Differential gene expression analysis using RNA-seq

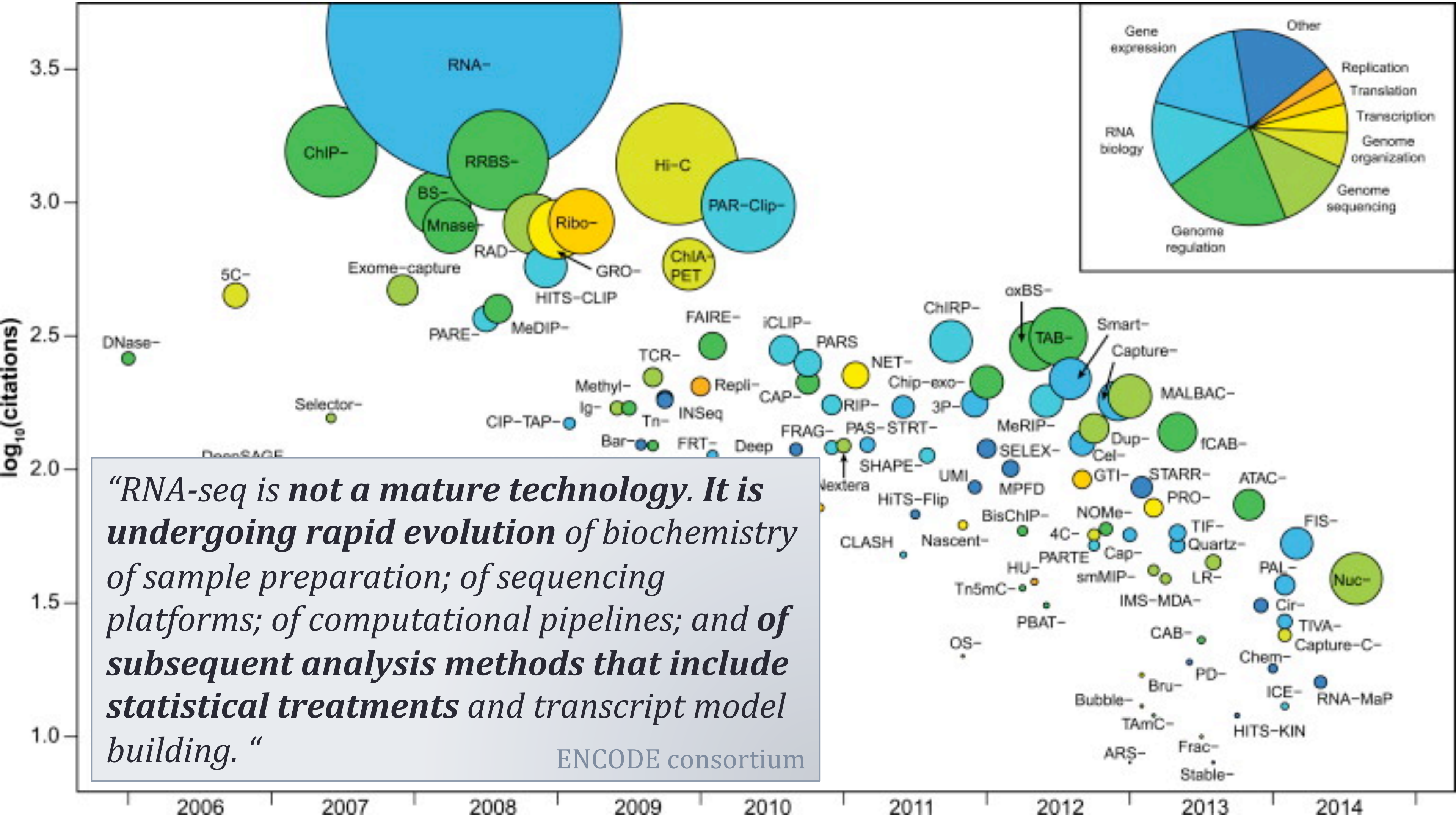
Applied Bioinformatics Core, March 2018

Friederike Dündar with Luce Skrabanek & Paul Zumbo

Day 1: Introduction into high-throughput sequencing [many general concepts!]

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

RNA-seq is popular, but still 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

“Analysis paralysis”

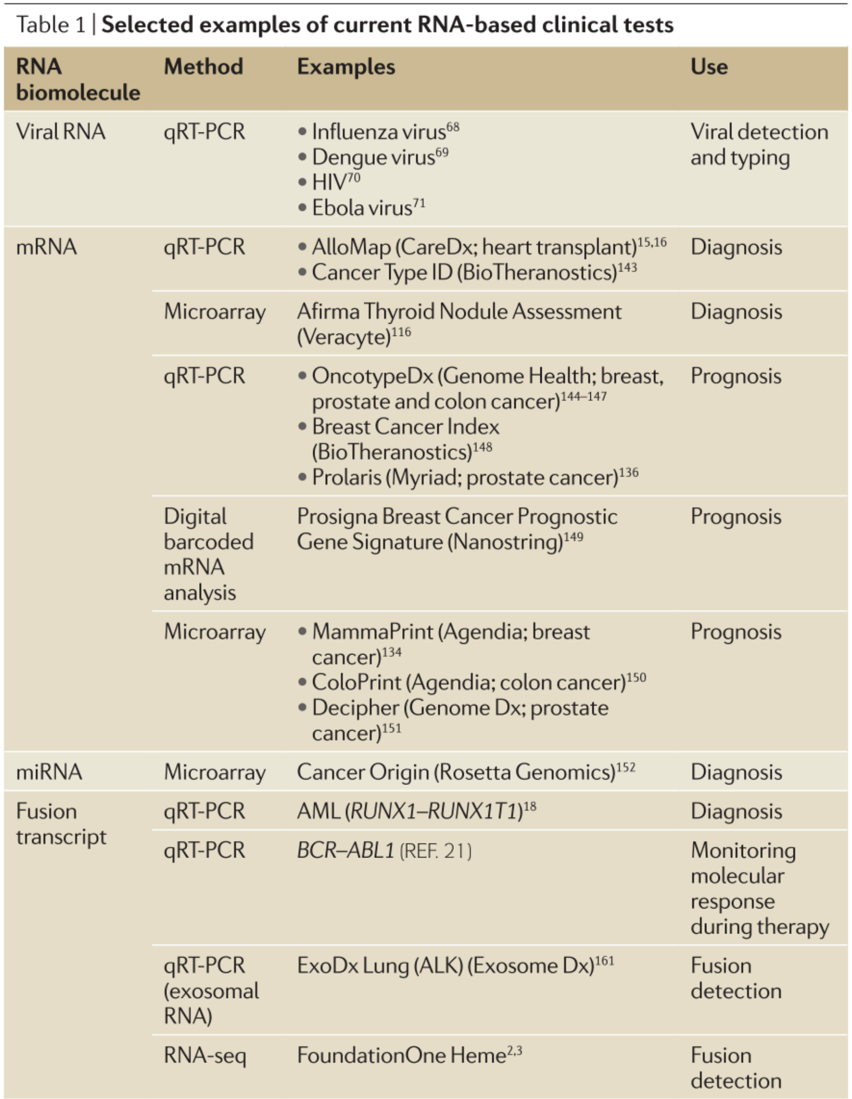


Table 1 | **Selected examples of current RNA-based clinical tests**

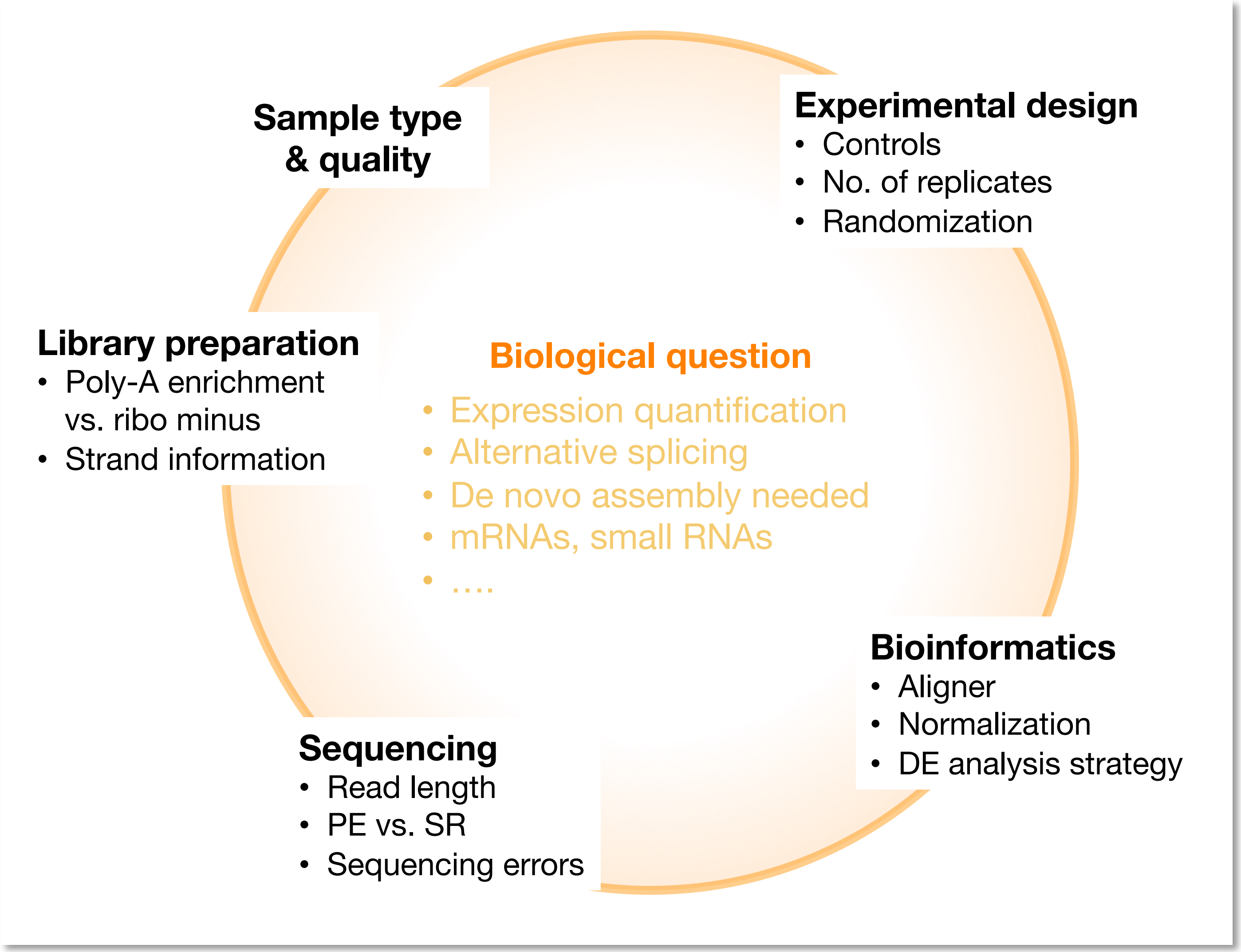
RNA biomolecule	Method	Examples	Use
Viral RNA	qRT-PCR	• Influenza virus[68] • Dengue virus[69] • HIV[70] • Ebola virus[71]	Viral detection and typing
mRNA	qRT-PCR	• AlloMap (CareDx; heart transplant)[15,16] • Cancer Type ID (BioTheranostics)[143]	Diagnosis
	Microarray	Afirma Thyroid Nodule Assessment (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 (tx 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)

What to expect from the class



Sample type & quality

Experimental design
- Controls
- No. of replicates
- Randomization

Library preparation
- Poly-A enrichment vs. ribo minus
- Strand information

Biological question
- Expression quantification
- Alternative splicing
- De novo assembly needed
- mRNAs, small RNAs
- ….

Bioinformatics
- Aligner
- Normalization
- DE analysis strategy

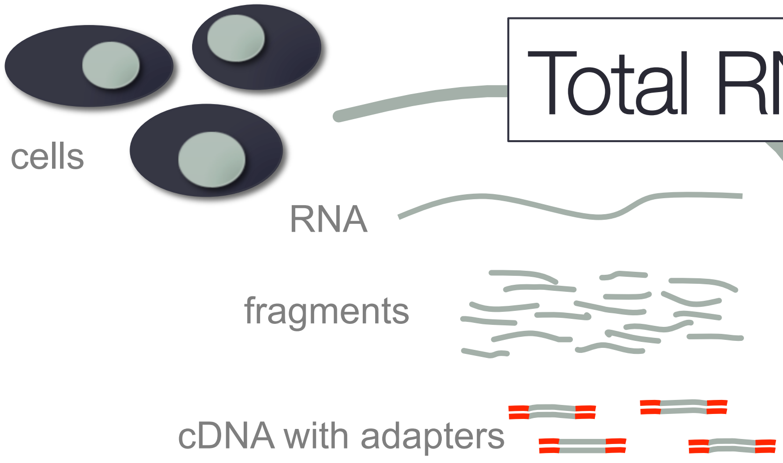
Sequencing
- Read length
- PE vs. SR
- Sequencing errors

NOT COVERED:

- novel transcript discovery
- transcriptome assembly
- alternative splicing analysis

(see the course notes for references to useful reviews)

RNA-seq workflow overview



Total RNA extraction

- mRNA enrichment
- fragmentation
- cDNA library

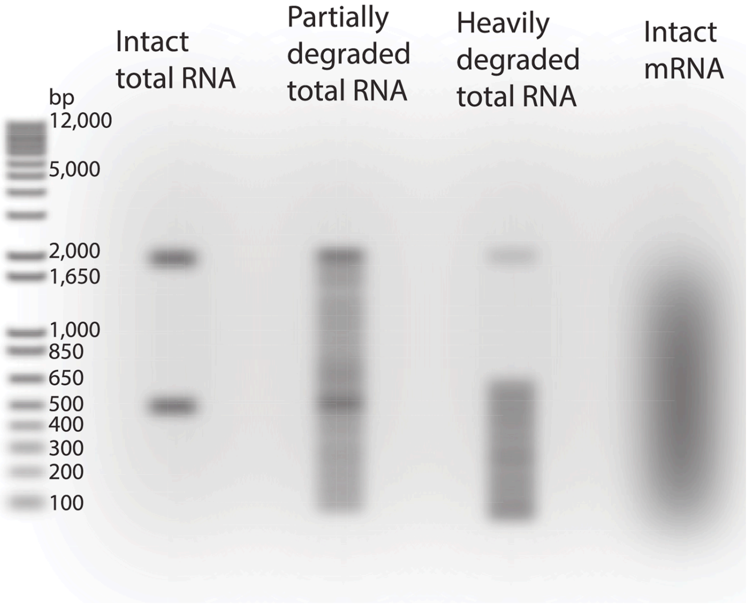
Sequencing

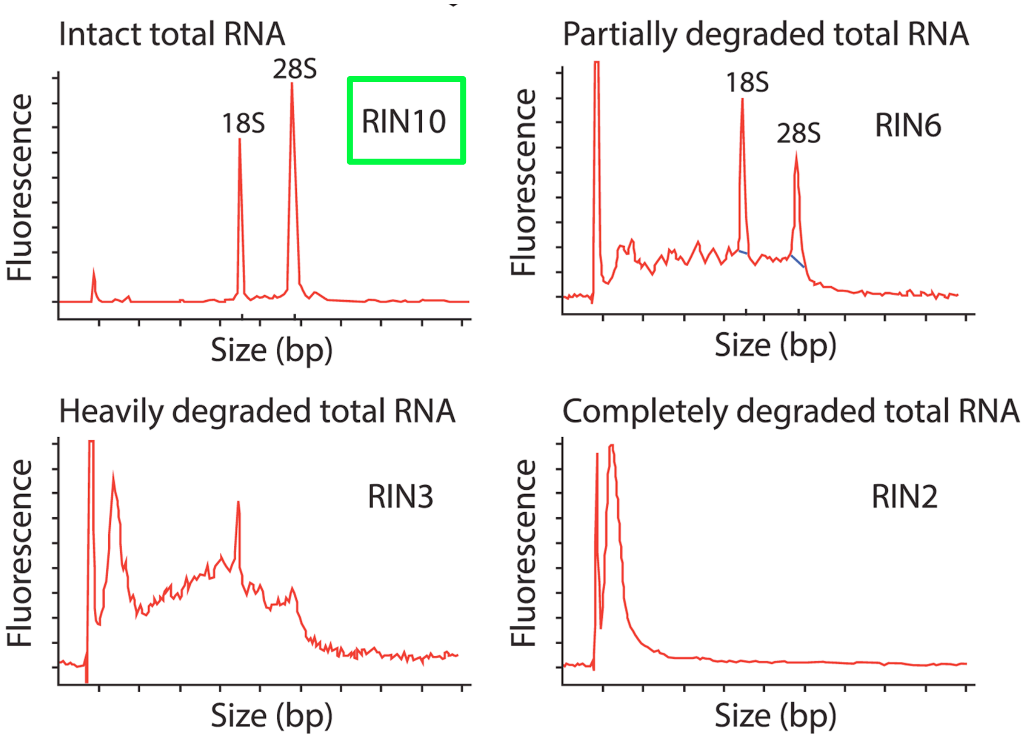
- cluster generation
- sequencing by synthesis
- image acquisition

Bioinformatics

Quality control of RNA extraction

Gel electrophoresis




Intact total RNA

Partially degraded total RNA

Heavily degraded total RNA

Completely degraded total RNA

RIN = 28S:18S ratio

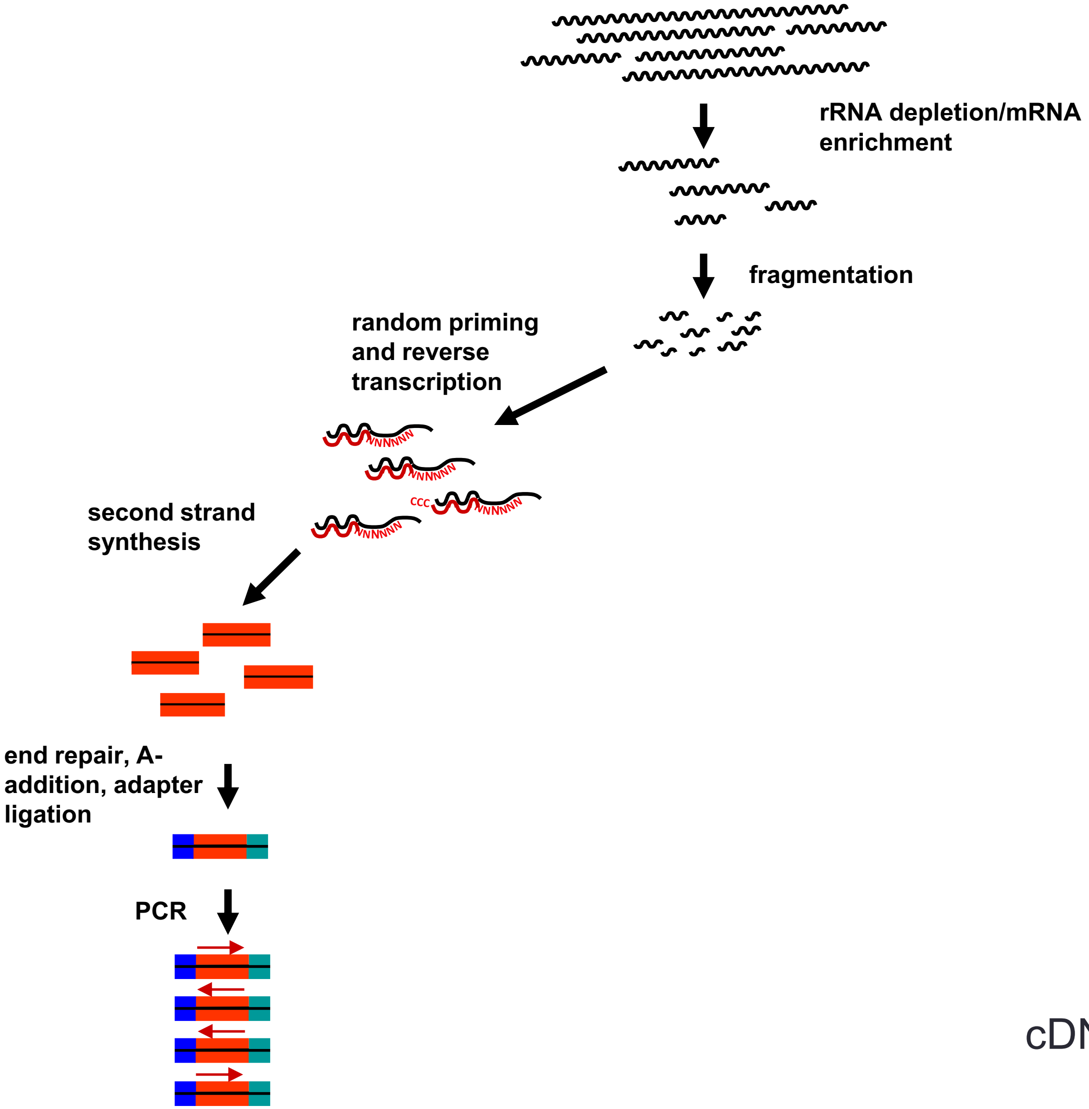
avoid degraded RNA junk

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

RNA-seq library preparation

QC!



rRNA depletion/mRNA enrichment

fragmentation

random priming and reverse transcription

second strand synthesis

end repair, A-addition, adapter ligation

PCR

classical Illumina protocol
(unstranded)

RNA extraction

↓

mRNA enrichment

↓

fragmentation (~200 bp)

↓

cDNA synthesis

↓

cDNA plus sequencing adapters

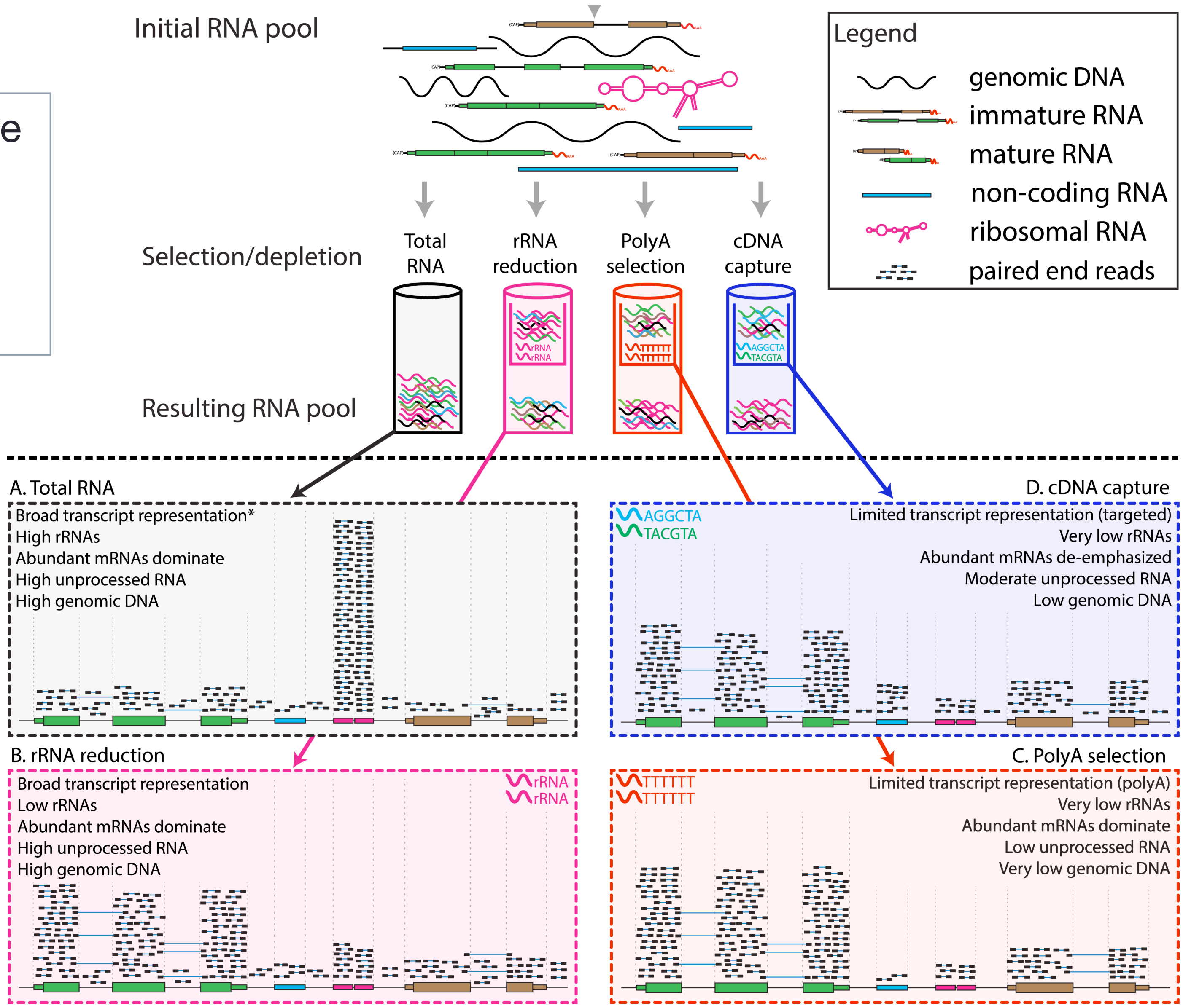
Van Dijk et al. (2014). Experimental Cell Research, 322(1), 12–20. doi:10.1016/j.yexcr.2014.01.008

Influence of the **RNA enrichment** strategy

which transcripts are you interested in?

what type of noise can you tolerate?

- Total RNA
- rRNA depletion
- mRNA selection
- cDNA capture



A. Total RNA
Broad transcript representation*
High rRNAs
Abundant mRNAs dominate
High unprocessed RNA
High genomic DNA

D. cDNA capture
Limited transcript representation (targeted)
Very low rRNAs
Abundant mRNAs de-emphasized
Moderate unprocessed RNA
Low genomic DNA

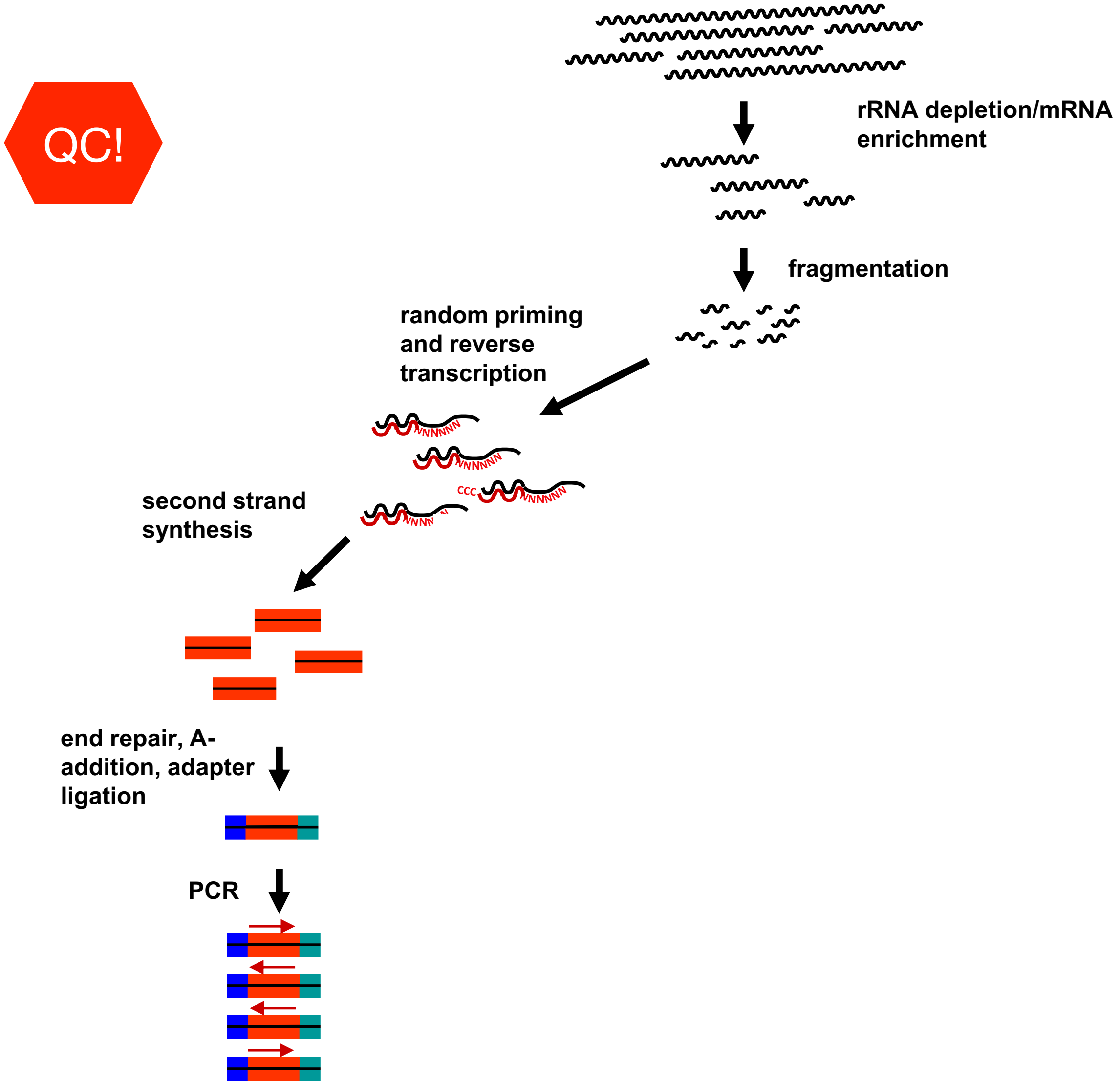
B. rRNA reduction
Broad transcript representation
Low rRNAs
Abundant mRNAs dominate
High unprocessed RNA
High genomic DNA

C. PolyA selection
Limited transcript representation (polyA)
Very low rRNAs
Abundant mRNAs dominate
Low unprocessed RNA
Very low genomic DNA

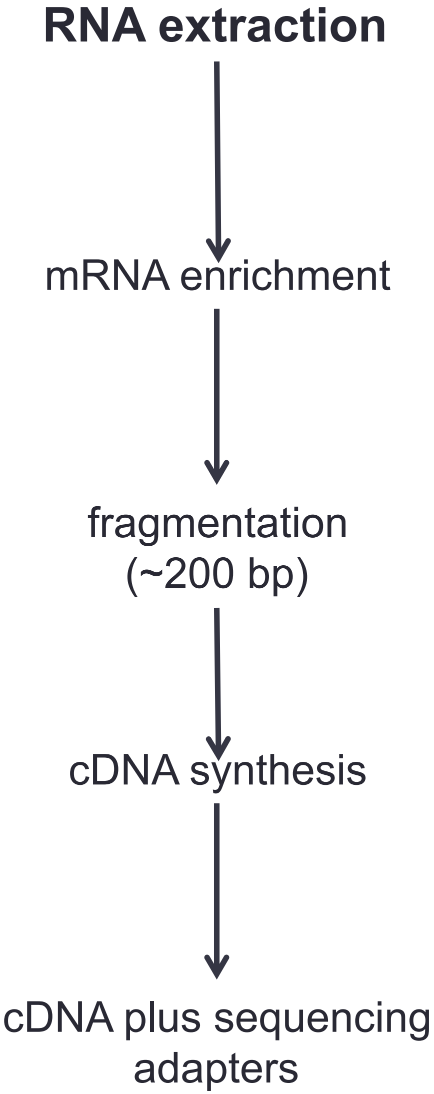
Figure from Griffith et al., PLoS Comp Biol (2015). doi: 10.1371/journal.pcbi.1004393

RNA-seq library preparation: pick one!

QC!


rRNA depletion/mRNA enrichment
fragmentation
random priming and reverse transcription
NNNNNN
CCC
second strand synthesis
end repair, A-addition, adapter ligation
PCR

classical Illumina proto (unstranded)



RNA extraction

mRNA enrichment

fragmentation (~200 bp)

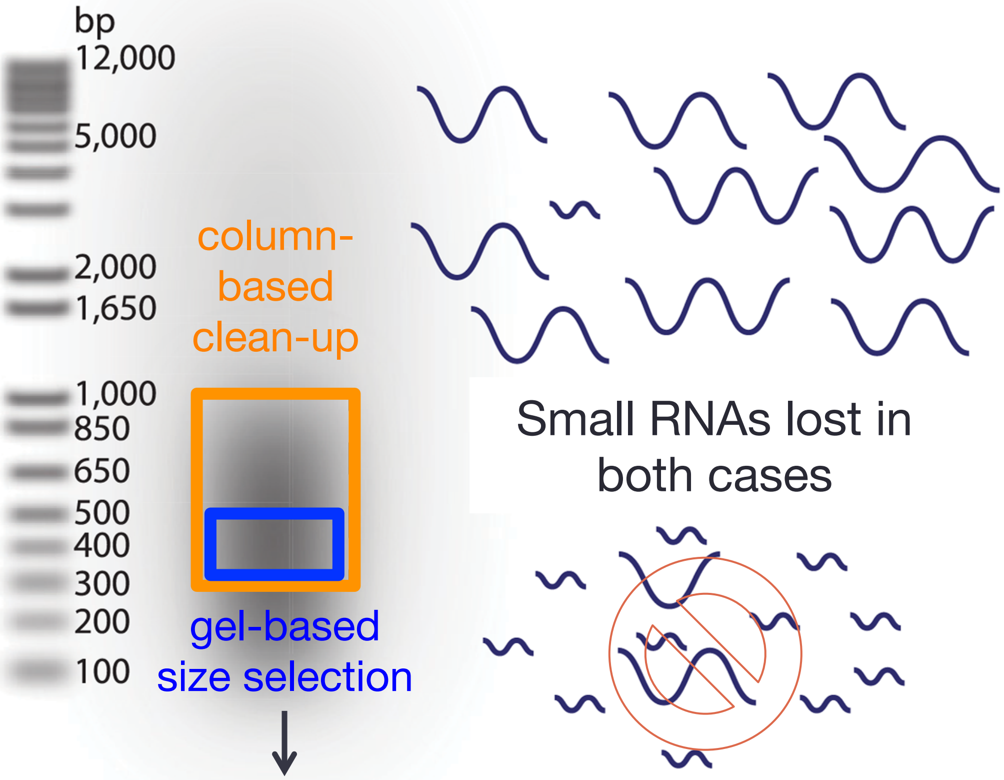
cDNA synthesis

cDNA plus sequencing adapters

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

Size selection

Size selection or exclusion
(e.g. PAGE, SPRI magnetics beads, etc.)



column-based clean-up

gel-based size selection

Small RNAs lost in both cases

more efficient sequencing

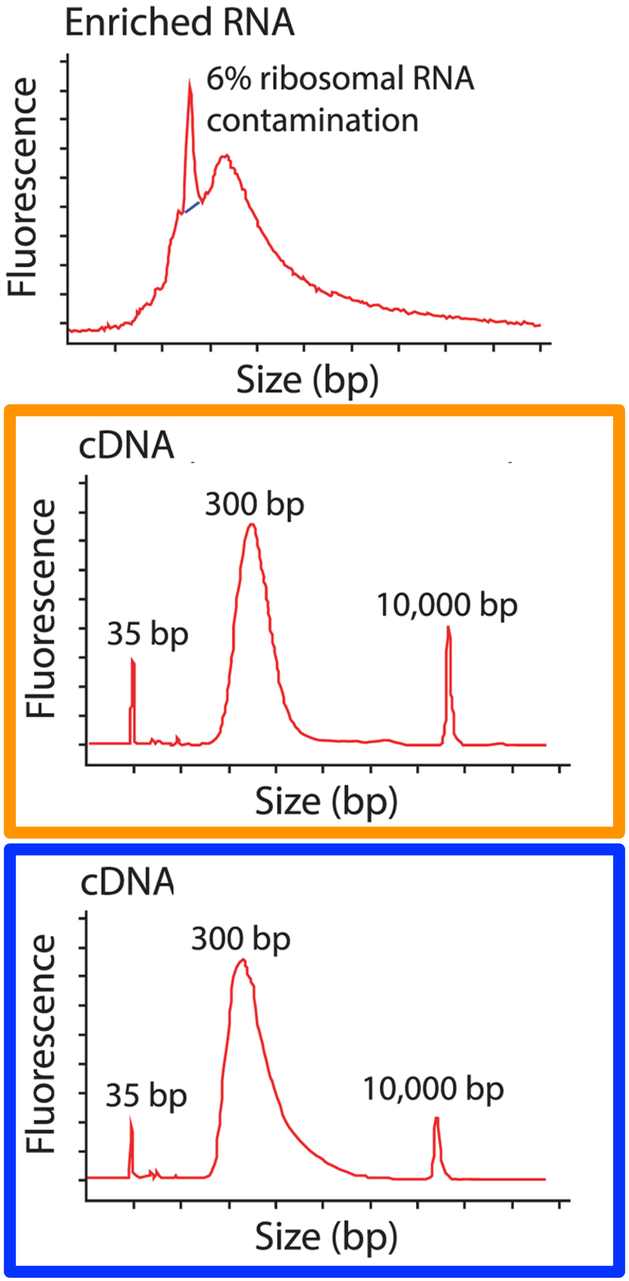
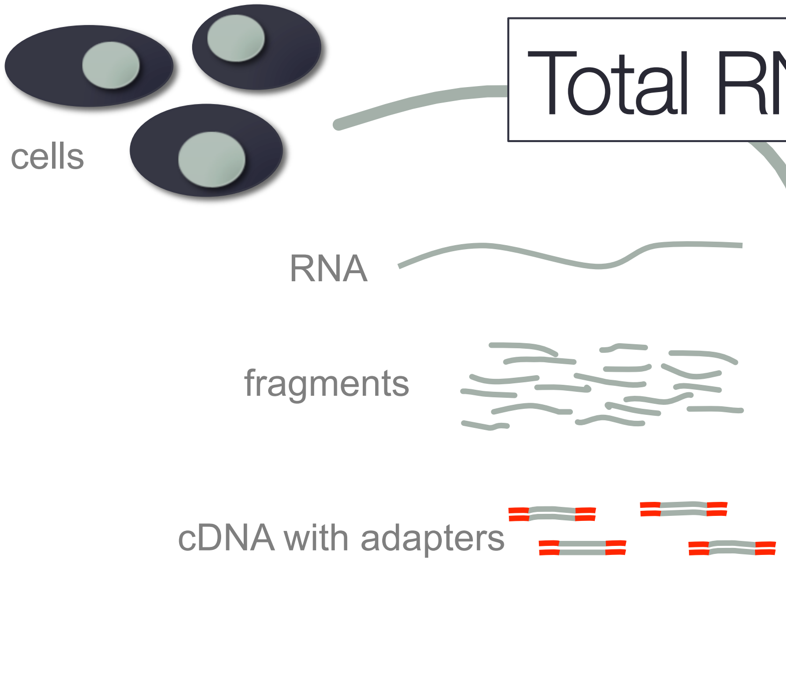


Figure modified from Griffith et al., PLoS Comp Biol (2015). doi: 10.1371/journal.pcbi.1004393

RNA-seq workflow overview



Total RNA extraction

cells

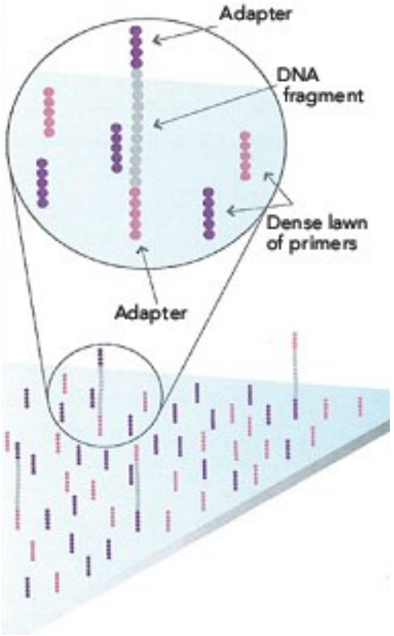
RNA

fragments

cDNA with adapters

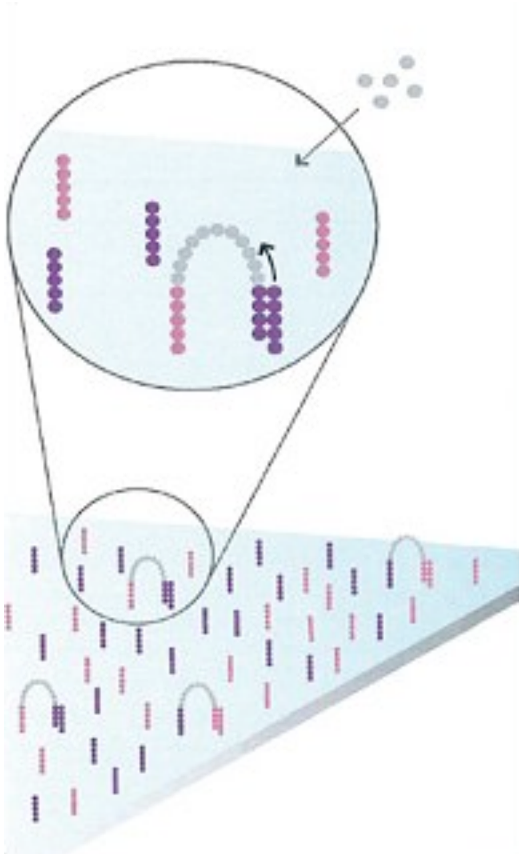
Sequencing

flowcell with primers

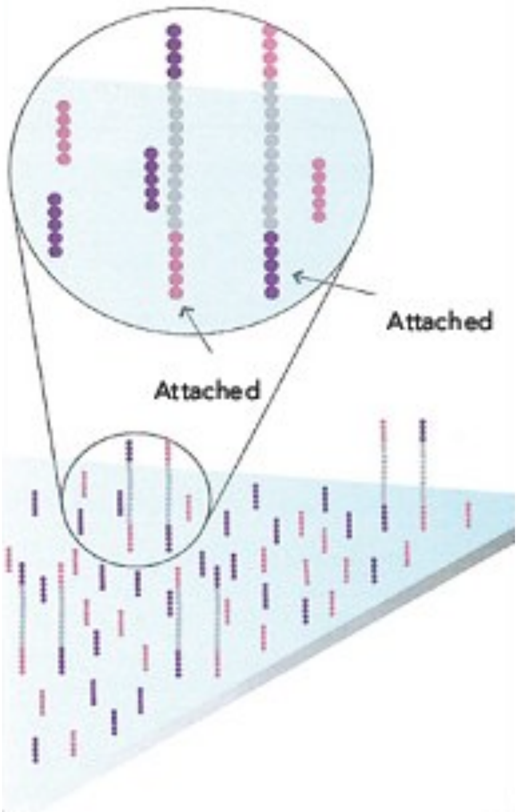


http://informatics.fas.harvard.edu/test-tutorial-page/

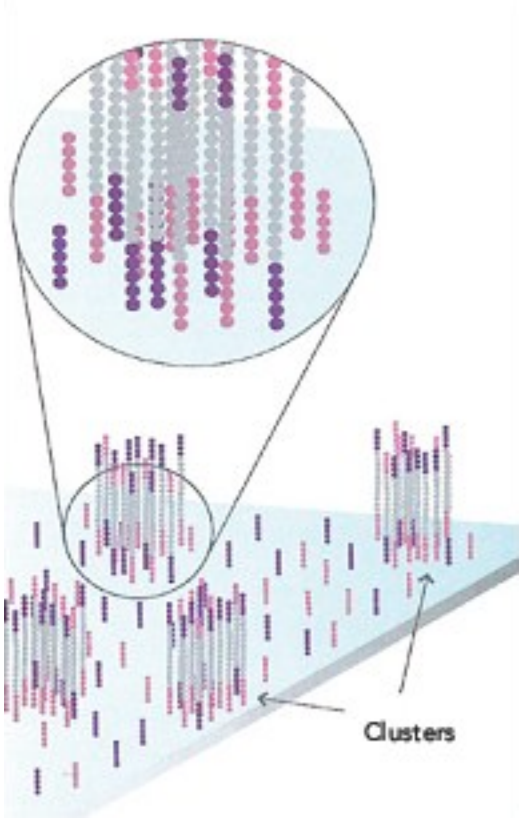
Cluster generation



bridge amplification



denaturation

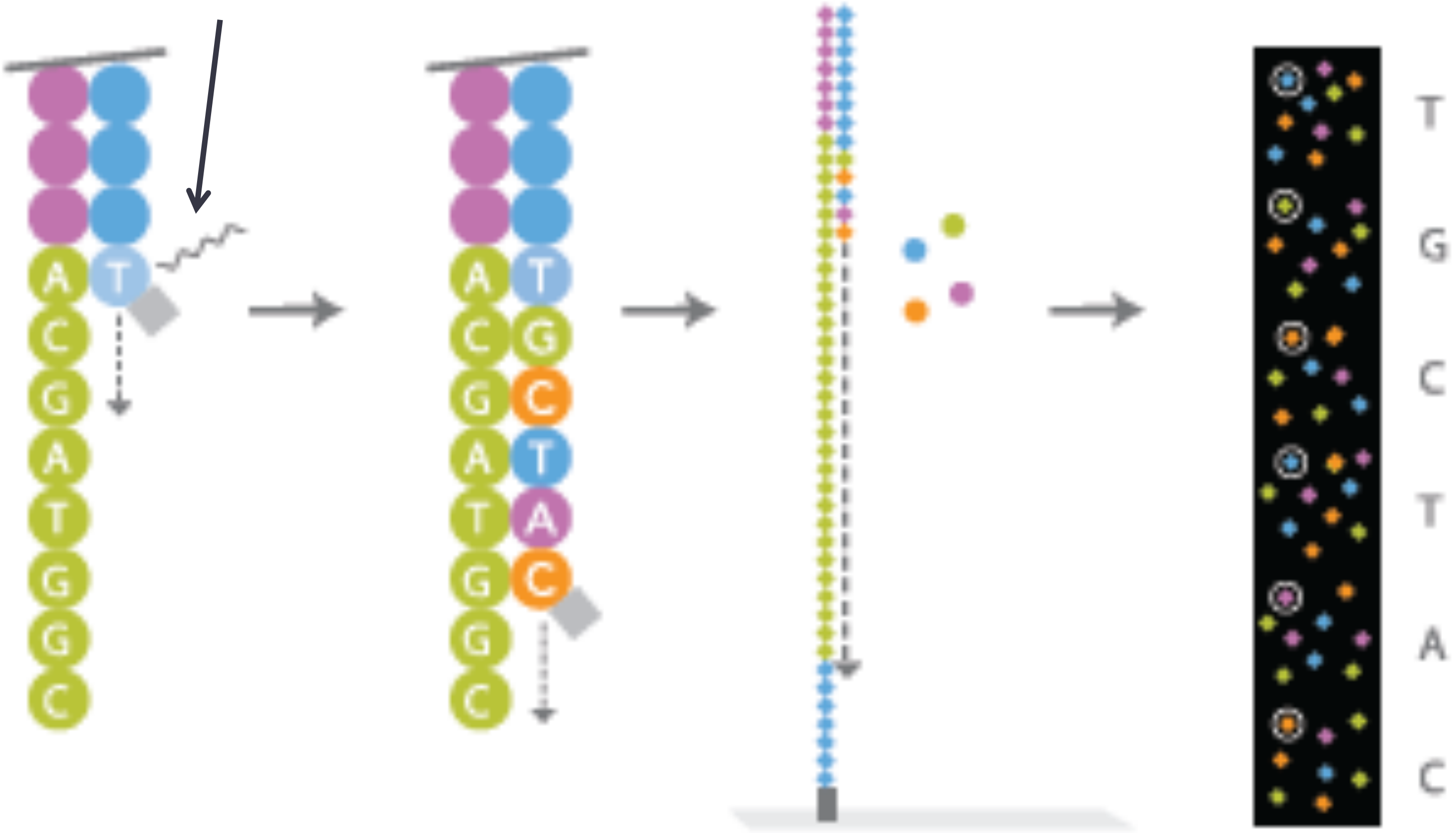


cluster generation

removal of complementary strands → identical fragment copies remain

http://informatics.fas.harvard.edu/test-tutorial-page/

Sequencing by synthesis



labelled dNTP

1. extend 1st base
2. read
3. deblock

repeat for 50 – 100 bp

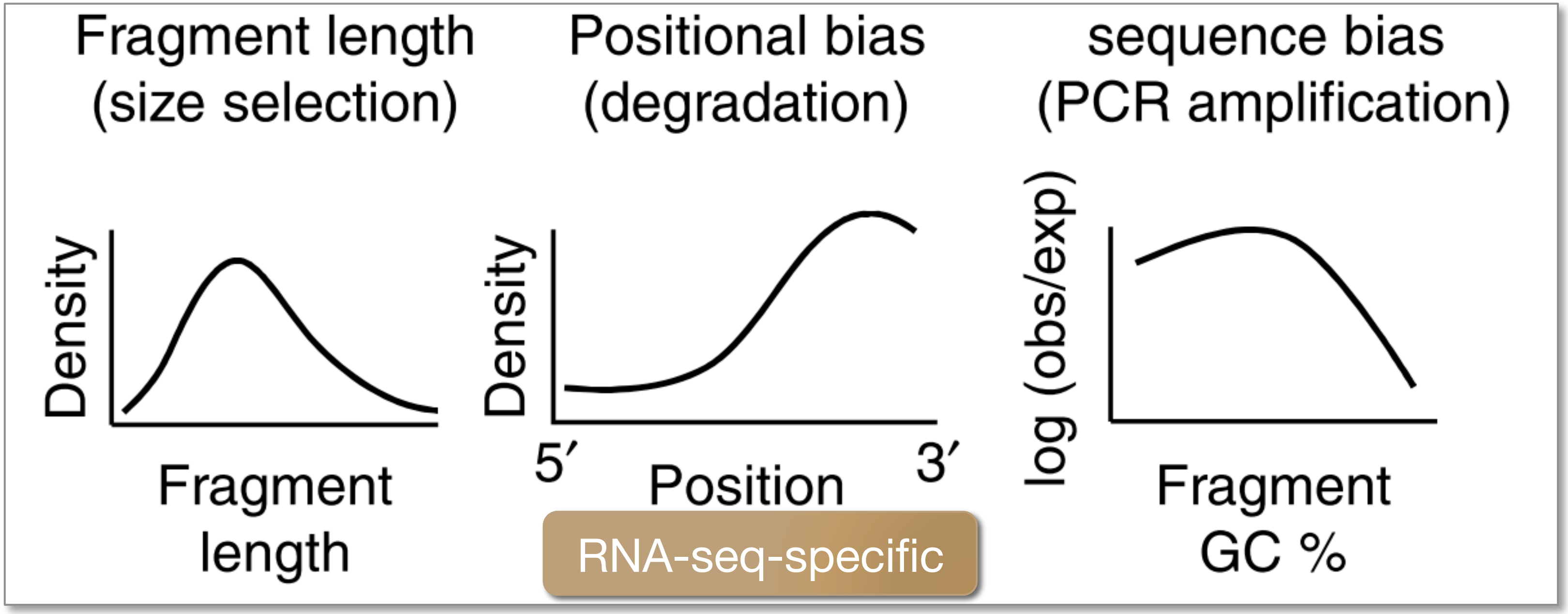
generate base calls

Image from Illumina

Typical biases of Illumina sequencing

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

sample-specific problems!



Fragment length (size selection)

Positional bias (degradation)

sequence bias (PCR amplification)

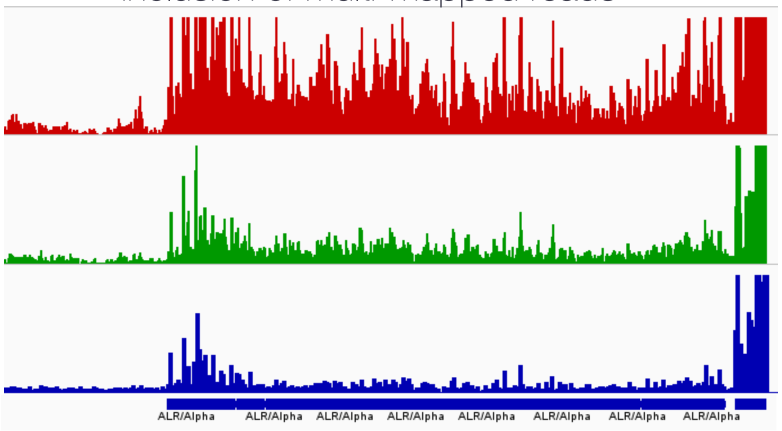
RNA-seq-specific

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

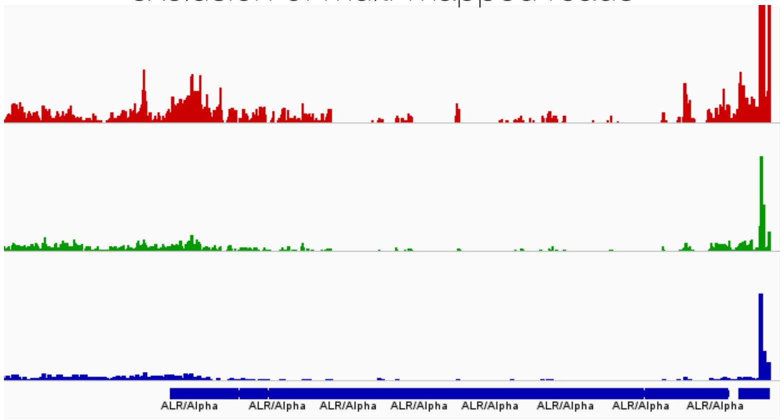
General sources of biases
(not inherently sample-specific)

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

- inappropriate **data processing**



inclusion of multi-mapped reads

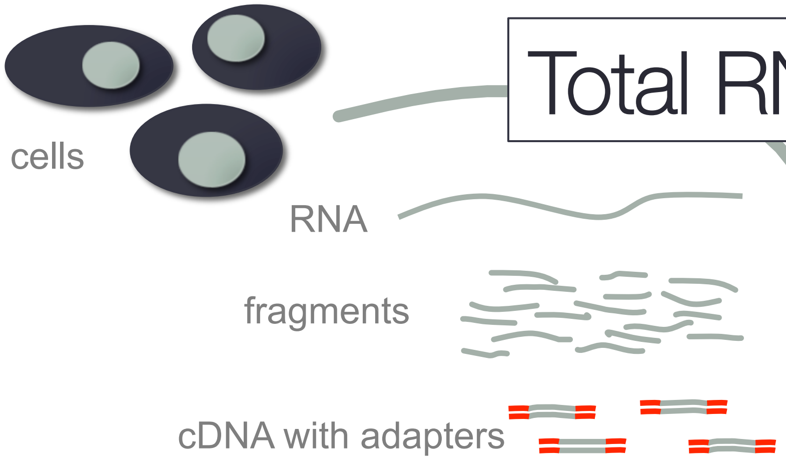


exclusion of multi-mapped reads

RAW SEQUENCING READS

Let the data wrangling begin!

RNA-seq workflow overview



Total RNA extraction

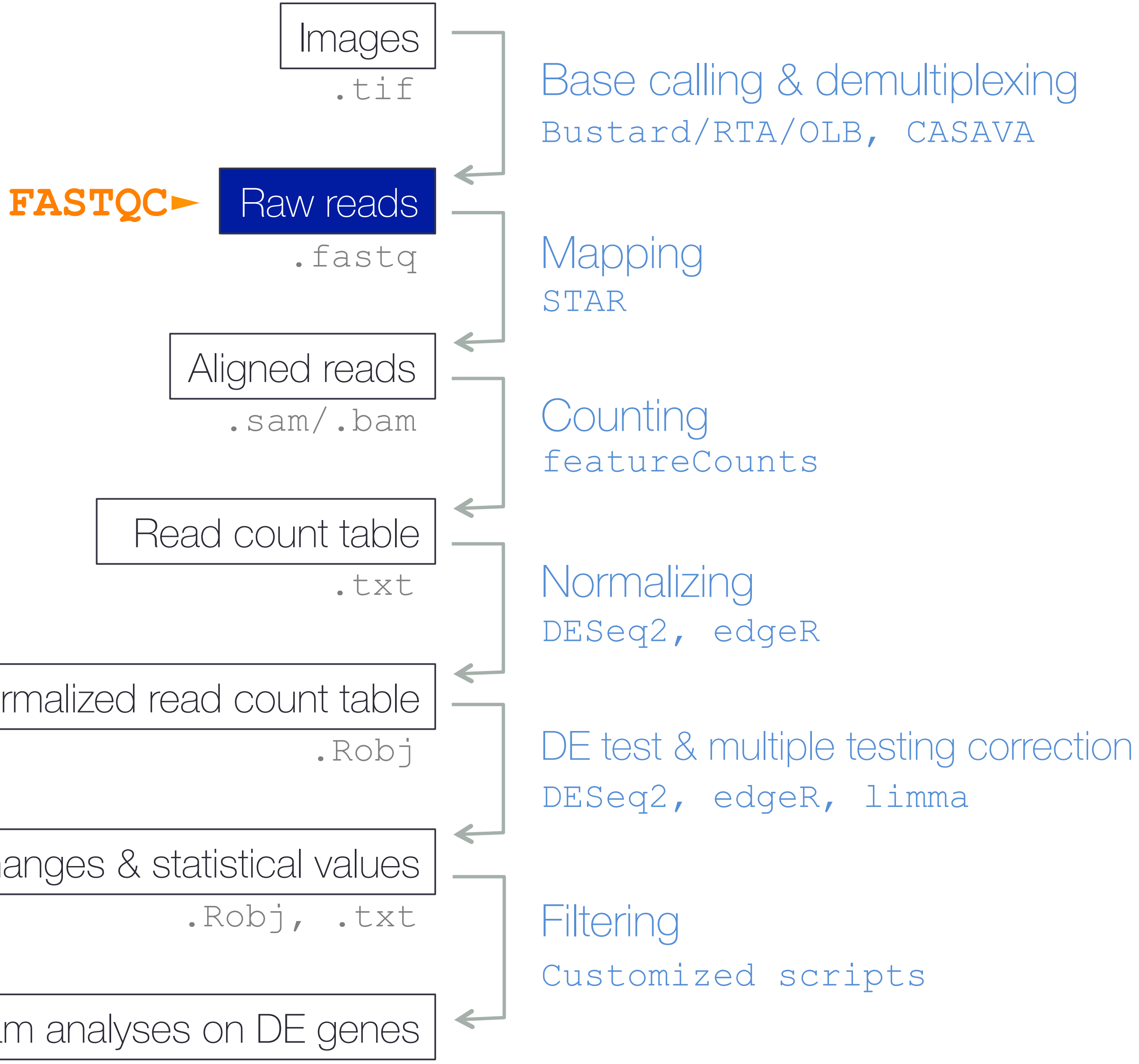
- mRNA enrichment
- fragmentation
- cDNA library

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

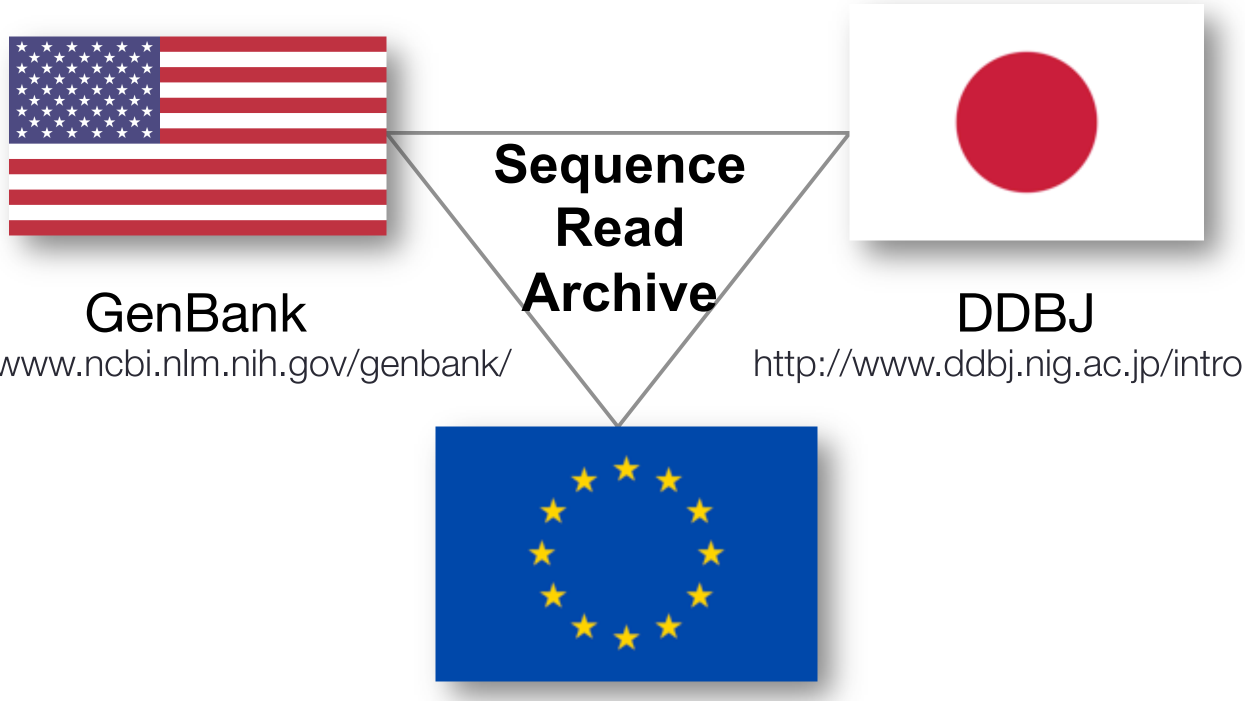
DE test & multiple testing correction
DESeq2, edgeR, limma

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

Filtering
Customized scripts

Downstream analyses on DE genes

Where are all the reads?



Sequence Read Archive

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 SRA is the main repository for publicly available DNA and RNA sequencing data of which three instances are maintained world-wide.

Let’s download!

- We will work with a data set submitted by Gierlinski et al.
- they deposited the sequence files with SRA – we will retrieve it via ENA (https://www.ebi.ac.uk/ena/)
- accession number: ERP004763

Course notes @ https://chagall.med.cornell.edu/RNASEQcourse/
of @ http://www.trii.org/courses/rnaseq.html

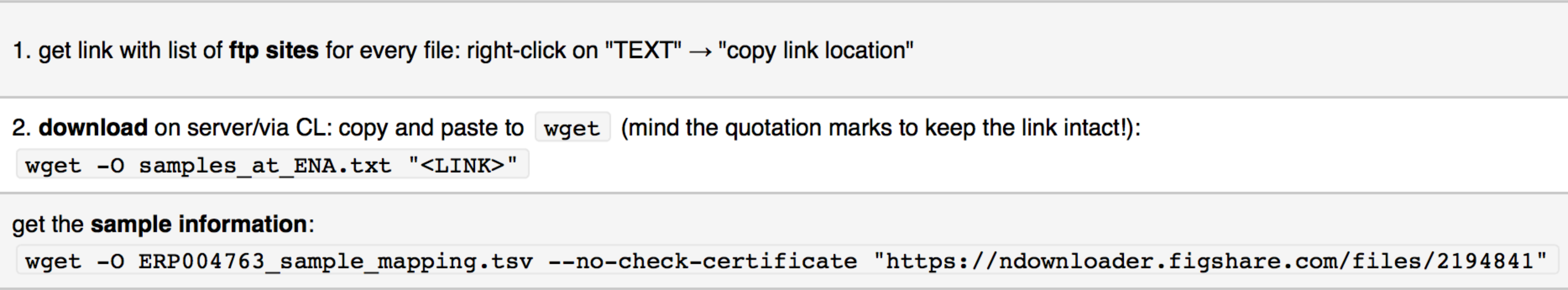
See **Section 2 (Raw Data)** for download instructions etc.

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

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

Downloading a batch of fastq files

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



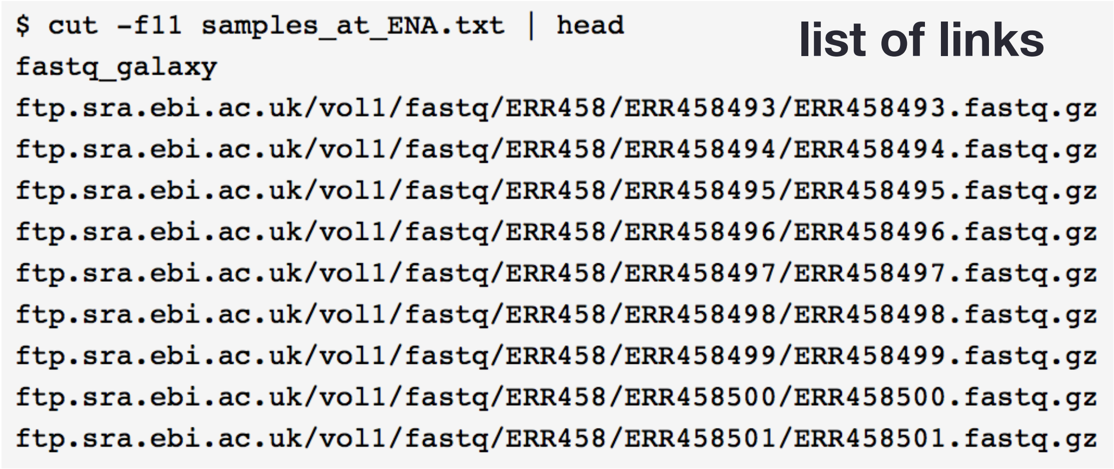
1. get link with list of **ftp sites** for every file: right-click on "TEXT" → "copy link location"

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>"
```

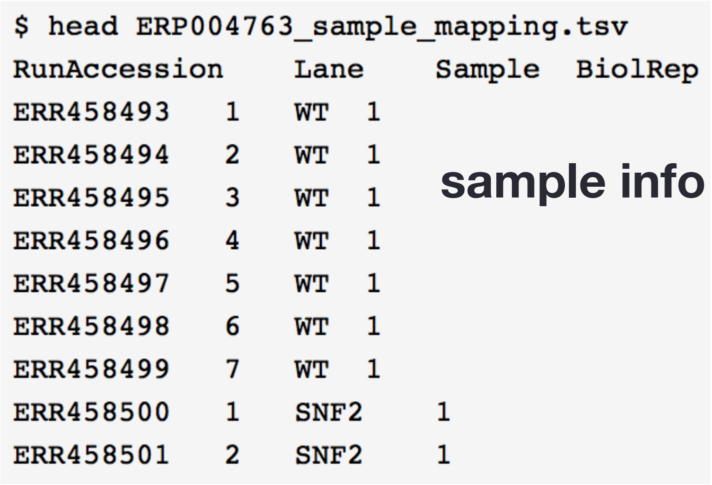
get the **sample information**:

```
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

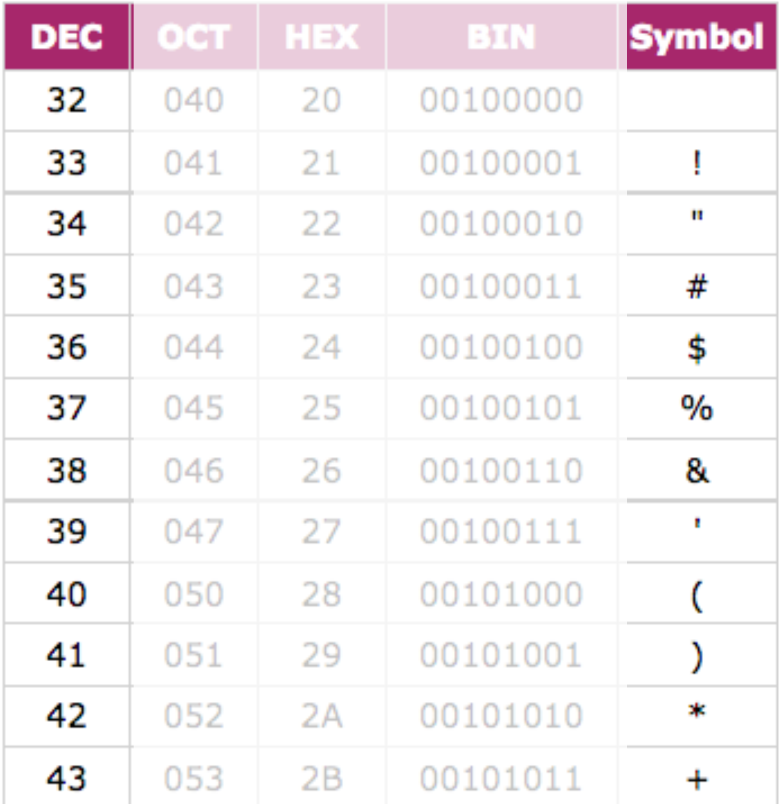
1 read ⇔ 4 lines!

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

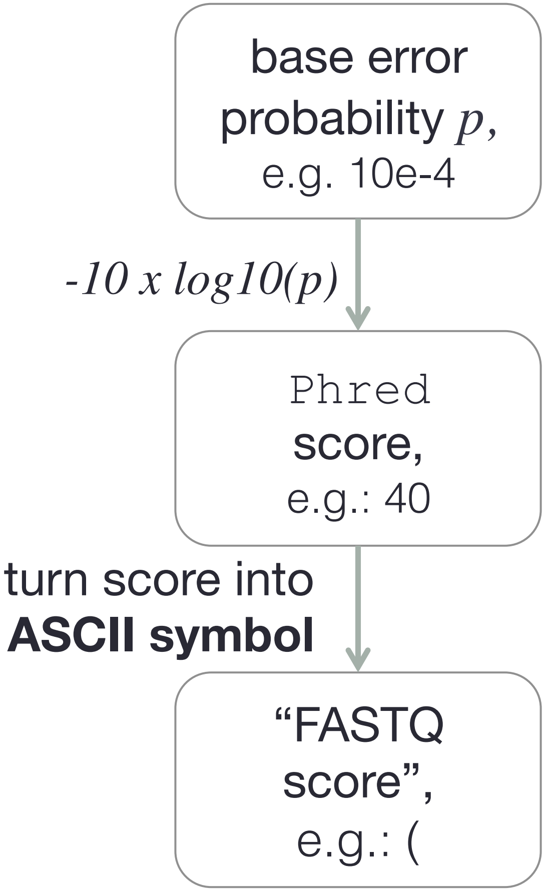
1. @Read ID and sequencing run information
2. sequence
3. + (additional description possible)
4. quality scores

Base quality score

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



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	+



base error probability p, e.g. 10e-4

-10 x log10(p)

Phred score, e.g.: 40

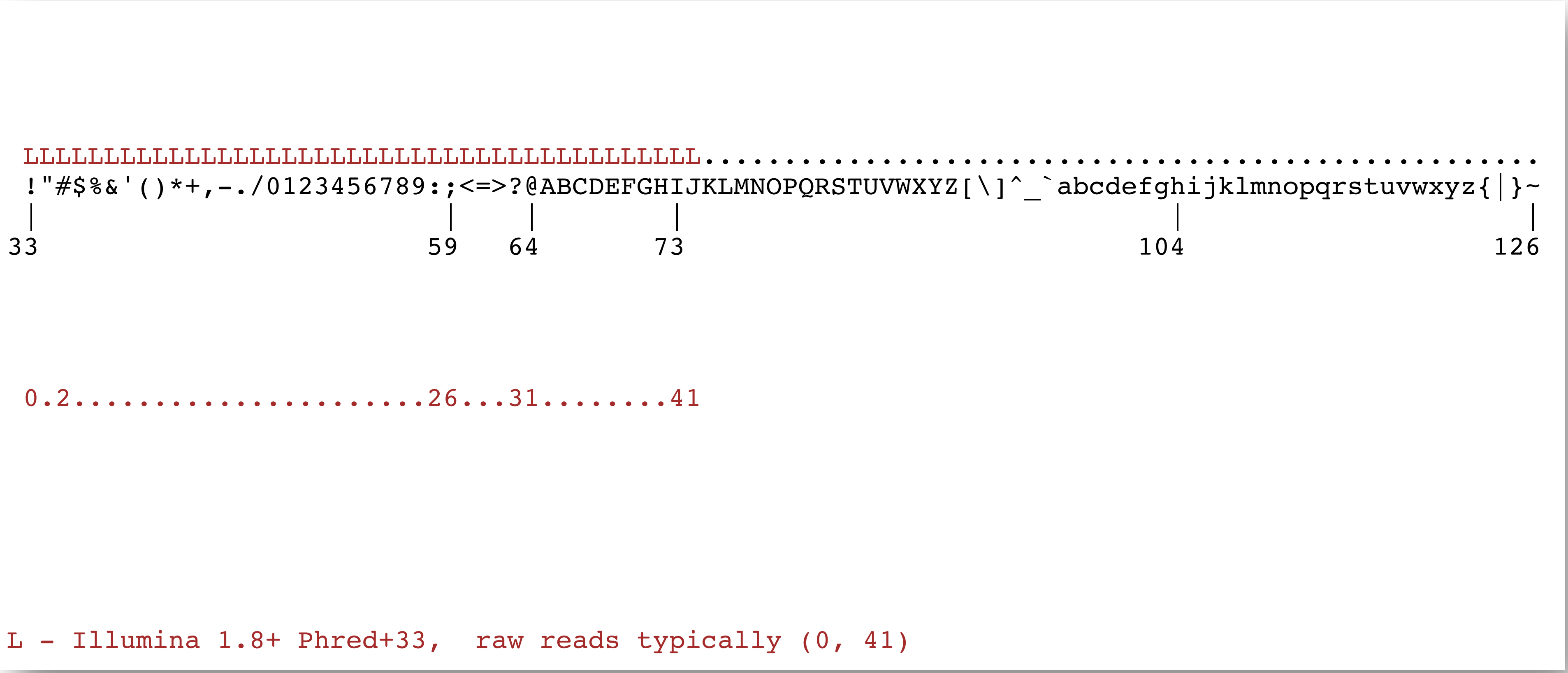
turn score into **ASCII symbol**

“FASTQ score”, e.g.: (

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



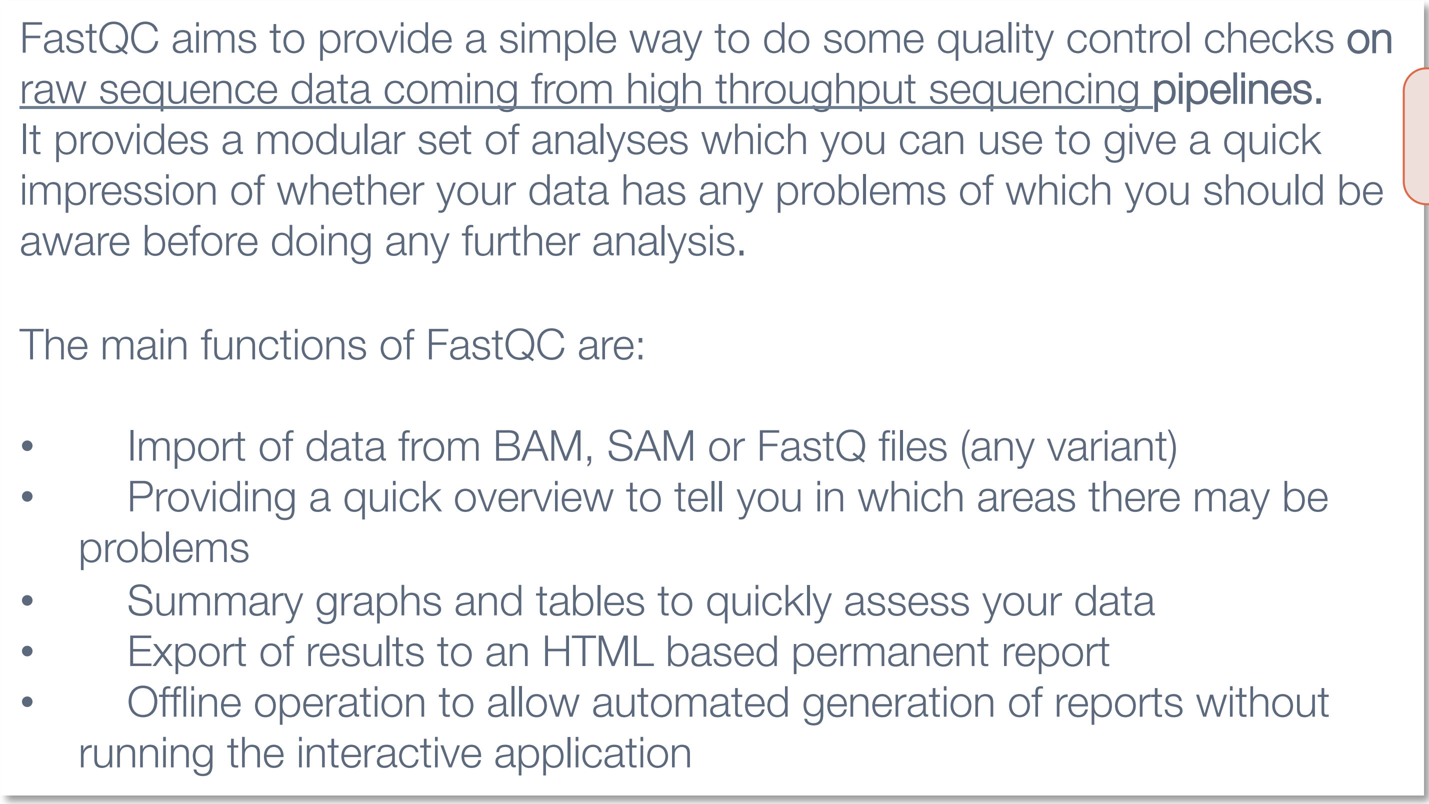
```
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

 0.2......................26...31........41

L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

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.

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

not specific for RNA-seq data!

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

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