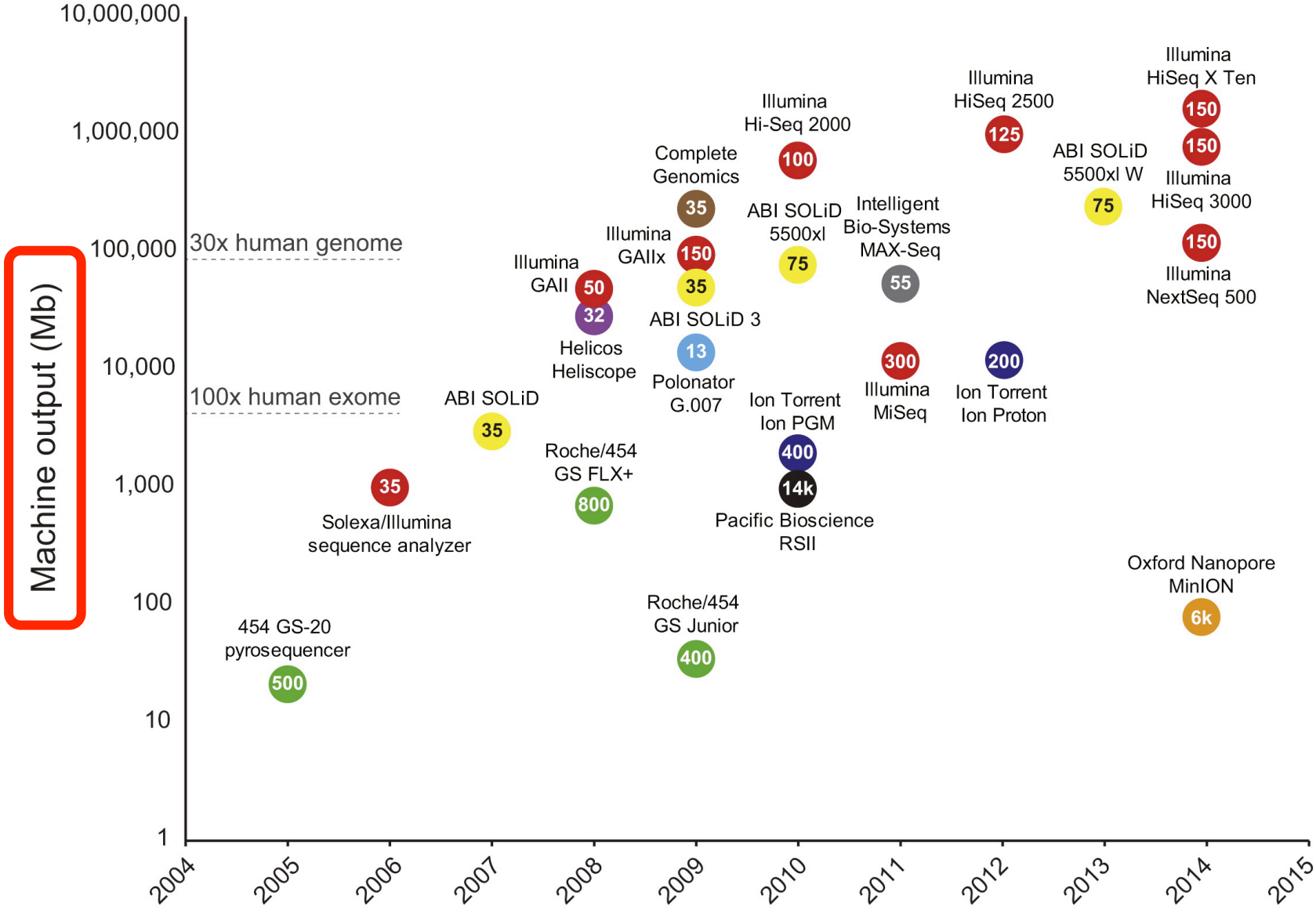


Fig. from Reuter et al. (2015). Molecular Cell, 58(4), 586–597.

The DNA sequencing revolution has boosted the need for bioinformatics

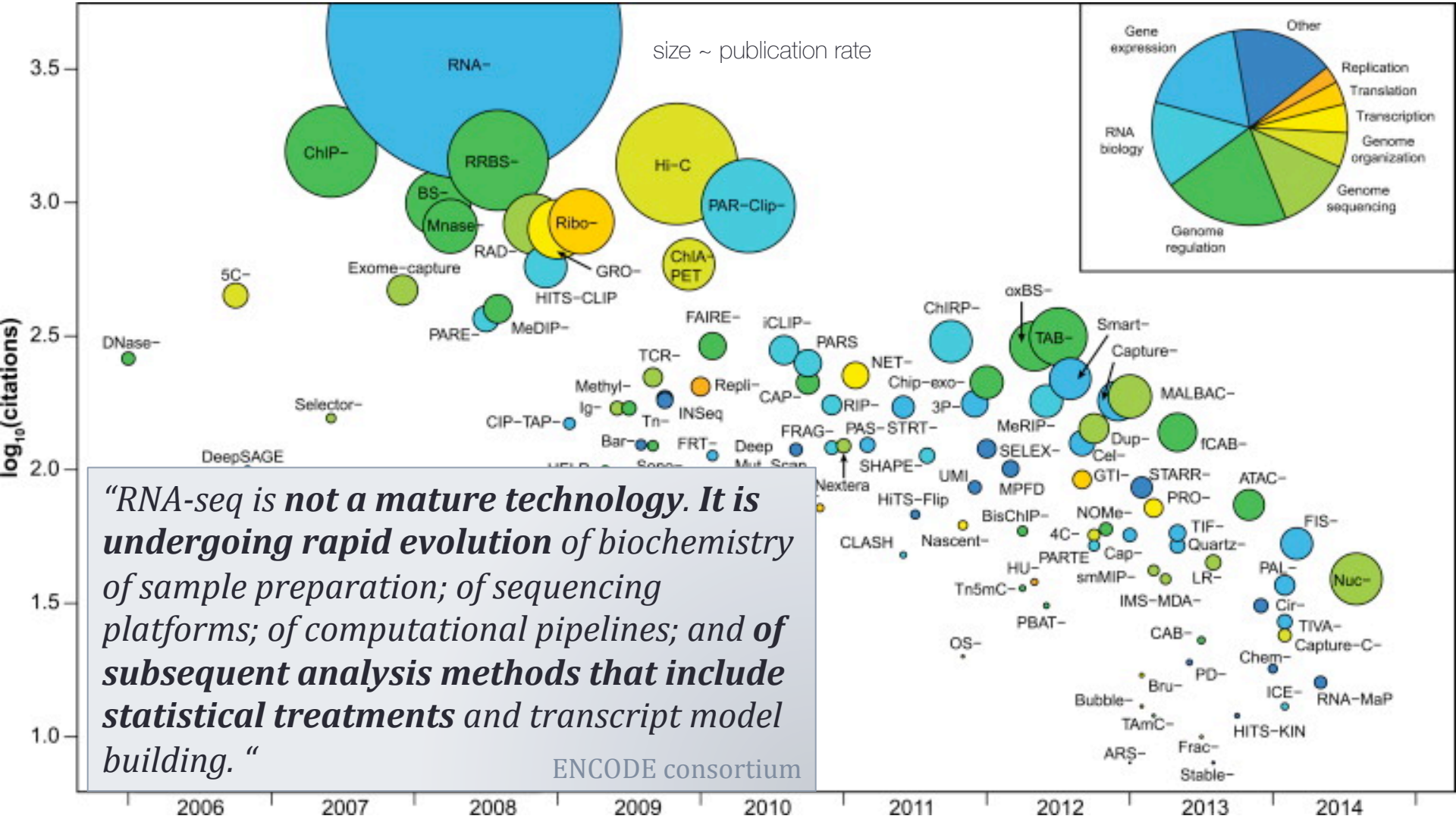
Next-generation sequencing	
High-throughput	millions of nucleotides can be sequenced at once
Relatively cheap*	experiments involving NGS have become abundant

relatively **large data files** are being generated on a regular basis

Bioinformatics
Processing
formatting, data wrangling
Alignment
Statistical analyses
Interpretation

*** The cost of analysis has remained high and is difficult to estimate!**

RNA-seq is popular, but constantly developing



*"RNA-seq is **not a mature technology. It is undergoing rapid evolution** of biochemistry of sample preparation; of sequencing platforms; of computational pipelines; and **of subsequent analysis methods that include statistical treatments** and transcript model building. "*

ENCODE consortium

Reuter et al. (2015). Mol Cell. Goodwin, McPherson & McCombie (2016). Nat Gen, 17(6), 333–351

RNA-seq "analysis paralysis"

Table 1 | **Selected examples of current RNA-based clinical tests**

RNA biomolecule	Method	Examples	Use
Viral RNA	qRT-PCR	• Influenza virus[68] • Dengue virus[69] • HIV[70] • Ebola virus[71]	Viral detection and typing
mRNA	qRT-PCR	• AlloMap (CareDx; heart transplant)[15,16] • Cancer Type ID (BioTheranostics)[143]	Diagnosis
	Microarray	Afirma Thyroid Nodule Assessment (Veracyte)[116]	Diagnosis
	qRT-PCR	• OncotypeDx (Genome Health; breast, prostate and colon cancer)[144–147] • Breast Cancer Index (BioTheranostics)[148] • Prolaris (Myriad; prostate cancer)[136]	Prognosis
	Digital barcoded mRNA analysis	Prosigna Breast Cancer Prognostic Gene Signature (Nanostring)[149]	Prognosis
	Microarray	• MammaPrint (Agendia; breast cancer)[134] • ColoPrint (Agendia; colon cancer)[150] • Decipher (Genome Dx; prostate cancer)[151]	Prognosis
miRNA	Microarray	Cancer Origin (Rosetta Genomics)[152]	Diagnosis
Fusion transcript	qRT-PCR	AML (*RUNX1–RUNX1T1*)[18]	Diagnosis
	qRT-PCR	*BCR–ABL1* (REF. 21)	Monitoring molecular response during therapy
	qRT-PCR (exosomal RNA)	ExoDx Lung (ALK) (Exosome Dx)[161]	Fusion detection
	RNA-seq	FoundationOne Heme[2,3]	Fusion detection

- basically no generally accepted standard reference (transcript definitions often change quarterly)
- myriad tools → highly complex & specialized "pipelines"

"The (…) flexibility and seemingly infinite set of options (…) have hindered its path to the clinic. (…) The **fixed nature of probe sets with microarrays or qRT-PCR offer an accelerated path** (…) without the lure of the latest and newest analysis methods."

Byron et al., 2016

Byron et al. Nat Rev Genetics (2016) doi: 10.1038/nrg.2016.10

RNA-seq platforms

Illumina almost has a *de facto* monopoly on high-throughput sequencing

Platform (Manufacturer)	Commercial release	Typical read length	Maximum throughput per run	Single read accuracy	RNA-Seq runs deposited in the NCBI SRA (Oct 2016) [74]
454 (Roche, Basel, Switzerland)	2005	700 bp	0.7 Gbp	99.9%	3548
Illumina (Illumina, San Diego, CA, USA)	2006	50–300 bp	900 Gbp	99.9%	362903
SOLiD (Thermo Fisher Scientific, Waltham, MA, USA)	2008	50 bp	320 Gbp	99.9%	7032
Ion Torrent (Thermo Fisher Scientific, Waltham, MA, USA)	2010	400 bp	30 Gbp	98%	1953
PacBio (Pacbio, Menlo Park, CA, USA)	2011	10,000 bp	2 Gbp	87%	160

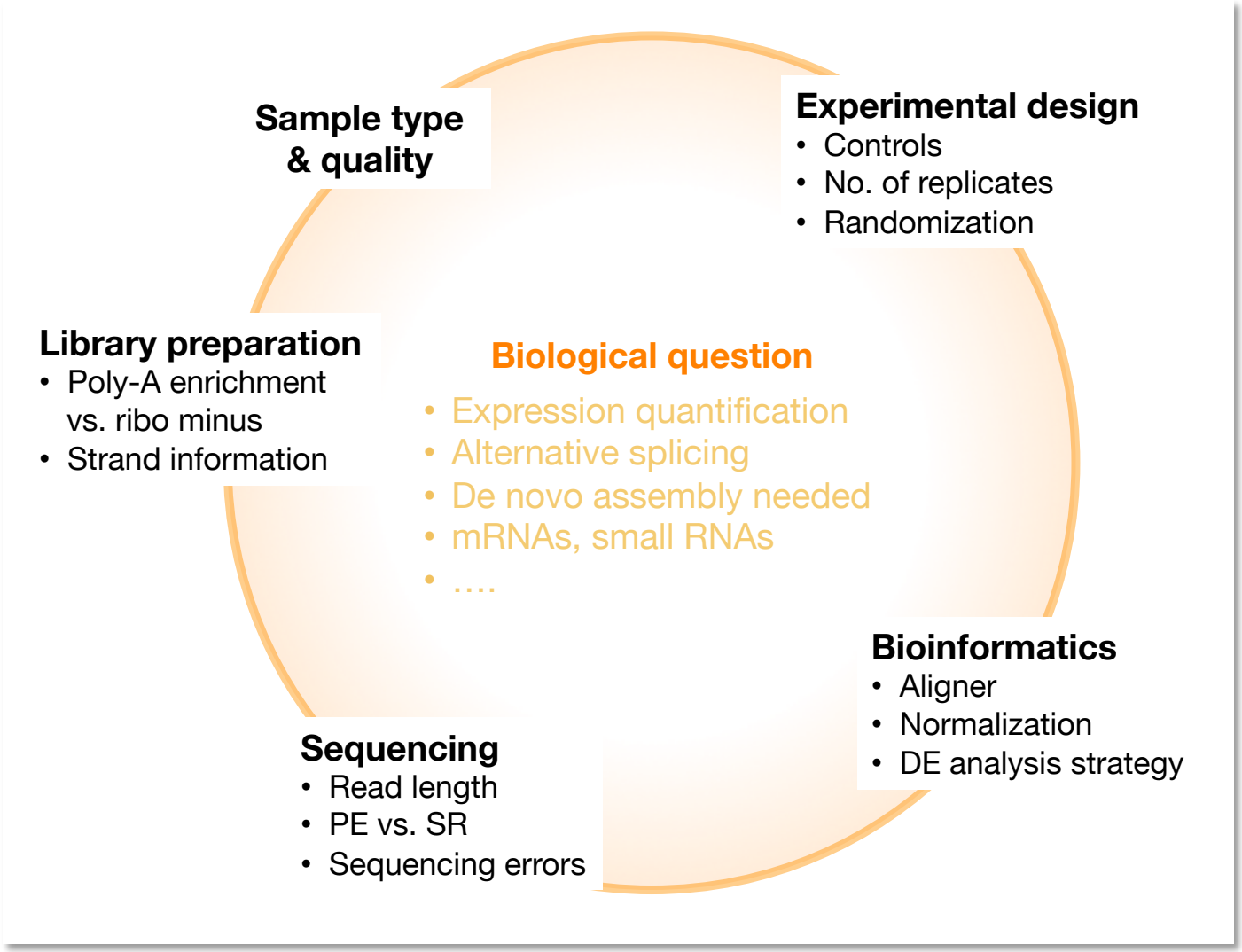
NCBI, National Center for Biotechnology Information; SRA, Sequence Read Archive; RNA-Seq, RNA sequencing.

https://doi.org/10.1371/journal.pcbi.1005457.t002

currently, all mainstream RNA-seq solutions rely on copying RNA into cDNA prior to sequencing

direct sequencing of RNA (Nanopore) is in its infancy, but will be useful to detect modified bases and avoid biases from the amplification steps

What to expect from the class



NOT COVERED:

- single-cell RNA-seq
- circular RNAs
- novel transcript discovery
- transcriptome assembly
- alternative splicing analysis

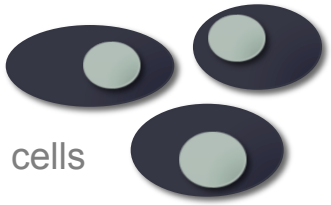
(see the course notes for references to useful reviews)

Day 1: Introduction into high-throughput sequencing

[many general concepts of NGS that are often not unique to RNA-seq]

1. RNA isolation & library preparation
2. Illumina's sequencing by synthesis
3. raw sequencing reads
    - download
    - quality control

RNA-seq workflow overview



Total RNA extraction

- mRNA enrichment
- fragmentation
- cDNA library preparation

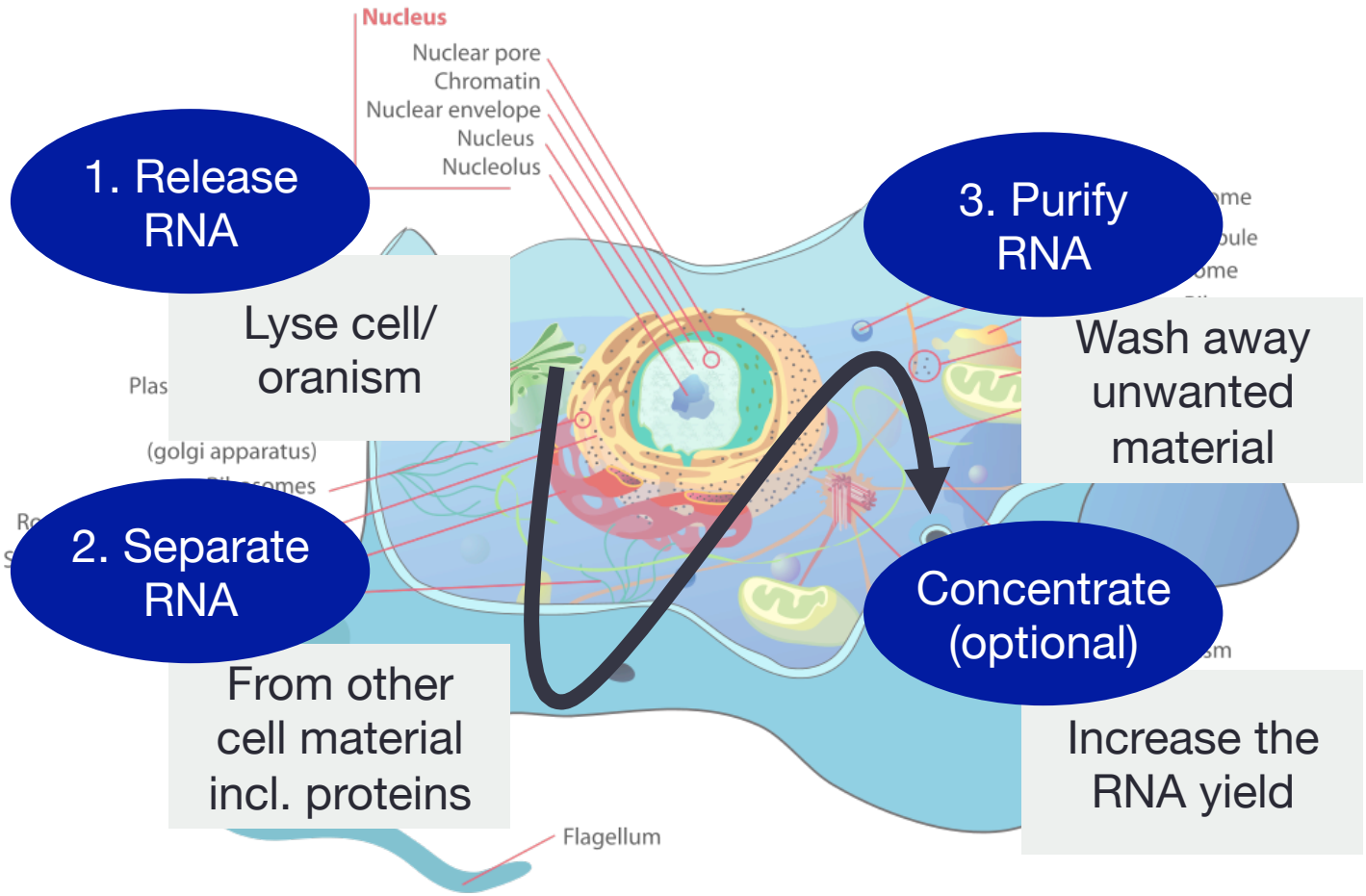
Sequencing

- cluster generation
- sequencing by synthesis
- image acquisition

Bioinformatics

RNA extraction

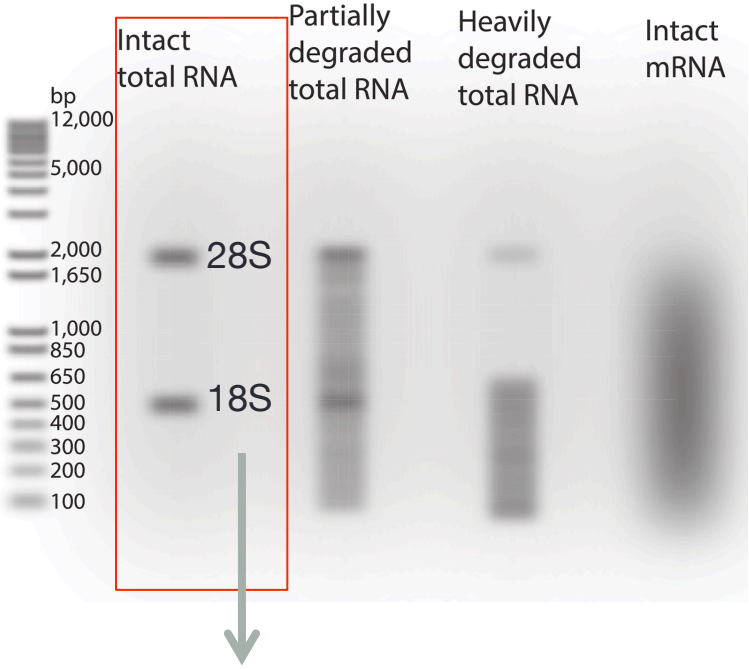
Goal: Extract **all** of the RNAs of the cells **without degrading** the molecules.



Thatcher (2015). DNA/RNA preparation for molecular detection. doi: 10.1373/clinchem.2014.221374

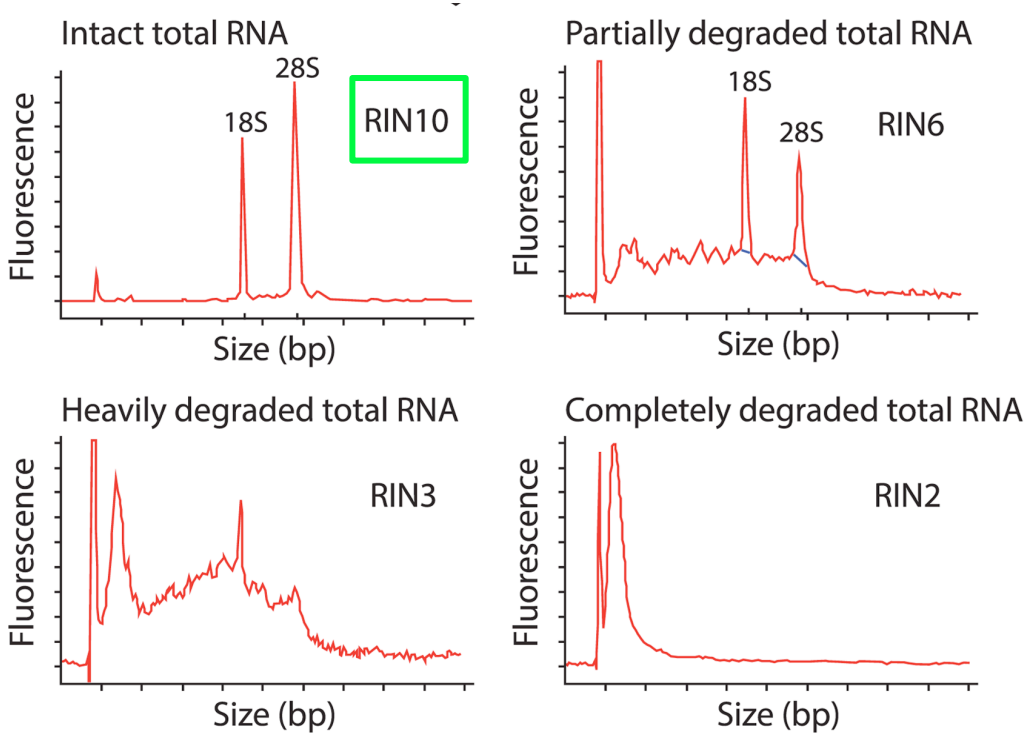
Quality control of extracted total RNA

Gel electrophoresis



RIN = 28S:18S ratio

avoid degraded RNA junk
optimum: RIN = 10



Use the expertise of the sequencing facility staff! They've seen it all!

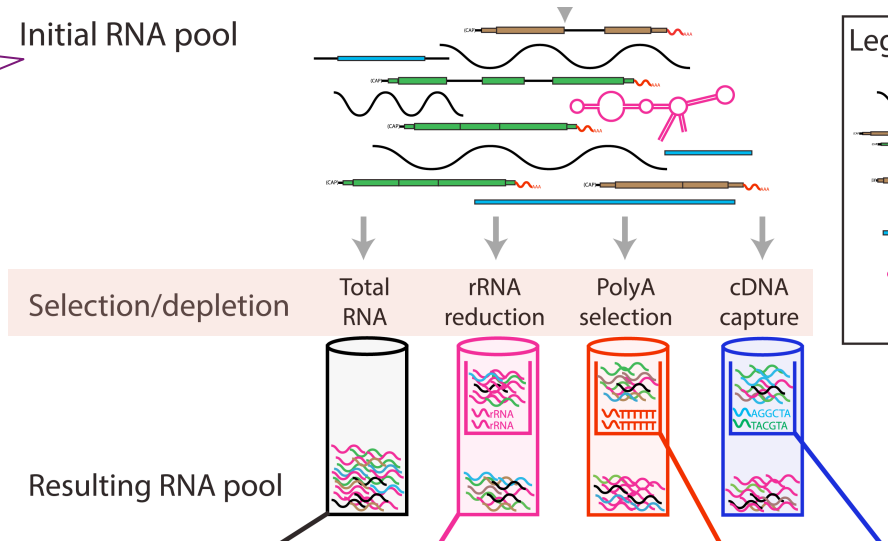
Figure from Griffith et al., PLoS Comp Biol (2015). doi: 10.1371/journal.pcbi.1004393

Influence of the **RNA enrichment** strategy

mostly
rRNA & tRNA
< 2% mRNA!

which transcripts are you interested in?

what type of noise can you tolerate?



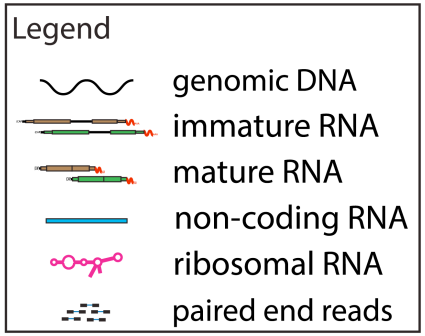
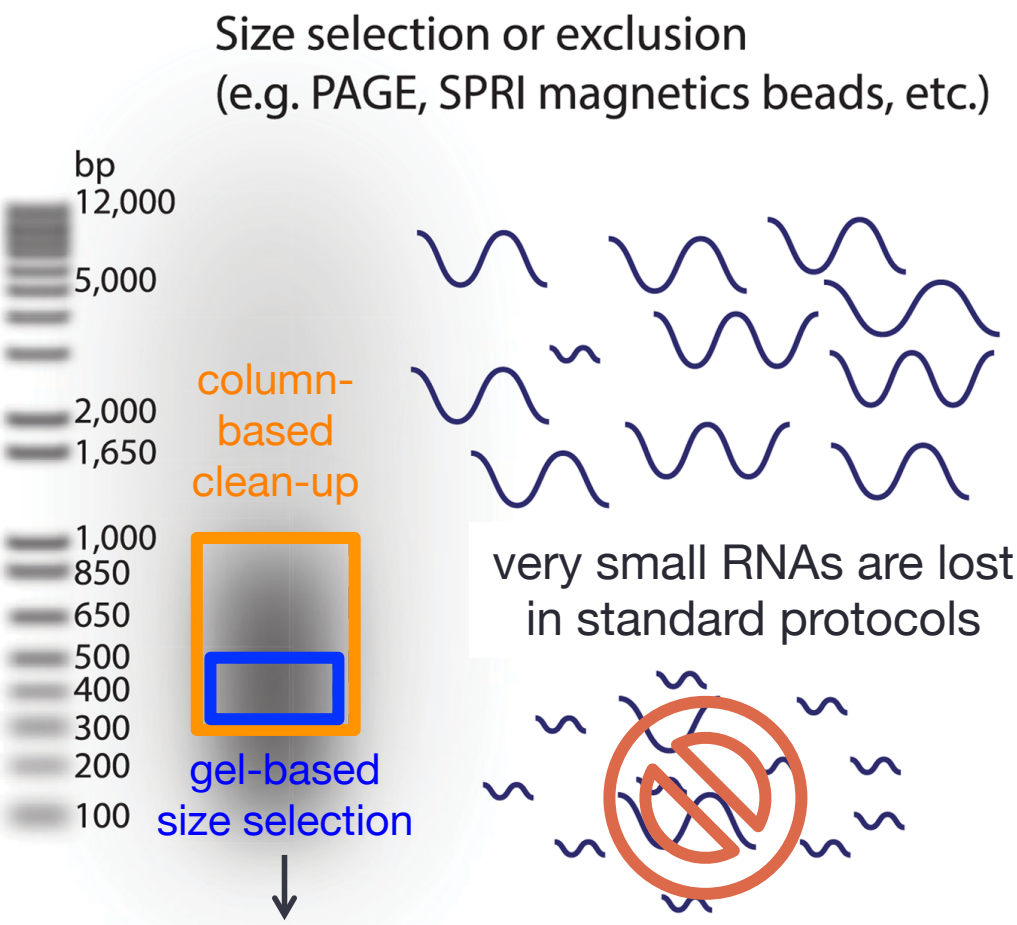
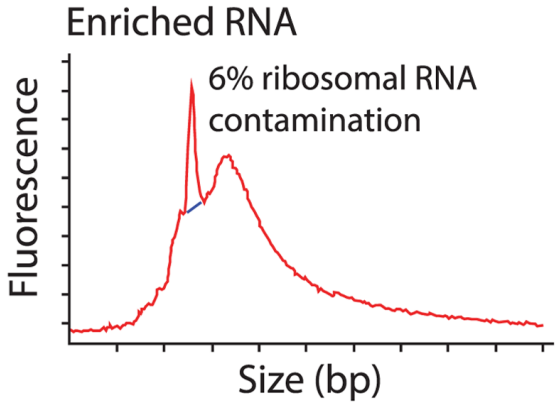
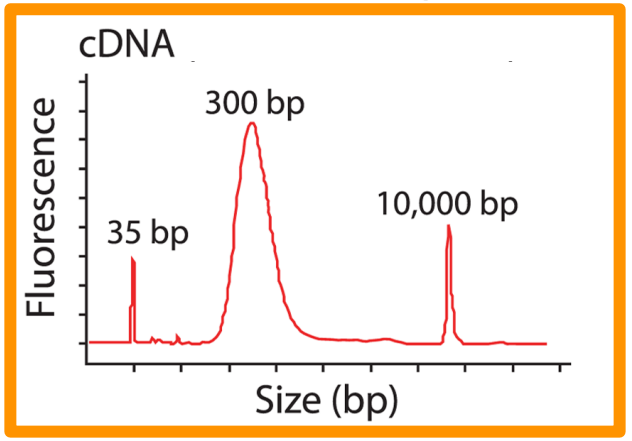


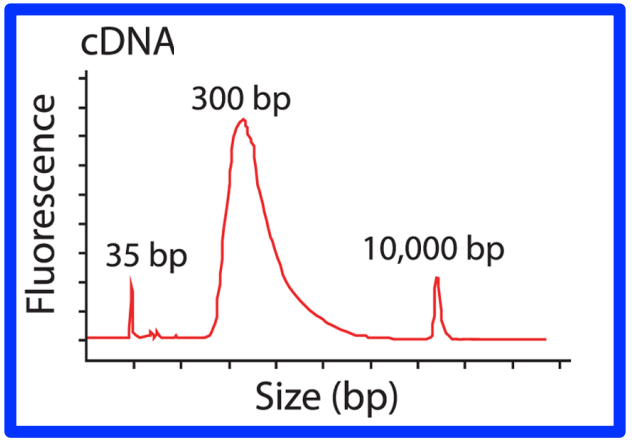
Figure from Griffith et al., PLoS Comp Biol (2015). doi: 10.1371/journal.pcbi.1004393

Size selection: Illumina needs short fragments! (but not too short)



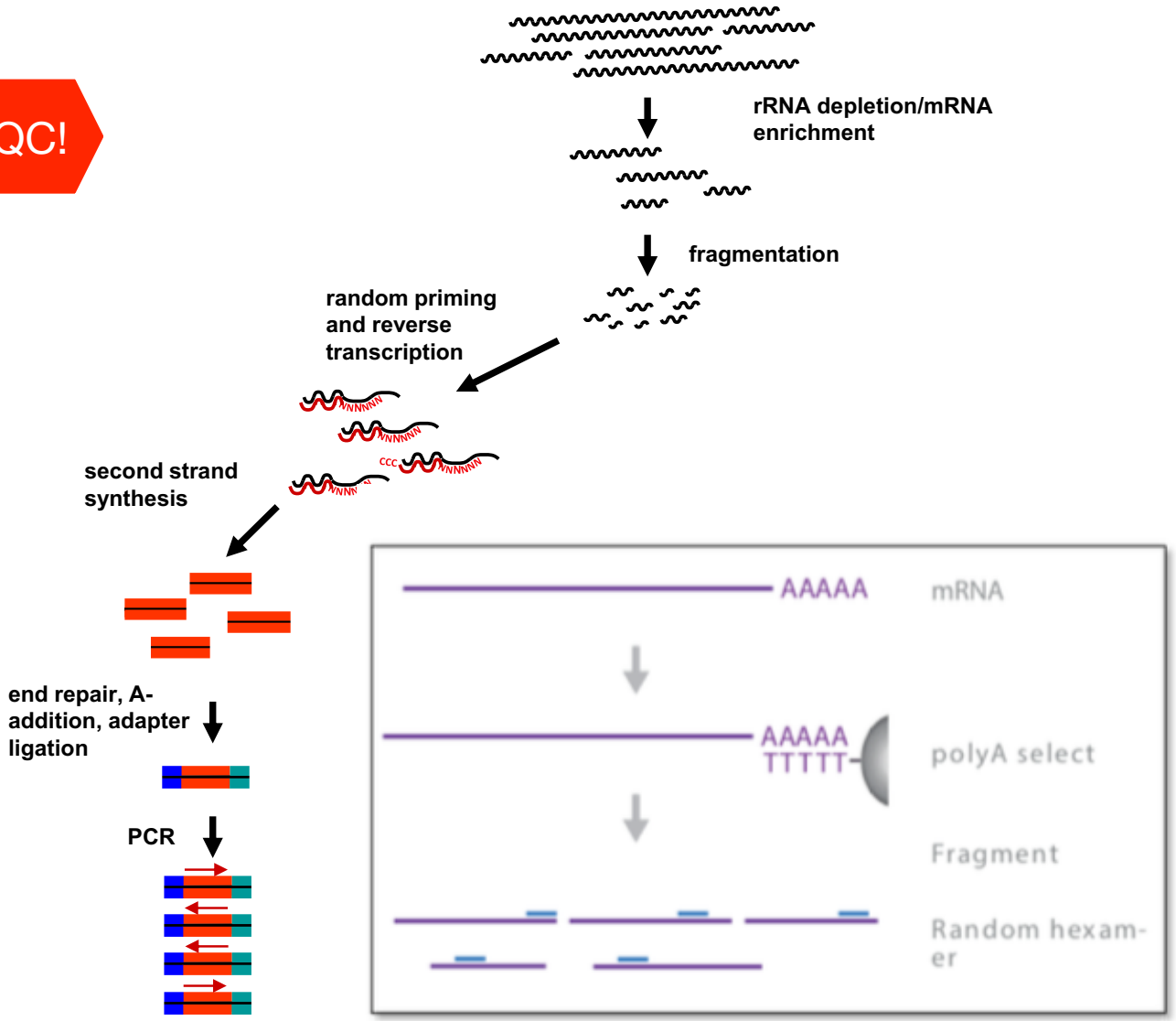






RNA-seq library preparation: pick one!

QC!



classical Illumina protocol (unstranded)

RNA extraction

↓

mRNA enrichment

↓

fragmentation (~200 bp)

↓

cDNA synthesis

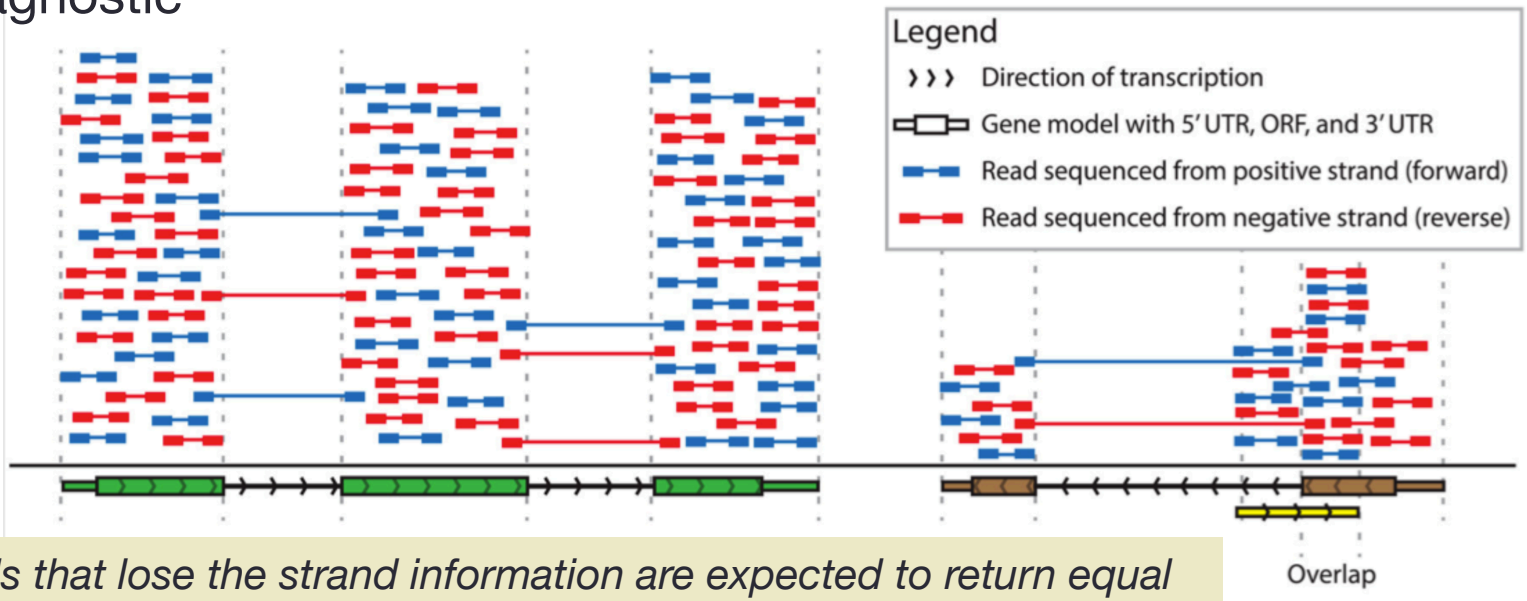
↓

cDNA plus sequencing adapters

Van Dijk et al. (2014). Experimental Cell Research, 322(1), 12–20. doi:10.1016/j.yexcr.2014.01.008

Unstranded vs. stranded protocols

strand-agnostic



protocols that lose the strand information are expected to return equal fractions of fwd/rev reads because they originate from ds-cDNA

strand-specific

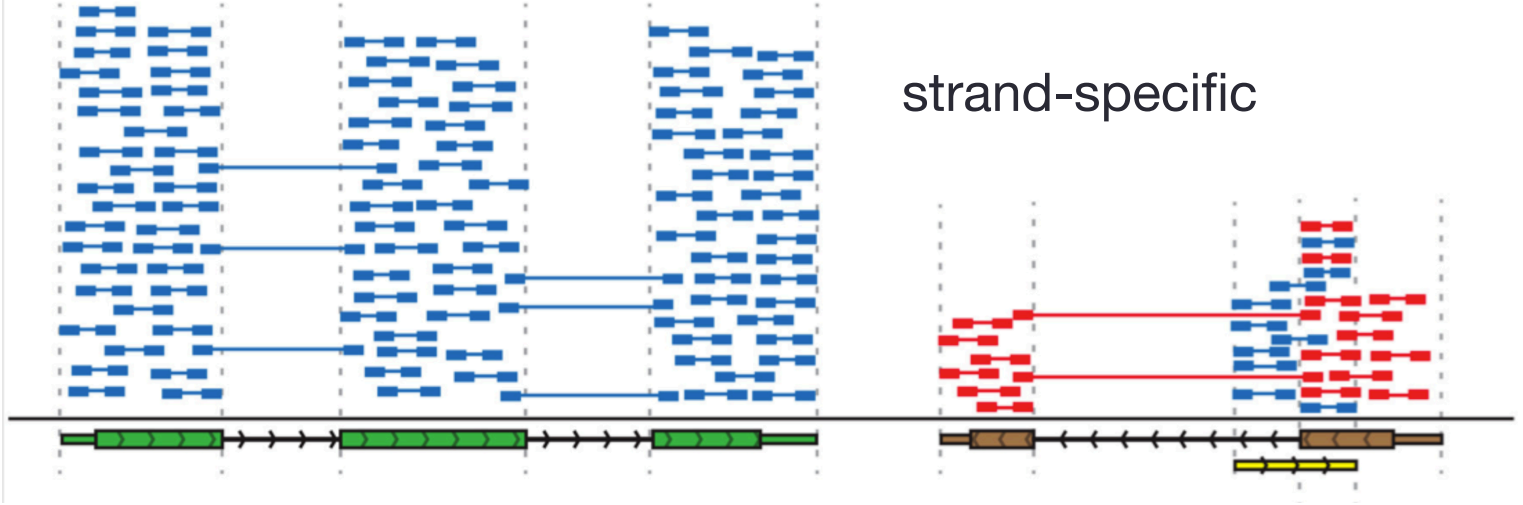
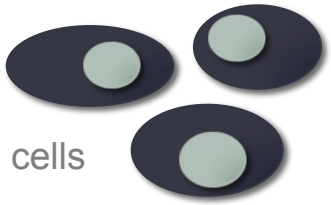


Fig. from Griffith et al. (2015) doi:10.1371/journal.pcbi.1004393

RNA-seq workflow overview



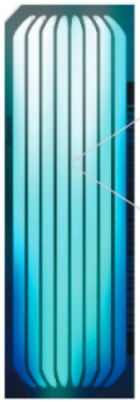
cells

Total RNA extraction

RNA

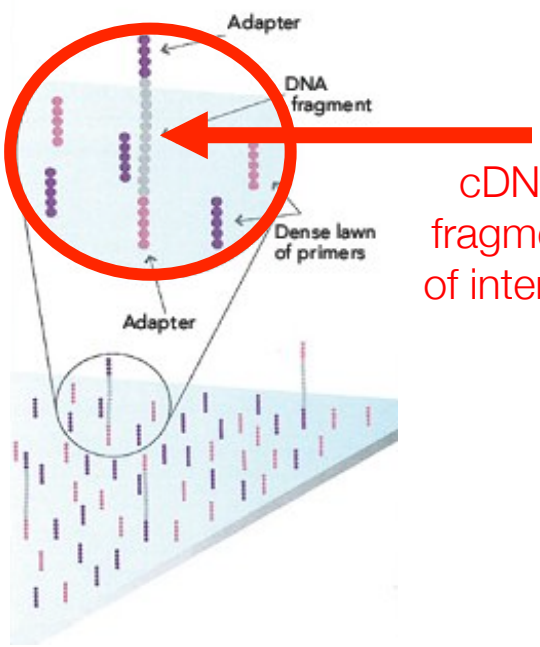
fragments of the RNAs of interest

cDNA with adapters



flowcell coated with universal adapter oligos

Sequencing

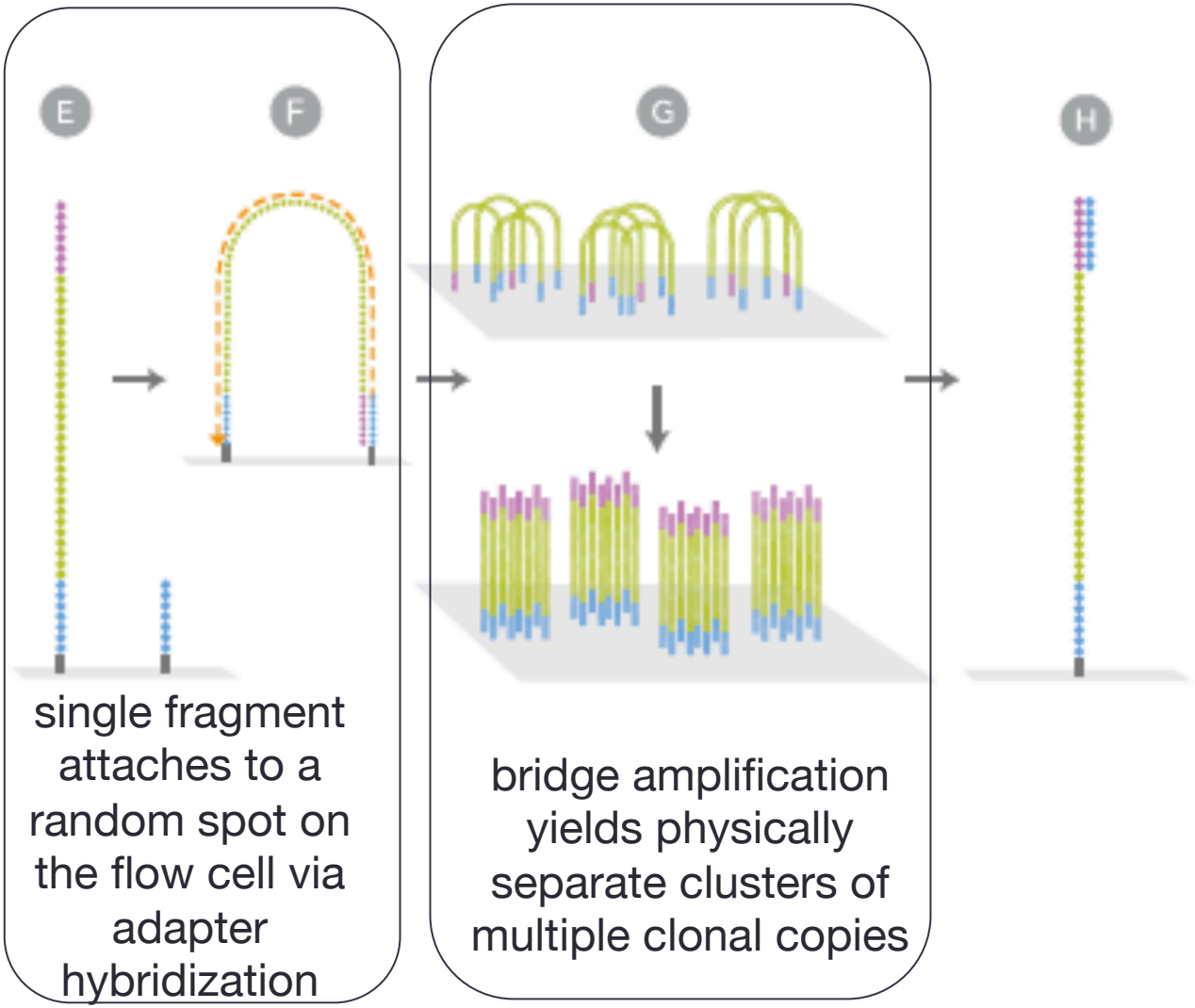


cDNA fragment of interest

flowcell with primers

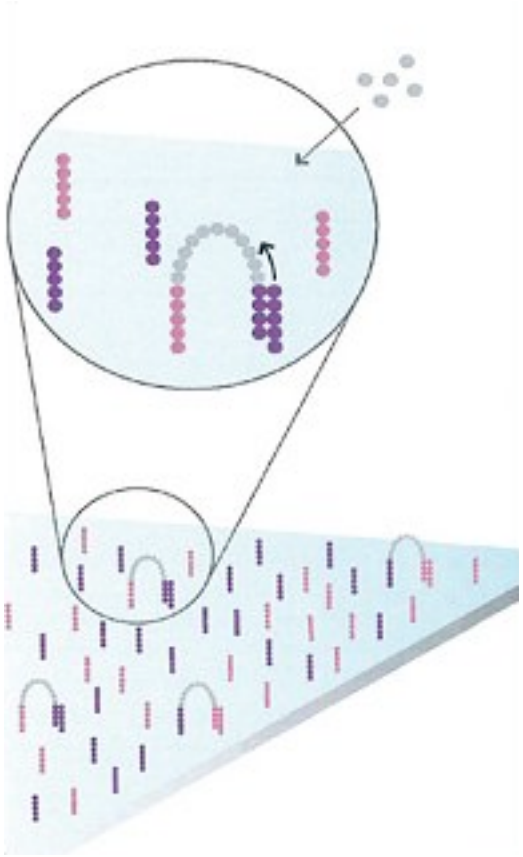
Cluster generation

= physically separate clusters of identical **clonal copies** of individual fragments

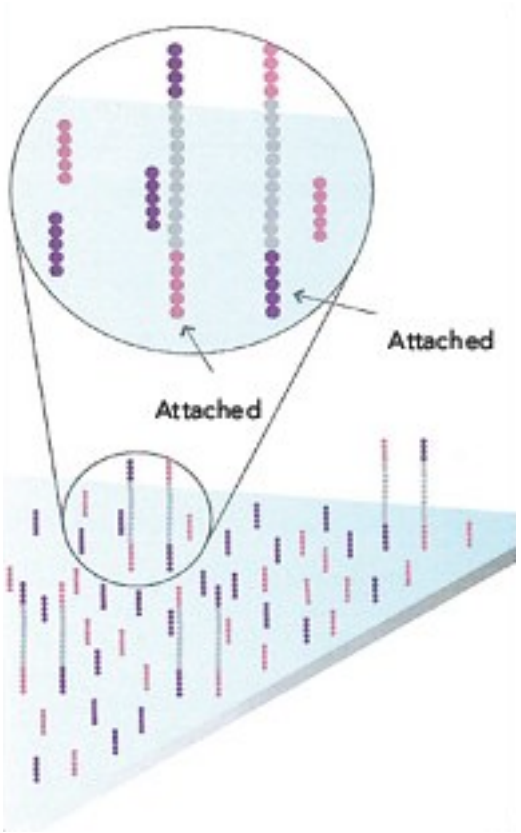


Cluster generation

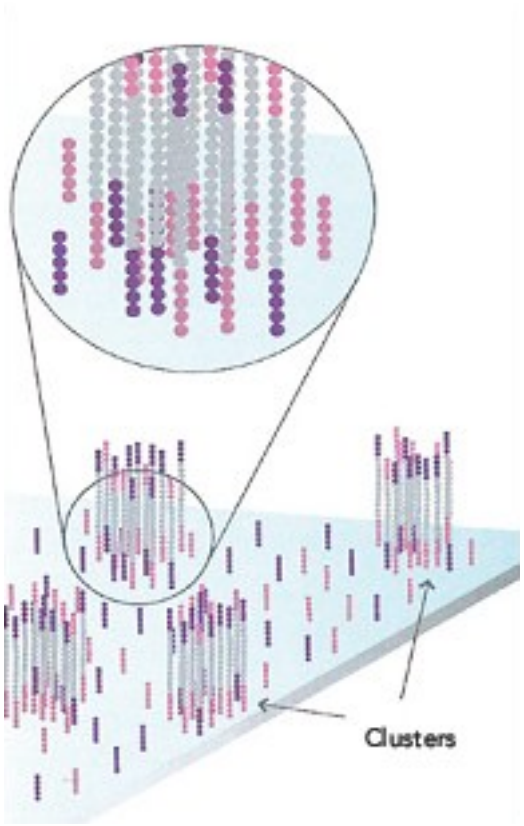
= physically separate clusters of identical **clonal copies** of individual fragments



bridge amplification

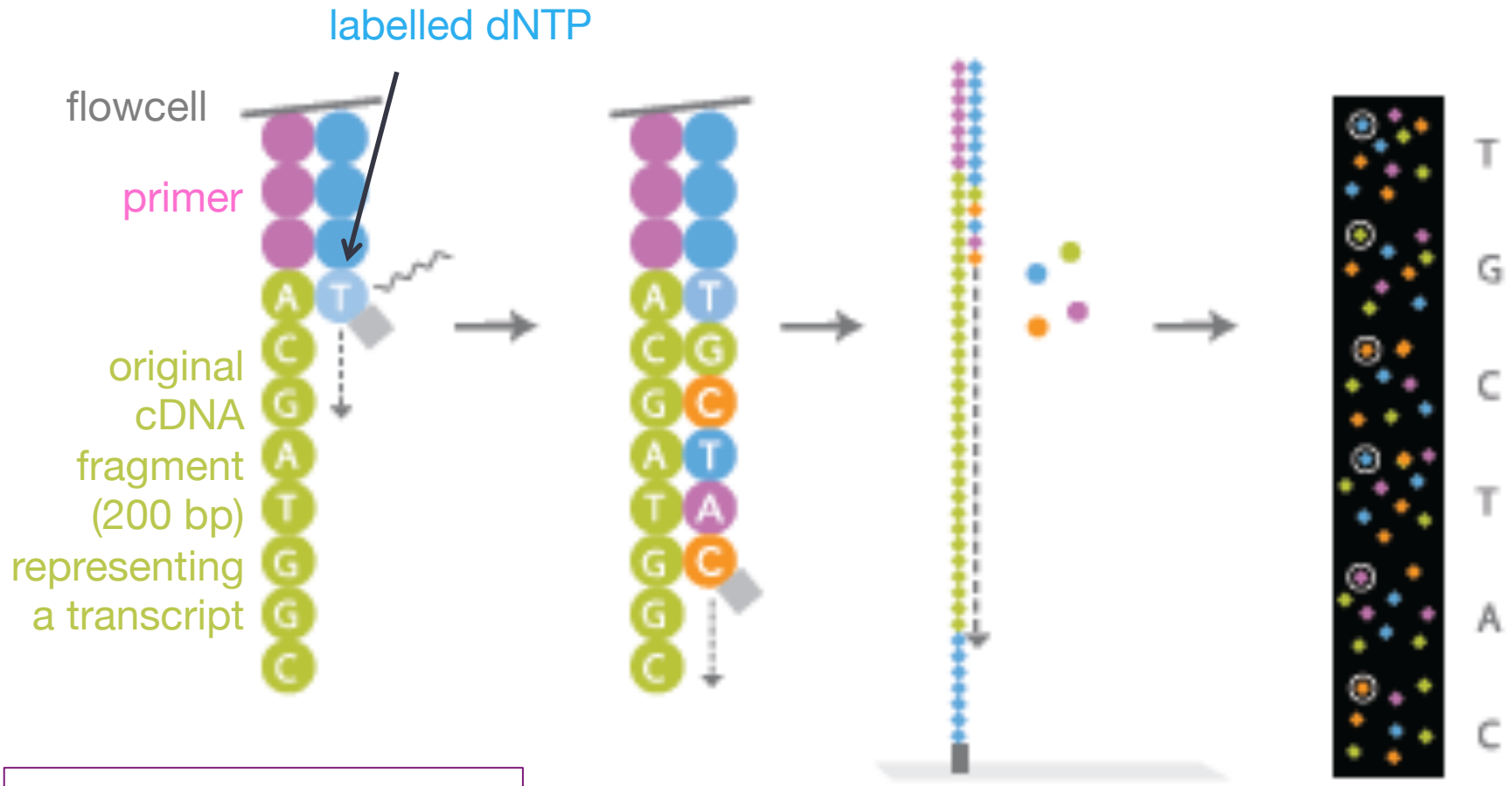


denaturation



cluster generation
removal of complementary strands → identical fragment copies remain

Sequencing by synthesis



1. extend 1st base
2. read (excite & capture image)
3. de-block

repeat for 50 – 100 cycles (= length of read)

generate base calls from the images

Image from Illumina documentation

Typical biases of Illumina sequencing

- sequencing errors
- miscalled bases

specific for the sequencing platform/ machine

- **PCR artifacts (library preparation)**
  - duplicates (due to low amounts of starting material)
  - length bias
  - GC bias

sample-specific problems!

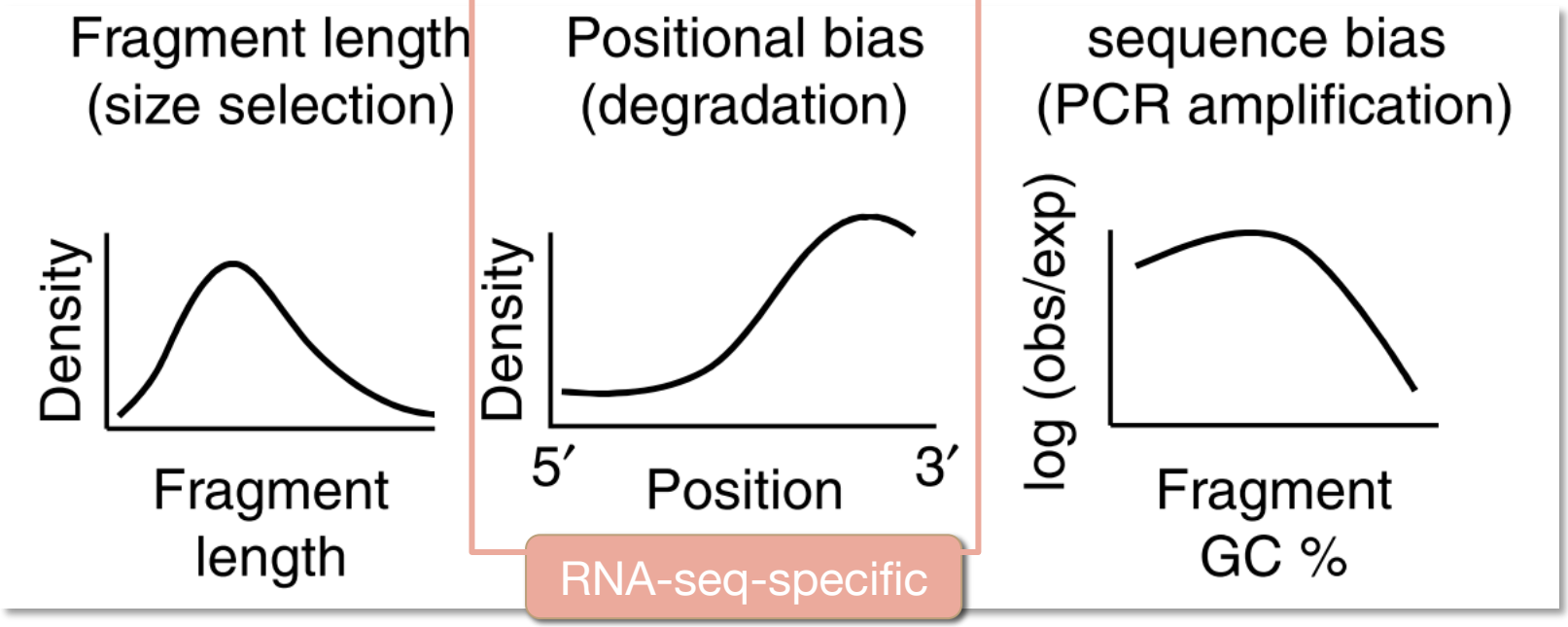


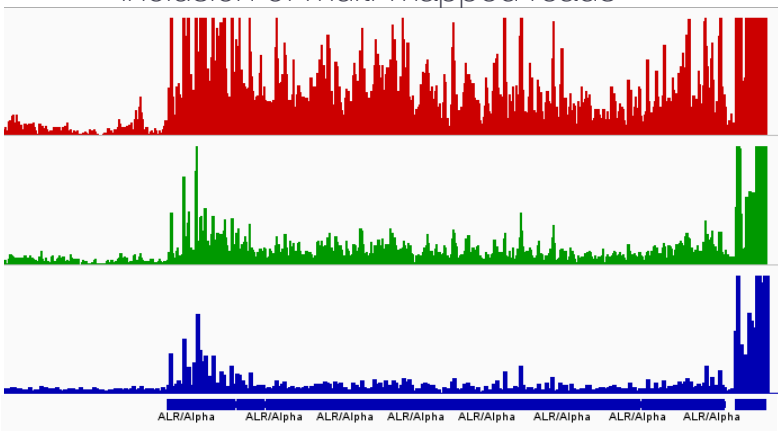
Figure from Love et al. (2016). Nat Biotech, 34(12). More details & refs in course notes (esp. Table 6).

General sources of biases (not inherently sample-specific)

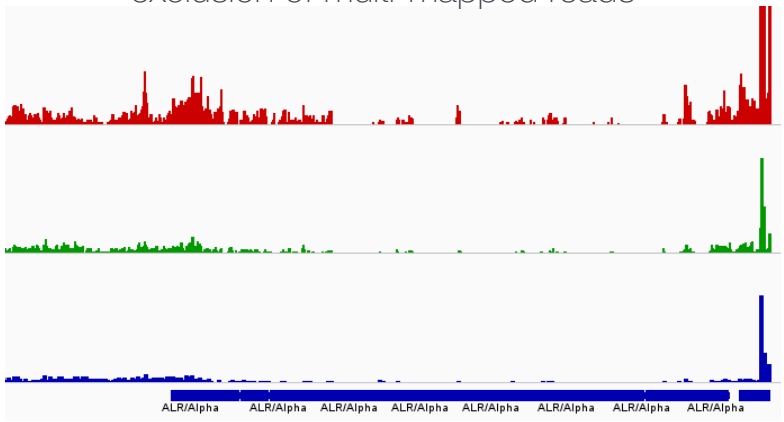
- issues with the **reference**
  - CNV
  - mappability

- inappropriate **data processing**

inclusion of multi-mapped reads



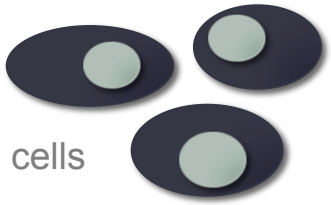
exclusion of multi-mapped reads



RAW SEQUENCING READS

Let the data wrangling begin!

RNA-seq workflow overview



Total RNA extraction

- mRNA enrichment
- fragmentation
- cDNA library

Sequencing

- cluster generation
- sequencing by synthesis
- image acquisition

Bioinformatics

Bioinformatics workflow of RNA-seq analysis

Images
`.tif`

Base calling & demultiplexing
`Bustard/RTA/OLB, CASAVA`

FASTQC► Raw reads
`.fastq`

Mapping
`STAR`

Aligned reads
`.sam/.bam`

Counting
`featureCounts`

Read count table
`.txt`

Normalizing
`DESeq2, edgeR`

Normalized read count table
`.Robj`

DE test & multiple testing correction
`DESeq2, edgeR, limma`

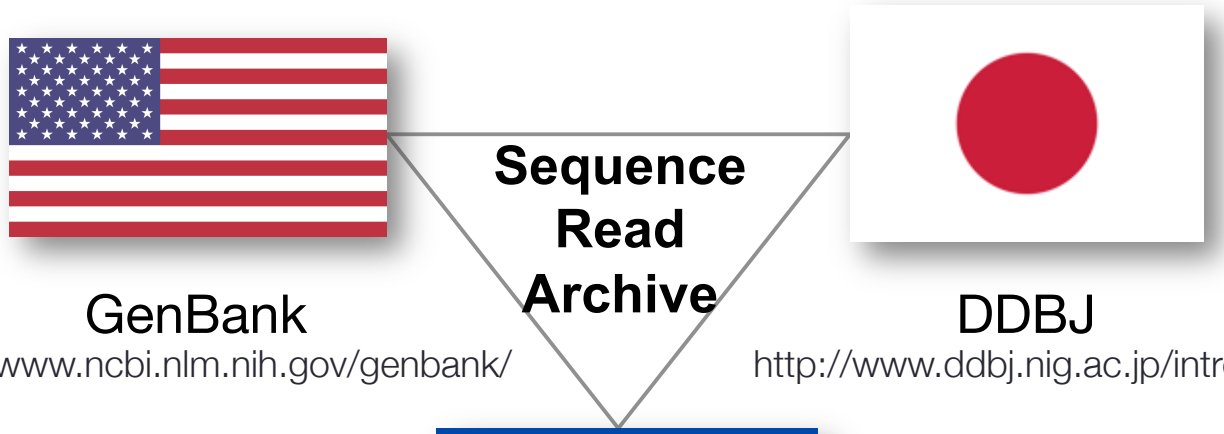
List of fold changes & statistical values
`.Robj, .txt`

Filtering
`Customized scripts`

Downstream analyses on DE genes

Where are all the reads?

The sequence read archive (**SRA)** is the main repository for publicly available DNA and RNA sequencing data of which 3 instances are maintained world-wide.



GenBank
http://www.ncbi.nlm.nih.gov/genbank/

DDBJ
http://www.ddbj.nig.ac.jp/intro-e.html

ENA
https://www.ebi.ac.uk/ena/

The different mirrors provide different routes for browsing & downloading the data.

Detailed information about the SRA: O'Sullivan et al. (2018) Managing Sequence Data. In *Bioinformatics: Data, Sequence Analysis, and Evolution*. doi: 10.1007/978-1-4939-6622-6_4

Public RNA-seq resources

Collection	Number of samples/ libraries	Reference
TCGA	12,000	Cancer Genome Atlas Research Network. 2013. PMID: 24071849
GTEX	11,000	Carither et al., 2015. PMID: 26484571
Human Protein Atlas	8,00	Uhlen et al. 2015. PMID: 25613900
ENCODE	2,300	David et al. 2018. PMID: 29126249
GEUVADIS	1,100	Lappalainen et al. 2013. PMID: 24037378
Cancer Cell Line Encyclopedia	650	Barretina et al. 2012. PMID: 22460905
Leucegene (AML focus)	550	Lavalée et al. 2018. PMID: 29550835

The **recount** resource offers **processed read counts** of >2,000 different studies!
`https://jhubiostatistics.shinyapps.io/recount/`

MetaSRA

http://metasra.biostat.wisc.edu/

Normalized metadata for the Sequence Read Archive

Find human RNA-seq samples

RESET

matching **all** of these terms: Find term

but **none** of these terms: Find term

Sample type: All | cell line | tissue | primary cells | stem cells | in vitro differentiated cells | iPS cell line

Examples

- **Find healthy liver tissue:** require liver, exclude disease and treatment. Sample type: tissue.
- **Find healthy, primary T-cells:** require T cell, exclude disease and treatment. Sample type: primary cells.
- **Find glioblastoma samples:** require glioblastoma multiforme and brain.

Key: Anatomy | Disease | Cell Line | Cell Type | Experimental Factor

Let's download!

- We will work with a data set submitted by Gierlinski et al.
  - 2 conditions: SNF2 (knock-out), WT
  - 48 biological replicates with 7 technical replicates each

- they deposited the sequence files with SRA – we will retrieve it via **ENA** (https://www.ebi.ac.uk/ena/)

- accession number of the Gierlinski data: ERP004763

Course notes @ **`https://chagall.med.cornell.edu/RNASEQcourse/`**

See **Section 2 (Raw Data)** for download instructions etc.

```
ls
mkdir
wget
cut
grep
awk
```

Excellent tutorial on speeding up the download from SRA:
`https://www.biostars.org/p/325010/`

Gierliński et al. (2015). Bioinformatics, 31(22), 3625–3630. & Schurch et al. (2016) RNA.

General workflow for the data download

1. Find the **samples** that belong to the study by Gierlinski et al. using the accession number given in the publication (ERP004763)
   - Figure out which sample name belongs to what type of sample. Gierlinski et al. provide that via a table on figshare.
2. Extract the web **links** to the individual files. (`cut`/`awk`/…)
3. **Download** each file. (`wget` or `curl`)
   - technical replicates of the same biological sample should be placed into the same folder

Downloading a batch of fastq files

https://www.ebi.ac.uk/ena/ → study ERP004763

1. get link with list of **ftp sites** for every file: right-click on "TEXT" → "copy link location"

2. **download** on server/via CL: copy and paste to `wget` (mind the quotation marks to keep the link intact!):

```
wget -O samples_at_ENA.txt "<LINK>"
```

get the **sample information**:

```
wget -O ERP004763_sample_mapping.tsv --no-check-certificate "https://ndownloader.figshare.com/files/2194841"
```

```
$ cut -f11 samples_at_ENA.txt | head
fastq_galaxy
ftp.sra.ebi.ac.uk/vol1/fastq/ERR458/ERR458493/ERR458493.fastq.gz
ftp.sra.ebi.ac.uk/vol1/fastq/ERR458/ERR458494/ERR458494.fastq.gz
ftp.sra.ebi.ac.uk/vol1/fastq/ERR458/ERR458495/ERR458495.fastq.gz
ftp.sra.ebi.ac.uk/vol1/fastq/ERR458/ERR458496/ERR458496.fastq.gz
ftp.sra.ebi.ac.uk/vol1/fastq/ERR458/ERR458497/ERR458497.fastq.gz
ftp.sra.ebi.ac.uk/vol1/fastq/ERR458/ERR458498/ERR458498.fastq.gz
ftp.sra.ebi.ac.uk/vol1/fastq/ERR458/ERR458499/ERR458499.fastq.gz
ftp.sra.ebi.ac.uk/vol1/fastq/ERR458/ERR458500/ERR458500.fastq.gz
ftp.sra.ebi.ac.uk/vol1/fastq/ERR458/ERR458501/ERR458501.fastq.gz
```

list of links

```
$ head ERP004763_sample_mapping.tsv
RunAccession    Lane    Sample  BiolRep
ERR458493   1   WT  1
ERR458494   2   WT  1
ERR458495   3   WT  1
ERR458496   4   WT  1
ERR458497   5   WT  1
ERR458498   6   WT  1
ERR458499   7   WT  1
ERR458500   1   SNF2    1
ERR458501   2   SNF2    1
```

sample info

1. find out which RunAccession numbers belong to the WT and SNF2 samples of BiolRep #1

```
awk '$4 == 1 {print $0}' ERP004763_sample_mapping.tsv
```

2. download individual sample

```
awk -F "\t" '$5 == "ERR458493" {print $11}' samples-overview.txt | xargs wget
```

3. either do this 6 more times individually or write a for-loop

```
for i in `seq 3 9`
do
SAMPLE=ERR45849${i}
egrep ${SAMPLE} samples_at_ENA.txt | cut -f11 | xargs wget
done
```

4. for-loop for SNF2 samples

```
 for i in `seq 0 6`
 do
    SAMPLE=ERR45850${i}
    egrep ${SAMPLE} samples_at_ENA.txt | cut -f11 | xargs wget
 done
```

5. sort reads into folders

```
$ mkdir raw_reads
$ mkdir WT_1
$ mkdir SNF2_1
$ mv ERR45849*gz WT_1/
$ mv ERR4585*gz SNF2_1/
```

FASTQ file format

= FASTA + **q**uality scores

1 read ⇔ 4 lines!

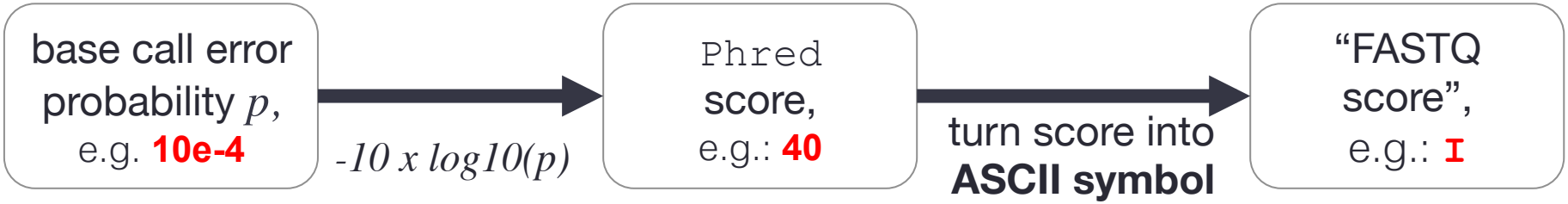
```
1 @ERR459145.1 DHKW5DQ1:219:D0PT7ACXX:2:1101:1590:2149/1
2 GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC
3 +
4 @7<DBADDDBH?DHHI@DH>HHHEGHIIIGGIFFGIBFAAGAFHA’5?B@D
```

1. @Read ID and sequencing run information
2. sequence
3. + (additional description possible)
4. quality scores

Base quality score: summarizing numerical values into single-character representations

Illumina's CASAVA pipeline:
Base calls are immediately recorded (based on the images of the fluorescent signals) together with an error probability (`BCL` files).

These error probability values are translated into ASCII symbols in the `FASTQ` files.




```
@ERR459145.1 DHKW5DQ1:219:D0PT7ACXX:2:1101:1590:2149/1
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC
+
@7<DBADDDBH?DHHI@DH>HHHEGHIIIGGIFFGIBFAAGAFHA'5?B@D
```

ASCII symbols

DEC	OCT	HEX	BIN	Symbol
32	040	20	00100000	
33	041	21	00100001	!
34	042	22	00100010	"
35	043	23	00100011	#
36	044	24	00100100	$
37	045	25	00100101	%
38	046	26	00100110	&
39	047	27	00100111	'
40	050	28	00101000	(
41	051	29	00101001	)
42	052	2A	00101010	*
43	053	2B	00101011	+
65	101	41	01000001	A
66	102	42	01000010	B
67	103	43	01000011	C
68	104	44	01000100	D
69	105	45	01000101	E

ASCII encodes 128 specified characters into integers.

The first 33 characters represent unprintable control codes (e.g. “Start of text”), which is why Phred scores were originally encoded with an **offset of +33**.

http://www.ascii-code.com/

Base quality scores

- each base has a certain error probability (p)
- Phred score = $-10 \times log10(p)$
- Phred scores are ASCII-encoded, e.g., "!" **COULD** represent Phred score 33

also see Table 2 in the course notes

image from https://en.wikipedia.org/wiki/FASTQ_format

Quality control of raw reads: FastQC

http://www.bioinformatics.babraham.ac.uk/projects/fastqc

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

not specific for RNA-seq data!

The main functions of FastQC are:

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

```
$ ~/mat/software/FastQC/fastqc

$ ~/mat/software/anaconda2/bin/multiqc
```